CERN–2013–001
8 January 2013

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

High Power Hadron Machines

Bilbao, Spain
24 May – 2 June 2011

Proceedings

Editor: R. Bailey

GENEVA
2013

ISBN 978–92–9083–384–0
ISSN 0007–8328
Copyright © CERN, 2013
Creative Commons Attribution 3.0
Knowledge transfer is an integral part of CERN's mission.
CERN publishes this report Open Access under the Creative Commons Attribution 3.0 license (http://creativecommons.org/licenses/by/3.0/) in order to permit its wide dissemination and use.

This report should be cited as:
Proceedings of the CAS–CERN Accelerator School: High Power Hadron Machines, Bilbao, Spain, 24 May – 2 June 2011, edited by R. Bailey, CERN-2013-001.

A contribution in this report should be cited as:
[Author name(s)], in Proceedings of the CAS–CERN Accelerator School: High Power Hadron Machines, Bilbao, Spain, 24 May – 2 June 2011, edited by R. Bailey, CERN-2013-001, pp. [first page] – [lastpage]

Abstract

These proceedings collate lectures given at the twenty-fifth specialized course organised by the CERN Accelerator School (CAS). The course was held in Bilbao, Spain from 24 May to 2 June 2011, in collaboration with ESS Bilbao.

The course covered the background accelerator physics, different types of particle accelerators and the underlying accelerator systems and technologies, all from the perspective of high beam power. The participants pursued one of six case studies in order to get "hands-on" experience of the issues connected with high power machines.

Foreword

The aim of the CERN Accelerator School (CAS) is to collect, preserve and disseminate the knowledge accumulated in the world's accelerator laboratories over the years. This applies not only to general accelerator physics, but also to related sub-systems and associated technologies, and to how these are adapted to particular requirements. This wider aim is achieved by means of specialized courses currently held yearly. The topic of the 2011 specialized course was High Power Hadron Machines and was held at the Hotel Barceló Nervión, Bilbao, Spain from 24 May to 2 June 2011.

The course was made possible by the productive collaboration with ESS Bilbao, in particular through the efforts of Javier Bermejo and Sira Cordon.

The backing of the CERN management and the guidance of the CAS Advisory and Programme Committees enabled the course to take place, while the attention to detail of the Local Organising Committee and the management and staff of the Hotel Barceló Nervión ensured that the school was held under optimum conditions.

Special thanks must go to the lecturers for the preparation and presentation of the lecturers, even more so to those who have written a manuscript for these proceedings.

For the production of the proceedings we are indebted to the efforts of Barbara Strasser and to the CERN IT Collaboration and Information Services, especially Jerome Caffaro for his positive collaboration as we move into new ways of publishing.

Finally, the enthusiasm of the participants of 22 nationalities, from institutes in 9 countries, provides convincing proof of the usefulness and success of the course.

Roger Bailey
CERN Accelerator School

PROGRAMME High Power Hadron Machines
24 May – 2 June 2011, Bilbao, Spain

Time	Wednesday 25 May	Thursday 26 May	Friday 27 May	Saturday 28 May	Sunday 29 May	Monday 30 May	Tuesday 31 May	Wednesday 1 June	Thursday 2 June
08:30	Introduction I K. Clausen	Multiparticle Beam Dynamics in Linacs II A. Letchford	RF Generation R. Carter	Tuners and Couplers G. Devanz		Specific Beam Diagnostics II K. Wittenburg	New Target Concepts I. Efthymiopoulos	Radio Protection H. Vincke	Departure to airport
09:30 10:30	Challenges and Beam Parameters of Machines I M. Lindroos	Linacs M. Vretenar	RF Basics I F. Gerigk	SC versus NC Cavities G. Clemente		Vacuum I G. Franchetti	Ion Sources I D. Faircloth	Activation & Radiation Damage of Components in the Environment of Proton Accelerators D. Kiselev	
	C O F F E E					**C O F F E E**			
11:00 12:00	Introduction II K. Clausen	Multiparticle Beam Dynamics in Rings I C. Prior	RF Basics II F. Gerigk	H^- Injection C. Prior		Vacuum II G. Franchetti	Ion Sources II D. Faircloth	Remote Handling M. Wohlmuther	
12:00 13:00	Challenges and Beam Parameters of Machines II M. Lindroos	Multiparticle Beam Dynamics in Rings II C. Prior	Beam Loading I A. Gamp	Lattice Design I B. Holzer		Fundamentals of Cryogenics I P. Pierini	Collimation S. Wronka	Comments on Case Study	
	L U N C H					**L U N C H**			
14:30 15:30	Beam Dynamics with Space Charge I C. Prior	Cyclotrons M. Seidel	Beam Loading II A. Gamp	Lattice Design II B. Holzer		Fundamentals of Cryogenics II P. Pierini	Case Study A. Jansson/ C. Oyon	Commissioning Strategies J. Galambos	
15:30 16:30	Multiparticle Beam Dynamics in Linacs I A. Letchford	Synchrotrons O. Boine-Frankenheim	RF Transport S. Choroba	RFQ M. Vretenar		Targets and Beam Dumps I M. Wohlmuther	Case Study A. Jansson/ C. Oyon	Reliability & Tolerance Case of ADS J.-L. Biarrotte	
	T E A					**T E A**			
17:00 18:00	Beam Dynamics with Space Charge II C. Prior	FFAGs S. Machida	HOMs H.-W. Glock	Specific Beam Diagnostics I K. Wittenburg		Targets and Beam Dumps II M. Wohlmuther	Case Study A. Jansson/ C. Oyon	Closing Talk R. Bailey	
19:00 20:00	**Welcome Drink Dinner**	**Dinner**	**Dinner**	**Dinner**	**Special Dinner**	**Dinner**	**Dinner**	**Dinner**	

E X C U R S I O N

Contents

Beam dynamics in linacs

Alan Letchford
STFC Rutherford Appleton Laboratory, Didcot, UK

Abstract
An introduction to beam dynamics in proton linear accelerators in the absence of space charge is given.

1 Introduction

Beam dynamics is the study of the collective behaviour of an ensemble of particles constituting the beam in a particle accelerator. For the purposes of this introduction to the subject, the particle motion in the transverse (x and y) and longitudinal (z) directions is treated entirely independently which is a perfectly good assumption in the vast majority of cases where space charge forces are absent or negligible.

Section 2 introduces transverse particle dynamics while Section 3 covers the longitudinal dynamics.

2 Transverse dynamics in linacs

Transverse dynamics studies the motion of particles in a plane perpendicular to the average direction of motion of the particles. In a linear accelerator there is a well-defined forward direction and the transverse plane is defined with reference to this. In the majority of cases the motion is studied independently in two orthogonal directions within this plane, usually the horizontal (x) direction and the vertical (y) direction. Occasionally, where the forces involved are axisymmetric, motion is considered simply in the radial direction.

In the absence of any other tendency for beam particles to move away from the beam axis, the very act of acceleration introduces a transverse perturbation to the motion. Some method of external focusing is necessary, with the most common method being the magnetic quadrupole lens.

2.1 Transverse radiofrequency defocusing

Acceleration within a linac is usually accomplished by means of radiofrequency (RF) electric fields. The details of this process will be expanded in Section 3. The RF fields are generated inside a structure generically referred to as a cavity. There are a great variety of types of cavity however the specific details are not important for analysing the effect on the transverse motion of the beam particles. For the purposes of this analysis the important aspect of an accelerating cavity is that it concentrates high-frequency electromagnetic fields in a relatively small region of space around the beam axis. This region is referred to as a RF gap or accelerating gap.

Figure 1 shows a schematic representation of a RF gap. The solid areas represent parts of the boundary of the cavity between which the electric field lines are shown. It is apparent from the shape of the electric field lines that off-axis particles will experience a radial force. The radial component of the electric field is an unavoidable consequence of concentrating the field into a relatively small longitudinal extent as can be shown by examining Gauss's law which states that in the absence of free charges the divergence of the electric field must be zero.

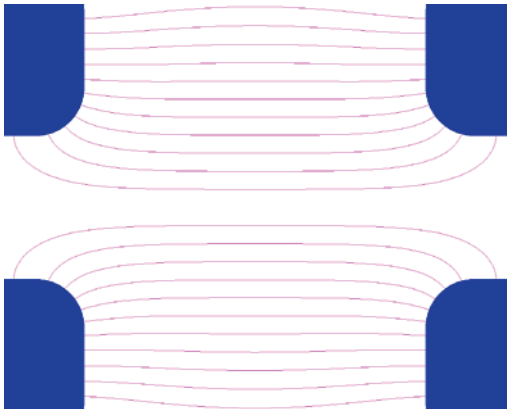


Fig. 1: Schematic representation of a generic accelerating gap

Gauss's law:

$$\nabla \cdot E = 0. \tag{1}$$

Any change in the magnitude of the longitudinal field will result in an off-axis radial field. This effect is shown in Fig. 2. The longitudinal accelerating component of the field is zero outside the gap, rising to some peak value within the gap. In the regions where the magnitude of the longitudinal field is changing a radial field results.

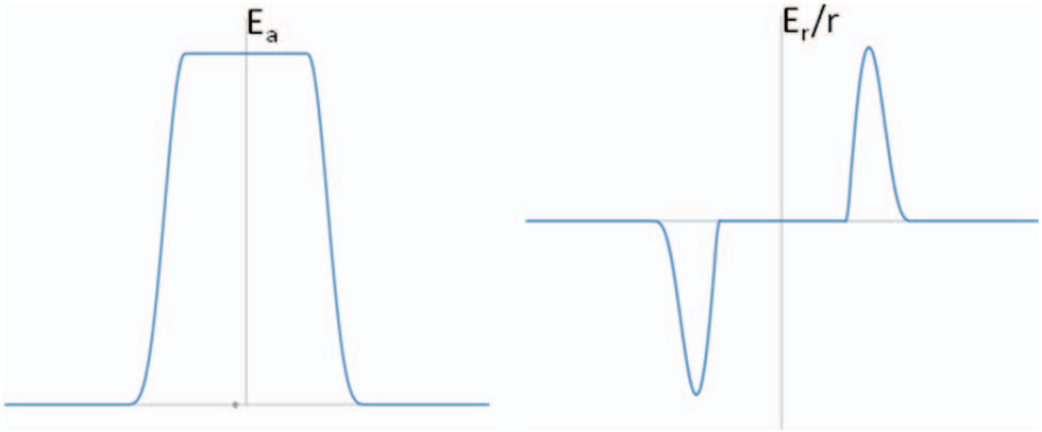


Fig. 2: Left, typical variation of the longitudinal field in a RF gap. Right, the resulting off-axis radial field component.

It might be supposed that due to the symmetrical nature of the gap the oppositely directed radial electric field in the two halves of the gap will cancel out to give no net radial impulse to the beam. However, for reasons which will be expanded in the section on longitudinal dynamics, the field is usually rising as the particles cross the gap resulting in the outwardly directed radial impulse in the second half of the gap exceeding the inwardly directed impulse in the first half. A net defocusing force is therefore experienced.

2.1.1 Radial RF impulse

A typical linac cavity is cylindrically symmetric and operated in a resonating mode that results in only three non-zero field components: E_z, E_r and B_θ where E is the electric field, B is the magnetic flux density and z, r and θ are the longitudinal, radial and azimuthal field components, respectively.

The resulting Lorentz force impulse leads to a change in the radial momentum of a particle of charge q and velocity βc in a RF gap of

$$\Delta p_r = q \int_{-L/2}^{L/2} (E_r - \beta \mathrm{c} B_\theta) \frac{\mathrm{d}z}{\beta c} \tag{2}$$

where $\pm L/2$ are the extents of the gap. If E_z is independent of r near to the axis, then the divergence and curl relationships of Maxwell's equations lead to

$$E_r = -\frac{r}{2} \frac{\partial E_z}{\partial z} \tag{3}$$

$$B_\theta = \frac{r}{2\mathrm{c}^2} \frac{\partial E_z}{\partial t} \tag{4}$$

which when substituted into Eq. (2) gives

$$\Delta p_r = -\frac{q}{2} \int_{-L/2}^{L/2} r \left(\frac{\partial E_z}{\partial z} + \frac{\beta}{c} \frac{\partial E_z}{\partial t} \right) \frac{\mathrm{d}z}{\beta c}. \tag{5}$$

By noting that

$$\frac{\mathrm{d}E_z}{\mathrm{d}z} = \frac{\partial E_z}{\partial z} + \frac{1}{\beta c} \frac{\partial E_z}{\partial t} \tag{6}$$

and replacing the longitudinal electric field E_z by

$$E_z = E_a(z) \cos(\omega t + \varphi) \tag{7}$$

where E_a is the accelerating field, ω is the RF angular frequency and φ is the phase as the particle passes the centre of the gap, the change in radial momentum in a RF gap is given by

$$\Delta p_r = -\frac{\mathrm{q}r\omega}{2\gamma^2 \beta^2 c^2} \sin\varphi \int_{-L/2}^{L/2} E_a(z) \cos(kz)\, \mathrm{d}z \tag{8}$$

where $k = 2\pi/\beta\lambda$, $\omega t = kz$.

2.1.2 Radial deflection in a RF gap

The momentum of a particle is given by

$$p_r = mc\beta\gamma r' \tag{9}$$

where m is the particle mass, $\beta\gamma$ the usual relativistic parameters and

$$r' = \frac{\mathrm{d}r}{\mathrm{d}z}. \tag{10}$$

Substituting Eq. (9) into Eq. (8) and replacing the integral $\int_{-L/2}^{L/2} E_a(z) \cos(kz)\, \mathrm{d}z$ by the effective accelerating voltage $E_0 TL$ (see Section 3) gives

$$\Delta(\beta\gamma r') = -\frac{\pi q E_0 TL r \sin(\varphi)}{mc^2 \beta^2 \gamma^2 \lambda}. \tag{11}$$

The radial deflection in a RF gap is proportional to the accelerating voltage, the radial position and the sine of the RF phase and inversely proportional to $(\beta\gamma)^2$ and the RF wavelength λ. Stable acceleration requires that $\varphi < 0$ resulting in RF defocusing.

2.2 Quadrupole focusing

In order to compensate for the transverse defocusing effect of a RF gap and any inherent divergence in the particle beam, some form of external focusing must be provided. In linacs the magnetic quadrupole lens is by far the most common. Figure 3 shows a cross-section of the poles and magnetic field lines of a quadrupole magnet.

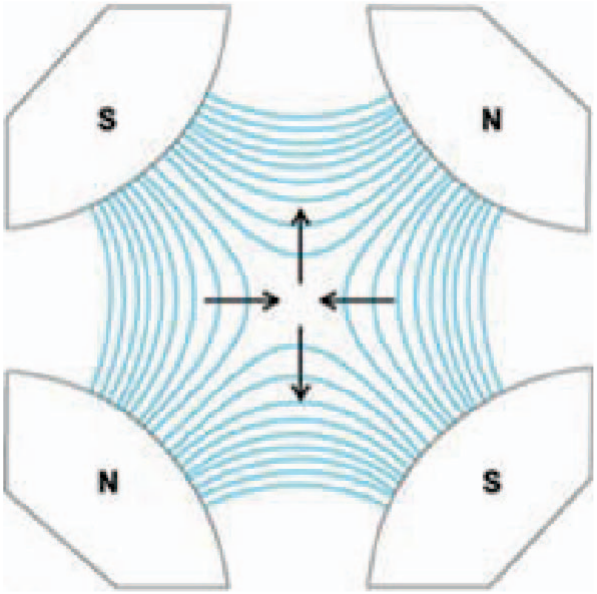


Fig. 3: Cross-section of a magnetic quadrupole showing the four poles and magnetic field lines

In an ideal quadrupole the vertical component of the magnetic field is linearly proportional to the horizontal position and the horizontal component of the magnetic field is linearly proportional to the vertical position. This results in particle forces which are linear functions of position. The quadrupole gradient G is defined as

$$G = \frac{\partial B_x}{\partial y} = \frac{\partial B_y}{\partial x} = \frac{B_0}{a} \tag{12}$$

where B_0 is the field at the pole tip and a is the distance from the pole tip to the central axis. For a particle moving in z with a velocity v the Lorentz force is

$$\begin{aligned} F_x &= -qvGx, \\ F_y &= qvGy. \end{aligned} \tag{13}$$

The effect is analogous to that of an optical lens except one which focuses in one direction but defocuses in the other. If qG is positive the lens focuses in x and defocuses in y. Despite this apparent deficiency the quadrupole lens can achieve a net focusing effect by combining magnets in sequences with both polarities.

2.2.1 Particle motion in a quadrupole

For a particle with charge q travelling with velocity βc the equations of motion in the two transverse coordinates x and y with axial position s are

$$\begin{aligned} \frac{d^2x}{\mathrm{d}s^2} + \kappa^2(s)x &= 0 \\ \frac{d^2y}{\mathrm{d}s^2} - \kappa^2(s)y &= 0 \end{aligned} \tag{14}$$

where

$$\kappa^2(s) = \frac{|qG(s)|}{m\beta\gamma c}. \tag{15}$$

In an ideal hard-edged quadrupole $G(s) = G_0$ and simple harmonic motion results. Equation (14) where the restoring force is a linear function of the transverse position is an example of Hill's equation. A convenient method to study the behaviour of Hill's equation is through matrix solutions.

2.2.2 Transfer matrix representation

Write Eq. (14) in one plane as

$$x'' + K(s)x = 0 \tag{16}$$

with

$$| K(s) | = \kappa(s)^2.$$

The solution to the linear second-order differential equation can be written in matrix form as

$$\begin{bmatrix} x_1 \\ x_1' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_0 \\ x_0' \end{bmatrix}$$
$$x' = \frac{\mathrm{d}x}{\mathrm{d}s}, \qquad x'' = \frac{d^2x}{\mathrm{d}s^2}. \tag{17}$$

The 2 × 2 matrix is called the transfer matrix $\boldsymbol{R}$:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \boldsymbol{R}. \tag{18}$$

For a sequence of elements the total transfer matrix is the product of the individual $\boldsymbol{R}$ matrices. The order of the multiplication is significant so if the beam is transported through elements 1, 2, 3, …, n in order, the total transfer matrix $\boldsymbol{R}_t = \boldsymbol{R}_n.[\ldots]\boldsymbol{R}_3.\boldsymbol{R}_2.\boldsymbol{R}_1$.

In order to study the behaviour of sequences of quadrupole magnets separated by field free drifts it is useful to derive the following transfer matrices.

2.2.2.1 Drift space

For a field free drift of length l

$$R = \begin{bmatrix} 1 & l \\ 0 & 1 \end{bmatrix}. \tag{19}$$

2.2.2.2 Focusing quadrupole

For a quadrupole of length l focusing in x

$$R = \begin{bmatrix} \cos\sqrt{K}l & \frac{\sin\sqrt{K}l}{\sqrt{K}} \\ -\sqrt{K}\sin\sqrt{K}l & \cos\sqrt{K}l \end{bmatrix}$$
$$K = \frac{qG}{mc\beta\gamma} > 0. \tag{20}$$

For sufficiently small values of $\sqrt{K}l$ the quarupole can be approximated as a thin lens where

$$R = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \tag{21}$$

$$\frac{1}{f} = \mid Kl \mid = \mid \frac{qGl}{mc\beta\gamma} \mid . \tag{22}$$

2.2.2.3 Defocusing quadrupole

For a quadrupole of length *l* defocusing in *x*

$$R = \begin{bmatrix} \cosh\sqrt{\mid K \mid}\, l & \frac{\sinh\sqrt{\mid K \mid}\, l}{\sqrt{K}} \\ \sqrt{\mid K \mid}\sinh\sqrt{\mid K \mid}\, l & \cosh\sqrt{\mid K \mid}\, l \end{bmatrix} \tag{23}$$
$$K = \frac{qG}{mc\beta\gamma} < 0.$$

The thin lens approximation is

$$R = \begin{bmatrix} 1 & 0 \\ \frac{1}{f} & 1 \end{bmatrix}, \tag{24}$$
$$\frac{1}{f} = \mid Kl \mid = \mid \frac{qGl}{mc\beta\gamma} \mid .$$

2.3 Periodic solutions to Hill's equation

It is common in linacs that the arrangement of drifts, gaps and quadrupoles is periodic. This means that *K(s)* in Eq. (16) is also periodic and the solution to Hill's equation is oscillatory. The general solution is the so-called phase-amplitude form of the solution:

$$x(s) = \sqrt{\varepsilon_1 \tilde{\beta}(s)}\cos[\tilde{\varphi}_1 + \tilde{\varphi}(s)] \tag{25}$$

where $\tilde{\beta}(s)$ is the amplitude function and $\tilde{\varphi}(s)$ is the phase function related to the amplitude function by

$$\tilde{\varphi}(s) = \int \frac{\mathrm{d}s}{\tilde{\beta}(s)}. \tag{26}$$

The constants $\tilde{\varphi}_1$ and ε_1 are given by the initial conditions. Defining two additional functions

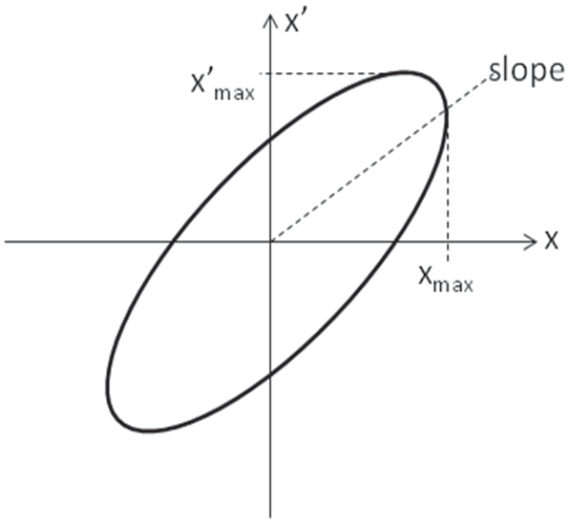
$$\tilde{\alpha}(s) = -\frac{1}{2}\frac{d\tilde{\beta}(s)}{\mathrm{d}s} \tag{27}$$

$$\tilde{\gamma}(s) = \frac{1+\tilde{\alpha}(s)^2}{\tilde{\beta}(s)} \tag{28}$$

gives the three well-known Courant–Snyder or Twiss parameters $\tilde{\alpha}(s)$, $\tilde{\beta}(s)$ and $\tilde{\gamma}(s)$ which are all periodic with the same period as *K(s)*. Making use of the Twiss parameters the transverse coordinates satisfy the equation

$$\tilde{\gamma}(s)x^2 + 2\tilde{\alpha}(s)xx' + \tilde{\beta}(s)x'^2 = \varepsilon_1 \tag{29}$$

which is the equation of an ellipse in the x–x' phase space, centred at the origin with an area equal to $A = \pi\varepsilon_1$ as shown in Fig. 4.



$$x_{\max} = \sqrt{\varepsilon\tilde{\beta}}$$
$$x'_{\max} = \sqrt{\varepsilon\tilde{\gamma}}$$
$$\text{slope} = \frac{\tilde{\alpha}}{\tilde{\beta}}$$

Fig. 4: Ellipse representing the trajectory of a particle in x–x' phase space

Equation (29) can be written in matrix form as

$$X^T \sigma^{-1} X = \varepsilon \tag{30}$$

where

$$X = \begin{bmatrix} x \\ x' \end{bmatrix}$$

and

$$\sigma^{-1} = \begin{bmatrix} \tilde{\gamma} & \tilde{\alpha} \\ \tilde{\alpha} & \tilde{\beta} \end{bmatrix}$$

and

$$\sigma = \begin{bmatrix} \tilde{\beta} & -\tilde{\alpha} \\ -\tilde{\alpha} & \tilde{\gamma} \end{bmatrix}. \tag{31}$$

Using subscript 1 to denote the beam coordinates at the beginning of a period and subscript 2 to denote the coordinates at the end, then

$$X_1^T \sigma_1^{-1} X_1 = \varepsilon$$
$$X_2^T \sigma_2^{-1} X_2 = \varepsilon$$
$$X_2 = \boldsymbol{R}\boldsymbol{X}_1$$

which leads to

$$\sigma_2 = \mathbf{R}\boldsymbol{\sigma}_1 R^T \tag{32}$$

$$\begin{bmatrix} \alpha_2 \\ \beta_2 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} R_{11}^2 & -2\mathrm{R}_{11}R_{12} & R_{12}^2 \\ -R_{11}R_{21} & 1 + R_{12}R_{21} & -R_{12}R_{22} \\ R_{21}^2 & -2\mathrm{R}_{21}R_{22} & R_{21}^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \end{bmatrix}. \tag{33}$$

For a periodic solution with the same Twiss parameters at the beginning and end of a sequence of elements

$$\sigma_2 = \sigma_1$$

$$\begin{bmatrix}\alpha_2\\ \beta_2\\ \gamma_2\end{bmatrix} = \begin{bmatrix}\alpha_1\\ \beta_1\\ \gamma_1\end{bmatrix}.$$

For such a periodic channel of length L the periodic transfer matrix $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = R(s \to s+L) = \begin{bmatrix}\cos\sigma + \tilde{\alpha}\sin\sigma & \tilde{\beta}\sin\sigma\\ -\tilde{\gamma}\sin\sigma & \cos\sigma - \tilde{\alpha}\sin\sigma\end{bmatrix} \tag{34}$$

where

$$\sigma = \Delta\tilde{\varphi} = \int^{L} \frac{\mathrm{d}s}{\tilde{\beta}(s)}$$

is the phase advance per period. Each particle in the beam lies on its own elliptical trajectory and undergoes the same phase advance σ. For stable solutions, $|\cos(\sigma)| < 1$.

By constructing the total transfer matrix or transporting two orthogonal particles $\begin{bmatrix}x_1\\ 0\end{bmatrix}$ and $\begin{bmatrix}0\\ x_1'\end{bmatrix}$ through the channel the elements of $\boldsymbol{P}$, the periodic Twiss parameters and the phase advance can be calculated:

$$\begin{bmatrix}x_2\\ x_2'\end{bmatrix} = P\begin{bmatrix}x_1\\ 0\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}x_3\\ x_3'\end{bmatrix} = P\begin{bmatrix}0\\ x_1'\end{bmatrix}$$
$$x_2 = P_{11}x_1 = (\cos\sigma + \tilde{\alpha}\sin\sigma)x_1$$
$$x_3' = P_{22}x_1' = (\cos\sigma - \tilde{\alpha}\sin\sigma)x_1'$$
$$2\cos\sigma = P_{11} + P_{22} = \frac{x_2}{x_1} + \frac{x_3'}{x_1'}. \tag{35}$$

The Twiss parameters follow directly from the phase advance from Eq. (34).

2.4 The smooth approximation

Equation (14) can be written to include the effects of RF defocusing by approximating the defocusing as a continuous force which leads to

$$\begin{aligned}\frac{\mathrm{d}^2x}{\mathrm{d}s^2} + \kappa^2(s)x - \frac{k_{lo}^2}{2}x = 0\\ \frac{\mathrm{d}^2y}{\mathrm{d}s^2} - \kappa^2(s)y - \frac{k_{lo}^2}{2}y = 0\end{aligned} \tag{36}$$

where

$$k_{lo}^2 = \frac{2\pi qE_0T\sin(-\phi)}{mc^2(\gamma\beta)^3\lambda}. \tag{37}$$

The most common focusing arrangement found in linacs is the FODO structure. Each period consists of a focusing quadrupole and a defocusing quadrupole of equal strength with accelerating gaps between them. Figure 5 shows the typical arrangement. The period length is $2L$.

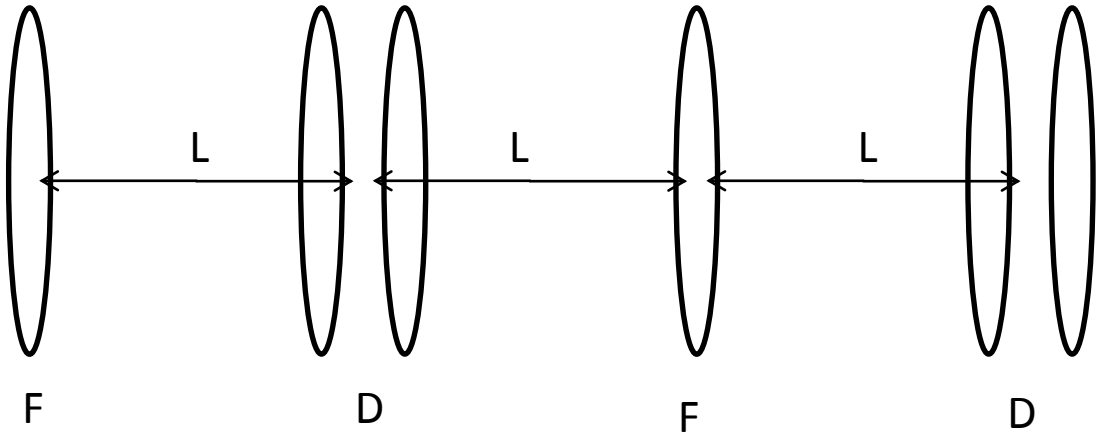


Fig. 5: One and half periods of the FODO structure

In the smooth approximation the particle trajectories become sinusoidal and the phase advance is given by

$$\sigma^2 \approx \left(\frac{L}{f_q}\right)^2 - \left(\frac{4L}{f_g}\right) \tag{38}$$

where

$$\frac{1}{f_q} = \frac{qGl}{mc\beta\gamma}$$

$$\frac{1}{f_g} = \frac{\pi q E_0 T L \sin(-\phi)}{mc^2(\beta\gamma)^3\lambda}.$$

Substituting for the focal lengths leads to

$$\sigma^2 \approx \left(\frac{qGlL}{mc\beta\gamma}\right)^2 - \frac{\pi q E_0 T \sin(-\phi)(2\mathrm{L})^2}{mc^2(\beta\gamma)^3\lambda}. \tag{39}$$

For stable behaviour the first term on the right-hand side, the quadrupole term, must be greater than the second, RF defocusing term. The $(\beta\gamma)^3$ in the RF defocus term makes it more important at low, typically not highly relativistic, velocities. The phase advance per unit length is given by

$$\left(\frac{\sigma}{2L}\right)^2 \tag{40}$$

and is independent of the period length. The period length is limited by the requirement that $\sigma < \pi$.

3 Longitudinal dynamics in linacs

A linac typically consists of a series of RF cavities designed to efficiently accelerate the particles. For the purposes of understanding the longitudinal dynamics the cavities are considered to be simple gaps of the type shown in Fig. 1. Only the on-axis component of the electric field is considered and any radial dependence is ignored.

3.1 Energy gain in a RF gap

For a generic RF gap of frequency ω and length L with an axial electric field $E_z(r, z, t)$ the field experienced by an on-axis particle is given by

$$E_z(r = 0, z, t) = E(0, z)\cos[\omega t(z)] + \phi \tag{41}$$

where

$$t(z) = \int_0^z \frac{dz}{v(z)}.$$

Taking the origin to be the centre of the gap when $t = 0$ and RF phase equal to ϕ, then the energy gain is

$$\Delta W = q \int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\cos[\omega t(z) + \phi]\,\mathrm{d}z. \tag{42}$$

Using a trigonometric identity the energy gain can be written as

$$\Delta W = q \int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)[\cos(\omega t)\cos(\phi) - \sin(\omega t)\sin(\phi)]\,\mathrm{d}z \tag{43}$$

which is expressed in the conventional form as

$$\Delta W = qE_0TL\cos\phi \tag{44}$$

where $E_0 = \frac{1}{L}\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\,\mathrm{d}z$ is the average on-axis electric field and

$$T = \frac{\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\cos\omega t(z)\,\mathrm{d}z}{\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\,\mathrm{d}z} - \tan(\phi)\frac{\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\sin\omega t(z)\,\mathrm{d}z}{\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\,\mathrm{d}z} \tag{45}$$

is the transit time factor.

3.2 Transit time factor

If $E(0,z)$ is symmetric about $z = 0$, then

$$\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\sin\omega t(z)\,\mathrm{d}z = 0$$

and

$$T = \frac{\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\cos\omega t(z)\,\mathrm{d}z}{\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\,\mathrm{d}z}. \tag{46}$$

Further, if the change in particle velocity across the gap is small

$$\omega t \approx \frac{\omega z}{\beta c} = \frac{2\pi z}{\beta\lambda}|$$

giving

$$T \approx \frac{\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\cos(2\pi z/\beta\lambda)\,\mathrm{d}z}{\int_{\frac{-L}{2}}^{\frac{L}{2}} E(0,z)\,\mathrm{d}z}. \tag{47}$$

3.3 Phase stability

In Eq. (43) the value of ϕ at which the cavity is designed to operate is called the synchronous phase. A particle arriving at the cavity with the synchronous energy and synchronous phase will also arrive at all subsequent cavities at the synchronous energies and phases. Acceleration only occurs when $\cos(\phi_s)$ is positive:

$$-\frac{\pi}{2} < \phi_s < \frac{\pi}{2}. \tag{48}$$

Phase stability occurs when the accelerating field is rising in time

$$-\pi < \phi_s < 0. \tag{49}$$

A particle that arrives earlier than the synchronous phase receives less acceleration than the synchronous particle. A later particle receives more acceleration. The effect is longitudinal focusing which drives the particles towards the synchronous phase. For stability and acceleration

$$-\frac{\pi}{2} < \phi_s < 0. \tag{50}$$

Figure 6 shows representative particles in the stable part of the RF cycle.

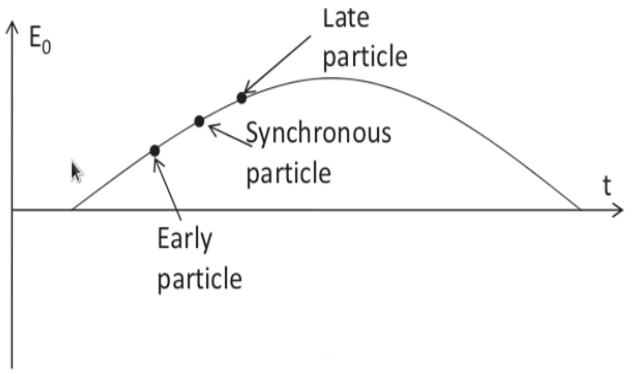


Fig. 6: Phase stable acceleration

3.4 Acceleration by a series of RF gaps

The longitudinal dynamics are studied by treating the linac as a series of thin gaps separated by field free drifts as shown in Fig. 7.

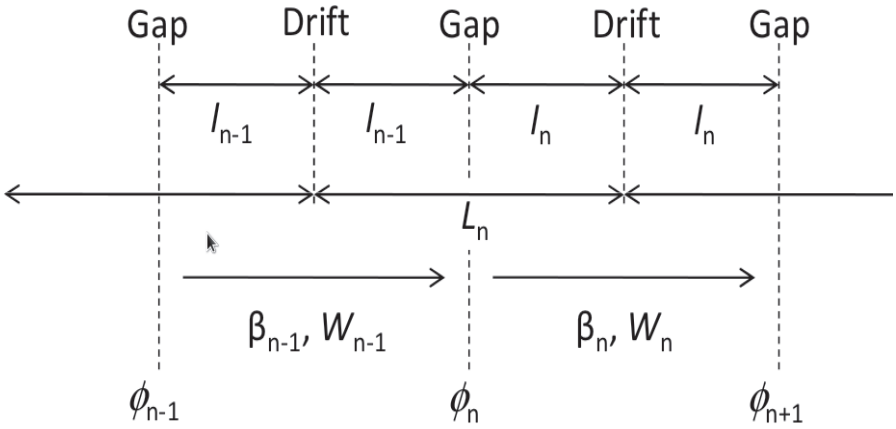


Fig. 7: The arrangement of gaps and drifts for analysing longitudinal dynamics

In Fig. 7, the parameters are

$$\phi_n = \phi_{n-1} + \frac{2\omega l_{n-1}}{\beta_{n-1}c} + \begin{Bmatrix} 0 & \text{for 0-mode} \\ \pi & \text{for } \pi\text{-mode} \end{Bmatrix}$$
$$l_{n-1} = \frac{N\beta_{s,n}\lambda}{2} \qquad \text{where} \quad \begin{Bmatrix} N = 1 & \text{for 0-mode} \\ N = \frac{1}{2} & \text{for } \pi\text{-mode} \end{Bmatrix}$$

If the synchronous velocity is given by $\beta_{s,n}$, then

$$L_n = \tfrac{N}{2}(\beta_{s,n-1} + \beta_{s,n})\lambda \tag{51}$$

and

$$\Delta(\phi - \phi_s) = 2\pi N\beta_{s,n-1}\left(\frac{1}{\beta_{n-1}} - \frac{1}{\beta_{s,n-1}}\right). \tag{52}$$

If

$$\beta - \beta_s = \tfrac{W - W_s}{mc^2\beta_s\gamma_s^3} \ll 1 \tag{53}$$

then

$$\frac{1}{\beta} - \frac{1}{\beta_s} \approx -\frac{\beta - \beta_s}{\beta_s^2} \tag{54}$$

and

$$\Delta(\phi - \phi_s)_n = -2\pi N\tfrac{W_{n-1} - W_{s,n-1}}{mc^2\beta_{s,n-1}^2\gamma_{s,n-1}^3}. \tag{55}$$

The difference in the particle energy is simply the difference in the effective voltage leading to a pair of coupled difference equations in relative energy and phase:

$$\Delta(W - W_s)_n = qE_0TL_n(\cos\phi_n - \cos\phi_{s,n}) \tag{56}$$
$$\Delta(\phi - \phi_s)_n = -2\pi N\frac{W_{n-1} - W_{s,n-1}}{mc^2\beta_{s,n-1}^2\gamma_{s,n-1}^3}.$$

3.4.1 Longitudinal equations of motion

The difference equations (55) and (56) can be converted into differential equations by noting that $s = nN\beta_s\lambda$. Then, by letting

$$\Delta(\phi - \phi_s) = \tfrac{d(\phi - \phi_s)}{dn}$$

and

$$\Delta(W - W_s) = \tfrac{d(W - W_s)}{dn}$$

leads to

$$\tfrac{d(\phi - \phi_s)}{ds} = -2\pi\tfrac{W - W_s}{mc^2\beta_s^3\gamma_s^3\lambda} \tag{57}$$

$$\frac{d(W-W_s)}{ds} = qE_0T(\cos\phi - \cos\phi_s). \tag{58}$$

Combining the two coupled differential equations (57) and (58) leads to a second-order differential equation:

$$\frac{d^2(\phi-\phi_2)}{ds^2} = -\frac{2\pi qE_0T}{mc^2\beta_s^3\gamma_s^3\lambda}(\cos\phi - \cos\phi_s). \tag{59}$$

Equation (59) leads in turn to the Hamiltonian of the longitudinal motion

$$\frac{Aw^2}{2} + B(\sin\phi - \phi\cos\phi_s) = H_\phi \tag{60}$$

where

$$A = \frac{2\pi}{\beta_s^3\gamma_s^3\lambda}$$

$$B = \frac{qE_0T}{mc^2}$$

$$w = \frac{W-W_s}{mc^2}.$$

The Hamiltonian, Eq. (60), is of the form *kinetic energy + potential energy = constant*. The potential energy term

$$V_\phi = B(\sin\phi - \phi\cos\phi_s) \tag{61}$$

indicates a potential well for $-\pi < \phi_s < 0$ as shown in Fig 8.

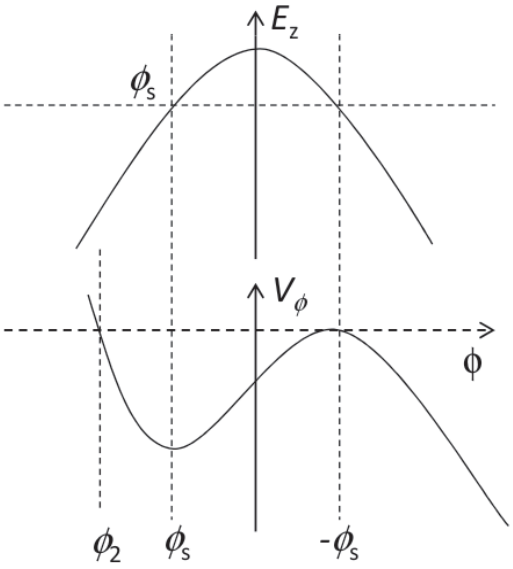


Fig. 8: The potential well for $-\pi < \phi_s < 0$

3.5 The separatrix

The separatrix is the name given to the boundary which separates longitudinal phase space into stable and unstable regions. It can be seen from Fig. 8 that the potential well creates a region of stable phase motion which covers

$$\phi_2 < \phi < -\phi_s. \tag{62}$$

The upper limit at $\phi = -\phi_s$ is a stationary point where

$$\frac{d\phi}{ds} = -Aw = 0 \tag{63}$$

which defines the Hamiltonian constant as

$$H_\phi = B(\sin(-\phi_s) - (-\phi_s\cos(\phi_s))) \tag{64}$$

and leads to the equation for the separatrix

$$\frac{Aw}{2} + B\big(\sin(\phi) - \phi\cos(\phi_s)\big) = -B(\sin\phi_s - \phi_s\cos\phi_s). \tag{65}$$

Figure 9 shows the separatrix in longitudinal phase space and its relationship to the accelerating field.

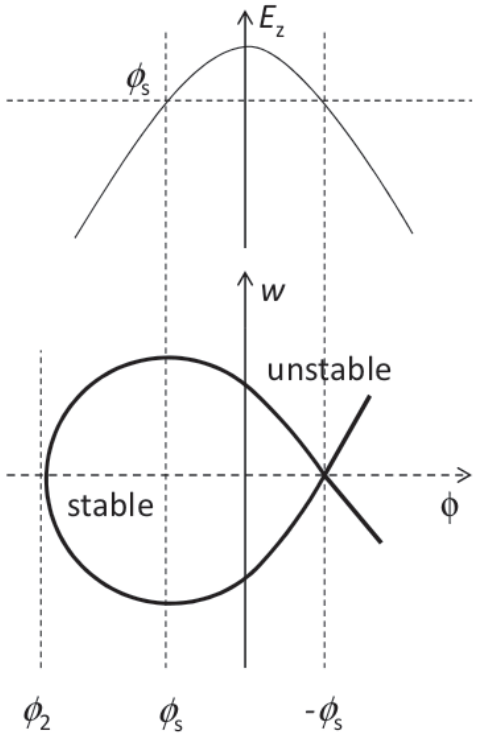


Fig. 9: The separatrix

The lower limit of the separatrix is at ϕ_2 which is given by

$$\sin\phi_2 - \phi_2\cos\phi_s = \phi_s\cos\phi_s - \sin\phi_s. \tag{66}$$

The total phase extent of the separatrix is therefore

$$\psi = |\phi_s| + |\phi_2| = -\phi_s - \phi_2. \tag{67}$$

Combining Eqs. (66) and (67) leads to the relationship

$$\tan\phi_s = \frac{\sin\psi - \psi}{1 - \cos\psi}. \tag{68}$$

The maximum energy extent of the separatrix occurs at $\phi = \phi_s$. Solving Eq. (65) at this point gives

$$w_{max} = \frac{(W-W_s)_{max}}{mc^2} = \sqrt{\frac{2qE_0T\beta_s^3\gamma_s^3\lambda}{\pi mc^2}(\phi_s\cos\phi_s - \sin\phi_s)}. \tag{69}$$

3.6 Ellipse representation

If $\phi - \phi_s$ is small compared with ϕ_s then trigonometric approximations allow the equation of phase motion, Eq. (59) can be written as

$$\frac{d^2\phi}{ds^2} + k_{0l}\left[(\phi - \phi_s) - \frac{(\phi-\phi_s)^2}{2\tan(-\phi_s)}\right] = 0 \tag{70}$$

where

$$k_{0l} = \frac{2\pi qE_0T\sin(-\phi_s)}{mc^2\beta_s^3\gamma_s^3\lambda}. \tag{71}$$

The quadratic term in Eq. (70) reduces the longitudinal focusing at large phase excursions. Similar approximations on the condition that $|\phi - \phi_s| << 1$ allows the Hamiltonian to be rewritten

$$Aw^2 + B\sin(-\phi_s)\sin(\phi - \phi_s)^2 = 2\big(H_\phi + \phi_s\cos\phi_s - \sin\phi_s\big). \tag{72}$$

For $\phi_s < 0$ this is the equation of an upright ellipse with its centre at $w = 0$ and $\phi = \phi_s$ as shown in Fig. 10.

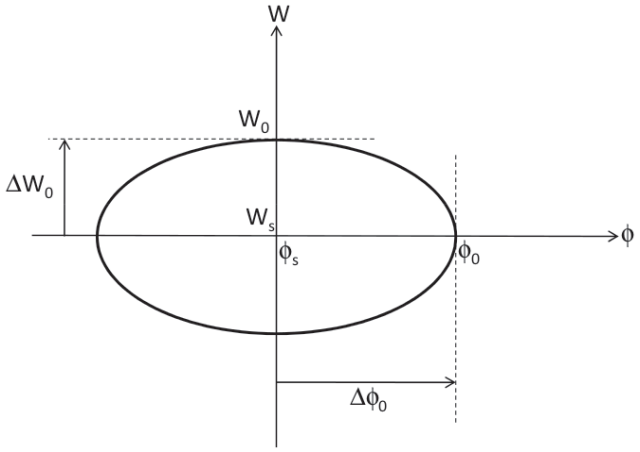


Fig. 10: Ellipse representing longitudinal motion for small phase oscillations

If

$$\phi_0 = \phi|_{w=0}$$

and

$$\Delta\phi_0 = \phi_0 - \phi_s$$

then

$$\frac{w^2}{w_0^2} + \frac{(\phi-\phi_s)^2}{\Delta\phi_0^2} = 1 \tag{73}$$

where

$$w_0 = w|_{\phi=\phi_s} = \sqrt{\frac{qE_0 T\beta_s^3\gamma_s^3\lambda\Delta{\phi_0}^2\sin(-\phi_s)}{2\pi mc^2}}. \tag{74}$$

The area of the ellipse is $\pi\varepsilon$ where ε is the longitudinal emittance

$$\Delta W_0 = W_0 - W_s$$

$$\varepsilon = \Delta\phi_0\Delta W_0 = \Delta\phi_0^2\sqrt{\frac{qE_0 Tmc^2\beta_s^3\gamma_s^3\lambda\sin(-\phi_s)}{2\pi}}. \tag{75}$$

Cyclotrons for high-intensity beams

Mike Seidel
Paul Scherrer Institute, Villigen, Switzerland

Abstract
This paper reviews the important physical and technological aspects of cyclotrons for the acceleration of high-intensity beams. Special emphasis is given to the discussion of beam loss mechanisms and extraction schemes.

1 Introduction

Cyclotrons have a long history in accelerator physics and are used for a wide range of medical, industrial, and research applications [1]. The first cyclotrons were designed and built by Lawrence and Livingston [2] back in 1931. The cyclotron represents a resonant-accelerator concept with several properties that make it well suited for the acceleration of hadron beams with high average intensity. In this paper, we concentrate on aspects of the high-intensity operation of cyclotrons. The electronic version of this document contains, in the references section, clickable links to many publications related to this theme.

2 The classical cyclotron

Although the classical cyclotron has major limitations and is practically outdated today, some fundamental relations are best explained within this original concept. In the classical cyclotron, an alternating high voltage at radio frequency (RF) is applied to two D-shaped hollow electrodes, the *dees*, for the purpose of acceleration. Ions from a central ion source are repeatedly accelerated from one dee to the other. The ions are kept on a piecewise circular path by the application of a uniform, vertically oriented magnetic field. On the last turn, the ions are extracted by applying an electrostatic field using an electrode. The concept is illustrated in Fig. 1.

The revolution frequency of the particle motion, called the *cyclotron frequency*, depends on the magnetic field B_z, and the charge q and the effective mass γm_0 of the particles:

$$\begin{aligned} f_\mathrm{c} = \frac{\omega_\mathrm{c}}{2\pi} &= \frac{qB_z}{2\pi\gamma m_0} \\ &\approx 15.2\ \mathrm{MHz} \cdot B(\mathrm{T})\ (\,\text{for protons}\,). \end{aligned} \tag{1}$$

The frequency of the accelerating voltage must be equal to the cyclotron frequency or an integer multiple of it, i.e., $\omega_\mathrm{RF} = h\omega_\mathrm{c}$. The harmonic number h equals the number of bunches that can be accelerated in one turn. With increasing velocity, particles travel at larger radii, so that $R \propto \beta$, and the revolution time remains constant and in phase with the RF voltage. The bending strength is given by $B_z R \propto p \propto \beta\gamma$. Thus, as long as $\gamma \approx 1$, the condition of *isochronicity* is fulfilled in a homogeneous magnetic field. However, for relativistic particles, the magnitude of the B-field has to be raised in proportion to γ at increasing radii, in order to keep the revolution time constant throughout the acceleration process. In summary, the condition of isochonicity in a cyclotron requires the following scaling of the orbit radius and bending field:

$$R \propto \beta, \quad B_z \propto \gamma\,. \tag{2}$$

This original cyclotron concept exhibits some essential properties that allow the high-intensity application of cyclotrons, which is the focus of this article. The acceleration process takes place continuously, and neither the RF frequency nor the magnetic bending field has to be cycled. The separation of subsequent turns allows continuous extraction of the beam from the cyclotron. Consequently, the production of a continuous-wave (CW) beam is a natural feature of cyclotrons. The so-called K-value

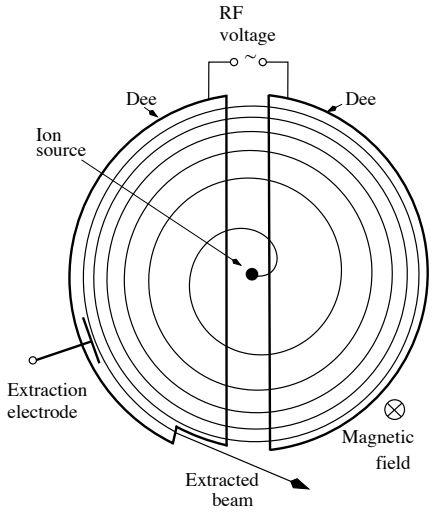


Fig. 1: Conceptual sketch of a classical cyclotron in plan view. In the non-relativistic approximation, the turn separation scales with the number of turns as $n_{\mathrm{t}}^{-1/2}$.

is a commonly used parameter for the characterization of the magnetic energy reach of a cyclotron design. This equals the maximum attainable energy for protons in the non-relativistic approximation. The K-value is proportional to the maximum squared bending strength, i.e., $K \propto (B\rho)^2$, and can be used to rescale the achievable kinetic energy per nucleon for varying charge-to-mass ratio:

$$\frac{E_{\mathrm{k}}}{A} = K \left(\frac{Q}{A}\right)^2. \tag{3}$$

The radial variation of the bending field in a classical cyclotron generates focusing forces. At a radius R, the slope of the bending field is described by the field index k, where

$$k = \frac{R}{B_z}\frac{\mathrm{d}B_z(R)}{\mathrm{d}R}. \tag{4}$$

Using Eq. (2), the scaling of the field index under isochronous conditions can be evaluated as follows:

$$\begin{aligned} \frac{R}{B}\frac{\mathrm{d}B}{\mathrm{d}R} &= \frac{\beta}{\gamma}\frac{\mathrm{d}\gamma}{\mathrm{d}\beta} \\ &= \gamma^2 - 1. \end{aligned} \tag{5}$$

The radial equation of motion of a single particle can be written as

$$m\ddot{r} = mr\dot{\varphi}^2 - qr\dot{\varphi}B_z. \tag{6}$$

We now consider small deviations around the central orbit R, namely $r = R + x, x \ll R$:

$$\begin{aligned} \ddot{x} + \frac{q}{m}vB_z(R+x) - \frac{v^2}{R+x} &= 0, \\ \ddot{x} + \frac{q}{m}v\left(B_z(R) + \frac{\mathrm{d}B_z}{\mathrm{d}R}x\right) - \frac{v^2}{R}\left(1 - \frac{x}{R}\right) &= 0, \\ \ddot{x} + \omega_{\mathrm{c}}^2(1+k)x &= 0. \end{aligned} \tag{7}$$

In this derivation, we have used the relations $\omega_{\mathrm{c}} = qB_z/m \approx v/R$ and $r\dot{\varphi} \approx v$. Thus, in the linear approximation, the horizontal 'betatron motion' is a harmonic oscillation around the central beam orbit,

$x(t) = x_{\max}\cos(\nu_r\omega_c t)$. The parameter ν_r is called the betatron tune. From Eq. (7), we see that the radial betatron frequency in a classical cyclotron is given by

$$\begin{aligned}\nu_r &= \sqrt{1+k} \\ &\approx \gamma. \end{aligned} \tag{8}$$

In the above, Eq. (5) has been used to derive the relation for γ. A similar calculation can be done for the vertical plane, using Maxwell's equation $\mathrm{rot}\,\vec{B} = 0$. This yields the following for the vertical betatron frequency:

$$\nu_z = \sqrt{-k}. \tag{9}$$

Beta functions can be defined in the sense of the Courant–Snyder theory [3] for a cyclotron. In the radial plane, the average beta function can be estimated via

$$\beta_r \approx \frac{R}{\nu_r} \approx \frac{R}{\gamma}. \tag{10}$$

A radial dispersion function can be defined as well:

$$\Delta R = D_r\frac{\Delta p}{p}, \quad D_r \approx \frac{R}{\gamma^2}. \tag{11}$$

The derivation of this relation is done in a similar way to the calculation of the radial step width in Eq. (15) in the next section. The two relations above can be used to establish rough matching conditions for beam injection into cyclotrons. As is obvious from Eq. (9), vertical focusing can be obtained only if the bending field decreases towards larger radii. However, a negative slope of the field would be inconsistent with the isochronicity condition stated above, which requires the field to increase in proportion to γ. Thus the classical cyclotron is limited to relatively low energies. As we will see in the next section, vertical focusing can in fact be achieved by an azimuthal variation of the bending field.

3 AVF and separated-sector cyclotrons

One way to overcome the problem of insufficient vertical focusing is the introduction of azimuthally varying fields, as is done in the Thomas, or AVF, cyclotron. The principle was proposed in 1938 by L.H. Thomas [4], but it took several decades before a cyclotron based on this principle was actually built (in Delft in 1958). The variation of the vertical bending field along the flight path leads to transverse forces on the particles that can be utilized to provide suitable focusing characteristics in both of the transverse planes. In a Thomas cyclotron, the average field strength can be increased as a function of radius without losing the vertical stability. As a result, this concept allows higher energies to be achieved, for example 1 GeV for protons. In fixed-focus alternating-gradient (FFAG) rings, the same type of edge focusing plays a role as well; however, in most FFAG designs, the focusing effect of the alternating gradients is dominant. The required field variation in the cyclotron can be achieved by special shaping of the poles of a compact single magnet. The focusing can be increased by the introduction of spiral sector shapes, and sector boundaries that have tilt angles with respect to the beam orbit. With such magnet configurations, the squared vertical betatron frequency is approximately

$$\nu_z^2 \approx -k + F^2(1+2\tan^2\delta), \quad F^2 = \frac{\overline{B_z^2} - \overline{B_z}^2}{\overline{B_z}^2}. \tag{12}$$

The so-called flutter factor F equals the relative root mean square (r.m.s.) variation of the bending field around the circumference of the cyclotron. The spiral angle δ is defined as shown in Fig. 2.

The next and most recent step in the history of cyclotron development was the introduction of separated-sector cyclotrons. Such cyclotrons have a modular structure consisting of several sector-shaped

dipole magnets and RF resonators for acceleration. The modular concept makes it possible to construct larger cyclotrons that can accommodate the bending radii of ions at higher energies.

For completeness, the synchrocyclotron, which represents another way to overcome the relativistic limit, should also be mentioned here. In the synchrocyclotron, the RF frequency is varied according to the variation in the speed of the accelerated particles. This means that pulses, i.e., trains of bunches of ions, are accelerated, which results in a drastic reduction in the average beam current that can be achieved. Historically, the synchrocyclotron was an important step in pushing the energy frontier, but this concept was superseded by the invention of the synchrotron. The sector-focused cyclotron is limited in energy, but it has retained its attractiveness for high-intensity applications owing to its advantage of CW operation. A comprehensive overview article of cyclotron concepts can be found in [1].

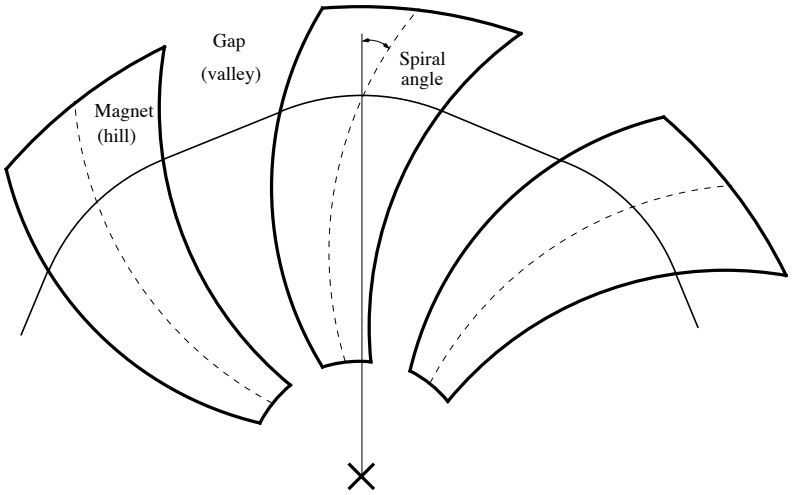


Fig. 2: Spiral magnet sectors and definition of the average spiral angle

For clean extraction with an extraction septum, the distance between the turns at the extraction radius must be maximized. It is therefore essential in the design of a high-intensity cyclotron to consider the transverse separation of the beam. To calculate the step width per turn, we start from the formula for the magnetic rigidity,

$$BR = \frac{p}{e} = \sqrt{\gamma^2 - 1}\,\frac{m_0 c}{e}. \tag{13}$$

By computing the total logarithmic differential, we obtain a relation between the changes in the radius, magnetic field, and energy of the particle beam:

$$\frac{\mathrm{d}B}{B} + \frac{\mathrm{d}R}{R} = \frac{\gamma\,\mathrm{d}\gamma}{\gamma^2 - 1}. \tag{14}$$

Using the field index k given in Eq. (4), we obtain

$$1 + k = \frac{\gamma R}{\gamma^2 - 1}\frac{\mathrm{d}\gamma}{\mathrm{d}R}.$$

Noting that the change in the relativistic quantity γ per turn is $\mathrm{d}\gamma/\mathrm{d}n_\mathrm{t} = U_\mathrm{t}/(m_0c^2)$, where U_t denotes the energy gain per turn, we finally obtain the step width in the radius,

$$\begin{aligned}
\frac{\mathrm{d}R}{\mathrm{d}n_\mathrm{t}} &= \frac{\mathrm{d}\gamma}{\mathrm{d}n_\mathrm{t}}\frac{\mathrm{d}R}{\mathrm{d}\gamma} \\
&= \frac{U_\mathrm{t}}{m_0c^2}\frac{\gamma R}{(\gamma^2-1)(1+k)} && (15) \\
&= \frac{U_\mathrm{t}}{m_0c^2}\frac{\gamma R}{(\gamma^2-1)\nu_r^2}. && (16)
\end{aligned}$$

In the outer region of the cyclotron, near the extraction radius, it is possible to violate the condition of isochronicity for a few turns. By reducing the slope of the field strength, which is related to the radial tune, it is possible to increase the turn separation locally. In the fringe field region of the magnets, the field decreases naturally. By going from Eq. (15) to Eq. (16) using Eq. (8), we can show the relation between the step width and the radial tune. If the condition of isochronicity remains valid, the dependence on the field index and the radial tune can be eliminated, and the step width is given by

$$\frac{\mathrm{d}R}{\mathrm{d}n_\mathrm{t}} = \frac{U_\mathrm{t}}{m_0 c^2} \frac{R}{(\gamma^2 - 1)\gamma}. \tag{17}$$

In this form, the equation shows the strong dependence of the step width on the beam energy. Above 1 GeV, it becomes very difficult to achieve clean extraction with an extraction septum. An effective way to increase the turn separation at the extraction element is the introduction of orbit oscillations by deliberately injecting the beam slightly off centre. When the phase and amplitude of the orbit oscillation are chosen appropriately, and also the behaviour of the radial tune is controlled in a suitable way, the beam separation can be increased by a factor of three. According to Eq. (15), this gain is equivalent to a cyclotron three times larger and is thus significant. Figure 3 illustrates how this scheme is used in the PSI Ring cyclotron. In [5], the beam profile in the outer turns was computed numerically for realistic conditions, and the results are in good agreement with measurements.

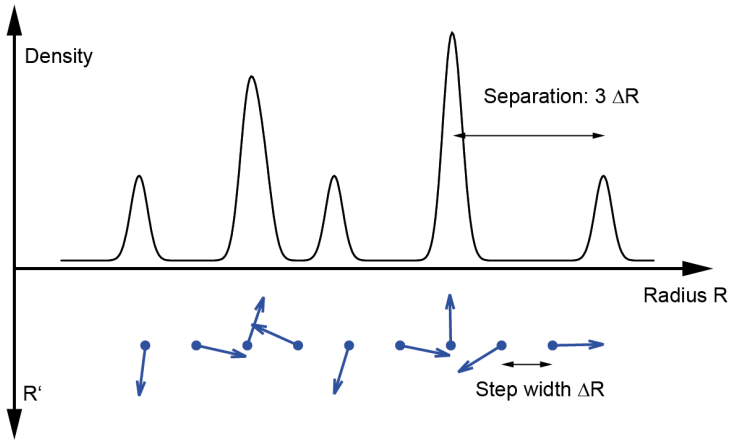


Fig. 3: Betatron oscillations of the centre of the beam around a closed orbit can be utilized to maximize the beam separation at extraction. The upper plot shows the calculated beam density, which is a superposition of Gaussian profiles. In the lower half, the clockwise-rotating phase space vector of the centroid of the beam is shown for each turn.

In summary, the clean extraction of the beam is of the utmost importance for high-intensity cyclotron operation. The turn separation at the extraction element can be maximized by the following measures.

- The extraction radius should be large, i.e., the overall dimensions of the cyclotron should not be chosen to be too small.
- The energy gain per turn should be maximized by installing a sufficient number of resonators with high performance.
- At relativistic energies, the turn separation diminishes quickly, and thus the final energy should be kept below approximately 1 GeV.
- In the extraction region, the turn separation can be increased by lowering the slope of the field index and by utilizing orbit oscillations resulting from controlled off-centre injection.

An alternative to the extraction method described here is extraction via charge exchange. More details of this method are given in Section 6.

4 Design aspects of separated-sector cyclotrons

Modern cyclotrons that are able to reach higher K-values, particularly those designed for high intensity, are typically realized as separated-sector cyclotrons. They employ a modular concept involving a combination of sector-shaped magnets, RF resonators, and empty sector gaps to form a closed circular accelerator. The modular concept simplifies the construction of cyclotrons with diameters significantly larger than those achieved with the classical single-magnet concept. The large orbit radius at maximum energy permits extraction with extremely low losses. The modularity also has significant advantages concerning the serviceability of the accelerator, especially in view of the need to handle activated components.

During the course of acceleration, the revolution time is kept constant, which leads to a significant variation in the average orbit radius. The lateral width of the elements in the ring is large in comparison with, for example, the elements of a synchrotron that uses strong focusing. The mechanical design of the vacuum chambers and sealed interconnections is thus challenging. On the other hand, the large variation in the radius makes it possible to separate the turns at the outer radius and to realize an extraction scheme for CW operation with very low losses. The close orbit spacing in FFAG rings, for example, makes continuous extraction difficult. The extraction loss is the limiting effect for high-intensity operation of cyclotrons.

The wide vacuum chambers (2.5 m for the PSI Ring cyclotron) require special sealing techniques. In a cyclotron, as in a single-pass accelerator, vacuum levels of 10^{-6} mbar are sufficient for the acceleration of protons. So-called inflatable seals are manufactured from thin steel sheets with two sealing surfaces per side and an intermittently evacuated volume between. To simplify installation, these seals are positioned on radial rails between two elements. Inflation with pressurized air seals the surfaces. This screwless scheme can tolerate small positioning errors and has the advantage of short mounting times.

The concept of the separated-sector cyclotron requires external injection of a beam of good quality. Both injection and extraction are often performed using an electrostatic deflection channel. In both cases the beam is deflected at a certain radius, while the neighbouring turns must not be affected. This is achieved by placing a thin electrode between the two turns. Particles in the beam tails that hit this electrode are scattered, and these generate losses and activation. A magnetic element would need much more material to be placed between the turns. A simplified view of the PSI Ring cyclotron is given in Fig. 4.

Some parameters of large cyclotrons operating today are listed in Table 1. The TRIUMF cyclotron [6] accelerates H^- ions and allows the extraction radius to be varied to adjust the final beam energy. The RIKEN Ring cyclotron [7] is not a high-intensity machine, but it allows a broad variety of ions to be accelerated. The special feature of the RIKEN cyclotron is the superconducting sector magnets, which deliver a very high bending strength, reflected by the corresponding K-value. The PSI Ring cyclotron was proposed in the 1960s by Willax [8]. It is specialized for high-intensity operation at the expense of reduced flexibility.

5 Space charge effects in cyclotrons

In high-intensity cyclotrons, space charge effects are of major importance in determining the maximum attainable intensity. In principle, the CW operation of cyclotrons results in low bunch charges, leading to moderate space charge effects in comparison with pulsed-accelerator concepts. On the other hand, the focusing forces are rather weak. In the transverse planes, space charge forces cause shifts in the focusing frequencies, and for large tune shifts this results in resonant losses. Strong defocusing space charge forces

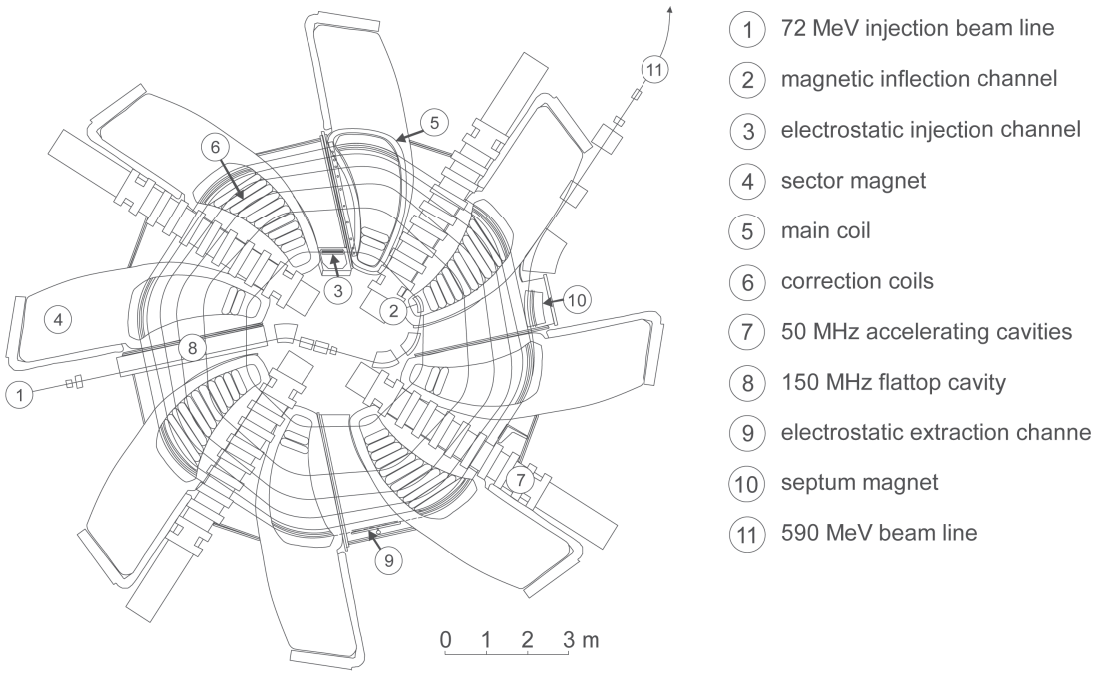


Fig. 4: Top view of the PSI Ring cyclotron. This separated-sector cyclotron contains eight sector magnets, four accelerating resonators (50 MHz), and one flat-top resonator (150 MHz).

Table 1: Selected parameters of large sector cyclotrons. The TRIUMF cyclotron uses a single magnet with sector poles. The maximum beam power of the RIKEN cyclotron was achieved in 2011, and may change with other ion species or operation modes.

Cyclotron	K (MeV)	N_{mag}	Harmonic number	R_{inj} (m)	R_{extr} (m)	Extraction method	Overall transmission	P_{max} (W)
TRIUMF	520	6 (sectors)	5	0.25	3.8–7.9	H^- stripping channel	0.70	110
PSI Ring	592	8	6	2.1	4.5	Electrostatic channel	0.9998	1400
RIKEN Ring	2600	6	6	3.6	5.4	Electrostatic channel	0.63	6.2 $\left({}^{18}\mathrm{O}\right)$

may even exceed the focusing forces generated by the cyclotron magnets. Longitudinal space charge forces lead to an increased energy spread, which is transformed into transverse beam tails. In cyclotrons with a deflecting extraction element, these beam tails are scattered at this element and limit the intensity. Analytical prediction of the beam dynamics under the influence of space charge forces is difficult, since the beams in different turns overlap, and therefore forces from neighbouring bunches cannot be neglected in general. With particle-tracking codes, the self-induced fields of a bunch under consideration and neighbouring bunches can be included in detailed predictions of the beam dynamics [9]. However, in order to gain some insight into the fundamental dynamics, it is sufficient to consider a simplified model of uniformly charged beam sectors resulting from completely overlapping turns.

For very short transversely separated bunches, the strong repelling space charge force results in a rapid motion of the particles around the centre of the bunch on cycloidal paths. In this specific regime, a compact, stable circular bunch shape is developed, despite the presence of a repelling central force within the bunch. In this case, complete coupling between the longitudinal and radial degrees of freedom is observed. Without going into too much detail, we will briefly summarize the transverse and longitudinal effects here, as well as the formation of a round beam for short bunches.

5.1 Transverse space charge forces

In the case of a cyclotron with overlapping turns, a current sheet model, assuming flat rotating sectors of charge, can be applied. The vertical force on a test particle at a distance y from the beam centre is

$$F_y = \frac{n_v e^2}{\epsilon_0 \gamma^2} \cdot y. \tag{18}$$

Here, n_v is the particle density in the centre of the bunch, given by

$$n_v = \frac{N}{(2\pi)^{3/2} \sigma_y D_\mathrm{f} R \, \Delta R}, \tag{19}$$

where D_f is the fraction of the circumference covered by the beam, i.e., the ratio of the average current to the peak current; σ_y is the vertical r.m.s. beam size; ΔR is the step width between the turns, which was discussed in Section 3; and N is the number of particles per turn, contained in h bunches. Assuming a Gaussian longitudinal distribution with an r.m.s. bunch length σ_z, we have

$$D_\mathrm{f} = \frac{h \sigma_z}{(2\pi)^{1/2} R}.$$

The focusing force generated by the magnet structure can be expressed as

$$F_y = -\gamma m_0 \omega_\mathrm{c}^2 \nu_{y0}^2 \cdot y. \tag{20}$$

Thus the resulting vertical equation of motion for a test particle can be written as

$$\ddot{y} + \left(\omega_\mathrm{c}^2 \nu_{y0}^2 - \frac{n e^2}{\epsilon_0 m_0 \gamma^3} \right) y = 0. \tag{21}$$

Obviously, an intensity limit is reached when the focusing term in the brackets vanishes. This condition was used by Blosser [10] to formulate a space charge limit for cyclotrons. In practice, an operating limit is reached somewhat earlier for a tune shift of approximately 0.4. The effective vertical tune can be deduced from Eq. (21) as follows:

$$\nu_y \quad = \quad \left(\nu_{y0} - \frac{4\pi c^2 r_\mathrm{p} n_\mathrm{v}}{\omega_\mathrm{c}^2 \gamma^3} \right)^{1/2}$$

$$\approx \quad \nu_{y0} - \frac{2\pi c^2 r_\mathrm{p} n_\mathrm{v}}{\omega_\mathrm{c}^2 \gamma^3 \nu_{y0}}. \tag{22}$$

Here, the classical proton radius $r_\mathrm{p} = 1.5 \times 10^{-18}$ m has been introduced. After some algebra, and by replacing the step width ΔR using Eq. (17), we finally obtain the following for the tune shift:

$$\Delta\nu_y = -\sqrt{2\pi}\frac{r_\mathrm{p} R}{e\beta c \nu_{y0}\sigma_z}\frac{m_0 c^2}{U_\mathrm{t}}\, I_\mathrm{avg}. \tag{23}$$

For typical high-intensity cyclotrons, this formula predicts a space charge limit at currents of several tens of milliamps.

5.2 Longitudinal space charge forces

Beam losses caused by longitudinal effects are important even at the milliamp level. The effect of longitudinal space charge forces can also be estimated in a sector model. Because of the shielding effect of the vacuum chamber, only charges within a radius approximately equal to the chamber height contribute to the force. Joho [11] estimated a quadratic dependence of the accumulated energy spread on the turn number n_t:

$$\begin{aligned} \Delta E_\mathrm{sc} &= \frac{16}{3}\frac{e g_\mathrm{1c} Z_0}{\beta_\mathrm{max}}\frac{I_\mathrm{avg}}{D_\mathrm{f}} n_\mathrm{t}^2 \\ &\approx 2800(\Omega)\frac{e I_\mathrm{avg} n_\mathrm{t}^2}{D_\mathrm{f}\beta_\mathrm{max}}. \end{aligned} \tag{24}$$

Here, $Z_0 = 377\ \Omega$ is the impedance of free space and $g_\mathrm{1c} \approx 1.4$ is a form factor. The calculation assumes non-relativistic conditions. In Ref. [12], the results of numerical simulations were compared with this simple analytical calculation and the agreement was satisfactory.

The energy spread generated by longitudinal space charge forces is transformed into transverse beam tails. Losses occur at the electrode of the extraction element owing to residual beam density between the orbits of the last two turns. The separation between these turns is proportional to n_t^{-1}. Consequently, under the constraint of constant losses, the maximum attainable beam current scales in inverse proportion to the third power of the turn number. Over the history of the PSI cyclotron accelerator, the beam current has been increased by a large factor. This has been achieved mainly by applying higher gap voltages in the resonators, thus reducing the number of turns. More powerful RF amplifiers and new resonators have been installed. In fact, the maximum beam current scaled according to the above third-power law [14].

5.3 Circular-bunch regime

For very short bunches, a self-focusing effect can be observed in a cyclotron, which leads to the formation of a circular stable bunch shape in the radial–longitudinal plane. Owing to the combination of strong space charge forces and the bending dipole field, the particles start to rotate rapidly around the bunch centre in cycloidal paths. Analytical descriptions of this effect were given by Chasman and Baltz [15] for the case of a potential due to a point-like central charge and by Bertrand and Ricaud [16] for the case of a potential due to a constant charge density. The bunches in the individual turns must be separated for most of the acceleration time in order to enter the circular-shape regime. This behaviour obviously does not occur in the sector model described earlier, involving overlapping turns. A circular bunch shape is observed in practice in the Injector II cyclotron at the PSI. Figure 5 shows the measured circular particle distribution at the exit of Injector II and, for comparison, the distribution measured roughly 20 m downstream. Over the relatively short drift length, the bunch shows a significant longitudinal increase, whereas over more than 200 m travel distance in the cyclotron it stays in the compact form shown. For a high-intensity cyclotron, it is of course desirable to enter the circular-bunch regime, since the longitudinal beam blow-up can be drastically reduced.

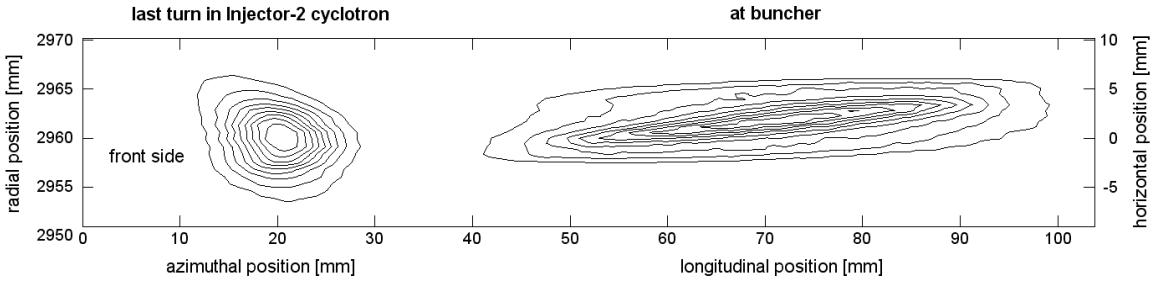


Fig. 5: Particle density for a current of 2.2 mA measured at the exit of the PSI Injector II cyclotron (72 MeV) and after a drift length of $\approx$ 20 m. The profile was measured by detecting protons scattered from a vertically oriented wire probe placed at varying horizontal positions. The distribution of such events was recorded as a function of time [13].

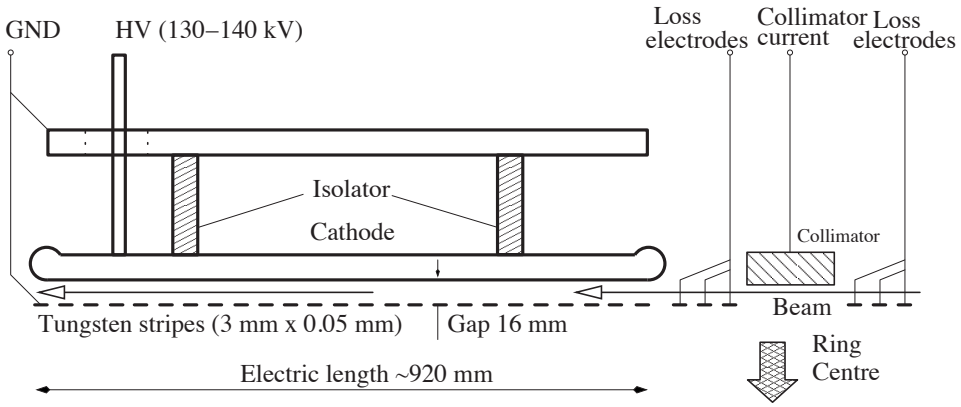


Fig. 6: Electrostatic extraction channel of the PSI Ring cyclotron

6 Injection and extraction

Extraction of the beam at high energy is one of the critical aspects of high-intensity cyclotrons. Two schemes are used for extraction. In the more classical scheme, the electrode of an electrostatic deflector is placed between the last and the second last turn in the cyclotron. The beam receives a kick angle of the order of 10 mrad, which is enough to separate the orbits to a distance that allows the insertion of a septum magnet. Although a thin electrode is typically used, some tail particles of the beam hit the electrode. The scattered particles may end up in the vacuum chamber in the extraction beam line. As described before, for this extraction scheme to work it is important to generate a large turn separation, resulting in a low beam density at the location of the electrode. A schematic drawing of an electrostatic deflector is shown in Fig. 6. The deflection angle can be calculated via the electrical rigidity,

$$\begin{aligned} E\rho &= \frac{\gamma+1}{\gamma}\frac{E_\mathrm{k}}{q} \\ &\approx 2U_\mathrm{gap} \ \ (\text{for } E_\mathrm{k} \ll E_0). \end{aligned} \tag{25}$$

In the low-energy approximation, U_gap denotes the gap voltage of the electrostatic deflector.

The other extraction scheme utilizes a stripping foil at the extraction point. Accelerated ions change their charge state when they pass through the foil. Owing to the change in the curvature of the orbit, the beam can then be extracted from the cyclotron bending field. This scheme can be applied to H^- or H_2^+ ions in order to produce a proton beam. However, the second electron of the H^- ion is weakly bound, and thus there exists a significant probability that the electron will be detached from the ion in a strong magnetic field. This effect generates unwanted losses from the beam. H^- ions are accelerated in the TRIUMF cyclotron; this provides versatility in the form of options to extract the beam at different

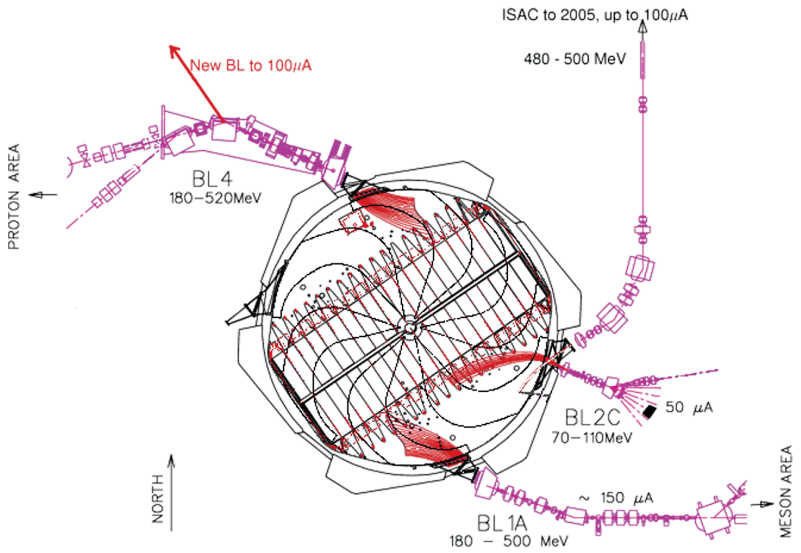


Fig. 7: Top view of the TRIUMF cyclotron, showing multiple extraction paths of stripped protons (drawing provided by R. Baartmann, TRIUMF, 2011)

energies, and even at several extraction points in parallel (Fig. 7). To limit the losses from unwanted ionization, a moderate average field of 0.46 T was chosen for this cyclotron. The H_2^+ ion has a stronger binding energy. However, the bending field of the cyclotron must be stronger because of the smaller charge-to-mass ratio of $1/2$ [17].

In summary, charge stripping represents an elegant method of extracting beams from a cyclotron. However, the following effects can potentially limit the efficiency of the stripping method and must be investigated.

- Scattering from residual gas molecules, and strong magnetic fields can cause dissociation and thus loss of the accelerated ions.
- The stripping process can also generate particles with unwanted charge-to-mass ratios, for example neutrals, and the transport and the locations of loss of these particles must be considered.
- The lifetime of the stripping foil is typically problematic; the foil is heated by power deposition from the beam and also the stripped electrons, which are bound in the presence of strong magnetic fields.

7 Magnets

As stated previously, the condition of isochronicity in a cyclotron requires the average magnetic field to be increased in proportion to γ as the radius increases. The cyclotron magnets have to be designed in such a way as to fulfil this requirement. The increase in average field can be achieved by introducing a slight vertical opening angle between the magnet poles. For fine-tuning of the isochronicity, cyclotron magnets are typically equipped with several correction coil circuits. Because of the variation in the radius of the beam, which is typically significant, cyclotron magnets have to cover a wide radial range. The mechanical design becomes large and heavy. In the case of the PSI Ring cyclotron, each magnet has a weight of 280 t. A field map and a photograph of the sector magnets of the PSI Ring cyclotron are shown as an example in Fig. 8. Besides the purpose of bending the beam orbit, the magnets must provide sufficient focusing in both planes. The magnets often have a spiral shape for this purpose (see Section 3). Most sector magnets for cyclotrons use normal-conducting coils, although some medical

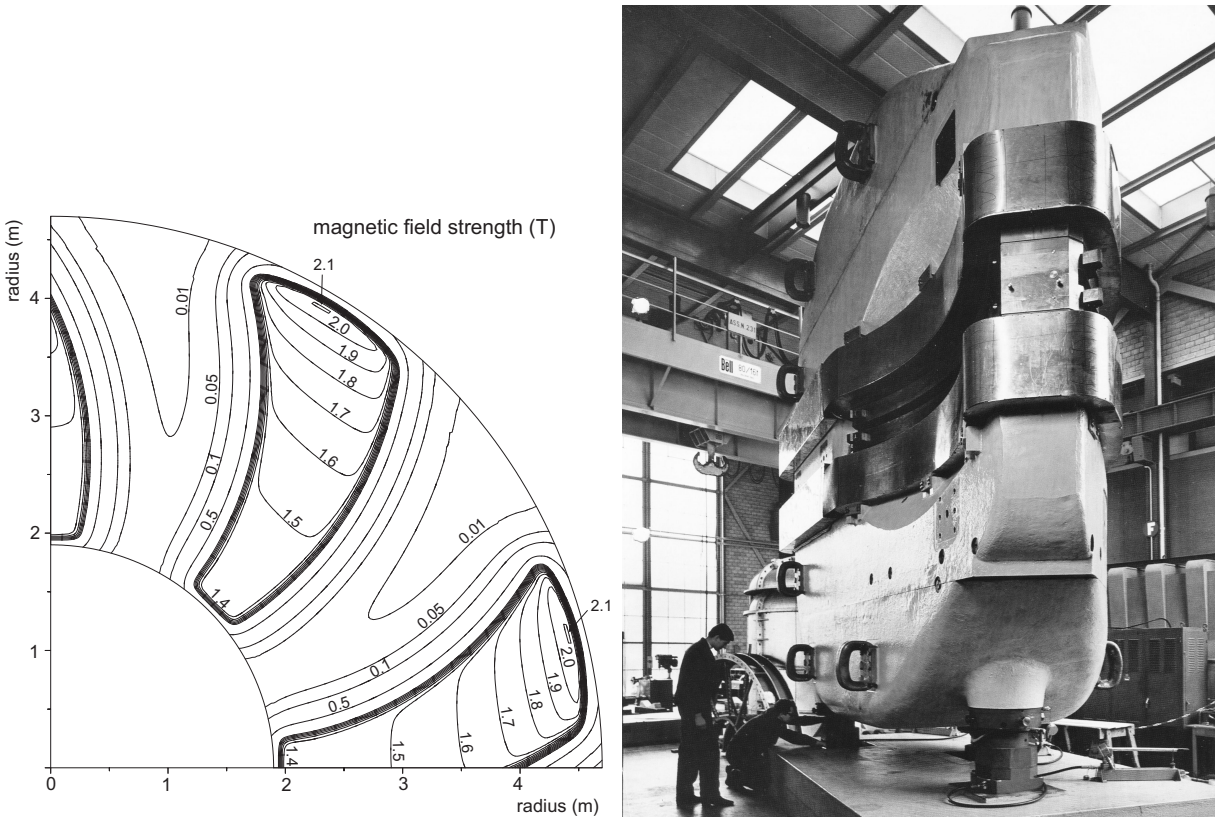


Fig. 8: Left: field map of the Ring cyclotron sector magnets at the PSI. The field increases towards larger radii to keep the particle revolution time constant. Right: photograph of a sector magnet before installation. Note the curved shape of the pole edges.

and industrial cyclotrons use superconducting magnets, to allow a compact and cost-effective design. Joho [18] compared the weight statistics of normal-conducting and superconducting cyclotrons. On average, those using superconducting coils were lighter by a factor of 15. The RIKEN cyclotron is unusual in this context, and employs superconducting magnets with a peak field strength of 3.8 T [19].

8 Radio frequency systems

In a classical cyclotron, an alternating voltage is applied across the gap between the dees (Fig. 1). In a separated-sector cyclotron, the space between the magnets allows separate resonators to be installed. These resonators function in principle like a rectangular cavity and can provide a significantly higher gap voltage, for example 1 MV. The beam passes through the resonator via a slit in the midplane. The electric field strength varies as a sine function along the radius (Fig. 9).

In this configuration, the resonance frequency depends on the radial length l and the height a of the cavity:

$$f_0 = \frac{c}{2}\sqrt{\frac{1}{a^2} + \frac{1}{l^2}}. \tag{26}$$

Thus the frequency is independent of the azimuthal width b of the cavity. In practice, the shape of the cavity is not made exactly rectangular; instead, the azimuthal width is reduced in the midplane (Fig. 10) to minimize the travel time of the particles in the field. In the case of the PSI cyclotrons, all accelerating resonators are operated at 50.6 MHz. In the Ring cyclotron, the resonators are made from copper and achieve a quality factor of 4.8×10^4. At the time of writing, the typical gap voltage is 830 kV. The design value is higher, about 1.2 MV. Each resonator can transfer 400 kW of power to the beam.

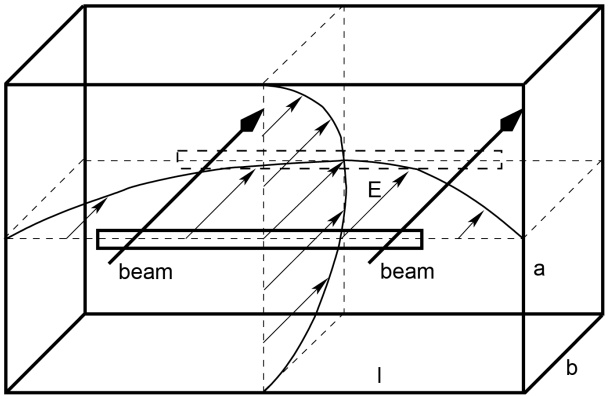


Fig. 9: Field distribution and orientation of the beam in a cyclotron box resonator

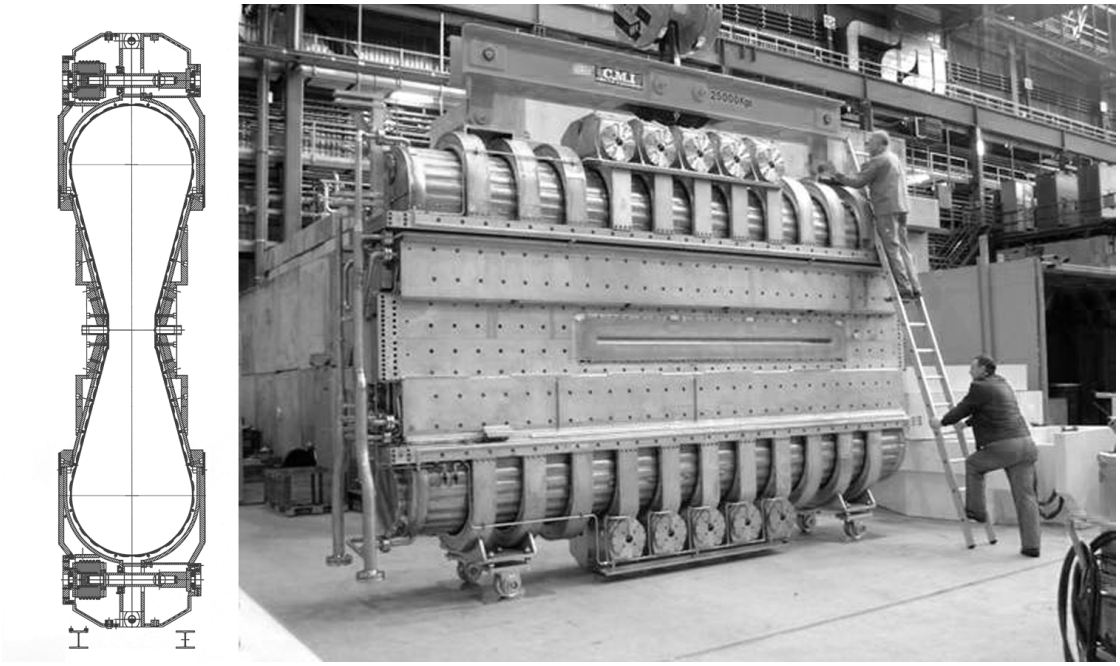


Fig. 10: Cross-section of PSI resonator (left), and photograph (right)

In order to minimize the variation of the energy over the bunch length, cyclotrons are often equipped with so-called flat-top resonators, which operate at the third harmonic. Using a decelerating flat-top voltage at 1/9 of the amplitude of the fundamental mode, the second derivative of the total voltage can be made zero at the nominal phase. In this way, the variation of the voltage as a function of the longitudinal position is minimized. For high-intensity cyclotrons, the power transfer from the electrical supply grid to the beam is a critical issue. Typically, staged tube amplifiers are used to generate the high-power RF signals required. The RF power is transferred to the resonators via coaxial lines and is coupled to the resonator volumes using loop couplers. The efficiency of the power transfer can be estimated from the product of the individual efficiencies of the components in the power transfer chain. The following values have been determined for the PSI: AC/DC conversion, 0.90; DC/RF conversion, 0.64; and RF to beam transfer, 0.55. Thus the total efficiency is about 32%, which is a relatively good number for a particle accelerator.

9 Performance of high-intensity cyclotrons, and discussion

As discussed in the previous sections, the major limitation on the operation of high-intensity cyclotrons is imposed by the extraction losses. With a world record beam power of 1.3 MW [20] in the PSI Ring cyclotron, the relative losses are at the level of 10^{-4}. The majority of the lost protons hit the vacuum chamber in the extraction beam line and activate magnets and other accelerator components. After many years of operation, the typical activation level is around 1 mSv/h, and hotspots of approximately 10 mSv/h are observed in the extraction beam line. In order to minimize the dose to personnel during servicing of critical components, special mobile shielding devices have been developed, for example for the exchange of the electrostatic extraction channel. The ultimate criterion for the activation problem in a high-intensity accelerator is the radiation dose that the service personnel receive during maintenance work. In the PSI facility, the total charge delivered per year has significantly increased over the years. Nevertheless, there exists no correlation with the dose received by personnel [21]. This fact demonstrates that it has been possible to keep the absolute beam losses at a constant level.

Another important aspect of the performance of a high-intensity accelerator concerns the efficiency of the power transfer from the grid to the beam. The PSI accelerator complex consumes 10 MW in total, and the beam power amounts to 1.3 MW. The total power includes experimental facilities and many magnets that are not essential for the production of the high-power beam. If one considers only the RF systems, the overall efficiency is 32%. Remarkably, the majority of the beam power is transferred through only four resonators in the Ring cyclotron. A potential application of high-intensity proton accelerators is in accelerator-driven systems (ADSs), which are subcritical reactors for burning thorium [22] or the transmutation of nuclear waste [23]. For such applications, the reliability and trip rate (the rate of short beam interruptions) of the accelerator are of the utmost importance. The typical trip rate of the PSI accelerator in recent years has been in the range of 20–50 trips per day. Most trips are caused by electrical breakdowns due to the voltage in electrostatic elements. After a trip, the beam current is ramped back up to its nominal value within 30 s. Trips of the RF systems occur much less frequently. A statistical analysis of trip durations has been given in Ref. [24]. ADS systems require a much lower trip rate, of the order of 0.01–0.1 trips per day. Although there exists a promising potential for improvement (see also [24]), it will be very difficult to achieve performance in this range starting from today's performance.

The following particular advantages and disadvantages of the cyclotron concept for high-intensity beam production can be stated.

- A cyclotron requires an extraction element with an electrode placed close to the beam. By comparison, an L-band superconducting linac has a large aperture, and thus it is potentially easier to achieve low losses in a linac. The electrostatic elements in cyclotrons are critical and fragile devices, causing relatively frequent beam trips and failures.
- Because of the concept of a circular accelerator, the beam dynamics in a cyclotron is more complicated, and it requires tedious tuning to achieve an optimized operational state with low losses.
- For fundamental reasons, the maximum energy of a cyclotron is limited to roughly 1 GeV.
- The edge focusing used in cyclotrons is weaker than the alternating-gradient focusing in linacs. In particular, the space charge forces in the vertical plane lead to a limitation on the maximum beam current. It is expected that maximum currents in the region of 10 mA can be achieved in sector cyclotrons [25].
- On the pro side, the circular-cyclotron concept allows the accelerating resonators to be re-used many times. The footprint of a cyclotron facility is smaller, allowing some savings with respect to the shielding and the building.
- The low-frequency resonators in a cyclotron are robust, simple devices with low trip rates and allow very high power throughput in the couplers.
- The efficiency of the power transfer from the grid to the beam is comparatively high, in the region of 30%.

In summary, the cyclotron concept is capable of delivering high-intensity beams with a beam power of up to 10 MW and an energy of 1 GeV and represents an effective alternative to other concepts in this range.

References

[1] L.M. Onishchenko, *Phys. Part. Nuclei* **39** (2008) 950.

[2] E.O. Lawrence and N.E. Edlefsen, *Science* **72** (1930) 376.

[3] E.D. Courant and H.S. Snyder, *Ann. Phys.* **3** (1958) 1.

[4] L.H. Thomas, *Phys. Rev.* **54** (1938) 580–598.

[5] Y.J. Bi *et al.*, *Phys. Rev. Spec. Top. Accel. Beams* **14** (2011) 054402.

[6] G. Dutto *et al.*, TRIUMF high intensity cyclotron development for ISAC, Proc. 17th Int. Conf. on Cyclotrons and Their Applications, Tokyo, 2004, pp. 82–86.

[7] M. Kase *et al.*, Present status of the RIKEN Ring cyclotron, Proc. 17th Int. Conf. on Cyclotrons and Their Applications, Tokyo, 2004, pp. 160–162.

[8] H. Willax, Proposal for a 500 MeV isochronous cyclotron with ring magnet, Proc. Int. Conf. on Sector-Focused Cyclotrons, Geneva, 1963, p. 386.

[9] J.J. Yang *et al.*, *Phys. Rev. Spec. Top. Accel. Beams* **13** (2010) 064201.

[10] A. Chao and M. Tigner (Eds.), *Handbook of Accelerator Physics and Engineering* (World Scientific, Singapore, 1999), Chapter 1.6.4.

[11] W. Joho, High intensity problems in cyclotrons, Proc. 5th Int. Conf. on Cyclotrons and Their Applications, Caen, 1981, pp. 337–347.

[12] E. Pozdeyev, A fast code for simulation of the longitudinal space charge effect in isochronous cyclotrons, Proc. 16th Int. Conf. on Cyclotrons and Their Applications, East Lansing, MI, 1981, p. 411.

[13] R. Dölling, Measurement of the time-structure of the 72 MeV proton beam in the PSI Injector-2 cyclotron, Proc. DIPAC2001, Grenoble, 2001, pp. 111–113.

[14] M. Seidel and P.A. Schmelzbach, Upgrade of the PSI Cyclotron Facility to 1.8 MW, Proc. 18th Int. Conf. on Cyclotrons and Their Applications, Giardini Naxos, Italy, 2007, pp. 157–162.

[15] C. Chasman and A.J. Baltz, *Nucl. Instrum. Methods* **219** (1984) 279.

[16] P. Bertrand and C. Ricaud, Specific cyclotron correlations under space charge effects in the case of a spherical beam, Proc. 16th Int. Conf. on Cyclotrons and Their Applications, Caen, 2001, p. 379.

[17] L. Calabretta *et al.*, A multi-megawatt cyclotron complex to search for CP violation in the neutrino sector, Proc. Cyclotrons and Their Applications 2010, Lanzhou, China, 2010, p. 299.

[18] W. Joho, Modern trends in cyclotrons, CERN Accelerator School: Accelerator Physics, Aarhus, Denmark, 1986, pp. 260–290.

[19] H. Okuno *et al.*, Magnets for the RIKEN superconducting RING cyclotron, Proc. 17th Int. Conf. on Cyclotrons and Their Applications, Tokyo, 2004, pp. 373–377.

[20] M. Seidel *et al.*, Production of a 1.3 MW proton beam at PSI, Proc. IPAC'10, Kyoto, 2010.

[21] M. Seidel, J. Grillenberger, and A. Mezger, Experience with the production of a 1.3 MW proton beam in a cyclotron based facility, Proc. TCADS, Karlsruhe, 2010, pp. 251–260.

[22] C. Rubbia *et al.*, An energy amplifier for cleaner and inexhaustible nuclear energy production driven by a particle beam accelerator, CERN/AT/93-47 (ET) (1993).

[23] R. Sheffield, Utilization of accelerators for transmutation and energy production, Proc. HB2010, Morschach, Switzerland, 2010, pp. 1–5.

[24] M. Seidel and A.C. Mezger, Performance of the PSI high power proton accelerator, IAEA, Int.

Topical Meeting on Nuclear Research Applications and Utilization of Accelerators, Vienna, 2009, AT/RD-10.

[25] T. Stammbach *et al.*, *Nucl. Instrum. Methods Phys. Res. B* **113** (1996) 1–7.

Fixed field alternating gradient

Shinji Machida
ASTeC, STFC Rutherford Appleton Laboratory, Didcot, United Kingdom

Abstract
The concept of a fixed field alternating gradient (FFAG) accelerator was invented in the 1950s. Although many studies were carried out up to the late 1960s, there has been relatively little progress until recently, when it received widespread attention as a type of accelerator suitable for very fast acceleration and for generating high-power beams. In this paper, we describe the principles and design procedure of a FFAG accelerator.

1 Introduction

The idea of a fixed field alternating gradient (FFAG) accelerator is not new. It was invented in the 1950s right after the alternating gradient (AG) synchrotron came out [1, 2]. Instead of using pulsed magnets as in a synchrotron, FFAG accelerators use constant field magnets like cyclotrons. Unlike cyclotrons, however, FFAGs rely on AG focusing so that the beam size can be much smaller. In the literature [1], there are accelerators described as FFAG betatrons, FFAG cyclotrons and FFAG synchrotrons. This is a little confusing and in these cases, FFAG refers merely to the focusing scheme based on so-called cardinal conditions, which we discuss later. Here we use the terms "FFAG accelerator" or "FFAG" to mean accelerators using the FFAG focusing scheme.

Although there was very active work at the Midwestern Universities Research Association (MURA) from the early 1950s to the late 1960s, the activity stopped when particle physics chose AG synchrotrons for its future tool. The idea of the FFAG was sound, but unfortunately it was not the best accelerator for energy frontier research. An AG synchrotron used more compact magnets and it was easier to obtain high output energy. Nevertheless, the research at MURA associated with the FFAG developments introduced a mathematical formalism and many concepts that became common later in accelerator physics. These include beam stacking, Hamiltonian theory of longitudinal motion, colliding beams, effects of non-linear forces, modelling collective instabilities, use of digital computers in design of orbits, proof of chaotic motion and synchrotron radiation rings. There are two especially informative publications [3, 4] for those interested in a historical view of FFAG developments. The former is free to download.

A new era of FFAG development started in the late 1990s in Japan (there were in fact some activities of FFAG design in 1980s and 1990s in the USA and in Europe; see, for example, Refs. [5, 6]), first in connection with muon acceleration in a neutrino factory [7, 8] and later for wider applications, especially for high-power beam production. The use of fixed field magnets enables very fast acceleration of a beam and is limited only by the available radio-frequency (RF) voltage because there is no restrictive magnet ramping cycle slower than the modulation of the RF frequency. As an extreme example, acceleration of a muon beam, whose lifetime in its rest frame is 2.2 μs, becomes possible although the required RF power would be huge. Such fast acceleration also means that beam acceleration can be repeated more often, for example in a pulsed source. Since beam power is a product of energy, the number of particles per pulse and the repetition rate, it is possible to produce high-power beam via the higher repetition possible in a FFAG accelerator.

A small-scale model of a proton FFAG was first commissioned in 1999 [9] and a prototype of a proton therapy accelerator of 150 MeV energy delivered a beam a few years later. Another FFAG accelerator with similar specifications was constructed as a proton driver test facility linked to the

Japanese Accelerator Driven Subcritical Reactor (ADSR) programme [10]. One of the breakthroughs that made construction of all of these accelerators possible was the success of a novel RF cavity with magnetic alloy material instead of conventional ferrite to modulate the RF frequency. High shunt impedance with very low Q factor is an ideal property for an RF cavity for a FFAG. Activities in Japan revived the potential of the FFAG principle not only as a tool for particle physics but for a variety of applications using state-of-the-art technology.

There was another initiative in FFAG accelerator development outside Japan, which tried to reduce the size of the lattice magnets, specifically for muon acceleration [11]. This line of enquiry led to a new concept being proposed in the late 1990s. Since this FFAG did not follow the so-called scaling law principle of a conventional FFAG, it became named a non-scaling FFAG. An accelerator based on this non-scaling concept was recently built and successfully commissioned in the UK [12].

In the following sections, we briefly explain the principles and design procedure of the different types of FFAG accelerators.

2 Scaling FFAG

The guiding field of a weak focusing cyclotron with a single pole decreases gradually with radius. In terms of the magnet field index n,

$$n = -\frac{r_0}{B_z(r_0)}\left(\frac{\partial B_z}{\partial r}\right)_{r=r_0} \tag{1}$$

where r_0 is the reference radius and $B_z(r_0)$ is the vertical field at the reference radius r_0, stable motion in both horizontal and vertical planes requires

$$0 < n < 1 \tag{2}$$

In an AG synchrotron, piecewise magnets of an accelerator lattice have large n, but the sign of n alternates, hence the name AG or strong focusing. AG magnets are either combined function type where dipole and quadrupole components co-exist in a single magnet or separated function type where pure dipole magnets are used for bending and quadrupole magnets for focusing. FFAG accelerators use the same the AG principle, but, whereas in an AG synchrotron the magnetic fields are ramped to match the increase in particle momentum under acceleration, in FFAGs the fields are held fixed

Consider the operation of an AG synchrotron without ramping the magnetic fields. It should be possible to increase the beam momentum by applying an RF voltage. The first thing one may observe is the shift of orbit outward or inward depending on the sign of the dispersion function. When the displaced orbit hits one side of the vacuum chamber, the beam is lost. In an ordinary AG synchrotron, this can happen with a momentum increase as small as 1 % or so.

One may also observe a reduction in focusing strength or equivalently transverse tune. The focal length of a quadrupole f is

$$\frac{1}{f} = \frac{B'L}{p/e} \tag{3}$$

where B' is the gradient of the quadrupole magnet, L is its length and p is the beam momentum. When the beam momentum increases, the focusing strength becomes weaker and the focal length becomes longer. This is called a chromaticity effect. Some synchrotrons have chromaticity-correcting sextuples

around the ring. The sextupole field profile adds to or subtracts from the field gradient of the lattice quadrupoles depending on the radial position of the beam. When a sextupole is located where the dispersion function is non-zero, off-momentum beams feel either a larger or smaller field gradient and by these means the tune dependence on momentum can be eliminated.

The original FFAG, called a *scaling FFAG*, exhibits a behaviour in terms of orbit and optics similar to a chromaticity-corrected synchrotron without ramping magnets, but designed for much wider momentum acceptance. To avoid the beam hitting an aperture limit, a scaling FFAG simply widens the horizontal size of the vacuum chamber. To increase the momentum range where chromaticity correction works, it uses a magnet whose field gradient depends on radial position as shown in Fig. 1.

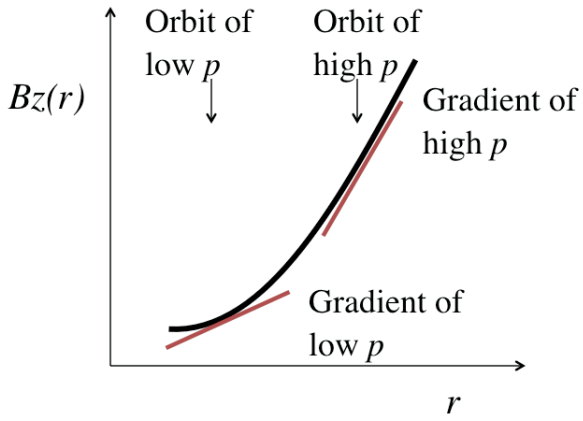


Fig. 1: Guiding field profile of a scaling FFAG magnet. The orbit of a high momentum beam sees a stronger field (black curve) as well as a larger field gradient (red curve).

Suppose that a beam with higher momentum circulates around an outer radius and a beam with lower momentum circulates around an inner radius. Since the magnetic field gradient becomes steeper as the beam is accelerated and migrates out, it feels effectively the same focusing force. We define the magnet field index locally as follows where ρ is the bending radius:

$$n = -\frac{\rho}{B}\frac{\partial B_z}{\partial x} \tag{4}$$

The constancy of n independent of beam momentum at each azimuthal position around the ring, expressed as

$$\left.\frac{\partial n}{\partial p}\right|_{\vartheta=const} = 0 \tag{5}$$

is called one of the *cardinal conditions* of a scaling FFAG.

This is not enough to ensure a constant tune independent of the beam momentum. The second cardinal condition, called geometrical similarity, is written as

$$\left.\frac{\partial}{\partial p}\left(\frac{\rho_0}{\rho}\right)\right|_{\vartheta=const}=0 \tag{6}$$

This ensures that the momentum-dependent orbits are similar and an exact photographic enlargement of each other.

The derivation above was based on physical intuition, but we can reach the same conclusion starting from the equations of horizontal and vertical betatron oscillations:

$$\frac{d^2x}{ds^2}+\frac{\rho_0^2}{\rho^2}(1-n)x=0 \tag{7a}$$

$$\frac{d^2z}{ds^2}+\frac{\rho_0^2}{\rho^2}nz=0 \tag{7b}$$

where x and z are the betatron oscillation amplitudes in the horizontal and vertical directions respectively, and s is the longitudinal coordinate. The requirement to make the coefficients of the restoring force, $\frac{\rho_0^2}{\rho^2}(1-n)$ for horizontal and $\frac{\rho_0^2}{\rho^2}n$ for vertical, independent of the beam momentum leads to the same results as Eqs. (5) and (6).

Now we need to find the magnetic field that satisfies the cardinal conditions. This should have the following form

$$B(r,\theta)=B_0\left(\frac{r}{r_0}\right)^k F(\vartheta) \tag{8}$$

where $F(\vartheta)$ is a periodic function of a generalized azimuthal angle ϑ. The use of a generalized angle becomes clearer later when a spiral sector type FFAG is discussed. The power k is called the *geometrical magnet field index*. Note that the radial and azimuthal dependence are separated. By expanding the radial dependence in a Taylor series, one can see that the field is a summation of multipoles of all orders:

$$\left(\frac{r}{r_0}\right)^k=1+\frac{k}{r_0}(r-r_0)+\frac{k(k-1)}{2!r_0^2}(r-r_0)^2+\frac{k(k-1)(k-2)}{3!r_0^3}(r-r_0)^3+\cdots \tag{9}$$

A magnet with this profile gives focusing in one plane, but not in both horizontal and vertical planes simultaneously. To form an AG lattice, we need another magnet with the opposite sign field gradient. In an AG synchrotron, this is done by changing the sign of k. In a scaling FFAG accelerator, the sign of B_0 is changed as shown in Fig. 2.

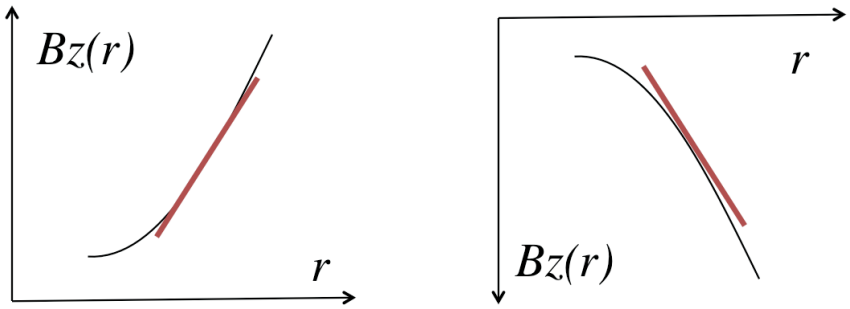


Fig. 2: In a FFAG accelerator, the sign of B_0 of Eq. (8) changes to realize AG (red lines)

This configuration of a FFAG is called *radial sector type*. The generalized angle is the same as the geometrical angle:

$$F(\vartheta) = F(\theta) \tag{10}$$

You may notice that there is an obvious drawback in doing this. One kind of magnet has the opposite gradient, but at the same time, it has the opposite bending angle as well as shown in Fig. 3. Owing to the cancellation caused by the normal and opposite bending angles, the ring circumference has to be larger than an AG synchrotron which has normal bending only.

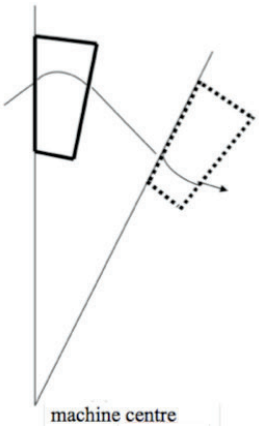


Fig. 3: In radial sector type FFAG, magnets with the opposite gradient have opposite bending angles as well

Another way of realizing AG focusing is to introduce edge focusing on one side of a magnet. If a beam does not go into the magnet at right angles to the face, a vertical kick is produced proportional to the vertical displacement:

$$\Delta p_z = -eB_z \tan\zeta \cdot z \tag{11}$$

where ζ is an injection angle with respect to the normal to the magnet edge. To make the focusing strength independent of momentum, the injection angle of the magnet should satisfy

$$\frac{rd\theta}{dr} = \tan\zeta \tag{12}$$

which then gives a spiral shape. The generalized angle can be defined as

$$\vartheta = \theta - \tan\zeta \cdot \ln\frac{r}{r_0} \tag{13}$$

In this case the bends can all be in the same direction with no need for alternating signs. This is called a *spiral sector type* as shown in Fig. 4.

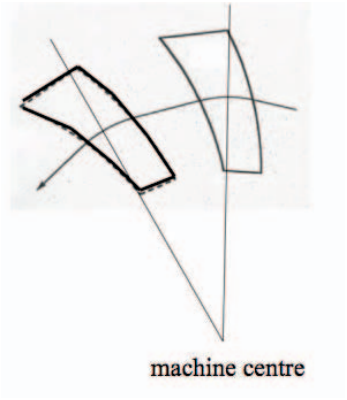


Fig. 4: In spiral sector type FFAG, all of the magnets are the same but AG focusing comes from the body and edge

The vertical focusing strength is limited by the need for a practical spiral angle. On the other hand, removing the opposite bending magnets from the lattice makes the circumference smaller than the radial sector type.

You may wonder what is the difference between a FFAG and the Thomas or AVF cyclotron. Apart from the fact that a FFAG does not have isochronous conditions, the Thomas cyclotron is essentially a radial sector FFAG [1]. In that sense, FFAG was not a new invention even when it was proposed in the 1950s. On the other hand, the design strategy is different from that of a cyclotron. A FFAG aims for a much higher momentum accelerator, for example more than 1 GeV, with a reasonable size of magnet by reducing the orbit shift during acceleration. This may not be compatible with the kind of isochronous condition normally found in a cyclotron.

3 Linear non-scaling FFAG

Let us consider the operation of an AG synchrotron without ramping magnets again. Remember there were two problems if we did this. One was that a beam would hit the wall due to the dispersion function and the other was that the focusing force would effectively decrease. In a scaling FFAG, we introduced the radial field profile so that the tune was constant for the entire momentum range during acceleration. For the orbit shift, we simply widen the aperture in the horizontal direction.

The orbit excursion in a scaling FFAG is smaller than in a cyclotron, but still not negligible, for example around 0.7 m compared with a 5 m radius for a few hundred megaelectronvolts [10]. This is because of the upper limit of the field index k. To squeeze the orbit shift during acceleration, the field index k should be as large as possible. On the other hand, it risks losing the stability inherent in AG focusing because too large a k leads to over-focusing. Also one may notice that most of the orbit shift happens in the lower momentum region where the field gradient is relatively small.

There is another way of designing a FFAG accelerator, which reduces the orbit shift as much as possible, without paying much attention to the tune excursion during acceleration. This is effectively a synchrotron lattice with as small a dispersion function as possible without chromaticity correction. If the dispersion function is small enough, the orbit shift caused by momentum changes can be

accommodated in a reasonably sized vacuum chamber. If we eliminate multipoles higher than quadrupole in Eq. (9), the dynamic aperture is expected to be large as well. We choose a quadrupole field gradient that gives a phase advance per focusing unit below 180° at the injection momentum and let it decrease when the beam is accelerated. This is called a *linear non-scaling FFAG*.

A way to design a synchrotron lattice with small dispersion function is well known from the design of synchrotron light sources [13]. The *H*-function is defined as

$$H = X_d^2 + P_d^2 \tag{14}$$

$$X_d = D/\sqrt{\beta_x} = \sqrt{2J_d}\cos\phi_d \tag{15}$$

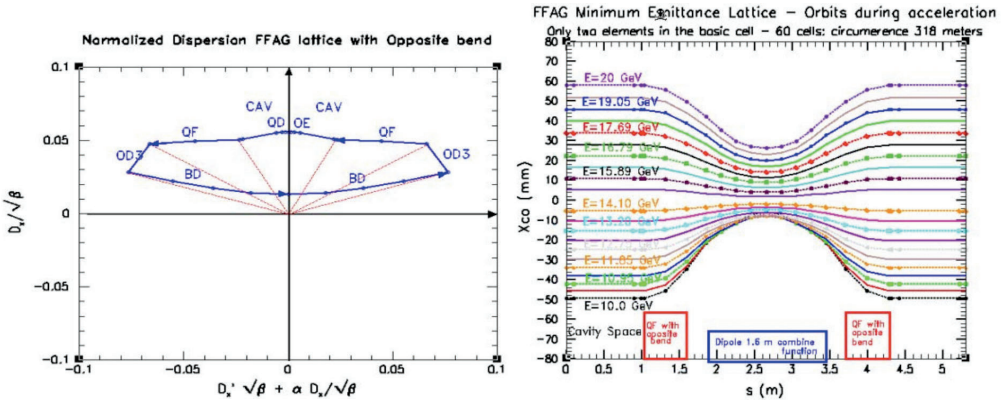
$$P_d = \left(\alpha_x D + \beta_x D'\right)/\sqrt{\beta_x} = -\sqrt{2J_d}\sin\phi_d \tag{16}$$

where D and D' are the dispersion function and its derivative, β_x and α_x are horizontal Twiss parameters, J_d is invariant in a region without dipole magnets and ϕ_d is identical to the betatron phase advance. With bending, the amplitude of J_d changes at a dipole as

$$\Delta X_d = 0 \tag{17}$$

$$\Delta P_d = \sqrt{\beta_x}\Delta D = \sqrt{\beta_x}\theta \tag{18}$$

With the help of the *H*-function, we can see that a dipole magnet where the beta function is small makes the dispersion function a minimum. A linear non-scaling FFAG lattice has therefore normal bending at a defocusing quadrupole and opposite bending at a focusing quadrupole [14]. Figure 5 shows the *H*-function and orbits of different momentum of the non-scaling FFAG designed in Ref. [14].



(a) (b)

Fig. 5: (a) Normalized dispersion space: the opposite bend at QF quadrupole, while the major bend is at BD. (b) Closed orbit offsets during acceleration from 10 GeV to 20 GeV. (Reproduced from Ref. [14].)

Although the design avoids at least a systematic resonance at 180° phase advance per unit cell, a potential problem is crossing higher-order resonances and non-systematic resonances during acceleration as shown in Fig. 6. In the early days, it was thought that keeping transverse tunes independent of momentum during acceleration was the essential design requirement of an accelerator with the AG principle. Resonances excited by repetitive actions to the beam eventually lead to beam loss. A linear non-scaling FFAG challenges this conventional wisdom.

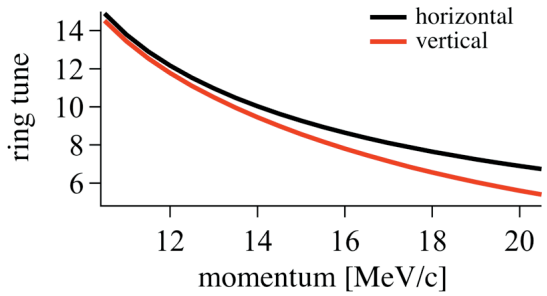


Fig. 6: Tune excursion in a linear non-scaling FFAG. Although the phase advance per unit cell is chosen below 180°, the total tune of the ring (in this example 42 unit cells) should cross several integer tunes during acceleration.

The effect of resonance crossing depends on the speed of crossing and the strength of the resonances. For some applications such as muon acceleration, the acceleration is very fast so that the fast tune excursion probably makes the resonance crossing harmless. In fact muon acceleration is the original application for which the linear non-scaling FFAG was designed. On the other hand, when we use a linear non-scaling FFAG for ordinary applications such as a particle therapy accelerator, the beam circulates for of the order of 1000 to 10 000 turns and therefore the speed of resonance crossing would be slow. We do not know how slow a resonance crossing could be tolerated in a linear non-scaling FFAG. This is an on-going research subject.

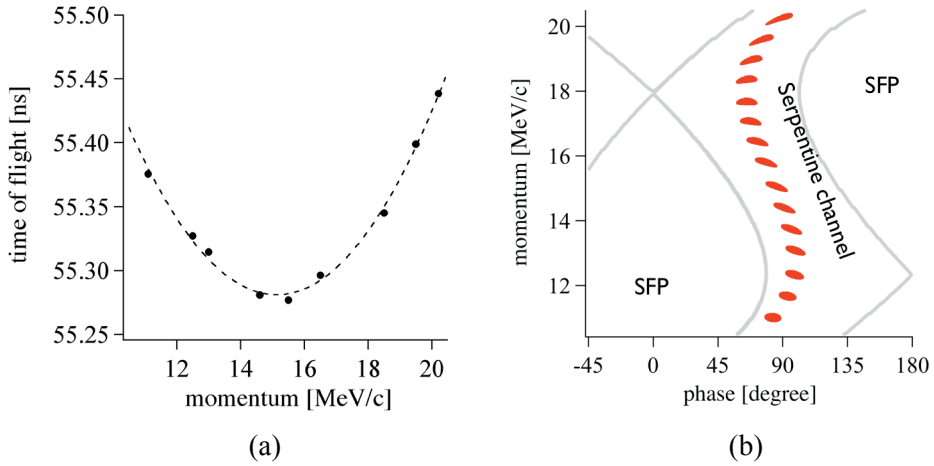


Fig. 7: (a) Orbital period (or time of flight) in a linear non-scaling FFAG as a function of momentum. (b) Serpentine channel created between two RF buckets with stable fixed points (SFP) shown. Red marks indicate a beam injected at the lower momentum and accelerated to the top.

A small orbit shift for the entire momentum range also means the orbital length does not vary much as a function of momentum. If a beam is in the ultra-relativistic regime, it also means that the orbital period is almost constant. This enables the use of a constant RF frequency cavity if the phase

slip during acceleration is small enough. In fact, in muon acceleration, which is completed in about 10 turns, a beam stays near the crest of the RF voltage throughout. By adjusting the lattice so that the orbital period during the acceleration has a parabolic shape as shown in Fig. 7(a) and choosing a RF frequency that corresponds to the orbital period within the parabola, we can create RF buckets close to each other with slightly different momentum centres. Instead of injecting a beam inside the RF bucket, the channel between the two RF buckets can be used for acceleration. This is called *serpentine channel acceleration*. This novel acceleration scheme with stability in a linear non-scaling FFAG was demonstrated recently in the UK [12].

4 Other types of FFAG

The non-scaling FFAG is a promising idea, but it is not clear whether the resonance crossing during acceleration is harmless. There are more concerns when we design an accelerator with moderate acceleration rate with moderate RF voltage such as one for particle therapy. In fact, a simulation study showed that, at the very least, more accurate alignment is required when we accelerate a beam at a slower rate in a linear non-scaling FFAG [15].

It is possible to fix a tune without following the cardinal conditions. One way is to start with the design of a scaling FFAG and relax the conditions. For example, we can truncate Eq. (9) at some particular order. Lattice magnets can be rectangular in shape instead of radial sector type. In these magnets, equipotential lines are straight rather than arcs of a circle. Within a unit cell, three magnets forming triplet focusing are aligned along a line instead of a circle. With careful design of the field profile in each magnet, it is still possible to ensure the tune variation is negligible over a wide range of momentum [16]. This design principle does not follow the scaling rule, but it does not use linear quadrupole magnets either. Therefore, it is called a *non-linear non-scaling* FFAG.

We can fix the tune, based on a linear non-scaling FFAG design, by introducing non-linear magnetic fields. It is like introducing sextupole magnets to correct chromaticity in a synchrotron lattice. Sextupole components are not enough, however, to correct the tune over a wide momentum range. Higher-order non-linearities such as octupole and/or decapole are required. This is another approach to designing a non-linear non-scaling FFAG.

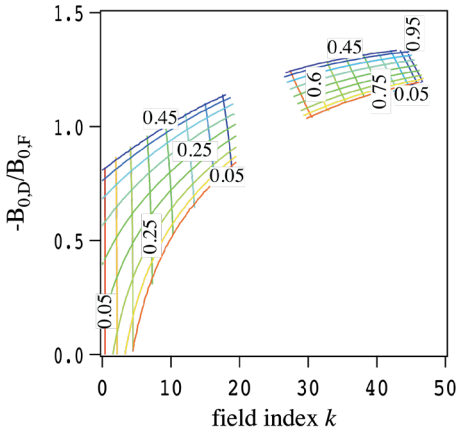


Fig. 8: Stability diagram with a practical lattice configuration with the geometrical field index k as abscissa and the ratio of focusing and defocusing magnet strengths as ordinate. Upright numbers indicate vertical cell tune and vertically aligned numbers indicate horizontal cell tune. Lines are drawn with 0.05 step. (Reproduced from Ref. [17].)

One drawback of a scaling FFAG compared with a non-scaling FFAG is its relatively large orbit shift during acceleration. Addressing this was the primary concern that led to the invention of the non-scaling FFAG. There is another way to reduce the orbit shift, however, while keeping the scaling principle. It was well known that there is another stability region in phase space when we increase the focusing strength and produce phase advances over 180° per cell, as shown in Fig. 8. In a synchrotron lattice, there is no advantage in choosing the second stability region because the beta function becomes larger and alignment tolerance and manufacturing specifications become tighter. On the other hand, the larger gradient needed to produce stronger focusing reduces the orbit shift in a FFAG. If we could maintain a reasonable tolerance in the design using the second stability region, it would be a big advantage. With a triplet focusing structure, it turns out that we can reduce the orbit shift by a factor of five if we use the second stability region [17]. The dynamic aperture is smaller and tolerance is tighter, but they are at manageable practical levels.

So far, we have discussed the type of optics for a FFAG used as a circulating accelerator. The optics that can accommodate large momentum spread can be used for beam transport as well. It has been found that such a beam transport line can be realized as the limit of a ring with infinitely large bending angle [18, 19] as shown in Fig. 9(a). A dispersion suppressor at the end of the transport line may add another advantage. Owing to its wide momentum acceptance, this system can be applied to the gantry of a particle therapy facility or a dump line for a synchrotron where the tripped beam momentum could be any momentum from injection to extraction.

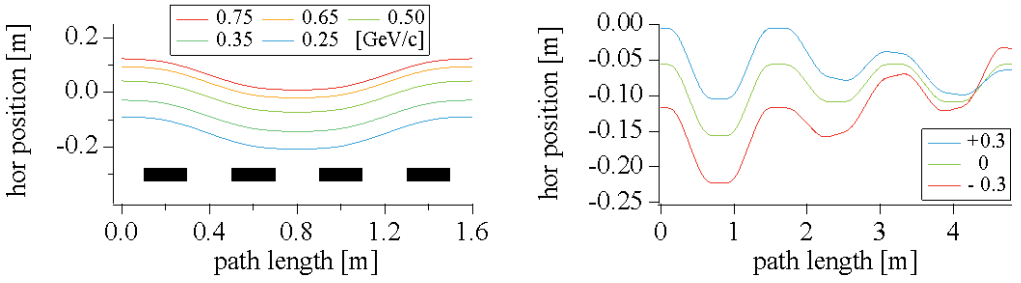


Fig. 9: (a) Different momentum orbits in a unit cell which satisfies the periodic boundary condition. Rectangles at the bottom show the position of FDDF magnets. (b) Different momentum orbits in a normal cell with dispersion suppressor. Numbers in the legend show the momentum deviation from the reference momentum. The normal cell is from 0 m to 1.6 m and the dispersion suppressor section from 1.6 m to 4.8 m. (Reproduced from Ref. [18].)

Acknowledgments

The author would like to thank C. R. Prior and D. J. Kelliher for their careful reading and comments on this lecture note.

References

[1] K.R. Symon, D.W. Kerst, L.W. Jones, L.J. Laslett and K.M. Terwilliger, *Phys. Rev.* **103** (1956) 1837.

[2] A.A. Kolomensky and A.N. Lebedev, *Theory of Cyclic Accelerators* (North-Holland, Amsterdam, 1966).

[3] F.T. Cole, O Camelot! Memoirs of the MURA years, http://accelconf.web.cern.ch/accelconf/c01/cyc2001/extra/Cole.pdf (1994).

[4] L. Jones, F. Mills, A. Sessler, K. Symon, D. Young, *Innovation Was Not Enough* (World Scientific, 2009).

[5] T.K. Khoe and R.L. Kustom, *IEEE Trans. Nucl. Sci.* **30** (1983) 2086.

[6] R. Kustom, S.A. Martin, P.F. Meads, G. Wuestefeld, E. Zaplatin and K. Ziegler, Proc. of the European Particle Accelerator Conference, 1994, p. 574.

[7] F. Mills and C. Johnstone, Proc. 4th International Conference Physics Potential and Development of $\mu^+\mu^-$ Colliders, Transparency Book (1997).

[8] S. Machida, *Nucl. Instrum. Methods Phys. Res., Sect. A* **503** (2003) 41.

[9] M. Aiba, *et al.*, Proc. of the European Particle Accelerator Conference, 2000, p. 581.

[10] M. Tanigaki, *et al.*, Proc. of the European Particle Accelerator Conference, 2006, p. 2367.

[11] C. Johnstone, W. Wan and A. Garren, Proc. of the Particle Accelerator Conference, 1999, p. 3068.

[12] S. Machida, *et al.*, *Nature Phys.* **8** (2012) 243.

[13] D. Trbojevic and E. Courant, Proc. of the European Particle Accelerator Conference, 1994, p. 1000.

[14] D. Trbojevic, J.S. Berg, M. Blaskiewicz, E.D. Courant, R. Palmer and A. Garren, Proc. of the Particle Accelerator Conference, 2003, p. 1816.

[15] S. Machida, *Phys. Rev. ST Accel. Beams* **11** 094003 (2008).

[16] S.L. Sheehy, K.J. Peach, H. Witte, D.J. Kelliher, S. Machida, *Phys. Rev. ST Accel. Beams* **13** 040101 (2010) .

[17] S. Machida, *Phys. Rev. Lett.* **103** 164801 (2009).

[18] S. Machida and R. Fenning, *Phys. Rev. ST Accel. Beams* **13** 084001 (2010).

[19] Y. Mori, T. Planche and J.B. Lagrange, Proc. of FFAG Workshop '08J, http://hadron.kek.jp/FFAG/FFAG08J_HP/ (2008).

Radio-frequency power generation

Richard G. Carter
Engineering Department, Lancaster University, Lancaster LA1 4YR, U.K.
and The Cockcroft Institute of Accelerator Science and Technology, Daresbury, UK

Abstract
This paper reviews the main types of radio-frequency power amplifiers which are, or may be, used for high-power hadron accelerators. It covers tetrodes, inductive output tubes, klystrons and magnetrons with power outputs greater than 10 kW continuous wave or 100 kW pulsed at frequencies from 50 MHz to 30 GHz. Factors affecting the satisfactory operation of amplifiers include cooling, matching and protection circuits are discussed. The paper concludes with a summary of the state of the art for the different technologies.

1 Introduction

All particle accelerators with energies greater than 20 MeV require high-power radio-frequency (RF) sources [1]. These sources must normally be amplifiers to achieve sufficient frequency and phase stability. The frequencies employed range from about 50 MHz to 30 GHz or higher. Power requirements range from 10 kW to 2 MW or more for continuous sources and up to 150 MW for pulsed sources. Figure 1 shows the main features of a generic RF power system. The function of the power amplifier is to convert d.c. input power into RF output power whose amplitude and phase is determined by the low-level RF input power. The RF amplifier extracts power from high-charge, low-energy electron bunches. The transmission components (couplers, windows, circulators, etc.) convey the RF power from the source to the accelerator, and the accelerating structures use the RF power to accelerate low-charge bunches to high energies. Thus, the complete RF system can be seen as an energy transformer which takes energy from high-charge, low-energy electron bunches and conveys it to low-charge, high-energy bunches of charged particles. When sufficient power cannot be obtained from a single amplifier then the output from several amplifiers may be combined. In some cases power is supplied to a number of accelerating cavities from one amplifier.

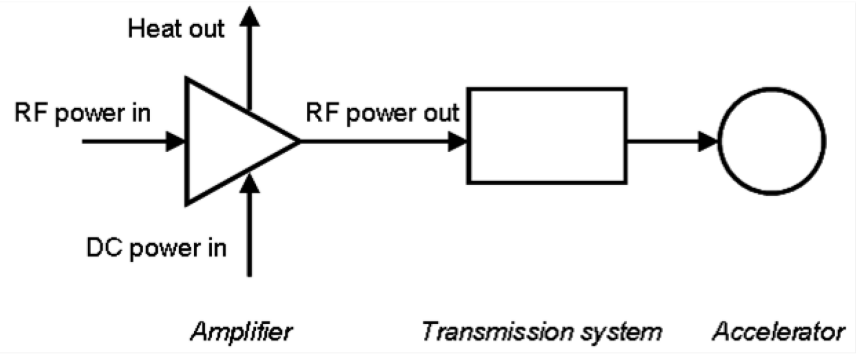


Fig. 1: Block diagram of the high-power RF system of an accelerator

Because the amplifier is never completely efficient there is always some conversion of energy into heat. The principle of conservation of energy requires that, in the steady state, the total input and output power must balance, that is

$$P_{RF\,in} + P_{DC\,in} = P_{RF\,out} + P_{Heat} \tag{1}$$

Strictly speaking, the input power should also include the power required for other purposes related to the amplifier including heaters, magnets and cooling systems. However, in many cases these are small compared with the RF output power and may be neglected to a first approximation. The efficiency (η_e) of the amplifier is the ratio of the RF output power to the total input power

$$\eta_e = \frac{P_{RF\ out}}{P_{DC\ in} + P_{RF\ in}} \approx \frac{P_{RF\ out}}{P_{DC\ in}} \tag{2}$$

In many cases the RF input power is small compared with the d.c. input power so that it may be neglected to give the approximation shown in Eq. (2). The efficiency is usually expressed as a percentage. The heat which must be dissipated is

$$P_{Heat} = (1 - \eta_e) P_{DC} \tag{3}$$

The other main parameter of the generic amplifier is its gain in decibels given by

$$G = 10 \log_{10} \left(\frac{P_{RF\ out}}{P_{RF\ in}} \right) \tag{4}$$

The physics of the energy exchange between the electron bunches and the RF power means that the size of the space in which the exchange takes place must be small compared with the distance an electron moves in one RF cycle. Thus, the size of an amplifier decreases with decreasing d.c. voltage and with increasing frequency. Hence, for a given RF output power and efficiency, the energy density within the amplifier increases with decreasing size. The need to keep the working temperature of the amplifier below the level at which it will cease to operate reliably means that the maximum possible output power from a single device is determined by the working voltage, the frequency and the technology employed. Other factors which are important in the specification of power amplifiers include reliability and, in some cases, bandwidth.

The capital and running cost of an accelerator is strongly affected by the RF power amplifiers in a number of ways. The capital cost of the amplifiers (including replacement tubes) is an appreciable part of the total capital cost of the accelerator. Their efficiency determines the electricity required and, therefore, the running cost. The gain of the final power amplifier determines the number of stages required in the RF amplifier chain. The size and weight of the amplifiers determines the space required and can, therefore, have an influence on the size and cost of the tunnel in which the accelerator is installed.

All RF power amplifiers for high-power hadron accelerators employ vacuum tube technology. Vacuum tubes use d.c., or pulsed, voltages from several kilovolts to hundreds of megavolts depending upon the type of tube, the power level and the frequency. The electron velocities can be comparable with the velocity of light and the critical tube dimensions are therefore comparable with the free-space wavelength at the working frequency. Vacuum tubes can therefore generate RF power outputs up to 1 MW continuous wave (c.w.) and 150 MW pulsed. The types employed in high-power hadron accelerators are gridded tubes (triodes and tetrodes) and klystrons. Table 1 shows the parameters of the RF power amplifiers used in some existing or projected high-power hadron accelerators. In the future it is possible that inductive output tubes (IOTs) and magnetrons could be used for this purpose. The magnetron is an oscillator rather than an amplifier and its use is currently restricted to medical linacs. This paper reviews the state of the art of these types of amplifier and discusses some of the factors affecting their successful operation.

Table 1a: RF power amplifiers for c.w. hadron accelerators

Lab	*Accelerator*	*Type*	*RF source*	*Frequency (MHz)*	*Power (kW)*
RIKEN	RIBF SRC	Cyclotron	Tetrode	18 to 42	150
TRIUMF	TRIUMF	Cyclotron	Tetrode	23.06	125
PSI	PSI	Cyclotron	Tetrode	50	850
IFMIF	IFMIF	Linac	Diacrode	175	1000
CERN	SPS (Philips)	Synchrotron	Tetrode	200	35
CERN	SPS (Siemens)	Synchrotron	Tetrode	200	125
CERN	LHC	Synchrotron	Klystron	400	300

Table 1b: RF power amplifiers for pulsed hadron accelerators

Laboratory	*Accelerator*	*Type*	*RF Source*	*Frequency (MHz)*	*Power (MW pk)*	*Duty*
RAL	ISIS Synchrotron	Synchrotron	Tetrode	1.3 to 3.1	1	50 %
GSI	FAIR UNILAC	Linac	Tetrode	36	2	50 %
GIST	FAIR UNILAC	Linac	Tetrode	108	1.6	50 %
RAL	ISIS Linac	Linac	Triode	202.5	5	2 %
GSI	FAIR Linac	Linac	Klystron	325	2.5	0.08 %
ESS	ESS DTL	Linac	Klystron	352.2	1.3 and 2.5	5 %
ORNL	SNS RFQ & DTL	Linac	Klystron	402.5	2.5	8 %
ESS	ESS Elliptical	Linac	Klystron	704.4	2	4 %
ORNL	SNS CCL	Linac	Klystron	805	5	9 %

Solid state RF power transistors operate at voltages from tens to hundreds of volts. The electron mobility is much less in semiconductor materials than in vacuum and the device sizes are therefore small and the power which can be generated by a single transistor is of the order of hundreds of Watts continuous and up to 1 kW pulsed. Large numbers of transistors must be operated in parallel to reach even the lowest power levels required for accelerators. At the present time solid-state amplifiers are not able to meet the final power amplifier requirements for high-power hadron accelerators [2].

2 Tetrode amplifiers

Tetrode vacuum tubes are well established as high-power RF sources in the very-high-frequency (VHF; 30 MHz to 300 MHz) band. The arrangement of a 150 kW, 30 MHz, tetrode is shown in Fig. 2 (see also.Ref. [3]). The construction is coaxial with the cathode inside and the anode outside. The output power available from such a tube is limited by the maximum current density available from the cathode and by the maximum power density which can be dissipated by the anode. The length of the anode must be much shorter than the free-space wavelength of the signal to be amplified to avoid variations in the signal level along it. The perimeter of the anode must likewise be much shorter than the free-space wavelength to avoid the excitation of azimuthal higher-order modes in the space between the anode and the screen grid. The spacings between the electrodes must be small enough for the transit time of an electron from the cathode to the anode to be much shorter than the RF period. If attempts are made to reduce the transit time by raising the anode voltage then there may be flashover between the electrodes. The bulk of the heat which must be dissipated arises from the residual kinetic energy of the electrons as they strike the anode. Thus, provision must be made for air or liquid cooling of the anode (see Section 7.1). Note the substantial copper anode with channels for liquid cooling shown in Fig. 2. For further information on gridded tubes, see Ref. [4].

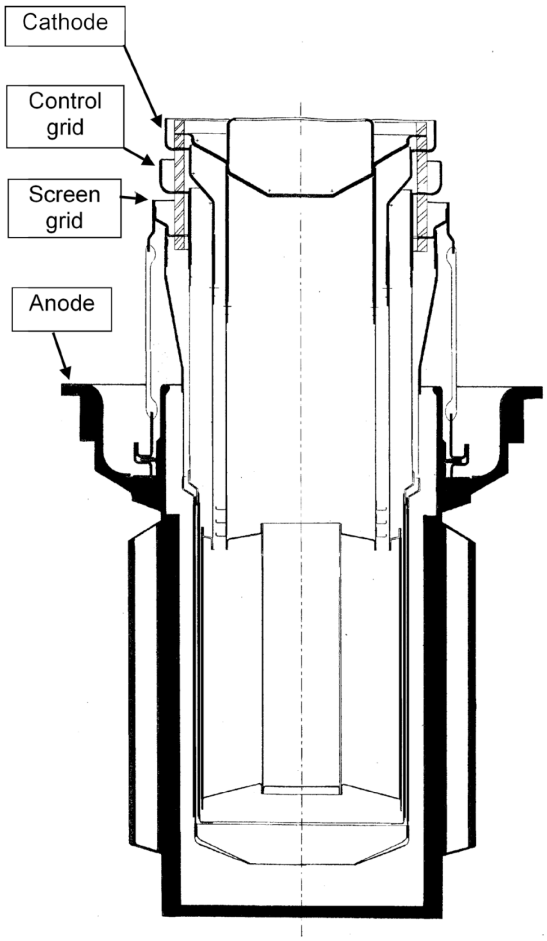


Fig. 2: Cross-sectional view of a high-power tetrode (courtesy of e2v technologies)

The current of electrons emitted from the cathode surface is controlled by the field of the control grid modified by those of the other two electrodes. The dependence of the anode current on the electrode voltages is given approximately by

$$I_a \approx C\left(V_{g1} + \frac{V_{g2}}{\mu_2} + \frac{V_a}{\mu_a}\right)^n \tag{5}$$

where V_{g1}, V_{g2} and V_a are, respectively, the potentials of the control grid, the screen grid and the anode with respect to the cathode and C, μ_1, μ_2 and n are constants [5]. Typically $\mu_2 \sim$ 5–10, $\mu_a \sim$ 100–200 and n is in the range 1.5–2.5.

Figure 3 shows the characteristic curves of a typical tetrode [6]. The control grid voltage is plotted against the anode voltage, both being referred to the cathode. The three sets of curves show the anode current (solid lines), control grid current (dashed lines) and the screen grid current (chain dotted lines). The voltages of the anode (known as the plate in the USA) and the screen grid are positive with respect to the cathode. The curves shown are for a fixed screen grid voltage of +900 V. It is clear that the anode current depends strongly on the control grid voltage and more weakly on the anode voltage. The control grid voltage is normally negative with respect to the cathode to prevent electrons being collected on the grid with consequent problems of heat dissipation. The screen grid, which is maintained at RF ground, prevents capacitive feedback from the anode to the control grid. The screen grid voltage is typically about 10 % of the d.c. anode voltage. If the anode voltage falls below that of the screen grid then any secondary electrons liberated from the anode are collected by the screen grid.

Thus, when tetrodes are operated as power amplifiers the anode voltage is always greater than the screen grid voltage.

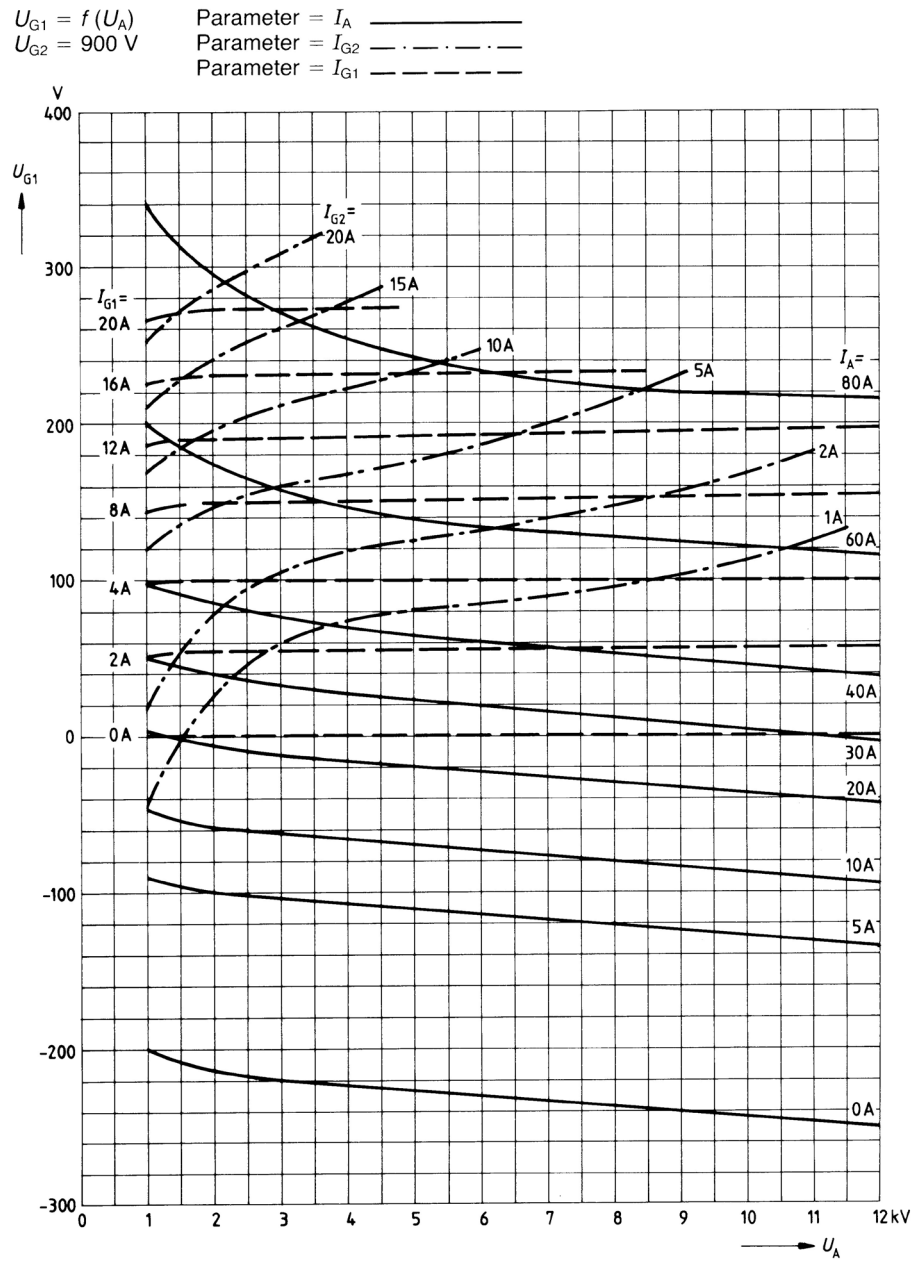


Fig. 3: Characteristic curves of the RS 2058 CJ tetrode for V_{g2} = 900 V (courtesy of Siemens AG)

2.1 Tetrode amplifier circuits

Figure 4 shows the circuit of a grounded-grid tetrode amplifier with a tuned anode circuit. At low frequencies, a resistive anode load may be used but this is unsatisfactory in the VHF band and above because of the effects of parasitic capacitance. The amplifiers used in accelerators are operated at a single frequency at any one time so the limited bandwidth of the tuned anode circuit is not a problem. At the resonant frequency the load in the anode circuit comprises the shunt resistance of the resonator (R_S) in parallel with the load resistance (R_L). If the load impedance has a reactive component then it merely detunes the resonator and can easily be compensated for. The d.c. electrode potentials are maintained by the power supplies shown and the capacitors provide d.c. blocking and RF bypass.

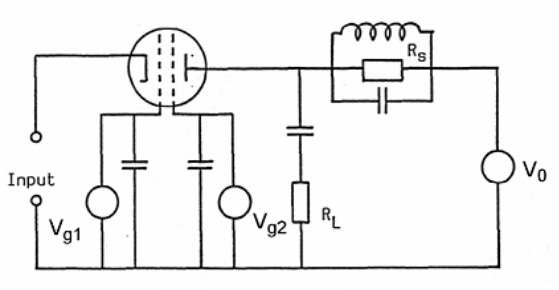


Fig. 4: Circuit of a grounded-grid tetrode amplifier

2.2 Class B operation

The operation of tuned amplifiers is designated as class A, B or C according to the fraction of the RF cycle for which the tube is conducting. Amplifiers for accelerators are normally operated in, or close to, class B in which the tube conducts for half of each cycle. The d.c. bias on the control grid is set so that the anode current is just zero in the absence of a RF input signal. This ensures that the tube only conducts during the positive half-cycle of the RF input voltage and the conduction angle is 180°. Figure 5 illustrates the current and voltage waveforms, normalized to their d.c. values, under the simplifying assumptions that the behaviour of the tube is linear and that the anode load is tuned to resonance.

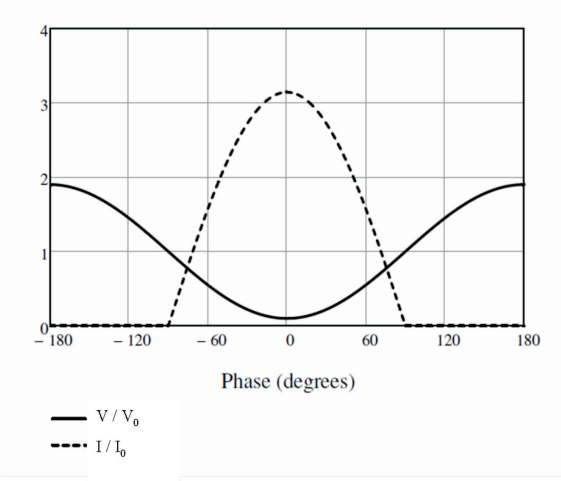


Fig. 5: Current and voltage waveforms in a tetrode amplifier operating in class B

If it is assumed that the tube is linear whilst it is conducting, the d.c. anode current is found by Fourier analysis of the current waveform in Fig. 5 to be

$$I_0 = \frac{I_{pk}}{\pi} = 0.318 \tag{6}$$

where I_{pk} is the peak current during the cycle. Similarly the RF anode current is given by

$$I_2 = 0.5\, I_{pk} \tag{7}$$

In an ideal class B amplifier the minimum RF voltage is zero so that the relationship between the RF and d.c. anode voltages is

$$V_2 = V_0 \tag{8}$$

The d.c power is given by

$$P_0 = V_0 I_0 = \frac{V_0 I_{pk}}{\pi} \tag{9}$$

and the RF output power by

$$P_2 = \frac{1}{2} V_2 I_2 = \frac{1}{4} V_0 I_{pk} \tag{10}$$

The theoretical efficiency is then

$$\eta_e = \frac{P_2}{P_0} = \frac{\pi}{4} = 78.5\% \tag{11}$$

In practice the efficiency of a tetrode amplifier is less than the theoretical limit for two reasons. First, the non-linear relationship between the sinusoidal control-grid voltage and the anode current shown by Eq. (5) means that the constants in Eqs. (6) and (7) become 0.278 and 0.458 when $n = 1.5$ and 0.229 and 0.397 when $n = 2.5$. Second, the need to ensure that the anode voltage always exceeds the screen grid voltage means that the amplitude of the RF anode voltage V_2 is limited to around 90 % of V_0. Substitution of these revised figures into Eqs. (9)–(11) shows that the practical efficiency may be expected to lie in the range 74 % to 78 % depending on the value of *n*.

2.3 Class A, AB and C operation

The proportion of the RF cycle during which the tetrode is conducting can be adjusted by changing the d.c. bias voltage on the control grid. The operation then falls into one of a number of classes as shown in Table 2. If the negative grid bias is reduced then anode current flows when there is no RF drive. The application of a small RF drive voltage produces class A amplification in which the tube conducts throughout the RF cycle. As the RF drive voltage is increased the tube becomes cut off for part of the cycle and the operation is intermediate between class A and class B and known as class AB. When the grid bias is made more negative than that required for class B operation the tube conducts for less than half the RF cycle and the operation is described as class C. Analysis similar to that given above shows that the ratio of the RF anode current to the d.c. anode current increases as the conduction angle is reduced and, therefore, the efficiency increases. However, the amplitude of the RF drive voltage required to produce a given amplitude of the RF anode current increases as the conduction angle is reduced so that the gain of the amplifier decreases. Finally, the harmonic content of the RF anode current waveform increases as the conduction angle increases. The properties of the different classes of amplifier are summarized in Table 2. The amplifiers used for particle accelerators should ideally have high efficiency, high gain and low harmonic output. For this reason it is usual to operate them in class B or in class AB with a conduction angle close to 180°.

Table 2: Classes of amplifier

Class	***Conduction angle***	***Maximum theoretical efficiency***	***Negative grid bias increasing***	***Gain increasing***	Harmonics increasing
A	360°	50 %	↓	↑	↓
AB	180° to 360°	50 % to 78 %			
B	180°	78 %			
C	< 180°	78 % to 100 %			

2.4 Tetrode amplifier design

The performance of a tetrode amplifier is best explained by means of an example. This is based upon a 62 kW, 200 MHz amplifier used in the CERN SPS [7]. The amplifier uses a single RS2058CJ tetrode [6] operating with a d.c. anode voltage of 10 kV and 900 V screen grid bias.

The actual amplifier is operated in class AB but quite close to class B. For simplicity class B operation is assumed in the calculations which follow. The first stage is to estimate the probable efficiency of the amplifier. We assume that the minimum anode voltage is 1.5 kV. Scaling the figures given above which take account of the non-linearity of the tetrode suggests that the efficiency of the amplifier will lie in the range 70 % to 74 %. Let us assume that the efficiency is 72 %. This figure can be adjusted later, if necessary, when the actual efficiency has been calculated. Then the d.c. power input necessary to obtain the desired output power is

$$P_0 = 62 / 0.72 = 86 \text{ kW} \tag{12}$$

The d.c. anode voltage was chosen to be 10 kV so the mean anode current is

$$I_0 = 86 / 10 = 8.6 \text{ A} \tag{13}$$

The theoretical value of I_{pk} is given by Eq. (6) but when the non-linearity of the tetrode is taken into account it is found that this figure lies in the range 3.6–4.4 depending upon the value of n. If we take the factor to be 4.0 then

$$I_{pk} = 4.0 \times 8.6 = 34 \text{ A} \tag{14}$$

Next we construct the load line on the characteristic curves for the tube shown in Fig. 6 by joining the point 1.5 kV, 34 A to the quiescent point (10 kV, 0 A) We note that this requires the control grid voltage to swing slightly positive with a maximum of +70 V.

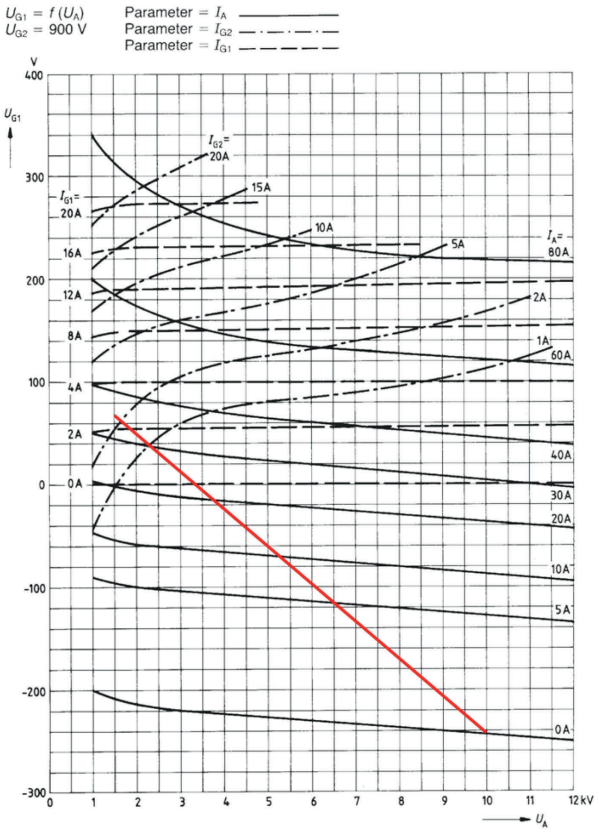


Fig. 6: Tetrode characteristic curves with load line (courtesy of Siemens AG)

The d.c. and RF currents are found by numerical Fourier analysis of the anode current waveform using values read from Fig. 7 at intervals of 15° with the results

$$I_0 = 8.9 \text{ A} \tag{15}$$

$$I_2 = 15.0 \text{ A} \tag{16}$$

Thus, the d.c. input power is

$$P_0 = I_0 V_0 = 89 \text{ kW} \tag{17}$$

The amplitude of the RF voltage is

$$V_2 = 10.0 - 1.5 = 8.5 \text{ kV} \tag{18}$$

and the RF output power is

$$P_2 = \frac{1}{2} V_2 I_2 = 64 \text{ kW} \tag{19}$$

which is very close to the desired value and gives an efficiency of 72 % as originally assumed. The effective load resistance is

$$R_L = V_2 / I_2 = 570 \, \Omega \tag{20}$$

The source impedance of the output of the amplifier can be found by noting that if the RF load resistance is zero the anode voltage is constant and the peak anode current is 46 A for the same RF voltage on the control grid. Thus, the short circuit RF current is 20 A and the anode source resistance (R_a) is 1.7 kΩ.

To find the input impedance of the amplifier we note that the amplitude of the RF control grid voltage is

$$V_1 = 245 + 70 = 315 \text{ V} \tag{21}$$

and that for grounded grid operation the amplitude of the RF input current is

$$I_1 = I_2 + I_{g1\,RF} \approx I_2 \tag{22}$$

The amplitude of the RF control grid current (I_{g1RF}) may be obtained by reading the control grid currents off Fig. 7 at 15° intervals and using numerical Fourier analysis. The result is 0.67 A which is small compared with the RF anode current and can be neglected in the first approxmation. The RF input resistance is

$$R_1 = V_1 / I_1 = 20 \, \Omega \tag{23}$$

Finally we note that the input power is

$$P_1 = \frac{1}{2} V_1 I_1 = 2.5 \text{ kW} \tag{24}$$

and that the power gain of the amplifier is

$$\text{Gain} = 10 \log (64 / 2.5) = 14 \text{ dB} \tag{25}$$

Table 3 shows a comparison between the figures calculated above and those reported in Ref. [7]. The differences between the two columns of Table 4 are attributable to the difference between the actual class AB operation and the class B operation assumed in the calculations.

Table 3: Comparison between actual and calculated parameters of the amplifier described in Ref. [7]

Parameter	*Actual*	*Calculated*	
V_0	10	10	kV
I_0	9.4	8.9	A
V_{g2}	900	900	V
V_{g1}	-200	-245	V
P_{out}	62	64	kW
P_{in}	1.8	2.5	kW
Gain	15.4	14	dB
η	64	72	%

2.5 Practical details

Figure 7 shows a simplified diagram of a tetrode amplifier. The tube is operated in the grounded grid configuration with coaxial input and output circuits. The outer conductors of the coaxial lines are at ground potential and they are separated from the grids by d.c. blocking capacitors. The anode resonator is a re-entrant coaxial cavity which is separated from the anode by a d.c. blocking capacitor. The output power is coupled through an impedance matching device to a coaxial line. The anode HT connection and cooling water pipes are brought in through the centre of the resonator.

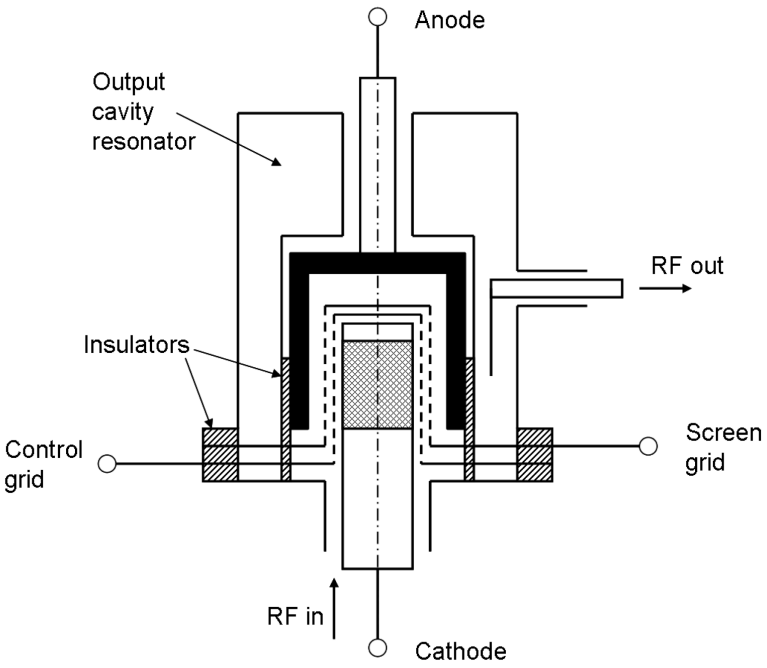


Fig. 7: Arrangement of a tetrode amplifier

The electrodes of the tube form coaxial lines with characteristic impedances of a few Ohms. We have seen above that the input impedance of the amplifier is typically a few tens of Ohms and the output impedance a few hundred Ohms. Thus, the terminations of both the input and output lines are close to open circuits. The anode resonator therefore has one end open circuited and the other short circuited and it must be an odd number of quarter wavelengths long at resonance. Typically the resonator is 3/4 of a wavelength long. In that case the point at which the output coaxial line is coupled into the resonator can be used to transform the impedance to provide a match.

2.6 Operation of tetrode amplifiers in parallel

When higher powers are required than can be obtained from a single tube then it is possible to operate several tubes in parallel. Two such systems are described in Ref. [8]. The original four 500 kW, 200 MHz power amplifiers for the CERN SPS each comprised four 125 kW tetrodes operating in parallel. Figure 8 shows the arrangement of one amplifier. The loads on the fourth arms of the 3 dB couplers normally receive no power. If one tube fails, however, they must be capable of absorbing the power from the unbalanced coupler. The amplifier also contains coaxial transfer switches (not shown in Fig. 8) which make it possible for a faulty tube to be completely removed from service. The remaining three tubes can then still deliver 310 kW to the load. A more recent design of 500 kW amplifier for the same accelerator employs sixteen 35 kW units operated in parallel with a seventeenth unit as the driver stage. Both types of amplifier operate at anode efficiencies greater than 55 % and overall efficiencies greater than 45 %.

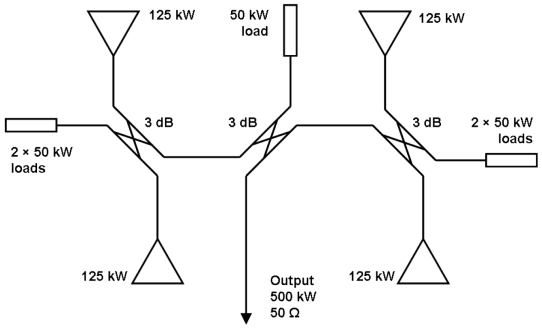


Fig. 8: Arrangement for operation of tetrode amplifiers in parallel (courtesy of Siemens AG)

2.7 The Diacrode ®

A recent development of the tetrode is the Diacrode [9]. In this tube the coaxial line formed by the anode and the screen grid is extended to a short circuit as shown in Fig. 9. The consequence of this change is that the standing wave now has a voltage anti-node, and a current node, at the centre of the active region of the tube. The tetrode, in contrast has a voltage anti-node and current node just beyond the end of the active region, as shown in Fig. 9. Thus, for the same RF voltage difference between the anode and the screen grid, the Diacrode has a smaller reactive current flow and much smaller power dissipation in the screen grid than a tetrode of similar dimensions. This means that, compared with conventional tetrodes, Diacrodes can either double the output power at a given operating frequency or double the frequency for a given power output. The gain and efficiency of the Diacrode are the same as those of a conventional tetrode. Table 4 shows the comparison between the TH 526 tetrode and the TH 628 Diacrode operated at 200 MHz.

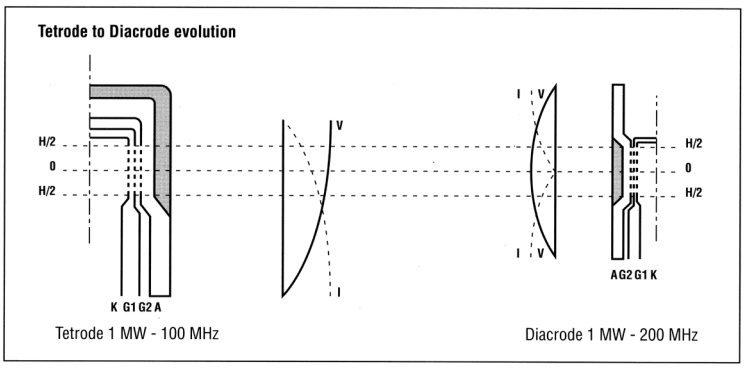


Fig. 9: Comparison between a tetrode and a Diacrode ® (courtesy of Thales Electron Devices)

Table 4: Comparison between the TH 526 tetrode and the TH 628 Diacrode at 200 MHz

Tube	*TH 526*		*TH 628*		
Pulse duration	2.2	c.w.	2.5	c.w.	ms
Peak output power	1600	–	3000	–	kW
Mean output power	240	300	600	1000	kW
Anode voltage	24	11.5	26	16	kV
Anode current	124	75	164	96	A
Peak input power	64.9	–	122.5	–	kW
Mean input power	–	21	–	32	kW
Gain	13.9	11.5	13.9	15	dB

3 Inductive output tubes

The tetrode suffers from the disadvantage that the same electrode, the anode, is part of both the d.c. and the RF circuits. The output power is, therefore, limited by screen grid and anode dissipation. In addition, the electron velocity is lowest when the current is greatest because of the voltage drop across the output resonator (see Fig. 5). To get high power at high frequencies it is necessary to employ high-velocity electrons and to have a large collection area for them. It is therefore desirable to separate the electron collector from the RF output circuit. The possibility that these two functions might be separated from each other was originally recognized by Haeff in 1939 but it was not until 1982 that a commercial version of this tube was described [10]. Haeff called his invention the 'inductive output tube' (IOT) but it is also commonly known by the proprietary name Klystrode ®.

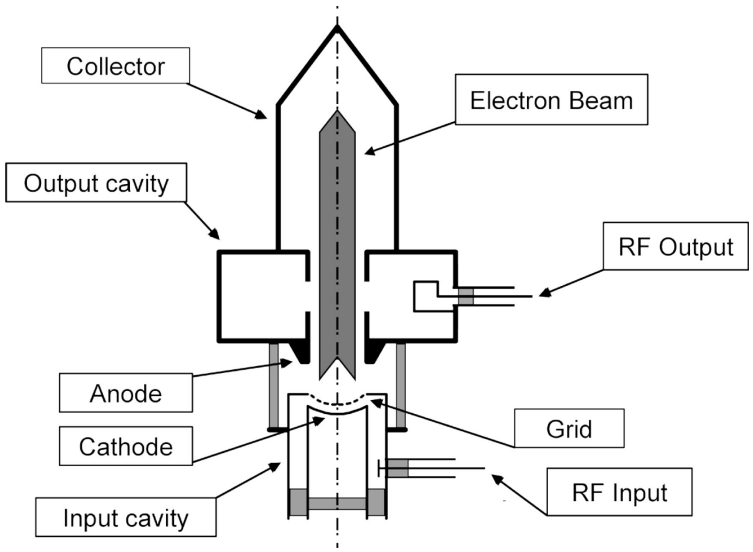


Fig. 10: Arrangement of an IOT (©2010 IEEE, reproduced with permission)

Figure 10 shows a schematic diagram of an IOT [11]. The electron beam is formed by a gridded, convergent flow, electron gun and confined by an axial magnetic field (not shown). The gun is biased so that no current flows except during the positive half-cycle of the RF input. Thus, electron bunches are formed and accelerated through the constant potential difference between the cathode and the anode. The bunches pass through a cavity resonator as shown in Fig. 11 so that their azimuthal magnetic field induces a current in the cavity (hence, the name of the tube). Because the cavity is

tuned to the repetition frequency of the bunches, the RF electric field in the interaction gap is maximum in the retarding sense when the centre of a bunch is at the centre of the gap. The interaction between the bunches and the cavity resonator is similar to that in a class B amplifier. Figure 12 shows a plot of the positions of typical electrons against time. The slopes of the lines are proportional to the electron velocities and they show how kinetic energy is extracted from the electron bunches as they pass through the output gap (indicated by dashed lines). The RF power transferred to the cavity is equal to the kinetic power given up by the electrons. Because the electron velocity is high it is possible to use a much longer output gap than in a tetrode. The RF power passes into and out of the vacuum envelope through ceramic windows.

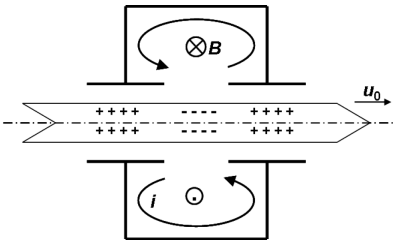


Fig. 11: Interaction between a bunched electron beam and a cavity resonator

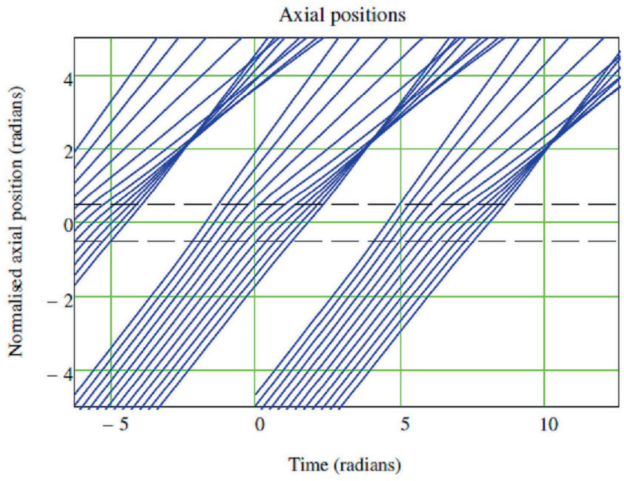


Fig. 12: Electron trajectories in an IOT

The efficiency of an IOT can be estimated by noting that the relationship between the RF and d.c. currents in the beam is, approximately, from Eqs. (6) and (7)

$$I_1 = \frac{\pi}{2} I_0 \tag{26}$$

The effective voltage of the output gap is the product of the RF gap voltage and a transit time factor. The effective gap voltage cannot be greater than about 90 % of the voltage used to accelerate the electrons because the electrons leaving the gap must have sufficient residual velocity to enable them to leave the gap and pass into the collector. The maximum RF output power is therefore given by

$$P_2 = \frac{1}{2} I_1 V_{gap,eff} = \frac{0.9}{2} \cdot \frac{\pi}{2} \cdot I_0 V_0 = 0.71 P_0 \tag{27}$$

so that the maximum efficiency is approximately 70 %.

The advantages of the IOT are that it does not need a d.c. blocking capacitor in the RF output circuit because the cavity is at ground potential and that it has higher isolation between input and output and a longer life than an equivalent tetrode. These advantages are offset to some extent by the

need for a magnetic focusing field. The typical gain is greater than 20 dB and is appreciably higher than that of a tetrode; high enough in fact for a 60 kW tube to be fed by a solid-state driver stage. IOTs have been designed for ultra-high-frequency (UHF) TV applications. The IOTs designed for use in accelerators are operated in class B or class C. Further information about the IOT can be found in Refs. [10, 12, 13]. Table 5 shows the parameters of some IOTs designed for use in accelerators.

Table 5: Parameters of IOTs for use in accelerators

Tube	***2KDW250PA***	***VKP-9050***	***VKL-9130A***	
Manufacturer	CPI/Eimac	CPI	CPI	
Frequency	267	500	1300	MHz
Beam voltage	67	40	35	kV
Beam current	6.0	3.5	1.3	A
RF output power	280	90	30	kW
Efficiency	70	>65	>65	%
Gain	22	>22	>20	dB

4 Klystrons

At a frequency of 1.3 GHz the continuous output power of an IOT is limited to around 30 kW by the need to use a control grid to modulate the electron beam. At higher frequencies and high powers it is necessary to modulate the beam in some other way. In the klystron this is achieved by passing an un-modulated electron beam through a cavity resonator which is excited by an external RF source. The electrons are accelerated or retarded according to the phase at which they cross the resonator and the beam is then said to be velocity modulated. The beam leaving the gap has no current modulation but, downstream from the cavity, the faster electrons catch up the slower electrons so that bunches of charge are formed as shown in Fig. 13.

When an output cavity, tuned to the signal frequency, is placed in the region where the beam is bunched, the result is the simple two-cavity klystron illustrated in Fig. 14. RF power is induced in the second cavity in exactly the same way as in an IOT. This cavity presents a resistive impedance to the current induced in it by the electron beam so that the phase of the field across the gap is in anti-phase with the RF beam current. Electrons which cross the gap within ±90° of the bunch centre are retarded and give up energy to the field of the cavity. Since more electrons cross the second gap during the retarding phase than the accelerating phase there is a net transfer of energy to the RF field of the cavity. Thus, the klystron operates as an amplifier by converting some of the d.c. energy input into RF energy in the output cavity.

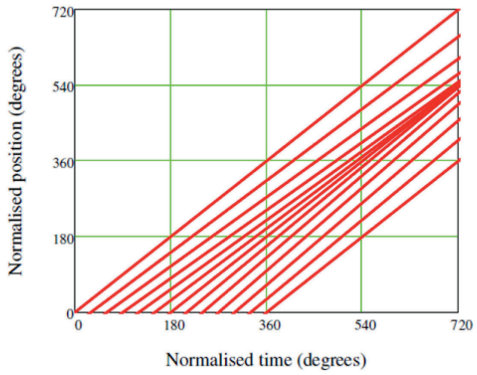


Fig. 13: Applegate diagram showing the formation of bunches in a velocity modulated electron beam

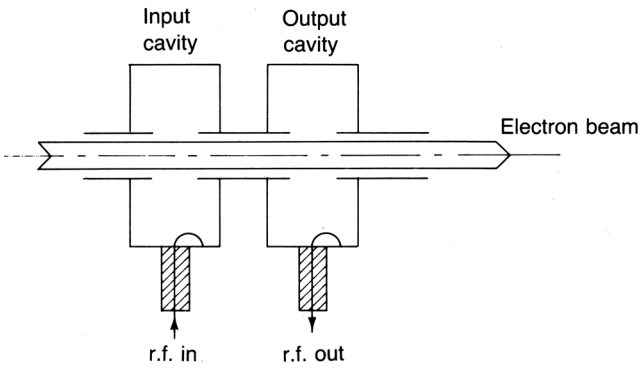


Fig. 14: Arrangement of a two-cavity klystron

In practice, the gain and efficiency of a two-cavity klystron are too low to be of practical value. It is therefore usual to add further cavity resonators to increase the gain, efficiency and bandwidth of the tube. Figure 15 shows the arrangement of a multicavity klystron. The electron beam is formed by a diode electron gun for which

$$I_0 = K V_0^{1.5} \tag{28}$$

where K is a constant known as the perveance which, typically, has a value in the range $0.5 \text{ to } 2.0 \times 10^{-6}$ $\text{AV}^{-1.5}$. The function of all of the cavities, except the last, is to form tight electron bunches from which RF power can be extracted by the output cavity. The first and last cavities are tuned to the centre frequency and have Q factors which are determined largely by the coupling to the input and output waveguides. The intermediate, or idler, cavities normally have high Q and are tuned to optimize the performance of the tube. The long electron beam is confined by an axial magnetic field to avoid interception of electrons on the walls of the drift tube. The spent electrons are collected by a collector in exactly the same way as in an IOT. The RF power passes into and out of the vacuum envelope through ceramic windows.

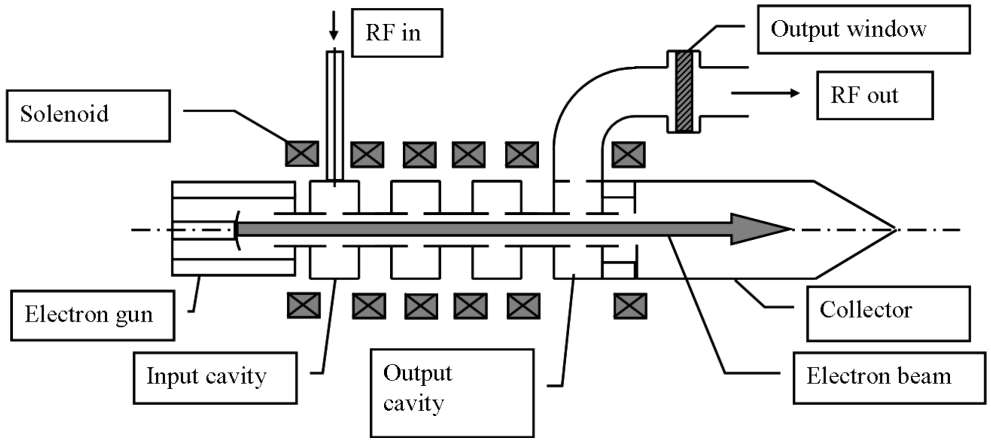


Fig. 15: Arrangement of a multicavity klystron

4.1 Electron bunching in klystrons

The Applegate diagram in Fig. 13 ignores the effect of space-charge on the bunching. The space-charge forces oppose the bunching and, under small-signal conditions, the beam has current modulation but no velocity modulation at the plane of the bunch. As the beam drifts further the space-charge forces cause the bunches to disperse and reform periodically. From the point of view of an observer travelling with the mean electron velocity, the electrons would appear to be executing

oscillations about their mean positions at the electron plasma frequency. The plasma frequency is modified to some extent by the boundaries surrounding the beam and by the presence of the magnetic focusing field. The electron plasma frequency is given by

$$\omega_p = \left(\eta\rho/\varepsilon_0\right)^{0.5} \tag{29}$$

where η is the charge to mass ratio of the electron and ρ is the charge density in the beam. The distance from the input gap to the first plane at which the bunching is maximum is then a quarter of a plasma wavelength $\left(\lambda_p\right)$ given by

$$\lambda_p = 2\pi\, u_0 / \omega_p \tag{30}$$

where u_0 is the mean electron velocity. Theoretically the second cavity should be placed at a distance $\lambda_p/4$ from the input gap so that the induced current in the second cavity is maximum. In practice, it is found that this would make a tube inconveniently long and the distance between the gaps is a compromise between the strength of interaction and the length of the tube.

The bunching length is independent of the input signal except at very high drive levels when it is found that it is reduced. If attempts are made to drive the tube still harder the electron trajectories cross over each other and the bunching is less. Figure 16 shows a typical Applegate diagram for a high-power klystron. It should be noted that, in comparison with the diagram in Fig. 13, the axes have been exchanged and uniform motion of the electrons at the initial velocity has been subtracted. The peak accelerating and retarding phases of the fields in the cavities are indicated by + and – signs. Those electrons which cross the input gap at an instant when the field is zero proceed without any change in their velocities and appear as horizontal straight lines. Retarded electrons move upwards and accelerated electrons move downwards in the diagram. Because the cavities are closely spaced space-charge effects are not seen until the final drift region.

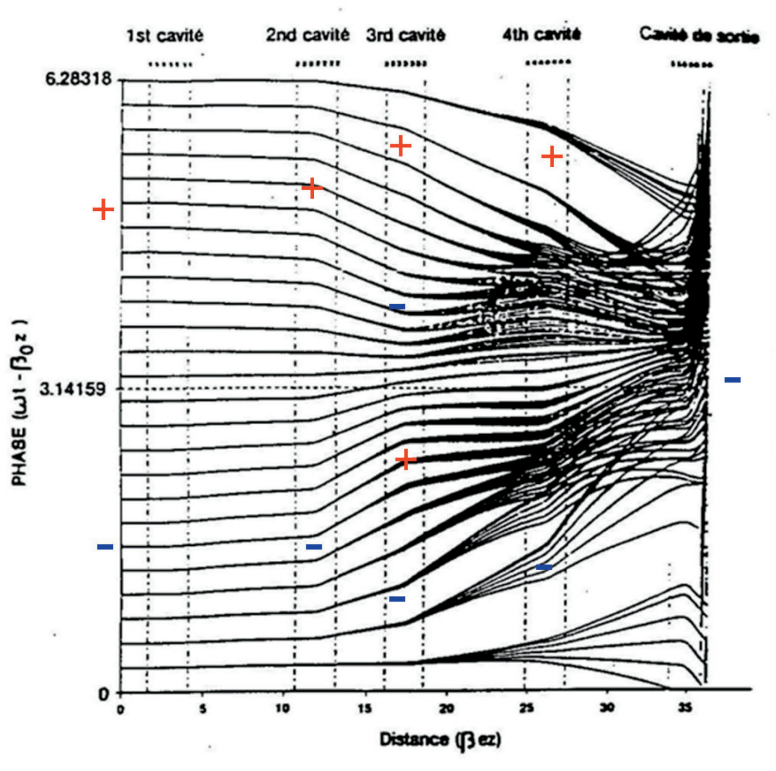


Fig. 16: Applegate diagram for a high-efficiency klystron (courtesy of Thales Electron Devices)

The tube illustrated has five cavities. The bunching produced by the first cavity is imperceptible on the scale of this diagram but it is sufficient to excite the RF fields in the second cavity. The second cavity is tuned to a frequency which is above the signal frequency so that it presents an inductive impedance to the beam current. As a result the bunch centre coincides with the neutral phase of the field in the cavity and further velocity modulation is added to the beam which produces much stronger bunching at the third cavity. The third cavity is tuned to the second harmonic of the signal frequency as can be seen from a careful examination of the diagram. The principal purpose of this cavity is to cause the electrons which lie farthest from the bunch centre to be gathered into the bunch. The use of a second harmonic cavity increases the efficiency of a klystron by at least ten percentage points. The splitting of the lines in the diagram which occurs at this plane is caused by a divergence in the behaviour of electrons in different radial layers within the electron beam. The fourth cavity is similar to the second cavity and produces still tighter bunching of the electrons. By the time they reach the final cavity nearly all of the electrons are bunched into a phase range which is ±90° with respect to the bunch centre. The output cavity is tuned to the signal frequency so that the electrons at the bunch centre experience the maximum retarding field and all electrons which lie within a phase range of ±90° with respect to the bunch centre are also retarded. If the impedance of the output cavity is chosen correctly then a very large part of the kinetic energy of the bunched beam can be converted into RF energy. It should be noted that space-charge repulsion ensures that the majority of trajectories are nearly parallel to the axis at the plane of the output gap so that the kinetic energy of the bunch is close to that in the initial unmodulated beam.

4.2 Efficiency of klystrons

The output power of a klystron is given by

$$P_2 = \frac{1}{2} I_1 V_{eff} \tag{31}$$

where I_1 is the first harmonic RF beam current at the output gap and V_{eff} is the effective output gap voltage. As in the case of the IOT the effective output gap voltage must be less than 90 % of V_0 to ensure that the electrons have sufficient energy to leave the gap and enter the collector. In the IOT the peak current in the bunch cannot exceed the maximum instantaneous current available from the cathode and the maximum value of I_1 is approximately equal to half the peak current. In the klystron, however, the d.c. beam current is equal to the maximum current available from the cathode and the bunches are formed by compressing the charge emitted in one RF cycle into a shorter period. In the theoretical limit the bunches become delta functions for which

$$I_1 = 2I_0 \tag{32}$$

Thus, the maximum possible value of I_1 in a klystron is four times that in an IOT with the same electron gun. The factor is actually greater than this because the current available from the triode gun in an IOT is less than that from the equivalent diode gun in a klystron. In practice, the effects of space charge mean that the limit given by Eq. (32) is not attainable, but computer simulations have shown that the ratio I_1/I_0 can be as high as 1.6 to 1.7 at the output cavity. Then, by substitution in Eq. (31), we find that efficiencies of up to 75 % should be possible.

It is to be expected that the maximum value of I_1/I_0 will decrease as the space-charge density in the beam increases. An empirical formula for the dependence of efficiency on beam perveance derived from studies of existing high-efficiency klystrons is given in Ref. [14]:

$$\eta_e = 0.9 - 0.2 \times 10^{-6} K \tag{33}$$

If it is assumed that the limit $K = 0$ corresponds to delta function bunches then it can be seen that Eq. (33) takes the maximum effective gap voltage to be $0.9 V_0$.

The maximum efficiency of klystrons decreases with increasing frequency because of increasing RF losses and of the design compromises which are necessary. This is illustrated by Fig. 17 which shows the efficiencies of c.w. klystrons taken from manufacturers' data sheets. It should be emphasized that the performance of most of these tubes will have been optimized for factors other than efficiency.

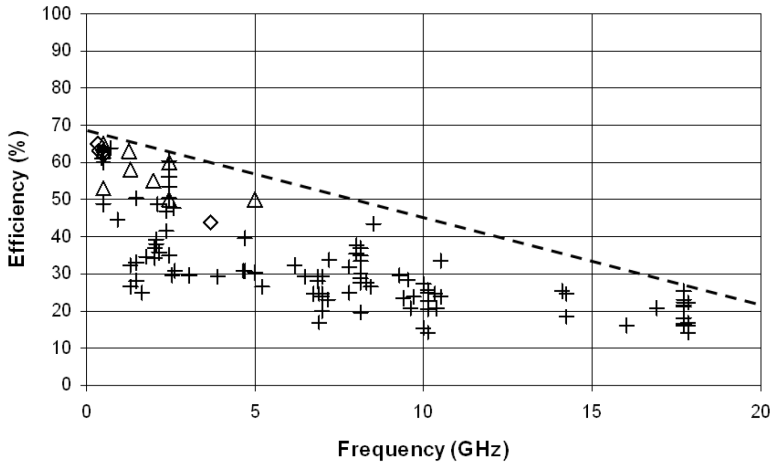


Fig. 17: Efficiencies of c.w. klystrons

4.3 Terminal characteristics of klystrons

The transfer characteristics of a klystron (Fig. 18) show that the device is a linear amplifier at low signal levels but that the output saturates at high signal levels. The performance of a klystron is appreciably affected by variations in the beam voltage, signal frequency and output match and we now examine these in turn.

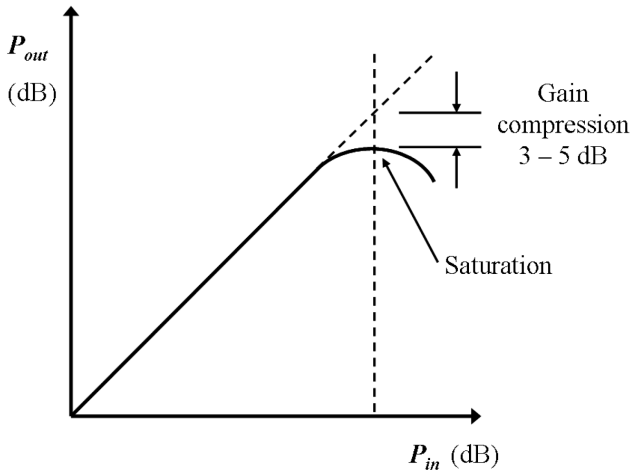


Fig. 18: Klystron transfer characteristics

Klystrons for use in accelerators are normally operated close to saturation to obtain the highest possible efficiency. Figure 18 shows that the output power is then insensitive to variations of input power and, by extension, to variations of beam voltage. The effects on the phase of the output signal are more serious because of the distance from the input to the output.

The output power and efficiency of a klystron are affected by the match of the load which is normally a circulator. This is usually represented by plotting contours of constant load power on a Smith chart of normalised load admittance. Figure 19 shows such a chart, known as a Rieke diagram, for a typical klystron. Care must be taken to avoid the possibility of voltage breakdown in the output gap. If the gap voltage becomes too high it is also possible for electrons to be reflected so reducing the efficiency of the tube and providing a feedback path to the other cavities which may cause the tube to become unstable. The forbidden operating region is shown by shading on the diagram. A further complication is provided by the effect of harmonic signals in the output cavity. Since the klystron is operated in the non-linear regime to obtain maximum efficiency it follows that the signal in the output waveguide will have harmonic components. These are incompletely understood but it is known that the reflection of harmonic signals from external components such as a circulator can cause the klystron output to behave in unexpected ways.

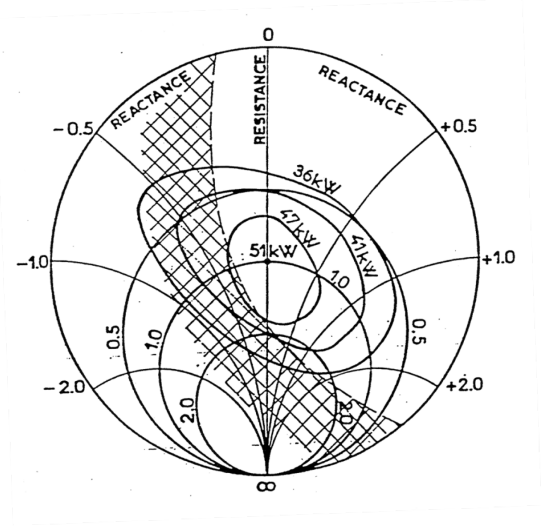


Fig. 19: Rieke diagram for a klystron (courtesy of Thales Electron Devices)

4.4 Typical super-power klystrons

Klystrons which have been developed specifically for use in accelerators are commonly known as super-power klystrons. Tables 6 and 7 summarize the state of the art for these tubes. The beam voltage is limited by the need to avoid voltage breakdown in the electron gun. It can be seen from the tables that the typical beam voltages are higher for pulsed tubes than for c.w. tubes because the breakdown voltage is higher for short pulses than for steady voltages. The beam current is limited by the current density which is available at the cathode and by the area of the cathode which decreases with frequency. The saturation current density of thermionic cathodes is greater for short (microsecond) pulses than for d.c. operation.

Table 6: Characteristics of typical c.w. super-power klystrons

Tube	***TH 2089***	***VKP-7952***	***TH 2103C***[a]	
Manufacturer	Thales	CPI	Thales	
Frequency	352	700	3700	MHz
Beam voltage	100	95	73	kV
Beam current	20	21	22	A
RF output power	1.1	1.0	0.7	MW
Gain	40	40	50	dB
Efficiency	65	65	44	%

[a]This tube was developed for heating plasmas for nuclear fusion experiments

Table 7: Characteristics of typical pulsed super-power klystrons

Tube	*Ref. [15]*	*Ref. [16]*	*Ref. [17]*	
Frequency	2.87	3.0	11.4	GHz
Pulse length	1.0	1.0	1.6	μs
Beam voltage	475	610	506	kV
Beam current	620	780	296	A
RF output power	150	213	75	MW
Gain	59	58	60	dB
Efficiency	51	44	50	%

4.5 Multiple-beam klystrons

We have seen that the efficiency of a klystron is determined by the perveance of the electron beam so that, to get high efficiency, it is necessary to use a high-voltage, low-current beam. The use of high voltages produces problems with voltage breakdown and it is therefore difficult to obtain very high power with high efficiency. One solution to this problem is to use several electron beams within the same vacuum envelope as shown in Fig. 20. A klystron designed in this way is known as a multiple-beam klystron (MBK). The individual beams have low perveance to give high efficiency whilst the output power is determined by the total power in all of the beams. The principle of the MBK has been known for many years [18] but, until recently, the only such tubes constructed were in the former Soviet Union for military applications. The first MBK designed specifically for use in particle accelerators was the Thales type TH1801, the performance of which is shown in Table 8; see also Ref. [19].

Table 8: Characteristics of a MBK

Type	*TH 1801*	
Frequency	1300	MHz
Beam voltage	115	kV
Beam current	133	A
Number of beams	7	
Power	9.8	MW
Pulse length	1.5	ms
Efficiency	64	%
Gain	47	dB

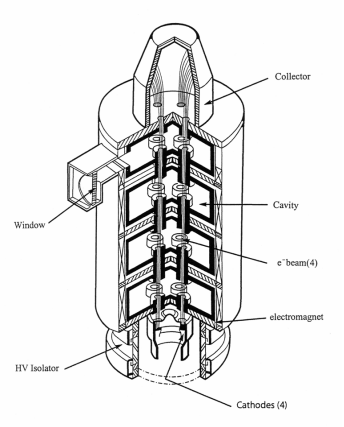


Fig. 20: Arrangement of a multiple beam klystron (courtesy of Thales Electron Devices)

5 Magnetrons

The principle of operation of the magnetron is illustrated in Fig. 21. The tube has a concentric cylindrical geometry. Electrons emitted from the cathode are drawn towards the surrounding anode by the potential difference between the two electrodes. The tube is immersed in a longitudinal magnetic field which causes the electron trajectories to become cycloidal so that, in the absence of any RF fields, the diode is cut off, no current flows, and the electrons form a cylindrical space-charge layer around the cathode. The anode is not a smooth cylinder but carries a number of equally spaced vanes such that the spaces between them form resonant cavities. The anode supports a number of resonant modes with azimuthal RF electric field. The one used for the interaction is the π mode in which the fields in adjacent cavities are in anti-phase with one another. The RF fields in the anode are initially excited by electronic noise and there is a collective interaction between the fields and the electron cloud which causes some electrons to be retarded. These electrons move outwards forming 'spokes' of charge, the number of which is half the number of the cavities in the anode. The spokes rotate in synchronism with the RF field of the anode and grow until electrons reach the anode and current flows through the device. The electron velocities are almost constant during the interaction and the energy transferred to the RF field comes from their change in potential energy. The magnetron is an oscillator whose power output grows until it is limited by non-linearity in the interaction. RF power is extracted from the anode via a coupler and vacuum window.

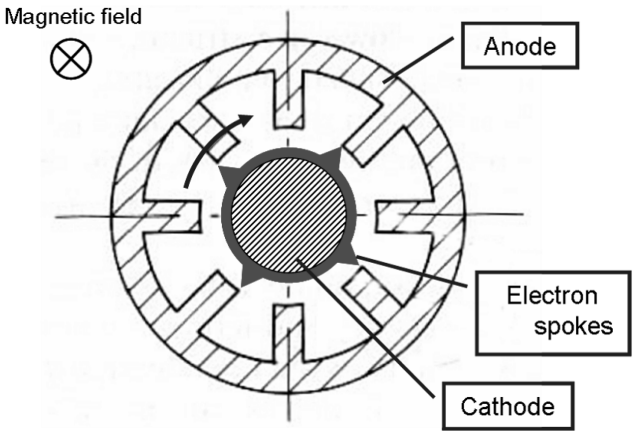


Fig. 21: Arrangement of a magnetron oscillator

The magnetron is a compact device which is capable of achieving efficiencies of up to 90 % and it has been recognized for many years that it would be an attractive alternative to other tubes for powering particle accelerators. However, because it is a free-running oscillator, the frequency is not stable enough for use in most accelerators. The frequency of a magnetron varies with the current flowing through the tube (known as frequency pushing) and it is possible to use this to provide a degree of control. On its own this is not sufficient. It is also possible to lock the phase of a free-running oscillator by injecting RF power at the desired frequency. The power required increases with the difference between the natural frequency and the locked frequency and it is found that the power required to lock the phase of a magnetron is typically about 10 % of the output power of the tube. This power is unacceptably high. Recent work has shown that, when the frequency of a magnetron is first stabilized by a control loop using frequency pushing, it is then possible to lock the phase with an injected RF signal which is less than 0.1 % of the output power of the tube [20]. Thus, locked magnetrons may be used in the future for powering accelerators [21].

6 Limitations of vacuum tubes

The performance of high-power vacuum tubes is limited by a number of factors which operate in much the same way for all devices. The chief of these are heat dissipation, voltage breakdown, output window failure and multipactor discharges.

The dimensions of the RF structures and the windows of microwave tubes generally scale inversely with frequency. The maximum continuous, or average, power which can be handled by a particular type of tube depends upon the maximum temperature that the internal surfaces can be allowed to reach. This temperature is independent of the frequency so the power that can be dissipated per unit area is constant.

The power of a klystron or IOT is also limited by the power in the electron beam. The beam diameter scales inversely with frequency and the beam current density is determined by the maximum attainable magnetic focusing field. Since that field is independent of frequency the beam current scales inversely with the square of the frequency. The beam voltage is related to the current by the gun perveance which usually lies in the range 0.5 to 2.0 for power tubes. The maximum gun voltage is limited by the breakdown field in the gun and so varies inversely with frequency for constant perveance. These considerations suggest that the maximum power obtainable from a tube of a particular type varies as frequency to the power –2.5 to –3.0 depending upon the assumptions made. For pulsed tubes the peak power is limited by the considerations in this paragraph and the mean power by those in the preceding one.

The efficiencies of tubes tend to fall with increasing frequency. This is partly because the RF losses increase with frequency and partly because of the design compromises which must be made at higher frequencies.

The maximum power obtainable from a pulsed tube is often determined by the power-handling capability of the output window. Very-high-power klystrons commonly have two windows in parallel to handle the full output power. Windows can be destroyed by excessive reflected power, by arcs in the output waveguide, by X-ray bombardment and by the multipactor discharges described in the following paragraph. The basic cause of failure is overheating and it is usual to monitor the window temperature and to provide reverse power and waveguide and cavity arc detectors.

Multipactor is a resonant RF vacuum discharge which is sustained by secondary electron emission [22]. Consider a pair of parallel metal plates in vacuum with a sinusoidally varying voltage between them. If an electron is liberated from one of the plates at a suitable phase of the RF field it will be accelerated towards the other plate and may strike it and cause secondary electron emission. If the phase of the field at the moment of impact is just 180° from that at the time when the electron left the first plate, then the secondary electrons will be accelerated back towards the first plate. These conditions make it possible for a stable discharge to be set up if the secondary electron emission coefficients of the surfaces are greater than unity. It is found that phase focusing occurs so that electrons which are emitted over a range of phases tend to be bunched together. It is also possible for multipactor discharges to occur on ceramic surfaces with surface charge providing a static field. It should be noted that this type of discharge is not resonant and does not require the presence of a RF electric field. The local heating of a window ceramic in this way can be sufficient to cause window failure. Signs of multipactor are heating, changed RF performance, window failure and light and X-ray emissions. A multipactor discharge can sometimes be suppressed by changing the shape of the surfaces, by surface coatings, and by the imposition of static electric and magnetic fields.

7 Cooling and protection

7.1 Cooling power tubes

The power tubes used in accelerators typically have efficiencies between 40 % and 70 %. It follows that a proportion of the d.c. input power is dissipated as heat within the tube. The heat to be dissipated

is between 40 % and 150 % of the RF output power provided that the tube is never operated without RF drive. If a linear beam tube is operated without RF drive then the electron collector must be capable of dissipating the full d.c. beam power. The greater part of the heat is dissipated in the anode of a tetrode or in the collector of a linear-beam tube. These electrodes are normally cooled in one of three ways: by blown air (at low power levels), by pumped liquid (usually de-ionized water) or by vapour phase cooling. The last of these may be less familiar than the others and needs a little explanation.

The electrode to be cooled by vapour phase cooling is immersed in a bath of the liquid (normally de-ionized water) which is permitted to boil. The vapour produced is condensed in a heat exchanger which is either within the cooling tank (see Fig. 22) or part of an external circuit. The cooling system therefore forms a closed loop so that water purity is maintained. In all water cooling systems it is important to maintain the water purity to ensure that the electrodes cooled are neither contaminated nor corroded. Either of these effects can degrade the effectiveness of the cooling system and cause premature failure of the tube. In blown air systems careful filtering of the air is necessary for the same reasons.

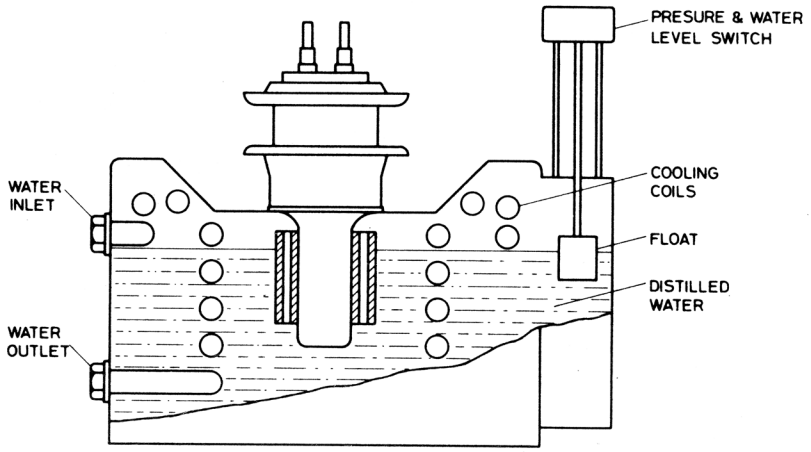


Fig. 22: Vapour phase cooling of a tetrode (courtesy of e2v technologies)

It is important to remember that, in a high-power tube, appreciable quantities of heat may be dissipated on parts of the tube other than the anode or collector especially if a fault occurs during operation. It is common to provide air or water cooling for these regions also. Inadequate cooling may lead to the internal distortion or melting of the tube and its consequent destruction. Further information on the cooling of tubes is given in Refs. [23, 24].

7.2 Tube protection

Power tubes are very expensive devices and it is vital that they are properly protected when in use. The energy densities in the tubes and their power supplies are so high that it is easy for a tube to be destroyed if it is not properly protected. Nevertheless, with adequate protection tubes are in fact very good at withstanding accidental overloads and may be expected to give long, reliable, service.

Two kinds of protection are required. First a series of interlocks must be provided which ensures that the tube is switched on in the correct sequence. Thus, it must be impossible to apply the anode voltage until the cathode is at the correct working temperature and the cooling systems are functioning correctly. The exact switch-on sequence depends upon the tube type and reference must be made to the manufacturer's operating instructions. The sequence must also be maintained if the tube has to be restarted after tripping off for any reason.

The second provision is of a series of trips to ensure that power is removed from the tube in the event of a fault such as voltage breakdown or excessive reflected power. Again the range of parameters to be monitored and the speed with which action must be taken varies from tube to tube. Examples are: coolant flow rate; coolant temperature; tube vacuum; output waveguide reverse power; and electrode over-currents. If a tube has not been used for some time it is sometimes necessary to bring it up to full power gradually to avoid repeated trips. The manufacturer's operating instructions should be consulted about this. If a tube trips out repeatedly it is best to consult the manufacturer to avoid the risk of losing it completely by unwise action taken in ignorance of the possible causes of the trouble. General information about tube protection and safe operation is given in Refs. [23, 24].

8 Conclusion

This paper has provided an introduction to the main types of RF power source which are, or may be, used in high-power hadron accelerators. Figure 23 shows the state of the art in terms of mean or c.w. power output as a function of frequency for the RF power sources currently used in accelerators. Solid-state sources can compete with tubes at the lower frequencies and power levels and are likely to become more commonly used. The fall-off in power output at high frequencies for each type of tube is related to the fundamental principles of its operation as discussed in Section 6. The power achieved by klystrons at low frequencies does not generally represent a fundamental limitation but merely the maximum which has been demanded to date. For tetrodes and solid-state devices the maximum power is probably closer to the theoretical limits for those devices. In any case higher powers can be produced by parallel operation.

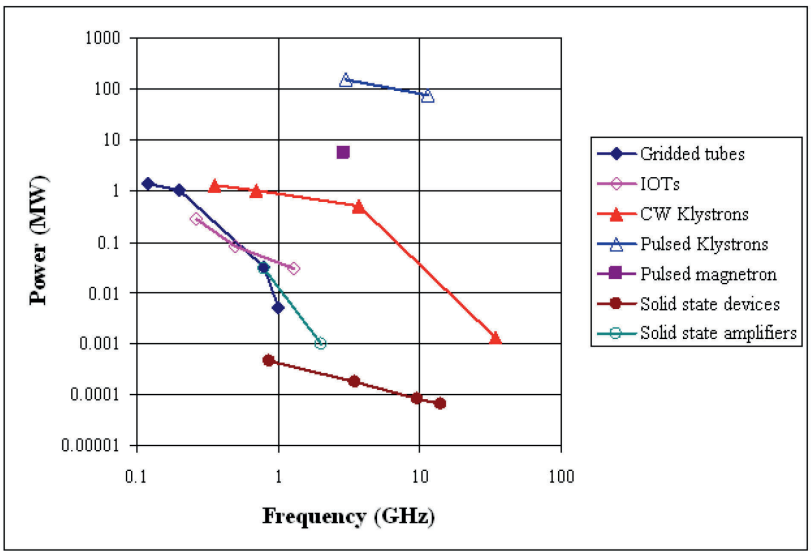


Fig. 23: State of the art of high-power RF sources

For further information on the theory of microwave tubes and for suggestions for background reading, see Refs. [24–26].

References

[1] R.G. Carter, *Contemp. Phys.* **52** (2011) 15–41.
[2] R.G. Carter, R.F. power generation, Proc. of the CERN Accelerator School 'RF for Accelerators', Ebeltoft, Denmark, 8–17 June 2010 (CERN, Geneva, 2011).

[3] CY1172 RF Power Tetrode Data Sheet, EEV Ltd, 1990.
[4] K. Spangenburg, *Vacuum Tubes* (McGraw-Hill, New York, 1948).
[5] A.H.W. Beck, *Thermionic Valves* (Cambridge University Press, Cambridge, 1953).
[6] Siemens, *Transmitting Tubes Data Book 1986/87* (Siemens AG, 1986).
[7] W. Herdrich and H.P. Kindermann, *IEEE Trans. Nucl. Sci.* **32** (1985) 2794–2796.
[8] H. P. Kindermann*, et al.*, *IEEE Trans. Nucl. Sci.* **30** (1983) 3414–3416.
[9] G. Clerc*, et al.*, A new generation of gridded tubes for higher power and higher frequencies, Proc. Particle Accelerator Conference, Vancouver, BC, 1997, pp. 2899–2901.
[10] D. Preist and M. Shrader, *Proc. IEEE* **70** (1982) 1318–1325.
[11] R.G. Carter, *IEEE Trans. Electron Dev.* **57** (2010) 720–725.
[12] H. Bohlen*, et al.*, Inductive output tubes for particle accelerators, Presented at the EPAC 2004, Lucerne, Switzerland, 2004.
[13] H. Bohlen*, et al.*, IOT RF power sources for pulsed and cw linacs, Presented at the LINAC 2004, Lübeck, Germany, 2004.
[14] R.S. Symons, Scaling laws and power limits for klystrons, Proc. International Electron Devices Meeting, 1986, pp. 156–159.
[15] T. Lee*, et al.*, *IEEE Trans. Plasma Sci*. **13** (1985) 545–552.
[16] S. Choroba*, et al.*, Performance of an S-band klystron at an output power of 200MW, Proc. XIX International Linac Conference, Chicago, IL, USA, 1998, pp. 917–919.
[17] D. Sprehn*, et al.*, Current status of the next linear collider X-band klystron development program, Proc. EPAC 2004, Lucerne, Switzerland, 2004, pp. 1090–1092.
[18] M.R. Boyd*, et al.*, *IRE Trans. Electron Dev.* **9** (1962) 247–252.
[19] A. Beunas and G. Faillon, 10 MW/1.5ms, L-band multi-beam klystron, Presented at the Displays and Vacuum Electronics Conference, Garmisch-Partenkirchen, Germany, 1998.
[20] I. Tahir*, et al.*, *IEEE Trans. Electron Dev*. **52** (2005) 2096–2103.
[21] R.G. Carter, Conceptual design of a 1MW 175MHz CW magnetron, Proc. IEEE International Vacuum Electronics Conference 2009 (IVEC '09), pp. 550–551.
[22] J.R.M. Vaughan, *IEEE Trans. Electron Dev*. **35** (1988) 1172–1180.
[23] EEV, Preamble - Tetrodes (EEV Ltd, 1976).
[24] L. Sivan, *Microwave Tube Transmitters* (Kluwer Academic Publishers, Dodrecht, 1994).
[25] R. J. Barker*, et al.*, *Modern Microwave and Millimetre-Wave Power Electronics* (IEEE, Piscataway, NJ, 2005).
[26] J.A. Eichmeier and M. Thumm, Eds., *Vacuum Electronics: Components and Devices* (Springer, Berlin, 2008).

RF Basics I and II

Frank Gerigk
CERN, Geneva, Switzerland

Abstract
Maxwell's equations are introduced in their general form, together with a basic set of mathematical operations needed to work with them. After simplifying and adapting the equations for application to radio frequency problems, we derive the most important formulae and characteristic quantities for cavities and waveguides. Several practical examples are given to demonstrate the use of the derived equations and to explain the importance of the most common figures of merit.

1 Introduction to Maxwell's equations

1.1 Maxwell's equations

Maxwell's equations were published in their earliest form in 1861–1862 in a paper entitled "On physical lines of force" by the Scottish physicist and mathematician James Clerk Maxwell. They represent a uniquely complete set of equations that covers all areas of electrostatic and magnetostatic problems, as well as electrodynamic problems, of which radio frequency (RF) engineering is only a subset. Surprisingly, they include the effects of relativity even though they were conceived much earlier than Einstein's theories.

In differential form, Maxwell's equations can be written as

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}\,, \tag{1}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}\,, \qquad \textbf{Maxwell's equations} \tag{2}$$

$$\nabla \cdot \mathbf{D} = q_v\,, \tag{3}$$

$$\nabla \cdot \mathbf{B} = 0\,, \tag{4}$$

where the field components and constants are defined as follows:

$\mathbf{E}$,	**electric field** (V/m)	
$\mathbf{D} = \varepsilon_0 \varepsilon_\mathrm{r} \mathbf{E}$,	**dielectric displacement** (A s/m^2)	(5)
$\mathbf{B}$,	**magnetic induction, magnetic flux density** (T)	
$\mathbf{H} = \frac{1}{\mu_0 \mu_\mathrm{r}} \mathbf{B}$,	**magnetic field strength or field intensity** (A/m)	(6)
$\mathbf{J} = \kappa \mathbf{E}$,	**electric current density** (A/m^2)	(7)
$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{D}$,	**displacement current** (A/m^2)	
$\varepsilon_0 = 8.854 \cdot 10^{-12}$,	**electric field constant** (F/m)	(8)
ε_r,	**relative dielectric constant**	
$\mu_0 = 4\pi \cdot 10^{-7}$,	**magnetic field constant** (H/m)	(9)
μ_r,	**relative permeability constant**	
κ.	**electrical conductivity** (S/m)	

In the following sections, we shall see that most of the important RF formulae can be derived in a few lines from Maxwell's equations.

1.2 Basic vector analysis and its application to Maxwell's equations

In order to make efficient use of Maxwell's equations, some basic vector analysis is needed, which is introduced in this section. More detailed introductions can be found in a number of textbooks, such as for instance the excellent *Feynman Lectures on Physics* [1].

Gradient of a potential

The gradient of a potential ϕ is the derivative of the potential function $\phi(x,y,z)$ in all directions of a particular coordinate system (e.g., x, y, z). The result is a vector that tells us how much the potential changes in different directions. Applied to the geographical profile of a mountain landscape, the gradient describes the slope of the landscape in all directions. The mathematical sign that is used for the gradient of a potential is the 'nabla operator'; applied to a Cartesian coordinate system, one can write

$$\nabla\Phi = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \Phi = \begin{pmatrix} \frac{\partial\Phi}{\partial x} \\ \frac{\partial\Phi}{\partial y} \\ \frac{\partial\Phi}{\partial z} \end{pmatrix}. \qquad \textbf{gradient of a potential} \tag{10}$$

The expressions for the gradient in cylindrical and spherical coordinate systems are given in Appendices B and C.

Divergence of a vector field

The divergence of a vector field **a** tells us if the vector field has a source. If the resulting scalar expression is zero, we have a 'source-free' vector field, as in the case of the magnetic field. From basic physics, we know that there are no magnetic monopoles, which is why magnetic field lines are always closed. In Maxwell's equations (4), this property is included by means of the fact that the divergence of the magnetic induction **B** equals zero.

In Cartesian coordinates, the divergence of a vector field is defined as

$$\nabla\cdot\mathbf{a} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \cdot \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix} = \frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial a_z}{\partial z}. \qquad \textbf{divergence of a vector} \tag{11}$$

The expressions for cylindrical and spherical coordinate systems are given in Appendices B and C.

Curl of a vector field

When we form the curl of a vector, we are interested in knowing if there are any curls or eddies in the field. Let us imagine that we are looking at the flow of water in a cooling pipe. To check for curls, we can use a stick around which a ball can rotate freely. We position a Cartesian coordinate system at an arbitrary origin and align the stick first with the x axis, and then with the y and z axes. If the ball starts rotating in any of these positions, then we know that the curl of the vector field describing the water flow is non-zero in the direction of the respective axis. The curl of a vector **a** is therefore also a vector, because its information is direction-specific. Its mathematical form in Cartesian coordinates is defined as

$$\nabla \times \mathbf{a} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \times \begin{pmatrix} a_x \\ a_y \\ a_z \end{pmatrix}$$
$$= \det \begin{pmatrix} \mathbf{u}_x & \mathbf{u}_y & \mathbf{u}_z \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ a_x & a_y & a_z \end{pmatrix} = \begin{pmatrix} \frac{\partial a_z}{\partial y} - \frac{\partial a_y}{\partial z} \\ \frac{\partial a_x}{\partial z} - \frac{\partial a_z}{\partial x} \\ \frac{\partial a_y}{\partial x} - \frac{\partial a_x}{\partial y} \end{pmatrix}. \qquad \textbf{curl of a vector} \tag{12}$$

The *unit* vectors $\mathbf{u}_n$ have no physical meaning and simply point in the x, y, and z directions. They have a constant length of 1. The expressions for cylindrical and spherical coordinate systems can be found in Appendices B and C.

Second derivatives

In some instances, we have to make use of second derivatives. One of the expressions that is used regularly in electrodynamics is the Laplace operator $\Delta = \nabla^2$, which—since the operator itself is scalar—can be applied to both scalar fields and vector fields:

$$\Delta\phi = \nabla \cdot (\nabla\phi) = \nabla^2\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2}. \qquad \textbf{Laplace operator} \tag{13}$$

The expressions for cylindrical and spherical coordinate systems can be found in Appendices B and C.

We also introduce two interesting identities:

$$\nabla \times (\nabla\phi) = 0, \tag{14}$$
$$\nabla \cdot (\nabla \times \mathbf{a}) = 0. \tag{15}$$

Equation (14) tells us that if the curl of a vector equals zero, then this vector can be written as the gradient of a potential. This feature can save us a lot of writing when we are dealing with complicated three-dimensional expressions for electric and magnetic fields, and we shall use this principle later on to define non-physical potential functions that can describe (via derivatives) complete three-dimensional vector functions.

In the same way, Eq. (15) can (and will) be used to describe divergence-free fields with simple 'vector potentials'.

1.3 Useful theorems by Gauss and Stokes

The theorems of Gauss and Stokes are some of the most commonly used transformations in this chapter, and therefore we shall take a moment to explain the concepts of them.

Gauss's theorem

Gauss's theorem not only saves us a lot of mathematics but also has a very useful physical interpretation when applied to Maxwell's equations. Mathematically speaking, we transform a volume integral over the divergence of a vector into a surface integral over the vector itself:

$$\int_V \underbrace{\nabla \cdot \mathbf{a}}_{\text{'sources'}} \, \mathrm{d}V = \oint_S \mathbf{a} \cdot \mathrm{d}\mathbf{S}. \qquad \textbf{Gauss's theorem} \tag{16}$$

The surface on the right-hand side of the theorem is the one that surrounds the volume on the left-hand side. If we remember that the divergence of a vector field is equal to its sources, Gauss's theorem tells us that:

- The vector flux through a closed surface equals the sources of flux within the enclosed volume.
- If there are no sources, the amounts of flux entering and leaving the volume must be equal.

These statements can be applied directly to Maxwell's equations. Using Eq. (3) and applying Gauss's theorem, we obtain

$$\int_V \nabla \cdot \mathbf{E}\, \mathrm{d}V = \oint_S \mathbf{E} \cdot \mathrm{d}\mathbf{S} = \frac{Q}{\varepsilon} \tag{17}$$

(Fig. 1), which means that one can calculate the amount of charge in a volume simply by integrating the electric flux lines over any closed surface that surrounds the charge, or vice versa.

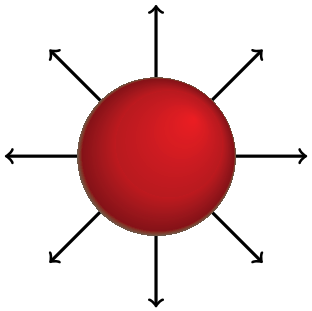


Fig. 1: Example of electric flux lines emanating from electric charge in the centre of a sphere

The same trick can be applied to the source-free magnetic field. Here, we use Eq. (4) and obtain

$$\int_V \nabla \cdot \mathbf{B}\, \mathrm{d}V = \oint_S \mathbf{B} \cdot \mathrm{d}\mathbf{S} = 0\,. \tag{18}$$

Equation (18) gives us the proof of what was already stated earlier: magnetic field lines have no sources ($\nabla \cdot \mathbf{B} = 0$), and therefore the magnetic flux lines are always closed and have neither sources nor sinks. If magnetic flux lines enter a volume, then they also have to leave that volume (Fig. 2).

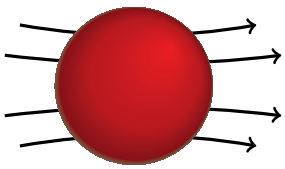


Fig. 2: Example of magnetic flux lines penetrating a sphere

Stokes's theorem

Whereas Gauss's theorem is useful for equations involving the divergence of a vector, Stokes's theorem offers a similar simplification for equations that contain the curl of a vector. With Stokes's theorem, we can transform surface integrals over the curl of a vector into closed line integrals over the vector itself:

$$\int_A (\nabla \times \mathbf{a}) \cdot \mathrm{d}\mathbf{A} = \oint_C \mathbf{a} \cdot \mathrm{d}\mathbf{l}\,. \qquad \textbf{Stokes's theorem} \tag{19}$$

One can interpret Stokes's theorem with the help of Fig. 3 as follows:

– the area integral over the curl of a vector field can be calculated from a line integral along its closed borders, or
– the field lines of a vector field with non-zero curl must be closed contours.

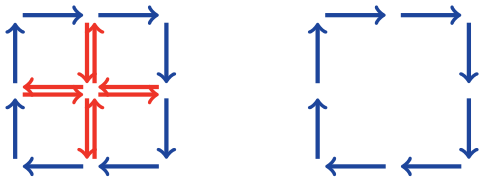
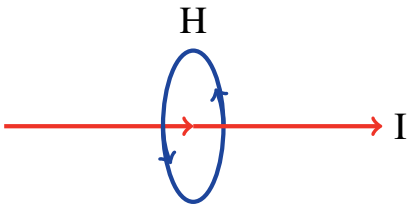


Fig. 3: Illustration of Stokes's theorem

The meaning of these statements becomes immediately clear when we apply Stokes's theorem to Maxwell's equation (1):

$$\int_A (\nabla \times \mathbf{H}) \cdot \mathrm{d}\mathbf{A} = \oint_C \mathbf{H} \cdot \mathrm{d}\mathbf{l} = \int_A \left(\mathbf{J} + \frac{\mathrm{d}\mathbf{D}}{\mathrm{d}t} \right) \cdot \mathrm{d}\mathbf{A} \,. \tag{20}$$

In the electrostatic case, the time derivative disappears and the area integral over the current density may, for instance, be the current flowing in an electric wire as shown in Fig. 4. This means that with a one-line manipulation of Maxwell's equations, we have derived Ampère's law, which tells us that every current induces a circular magnetic field around itself, whose strength can be be calculated from a simple closed line integral along a circular path with the current at its centre.



$$\oint_C \mathbf{H} \cdot \mathrm{d}\mathbf{l} = I \qquad \textbf{Ampère's law} \tag{21}$$

Fig. 4: Illustration of Ampère's law

With similar ease, we can derive Faraday's induction law, which is the basis of every electric motor and generator. All we have to do is apply Stokes's theorem to Maxwell's equation (2):

$$\int_A (\nabla \times \mathbf{E}) \cdot \mathrm{d}\mathbf{A} = \underbrace{\oint_c \mathbf{E} \cdot \mathrm{d}\mathbf{l}}_{V_i} = \underbrace{-\frac{\mathrm{d}}{\mathrm{d}t} \int_A \mathbf{B} \cdot \mathrm{d}\mathbf{A}}_{\frac{\mathrm{d}\psi_m}{\mathrm{d}t}} \,, \tag{22}$$

and again, after one line, we obtain one of the fundamental laws of electrical engineering.

Faraday's law tells us that an electric voltage is induced in a loop if the magnetic flux ψ penetrating the loop changes over time, as shown in Fig. 5. Alternatively, one can change the flux by moving the loop in or out of a static magnetic field.

I hope that these examples have convinced you that Maxwell's equations are indeed very powerful, and that with a bit of vector analysis we really can derive everything we need for RF engineering (although maybe not always in one line . . .).

1.4 Displacement current

Although most people have an idea of what electric and magnetic fields are, the displacement current $\mathrm{d}\mathbf{D}/\mathrm{d}t$ is often not so well understood. Since it is vital for the propagation of electromagnetic waves,

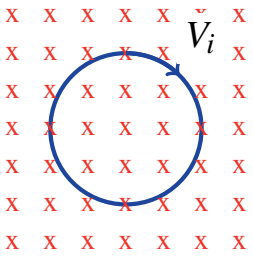


Fig. 5: Illustration of Faraday's induction law

$$V_i = -\frac{\mathrm{d}\psi_m}{\mathrm{d}t} \qquad \textbf{Faraday's induction law} \tag{23}$$

we shall spend a few lines studying this quantity. We start by deriving and interpreting the continuity equation, and then look at a simple practical example.

We apply the divergence to Maxwell's equation (1):

$$\underbrace{\nabla\cdot(\nabla\times\mathbf{H})}_{\equiv 0} = \nabla\cdot\mathbf{J} + \underbrace{\nabla\cdot\frac{\mathrm{d}\mathbf{D}}{\mathrm{d}t}}_{\frac{\mathrm{d}}{\mathrm{d}t}\rho_v}. \tag{24}$$

Using Maxwell's equation (3), we have made an association between the 'sources of the displacement current' $\nabla\cdot(\mathrm{d}\mathbf{D}/\mathrm{d}t)$ and the 'rate of change of electric charge' $(\mathrm{d}/\mathrm{d}t)\rho_v$. Using the identity (15), we obtain the continuity equation

$$\nabla\cdot\mathbf{J} = -\frac{\mathrm{d}}{\mathrm{d}t}\rho_v\,, \qquad \textbf{continuity equation} \tag{25}$$

to which we apply a volume integral and Gauss's theorem (16):

$$\int_V \nabla\cdot\mathbf{J}\,\mathrm{d}V = \oint_S \mathbf{J}\cdot\mathrm{d}\mathbf{S} = \sum I_n = -\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho_v\,\mathrm{d}V\,. \qquad \textbf{continuity equation} \tag{26}$$

In this form, the interpretation is very straightforward, and we can state that:

– if the amount of electric charge in a volume is changing over time, a current needs to flow; or, more poignantly, electric charges cannot be destroyed.

Now, it is good to know that electric charges cannot be destroyed, but that does not yet help us to understand the displacement current. For this purpose, we go back to Eq. (24) and this time we do not replace the expression for the displacement current. Instead, we apply a volume integral and Gauss's theorem and obtain

$$\oint_S \mathbf{J}\cdot\mathrm{d}\mathbf{S} = \sum I_n = -\frac{\mathrm{d}}{\mathrm{d}t}\oint_S \mathbf{D}\,\mathrm{d}S = -\frac{\mathrm{d}}{\mathrm{d}t}\int_V \rho_v\,\mathrm{d}V\,, \tag{27}$$

which we can apply to the simple geometry of a capacitor shown in Fig. 6, which is charged by a static current I.

If we assume a small volume with a surface S around one of the capacitor plates, then we can directly interpret Eq. (27): the current I, which enters the volume V on the left, equals the flux integral of the displacement current $-(\mathrm{d}/\mathrm{d}t)\mathbf{D}$, which leaves the volume V on the right. This means that the displacement current can be understood as a 'current without charge transport', which in this case can only exist because of the rate of change of the electric charge $(-(\mathrm{d}/\mathrm{d}t)\int_V \rho_v\,\mathrm{d}V)$ on the left capacitor plate.

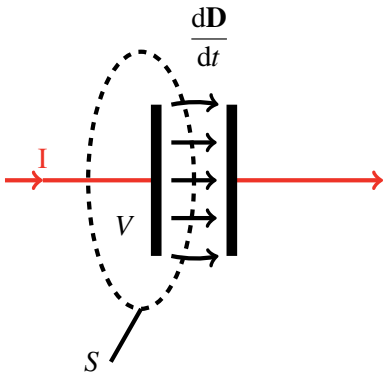


Fig. 6: Example of a displacement current: charging of a capacitor

1.5 Boundary conditions

Before we try to calculate electromagnetic fields in accelerating cavities, we need to understand how these fields behave close to material boundaries, for example the electrically conducting walls of a cavity. Using Stokes's and Gauss's theorems, we can quickly derive these boundary conditions.

Field components parallel to a material boundary

We start with the field components ($E_\parallel$, $H_\parallel$) parallel to a surface between two materials, as depicted in Fig. 7. We define a small surface ΔA, which is perpendicular to the boundary and encloses a small cross-section of the boundary area. Then we integrate Maxwell's equations (1) and (2) over this area and apply Stokes's theorem:

$$\int_A \nabla \times \mathbf{H} \cdot \mathrm{d}\mathbf{A} = \oint_C \mathbf{H} \cdot \mathrm{d}\mathbf{l} = \underbrace{\int_A \mathbf{J} \cdot \mathrm{d}\mathbf{A}}_{=i'\Delta l} + \underbrace{\frac{\mathrm{d}}{\mathrm{d}t} \int_A \mathbf{D} \cdot \mathrm{d}\mathbf{A}}_{\to 0 \text{ for } \mathbf{A}\to 0}, \tag{28}$$

$$\int_A \nabla \times \mathbf{E} \cdot \mathrm{d}\mathbf{E} = \oint_C \mathbf{E} \cdot \mathrm{d}\mathbf{l} = \underbrace{-\frac{\mathrm{d}}{\mathrm{d}t} \int_A \mathbf{B} \cdot \mathrm{d}\mathbf{A}}_{\to 0 \text{ for } \mathbf{A}\to 0}. \tag{29}$$

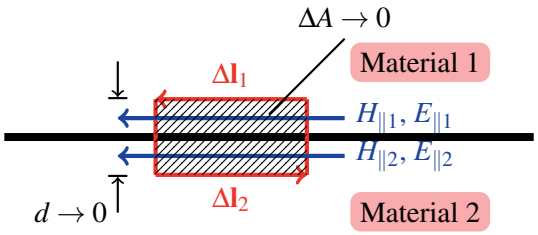


Fig. 7: Boundary conditions parallel to a material boundary

Using Stokes's theorem, the area integrals over A are transformed into line integrals around the contour C of the area. If the width d of the area (see Fig. 7) is now reduced to zero, the calculation of the contour integral simplifies to a multiplication of the field components $E_\parallel$ and $H_\parallel$ by the path elements Δl. The area integrals over $\mathbf{D}$ and $\mathbf{B}$ vanish and the area integral over the current density $\mathbf{J}$ is replaced by a surface current, which may flow in the boundary plane between the two materials, times the path

element Δl. This results in the following boundary conditions:

$$\begin{aligned} H_{\parallel 1} - H_{\parallel 2} &= i' \,, \\ E_{\parallel 1} &= E_{\parallel 2} \,. \end{aligned} \qquad \begin{array}{l} \textbf{conditions for magnetic and} \\ \textbf{electric fields parallel to a} \\ \textbf{material boundary} \end{array} \tag{30}$$

In the case of a waveguide or an accelerator cavity, we generally assume one of the materials (e.g., material 2) to be an ideal electrical conductor, and in that case the electric and magnetic field components in this material vanish, so that we obtain

$$\begin{aligned} H_{\parallel 1} &= i' \,, \\ E_{\parallel 1} &= 0 \,. \end{aligned} \qquad \begin{array}{l} \textbf{conditions for magnetic and} \\ \textbf{electric fields parallel to ideal} \\ \textbf{electric surfaces} \end{array} \tag{31}$$

Field components perpendicular to a material boundary

In a very similar way, we can derive the boundary conditions for fields ($D_\perp$, $B_\perp$) that are perpendicular to a boundary surface between two materials. This time, however, we do not define an area but a small cylinder with a volume ΔV around the boundary, as shown in Fig. 8. We form a volume integral from Maxwell's equations (3) and (4) over the volume of the cylinder and apply Gauss's theorem to transform the volume integrals into surface integrals:

$$\int_V \nabla \cdot \mathbf{D} \, \mathrm{d}V = \oint_S \mathbf{D} \cdot \mathrm{d}\mathbf{S} = \int_V q_v \, \mathrm{d}V \,, \tag{32}$$

$$\int_V \nabla \cdot \mathbf{B} \, \mathrm{d}B = \oint_S \mathbf{B} \cdot \mathrm{d}\mathbf{S} = 0 \,. \tag{33}$$

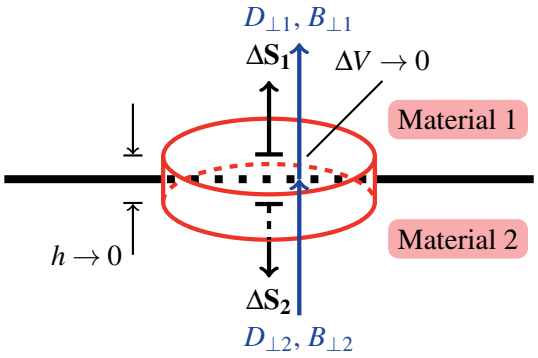


Fig. 8: Boundary conditions perpendicular to a material boundary

In the following step, we reduce the height of the cylinder to zero, so that we end up with two surfaces, one on each side of the boundary. And now it becomes clear why we have to start with a volume integral. Since the surface element $d\mathbf{S}$ is perpendicular to the surface of the cylinder, the 'dot product' in the integrals basically reduces the vector fields $\mathbf{D}$ and $\mathbf{B}$ to the components perpendicular to the surface of the cylinder. This means that above equations can now be written as

$$\begin{aligned} D_{\perp 1} - D_{\perp 2} &= q_s \,, \\ B_{\perp 1} &= B_{\perp 2} \,, \end{aligned} \qquad \begin{array}{l} \textbf{conditions for dielectric} \\ \textbf{displacement and magnetic} \\ \textbf{induction perpendicular to a} \\ \textbf{material boundary} \end{array} \tag{34}$$

where q_s is a surface charge (measured in units of C/m^2) that may exist on the boundary surface. In the

case where material 2 is an ideal conductor, we obtain

$$\begin{aligned} D_{\perp 1} &= q_s\,, \\ B_{\perp 1} &= 0\,. \end{aligned} \qquad \textbf{conditions for dielectric displacement and magnetic induction perpendicular to ideal electric surface} \tag{35}$$

We note that when the fields are parallel to a boundary surface, the electric and magnetic fields are used in the boundary conditions, whereas when they are perpendicular to the boundary surface, we have a condition for the dielectric displacement and the magnetic induction. This means that, for instance, the tangential electric field $E_\parallel$ may be smooth across a boundary but there will be a jump in the dielectric displacement $D_\parallel$ if there are different relative dielectric constants ε_r in the two materials. Similarly, the component of the magnetic induction $B_\perp$ perpendicular to a surface may be smooth, whereas the magnetic field $H_\perp$ will jump if the two materials have different relative magnetic field constants μ_r.

2 Electromagnetic waves

In this section, we shall derive the general form of the wave equation and then restrict ourselves to phenomena that are harmonic in time. Since RF systems mostly deal with sinusoidal waves, we shall be able to explain and understand most of the relevant phenomena with this approach. This includes the 'skin effect', the propagation of energy, RF losses, and acceleration via travelling waves.

2.1 The wave equation

We start with the simplification of looking only at homogeneous, isotropic media, meaning we assume that the electromagnetic fields 'see' the same material constants (μ, ε, κ) in all directions. With this assumption, Maxwell's equations can be conveniently expressed in terms of only E and H:

$$\nabla \times \mathbf{H} = \kappa \mathbf{E} + \varepsilon \frac{\partial \mathbf{E}}{\partial t}\,, \tag{36}$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}\,, \qquad \textbf{Maxwell's equations} \tag{37}$$

$$\nabla \cdot \mathbf{E} = \frac{q_v}{\varepsilon}\,, \tag{38}$$

$$\nabla \cdot \mathbf{H} = 0\,. \tag{39}$$

The curl of Eq. (37) together with Eq. (36), and the curl of Eq. (36) together with Eqs. (37) and (38) result in the general wave equations for a homogeneous medium

$$\begin{aligned} \nabla^2 \mathbf{E} - \nabla(\nabla \cdot \mathbf{E}) &= \mu\kappa \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E} + \mu\varepsilon \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{E}\,, \\ \nabla^2 \mathbf{H} &= \mu\kappa \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{H} + \mu\varepsilon \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{H}\,. \end{aligned} \qquad \textbf{wave equations in a homogeneous medium} \tag{40}$$

In the case of waveguides and cavities, we can simplify these equations even further by considering only the fields inside the waveguide or cavity, which exist in a non-conducting medium ($\kappa = 0$) and a charge-free volume ($\nabla \cdot E = 0$):

$$\begin{aligned} \nabla^2 \mathbf{E} &= \mu\varepsilon \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{E}\,, \\ \nabla^2 \mathbf{H} &= \mu\varepsilon \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{H}\,. \end{aligned} \qquad \textbf{wave equations in a non-conducting, charge-free homogeneous medium} \tag{41}$$

2.2 Complex notation for time-harmonic fields

The already compact wave equations in Eq. (41) can be simplified even further by taking into account the fact that in RF engineering one usually deals with time-harmonic signals, which are sometimes modulated in phase or amplitude. We can therefore introduce the complex notation for electric and magnetic fields. We start by assuming a time-harmonic electric field with amplitude E_0 and phase φ,

$$E(t) = E_0 \cos(\omega t + \varphi) , \tag{42}$$

which we can interpret as the real part of a complex expression,

$$E(t) = \Re\left\{E_0 e^{i\varphi} e^{i\omega t}\right\} = \Re\left\{E_0 \cos(\omega t + \varphi) + iE_0 \sin(\omega t + \varphi)\right\} . \tag{43}$$

In this form, we can easily separate the harmonic time dependence ωt from the phase information φ. The phase information can be merged into the amplitude by defining a 'complex amplitude' or 'phasor'

$$\tilde{E} = E_0 e^{i\varphi} . \tag{44}$$

We keep in mind that the real physical fields are obtained as the real part of the complex amplitude times $e^{i\omega t}$:

$$E_0 \cos(\omega t + \varphi) = \Re\left\{\tilde{E} e^{i\omega t}\right\} . \tag{45}$$

To simplify our writing, we skip the part with the harmonic time dependence and omit the tilde, which means that from now on all field quantities are written as complex amplitudes. In order to convince you that this really is a simplification, let us consider what happens to time derivatives when complex notation is used:

$$\frac{\mathrm{d}}{\mathrm{d}t} \tilde{E} e^{i\omega t} = i\omega \tilde{E} e^{i\omega t} . \tag{46}$$

This means that all time derivatives in Maxwell's equations and also in the wave equations can simply be replaced by a multiplication by $i\omega$, and we are able to do this because the time dependence is always harmonic. Only when we have to deal with transient events, such as the switching on of an RF amplifier or the sudden arrival of a beam in a cavity, do we have to go back the non-harmonic general equations.

As our first application of the complex notation, we rewrite Maxwell's equations as follows:

$$\nabla \times \mathbf{H} = i\omega \underline{\varepsilon} \mathbf{E} , \tag{47}$$

$$\nabla \times \mathbf{E} = -i\omega \mu \mathbf{H} , \tag{48}$$

Maxwell's equations in complex notation

$$\nabla \cdot \mathbf{E} = \frac{\rho_V}{\varepsilon} , \tag{49}$$

$$\nabla \cdot \mathbf{H} = 0 , \tag{50}$$

where the complex dielectric constant $\underline{\varepsilon}$ is defined as

$$\underline{\varepsilon} = \varepsilon' - i\varepsilon'' = \varepsilon \left(1 - i\frac{\kappa}{\omega\varepsilon}\right) . \tag{51}$$

complex dielectric constant

We note that $\underline{\varepsilon}$ is complex only in a conducting medium. We can now proceed to write the general wave equations in complex form:

$$\nabla^2 \mathbf{E} - \nabla(\nabla \cdot \mathbf{E}) = -\underline{k}^2 \mathbf{E} , \tag{52}$$

$$\nabla^2 \mathbf{H} = -\underline{k}^2 \mathbf{H} . \tag{53}$$

general complex wave equations

Here also, we note that the complex wavenumber $\underline{k}$ becomes real in the case of a non-conducting medium:

$$\underline{k}^2 = \omega^2 \mu \underline{\varepsilon} = \omega^2 \mu \varepsilon \left(1 - i\frac{\kappa}{\omega\varepsilon}\right). \qquad \textbf{complex wavenumber} \qquad (54)$$

Finally, we simplify the wave equations again for the case of a non-conducting, charge-free medium and obtain

$$\nabla^2 \mathbf{E} = -k^2 \mathbf{E}, \qquad \textbf{complex wave equations in a non-conducting, charge-free medium} \qquad (55)$$

$$\nabla^2 \mathbf{H} = -k^2 \mathbf{H}, \qquad (56)$$

with

$$k^2 = \omega^2 \mu \varepsilon = \frac{\omega^2}{c^2}. \qquad \textbf{free-space wavenumber} \qquad (57)$$

On the way, we have also introduced a simple definition for the speed of light, $c = 1/\sqrt{\mu\varepsilon}$, in Eq. (57).

2.3 Plane waves

As an introduction to the theory of electromagnetic waves, we look at a very simple case, that of so-called plane waves. We assume again that we are in a homogeneous, isotropic, linear medium and that there are no charges or currents, which means that Eqs. (52) and (53) apply. Furthermore—for a plane wave—we assume that the field components depend only on one coordinate (e.g., z). The solution of the harmonic wave equations (52) and (53) can then be written as a superposition of two waves

$$\begin{aligned} E_x(z) &= \underline{C}_1 e^{-\underline{\gamma} z} + \underline{C}_2 e^{+\underline{\gamma} z}, \\ H_y(z) &= \frac{1}{\underline{Z}} \left(\underline{C}_1 e^{-\underline{\gamma} z} + \underline{C}_2 e^{+\underline{\gamma} z}\right), \end{aligned} \qquad (58)$$

one of which propagates in the positive and one in the negative z direction. The complex propagation constant $\underline{\gamma}$ has a real component α, which describes the damping in a lossy material, and a complex component $i\beta$, which describes the propagation of the wave. The relation between the propagation constant γ and the wavenumber k is

$$\underline{\gamma} = \alpha + i\beta = i\underline{k} = i\omega\sqrt{\mu\underline{\varepsilon}}. \qquad \textbf{propagation constant} \qquad (59)$$

We already know that time-harmonic electric and magnetic fields are linked via Maxwell's equations, which means that their amplitudes have a certain fixed ratio to each other. This ratio has been introduced in Eq. (58) as the wave impedance $\underline{Z}$, the ratio between the electric and magnetic field amplitudes

$$\underline{Z} = \frac{E_y}{H_z} = \sqrt{\frac{\mu}{\underline{\varepsilon}}}, \qquad \textbf{complex wave impedance} \qquad (60)$$

which becomes real in the absence of lossy material. The wave impedance of free space is given by

$$Z_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} \approx 377\,\Omega. \qquad \textbf{free-space wave impedance} \qquad (61)$$

2.4 Skin depth

When electromagnetic waves encounter a conducting (lossy) material, we have to evaluate the boundary conditions (see Section 1.5), and we find that the wave amplitudes are attenuated suddenly by an attenuation constant α. In the RF case we can assume that

$$\frac{\kappa}{\omega\varepsilon} \gg 1\,, \tag{62}$$

which means that the complex wavenumber (54) and, obviously, also the complex dielectric constant (51) are dominated by their imaginary parts, so that we can write

$$\underline{\varepsilon} \approx -i\varepsilon'' = -i\frac{\kappa}{\omega} \quad \text{or} \quad \underline{k}^2 = -i\omega\mu\kappa\,, \tag{63}$$

which is actually equivalent to neglecting the displacement current. Using Eq. (59), we can then write the propagation constant as

$$\gamma = \alpha + i\beta = i\underline{k} = i\omega\sqrt{\frac{-i\mu\kappa}{\omega}} = (1+i)\sqrt{\frac{\kappa\mu\omega}{2}}\,, \tag{64}$$

which defines the attenuation constant α. The 'skin depth' is then defined as the distance after which the wave amplitudes have been attenuated by a factor $1/e \approx 36.8\%$:

$$\delta_\mathrm{s} = \frac{1}{\alpha} = \sqrt{\frac{2}{\omega\mu\kappa}}\,. \qquad \textbf{skin depth} \tag{65}$$

Knowing the value of the skin depth is crucial for the design of RF equipment. Let us assume that we want to build an accelerating cavity that resonates at 500 MHz. Since high-quality copper is quite expensive, we consider the possibility of constructing the cavity out of steel and then copper-plating the interior in order to obtain a good quality factor and reduce the losses in the surface. From Eq. (65), we calculate that the skin depth in copper is approximately 3 µm. Depending on how well the copper plating is done by the plating company, we can now define the thickness of the copper layer that is needed on the inside of the cavity. Typically, around 10–20 times the skin depth is chosen as the plating thickness. Figure 9 shows the dependence of the skin depth on the RF frequency.

Furthermore, the skin depth allows us to calculate the losses in the surface easily. For a wave travelling parallel to a conducting surface, one can define a surface resistance by assuming a constant current density in a layer of the surface material equivalent to the skin depth, as shown in Fig. 10:

$$R_\mathrm{surf} = \frac{1}{\kappa\delta_\mathrm{s}}\,[\Omega]\,. \qquad \textbf{surface resistance} \tag{66}$$

This value has to be multiplied by l/w to obtain the full RF resistance, where l is the length of the conducting wall and w is its width.

2.5 Energy and transport of energy

We start this section by presenting Poynting's law, and then explain its components. Poynting's law states nothing more than the conservation of electromagnetic energy:

$$-\frac{\mathrm{d}}{\mathrm{d}t}\int_V w\,\mathrm{d}V = \int_A \mathbf{S}\cdot\mathrm{d}\mathbf{A} + \int_V \mathbf{E}\cdot\mathbf{J}\,\mathrm{d}V\,. \qquad \textbf{Poynting's law} \tag{67}$$

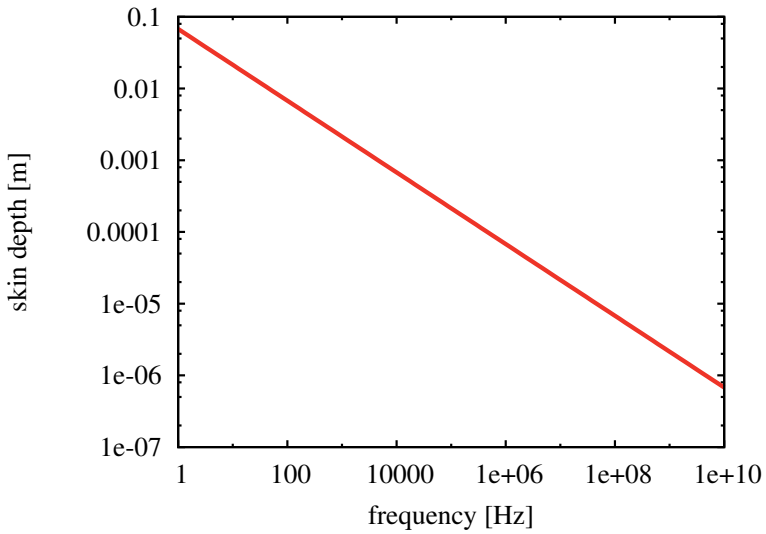


Fig. 9: Skin depth versus RF frequency

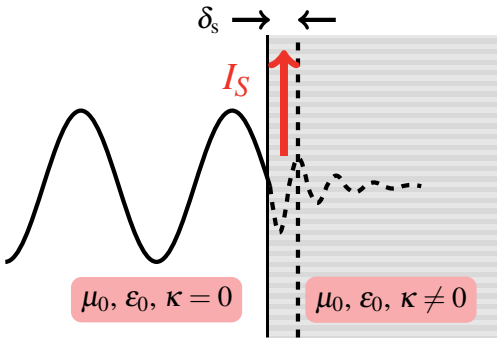


Fig. 10: Skin depth and surface resistance

This equation, read from left to right, states that 'the rate of change of stored energy in a volume equals the energy flow out of the volume (through a surface **A** surrounding the volume) plus the losses within the volume (the work performed on charges per unit time)'. In the following lines, we shall see that the components of Poynting's law do indeed correspond to what is stated in the previous sentence.

What is* $\mathbf{E}\cdot\mathbf{J}$*?

In order to understand the expression $\mathbf{E}\cdot\mathbf{J}$, we follow [1] and start with the force acting on a charge moving in an electromagnetic field,

$$\mathbf{F} = q\left(\mathbf{E} + \mathbf{v}\times\mathbf{B}\right). \qquad \textbf{Lorentz force} \tag{68}$$

Multiplying this equation by $\mathbf{v}$ and knowing that $\mathbf{a}\cdot(\mathbf{a}\times\mathbf{b}) \equiv 0$, we obtain an expression for the work done on a charge per unit time,

$$\mathbf{v}\cdot\mathbf{F} = q\mathbf{v}\cdot\mathbf{E}. \tag{69}$$

Assuming N particles per unit volume, we can write

$$N\mathbf{v}\cdot\mathbf{F} = Nq\mathbf{v}\cdot\mathbf{E} = \mathbf{J}\cdot\mathbf{E}. \tag{70}$$

Therefore the expression $\mathbf{J}\cdot\mathbf{E}$ must be equal to the work done on charges per unit time and unit volume, or, in other words, the loss of electromagnetic energy per unit volume.

***The Poynting vector* $\mathbf{S}$ *and the energy density* w**

These quantities can be understood by manipulating Maxwell's equations (compare, e.g., [2]). We multiply Eq. (1) by $\mathbf{E}$:

$$\mathbf{E}\cdot\mathbf{J} = \mathbf{E}\cdot(\nabla\times\mathbf{H}) - \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}\,. \tag{71}$$

Using Eq. (D.1), this can be rewritten as

$$\mathbf{E}\cdot\mathbf{J} = \mathbf{H}\cdot(\nabla\times\mathbf{E}) - \nabla\cdot(\mathbf{E}\times\mathbf{H}) - \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t}\,. \tag{72}$$

Using the second of Maxwell's equations (2) and assuming time-invariant μ and ε, we can write

$$\mathbf{E}\cdot\mathbf{J} = -\nabla\cdot(\mathbf{E}\times\mathbf{H}) - \frac{\partial}{\partial t}\left(\frac{1}{2}\mathbf{E}\cdot\mathbf{D} + \frac{1}{2}\mathbf{H}\cdot\mathbf{B}\right)\,. \tag{73}$$

Applying a volume integral together with Gauss's theorem (16) and rearranging the elements of the equation, we end up with

$$\begin{aligned} -\frac{\partial}{\partial t}\int_V \left(\frac{1}{2}\mathbf{E}\cdot\mathbf{D} + \frac{1}{2}\mathbf{H}\cdot\mathbf{B}\right)\mathrm{d}V \\ = \int_A (\mathbf{E}\times\mathbf{H})\cdot\mathrm{d}A + \int_V \mathbf{E}\cdot\mathbf{J}\,\mathrm{d}V\,, \end{aligned} \qquad \textbf{Poynting's law} \tag{74}$$

which can be compared directly with Eq. (67). On the left-hand side we have the definition of the energy density,

$$w = w_{\mathrm{el}} + w_{\mathrm{mag}} = \frac{1}{2}\mathbf{E}\cdot\mathbf{D} + \frac{1}{2}\mathbf{B}\cdot\mathbf{H}\,, \qquad \textbf{electric and magnetic energy density} \tag{75}$$

and from the right-hand side we obtain the definition of the energy flux density, or the Poynting vector $\mathbf{S}$,

$$\mathbf{S} = \mathbf{E}\times\mathbf{H}\,. \qquad \textbf{Poynting vector} \tag{76}$$

The Poynting vector gives us the direction in which an electromagnetic wave transports energy, and from the cross product we understand that this direction is always perpendicular to the electric and magnetic field components. This is consistent with Section 2.3, where we found that the field components (E_x, H_y) of a plane wave (see Eq. (58)) are perpendicular to the direction of propagation (z).

In the above derivation, we have used Maxwell's equations in their general form, meaning with time derivatives. In the case of the complex notation, the definitions of the energy density and Poynting vector have to be modified as follows (for a proof, see [2] or [3]):

$$w = w_{\mathrm{el}} + w_{\mathrm{mag}} = \frac{1}{4}\mathbf{E}\cdot\mathbf{D}^* + \frac{1}{4}\mathbf{B}\cdot\mathbf{H}^*\,, \qquad \textbf{electric and magnetic energy density in complex notation} \tag{77}$$

$$\mathbf{S} = \frac{1}{2}\left(\mathbf{E}\times\mathbf{H}^*\right)\,. \qquad \textbf{complex Poynting vector} \tag{78}$$

3 Electromagnetic waves in waveguides

In this section, we derive the field components of electromagnetic waves that propagate in waveguides. The same principle can then be used to calculate the standing-wave pattern in an accelerating cavity, which is nothing more than a superposition of two waves travelling in opposite directions.

3.1 Classification of modes in waveguides and cavities

Before we start to solve the wave equation, we need to introduce a classification of the field patterns that can be found in waveguides and cavities.

TM_{mnp} modes, or E_{mnp} modes

These modes have no magnetic field in the direction of propagation (z) and are therefore often called *transverse magnetic*, or TM, modes. On the other hand, they have an electric field component that is parallel to z, hence the equivalent name *E modes*.

The indices m, n, p indicate the number of zeros or variations in the three directions of a coordinate system. In the case of a waveguide, only the first two indices are used, whereas in the case of a cavity, owing to the standing-wave pattern along z, all three are needed for a complete description. In the case of a circular waveguide or cavity, the indices indicate the following:

- m, number of full-period variations of the field components in the azimuthal direction. For circularly symmetric geometries, $\mathbf{E}$, $\mathbf{B} \propto \cos(m\varphi)$, $\sin(m\varphi)$.
- n, number of zeros (x_{mn}) of the axial field component in the radial direction. For circularly symmetric geometries, E_z, $B_z \propto J_m(x_{mn} r/R_\mathrm{c})$.
- p, number of half-period variations of the field components in the longitudinal direction, with $\mathbf{E}$, $\mathbf{B} \propto \cos(p\pi z/l)$, $\sin(p\pi z/l)$.

The functions J_m introduced above are Bessel functions of the first kind and of mth order, and can be found in mathematical textbooks. The first three orders are shown in Fig. 11.

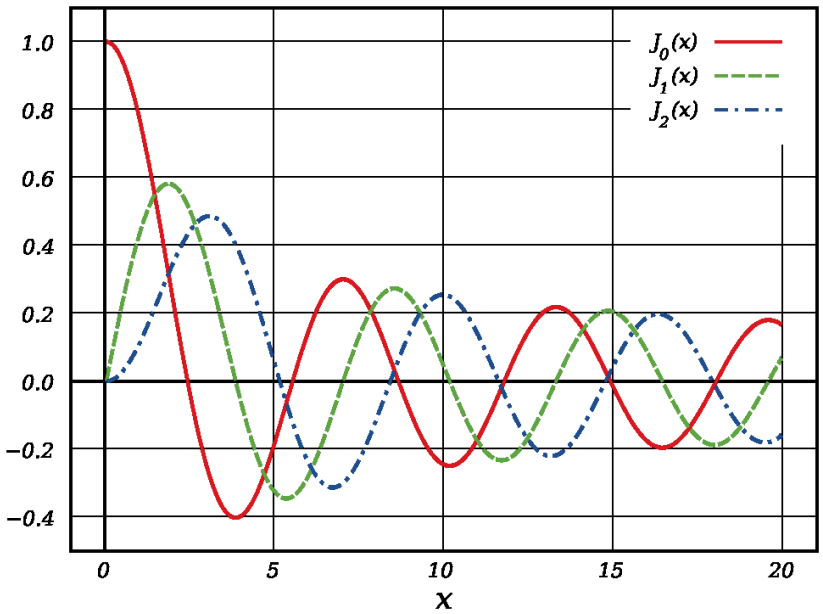


Fig. 11: Bessel functions of the first kind up to order 2

TE_{mnp} modes, or H_{mnp} modes

Here, there is no electric field in the direction of propagation z, hence the name *transverse electric*, or TE, modes. In analogy to the E modes, H modes have a magnetic field component parallel to z. The indices have the same meaning as above.

TEM modes

This class of modes has neither an electric nor a magnetic field component in the direction of propagation. They can exist between two isolated conductors, for example in a coaxial line. The advantage of TEM modes is that waves of any frequency can propagate, whereas TE and TM modes always have a cut-off frequency, below which they are damped exponentially (more on this later). However, the disadvantage of coaxial lines is that the losses in the two conductors are generally higher than in rectangular or circular waveguides.

3.2 Solution of the wave equation in a cylindrical waveguide

Instead of trying to find solutions for all six vector components of the electric and magnetic fields, one can simplify the problem by using a vector potential $\mathbf{A}$ (without any physical meaning) that has only one component. One can then quickly derive all six field components from this vector potential.

It can be shown that only two types of modes can exist in waveguides: TM and TE modes, as introduced above. For each mode type, we introduce a vector potential $\mathbf{A}$ as follows. Since $\mathbf{H}$ and $\mathbf{E}$ are divergence-free, and since $\nabla \cdot (\nabla \times \mathbf{a}) \equiv 0$, we can write

$$\mathbf{H}^{\mathrm{TM}} = \nabla \times \mathbf{A}^{\mathrm{TM}} \quad \text{with} \quad \mathbf{E}^{\mathrm{TM}} = -\frac{i}{\omega\varepsilon} \nabla \times (\nabla \times \mathbf{A}^{\mathrm{TM}}), \quad \textbf{vector potential for TM waves} \tag{79}$$

$$\mathbf{E}^{\mathrm{TE}} = \nabla \times \mathbf{A}^{\mathrm{TE}} \quad \text{with} \quad \mathbf{H}^{\mathrm{TE}} = \frac{i}{\omega\mu} \nabla \times (\nabla \times \mathbf{A}^{\mathrm{TE}}). \quad \textbf{vector potential for TE waves} \tag{80}$$

In both cases the vector potential obeys the wave equation

$$\nabla^2 \mathbf{A} = -k^2 \mathbf{A} \quad \text{with} \quad k^2 = \omega^2 \mu\varepsilon, \tag{81}$$

which can then be solved for various coordinate systems and has only one vector component, in the direction of propagation:

$$\mathbf{A} = A_z \mathbf{e}_z. \tag{82}$$

Fig. 12: Geometry of a circular waveguide

Circular waveguides

In a circular waveguide, as shown in Fig. 12, the vector potentials for the TE and TM modes are identical:

$$A_z^{\mathrm{TM/TE}} = C J_m(k_c r) \cos(m\varphi) e^{\pm i k_z z}, \quad \textbf{vector potential for circular waveguide} \tag{83}$$

with

$$k_z = \sqrt{k^2 - k_\mathrm{c}^2}\,. \qquad \textbf{wavenumber in } z \textbf{ direction} \tag{84}$$

Using Eq. (79), we can derive the field components for the TM modes:

$$\left.\begin{aligned}
E_r &= \frac{i}{\omega\varepsilon}\frac{\partial H_\varphi}{\partial z} &&= -C\frac{k_z k_\mathrm{c}}{\omega\varepsilon}J'_m(k_\mathrm{c}r)\cos(m\varphi)\\
E_\varphi &= -\frac{i}{\omega\varepsilon}\frac{\partial H_r}{\partial z} &&= C\frac{mk_z}{\omega\varepsilon r}J_m(k_\mathrm{c}r)\sin(m\varphi)\\
E_z &= \frac{ik_\mathrm{c}^2}{\omega\varepsilon}A_z &&= C\frac{ik_\mathrm{c}^2}{\omega\varepsilon}J_m(k_\mathrm{c}r)\cos(m\varphi)\\
H_r &= \frac{1}{r}\frac{\partial A_z}{\partial\varphi} &&= -C\frac{m}{r}J_m(k_\mathrm{c}r)\sin(m\varphi)\\
H_\varphi &= -\frac{\partial A_z}{\partial r} &&= -Ck_\mathrm{c}J'_m(k_\mathrm{c}r)\cos(m\varphi)
\end{aligned}\right\} e^{\pm ik_z z}\,. \qquad \begin{array}{l}\textbf{field components for TM}\\ \textbf{modes in a circular waveguide}\end{array} \tag{85}$$

Now we can use the boundary conditions to specify the cut-off wavenumber k_c. From Section 1.5, we know that the electric field components parallel to the waveguide surface have to vanish at the surface, which means

$$\left.\begin{aligned} E_\varphi(r=a) &= 0\\ E_z(r=a) &= 0 \end{aligned}\right\} \Rightarrow J_m(k_\mathrm{c}a) = 0 \quad \Rightarrow \quad k_\mathrm{c} = \frac{j_{mn}}{a}\,. \tag{86}$$

The nth zeros j_{mn} of the Bessel functions of order m are tabulated in mathematical textbooks (e.g., [4]).

Using

$$k_\mathrm{c} = \frac{2\pi}{\lambda_\mathrm{c}} = \frac{\omega_\mathrm{c}}{c}\,, \tag{87}$$

we can define the cut-off frequency of the waveguide,

$$\omega_\mathrm{c} = c\frac{j_{mn}}{a}\,. \qquad \begin{array}{l}\textbf{cut-off frequency for TM}\\ \textbf{modes in a circular waveguide}\end{array} \tag{88}$$

The mode that is most commonly used in circular waveguides is the TM_{01} mode, which has only three field components. By inserting $m = 0$ and $n = 1$ into Eq. (85) and using $J'_0(r) = -J_1(r)$, we obtain

$$\left.\begin{aligned}
E_r &= C\frac{k_z k_\mathrm{c}}{\omega\varepsilon}J_1(k_\mathrm{c}r)\\
E_z &= -C\frac{ik_\mathrm{c}^2}{\omega\varepsilon}J_0(k_\mathrm{c}r)\\
H_\varphi &= Ck_\mathrm{c}J_1(k_\mathrm{c}r)
\end{aligned}\right\} e^{\pm ik_z z}, \qquad \begin{array}{l}\textbf{field components of TM}_{01}\\ \textbf{mode in a circular waveguide}\end{array} \tag{89}$$

with a cut-off frequency $\omega_\mathrm{c} \approx c \times (2.405/a)$.

The field pattern of the TM_{01} mode is shown in Fig. 13 for a mode frequency 15% above the cut-off frequency. The distance between the minima or between the maxima of the field corresponds to 0.5 times the propagation wavelength λ_z. With decreasing mode frequency, $\lambda_z = 2\pi/k_z$ becomes longer, and finally becomes infinite when the mode frequency equals the cut-off frequency ω_c. This effect is shown in Fig. 14, where the TM_{01} mode propagates at a frequency just 0.5% above the cut-off frequency.

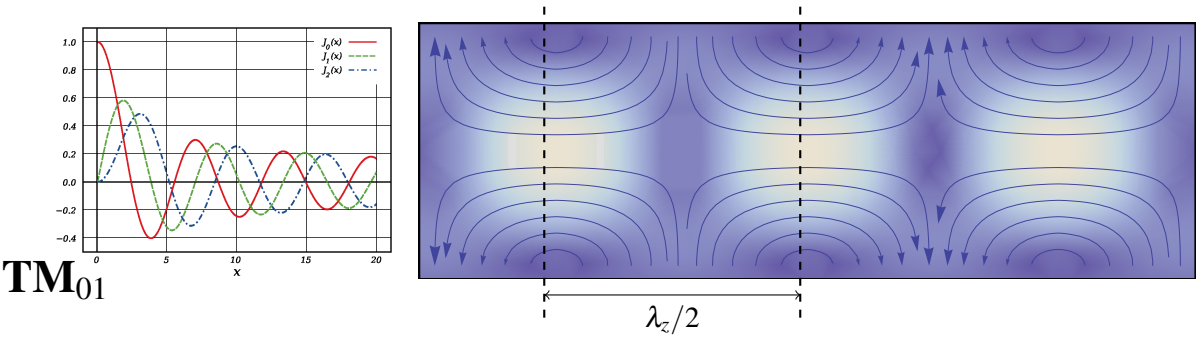


Fig. 13: Field lines of a TM_{01} mode in a circular waveguide with $\omega = 1.15\omega_c$. Solid lines, electric field lines; dashed lines, magnetic field lines. The brightness of the background is proportional to the norm of the field vector: light areas indicate high-field regions of the magnetic field in the left plot and of the electric field in the right plot.

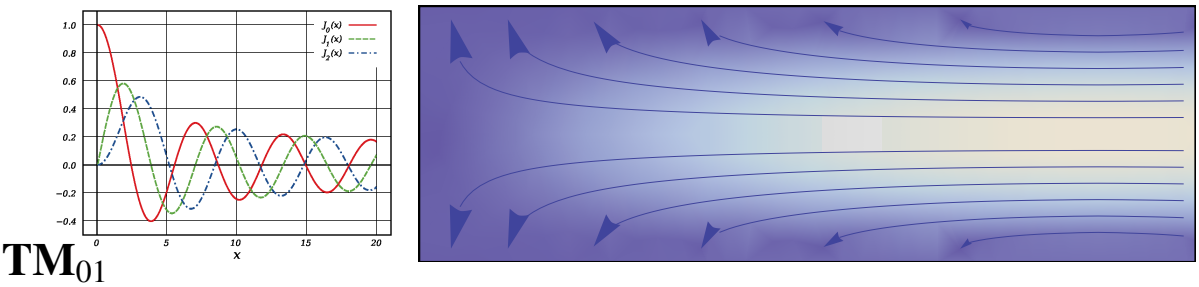


Fig. 14: Field lines of a TM_{01} mode in a circular waveguide with $\omega = 1.005\omega_c$. Solid lines, electric field lines; dashed lines, magnetic field lines. The brightness of the background is proportional to the norm of the field vector: light areas indicate high-field regions of the magnetic field in the left plot and of the electric field in the right plot.

Rectangular waveguides

The derivation of the fields in a rectangular waveguide follows the same principle as that used in the previous section for circular waveguides. In a rectangular waveguide, as shown in Fig. 15, two different vector potentials are needed to describe the TE and TM modes:

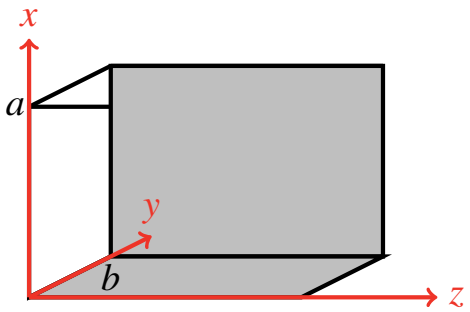


Fig. 15: Geometry of a rectangular waveguide with transverse dimensions a and b

$$A_z^{\mathrm{TM}} = C\sin(k_x x)\sin(k_y y)e^{\pm ik_z z}, \qquad \begin{matrix}\textbf{vector potential for TM waves}\\ \textbf{in a rectangular waveguide}\end{matrix} \tag{90}$$

$$A_z^{\mathrm{TE}} = C\cos(k_x x)\cos(k_y y)e^{\pm ik_z z}, \qquad \textbf{vector potential for TE waves in a rectangular waveguide} \tag{91}$$

where

$$k_z = \sqrt{k^2 - k_\mathrm{c}^2}, \quad \text{with} \quad k_\mathrm{c}^2 = k_x^2 + k_y^2. \qquad \textbf{wavenumber in z direction} \tag{92}$$

We note that the position of the origin of the coordinate system is linked to the sine and cosine terms in Eqs. (90) and (91). The fields derived from the vector potentials have to fulfil the boundary conditions on the waveguide walls. So if, for instance, we were to choose the origin in the centre of the waveguide, then the sine and cosine expressions would have to be exchanged to account for the changed symmetries with respect to the coordinate axes. Using Eq. (79) again, we derive the field components for the TM modes:

$$\left.\begin{aligned}
E_x &= \frac{i}{\omega\varepsilon}\frac{\partial H_y}{\partial z} &&= \pm C\frac{k_z}{\omega\varepsilon}\cos(k_x x)\sin(k_y y)\\
E_y &= -\frac{i}{\omega\varepsilon}\frac{\partial H_x}{\partial z} &&= \pm C\frac{k_z}{\omega\varepsilon}\sin(k_x x)\cos(k_y y)\\
E_z &= \frac{i(k_z^2-k^2)}{\omega\varepsilon}A_z^{\mathrm{TM}} &&= C\frac{i(k_z^2-k^2)}{\omega\varepsilon}\sin(k_x x)\sin(k_y y)\\
H_x &= \frac{\partial A_z^{\mathrm{TM}}}{\partial y} &&= Ck_y\sin(k_x x)\cos(k_y y)\\
H_y &= -\frac{\partial A_z^{\mathrm{TM}}}{\partial x} &&= -Ck_x\cos(k_x x)\sin(k_y y)
\end{aligned}\right\} e^{\pm ik_z z}. \qquad \begin{array}{l}\textbf{field components for}\\ \textbf{TM modes in a}\\ \textbf{rectangular waveguide}\end{array} \tag{93}$$

Using the boundary conditions, we can specify the wavenumbers k_x and k_y:

$$\left.\begin{aligned} E_y(x=a) &= 0\\ E_z(x=a) &= 0\end{aligned}\right\} \Rightarrow k_x = \frac{m\pi}{a} \quad \text{and} \quad m = 0,1,2,\ldots, \tag{94}$$

$$\left.\begin{aligned} E_x(y=b) &= 0\\ E_z(y=b) &= 0\end{aligned}\right\} \Rightarrow k_y = \frac{n\pi}{b} \quad \text{and} \quad n = 0,1,2,\ldots, \tag{95}$$

and the cut-off frequency for a rectangular waveguide is

$$\omega_\mathrm{c} = ck_\mathrm{c} = c\sqrt{k_x^2+k_y^2} = c\pi\sqrt{\left(\frac{m}{a}\right)^2+\left(\frac{n}{b}\right)^2}. \qquad \begin{array}{l}\textbf{cut-off frequency for TM}\\ \textbf{modes in a rectangular}\\ \textbf{waveguide}\end{array} \tag{96}$$

The usual convention is to have $a > b$, and in this case the TE_{10} mode is the mode with the lowest cut-off frequency. It is also the only mode that propagates in a relatively large frequency band, from $f_\mathrm{c}^{\mathrm{TE},10}$ to $2f_\mathrm{c}^{\mathrm{TE},10}$, which is why it is the mode most commonly used in rectangular waveguides. The fields of the TE modes can be derived from the TE vector potential using the same procedure.

3.3 Wave propagation and dispersion relation

In Figs. 13 and 14, we have seen that the propagation wavelength λ_z of a waveguide mode is determined by its frequency and by how far the mode frequency is above the cut-off frequency of the waveguide. If the propagation wavelength depends on the mode frequency, we can assume that the phase velocity of a particular mode also depends on the mode frequency. This relationship is called the dispersion relation and, using the definition of the wavenumber in Eq. (84), we can write

$$k_z^2 = k^2 - k_c^2 = \frac{\omega^2 - \omega_c^2}{c^2} = \frac{\omega^2}{v_{ph}^2}, \qquad \textbf{dispersion relation} \tag{97}$$

from which we can immediately see that:

- k_z can be real only if the mode frequency ω is above the cut-off frequency ω_c;
- for $\omega < \omega_c$, the mode cannot propagate and the fields are exponentially damped.

We also have a definition of the phase velocity, which is the speed at which the maxima and minima of the field patterns move along the waveguide:

$$v_{ph} = \frac{\omega}{k_z} = c^2 \frac{\omega^2}{\omega^2 - \omega_c^2}. \qquad \textbf{phase velocity} \tag{98}$$

This is not to be confused with the speed with which the wave actually propagates in the waveguide. The dispersion relation is usually plotted in the form of a 'Brillouin diagram', as shown in Fig. 16.

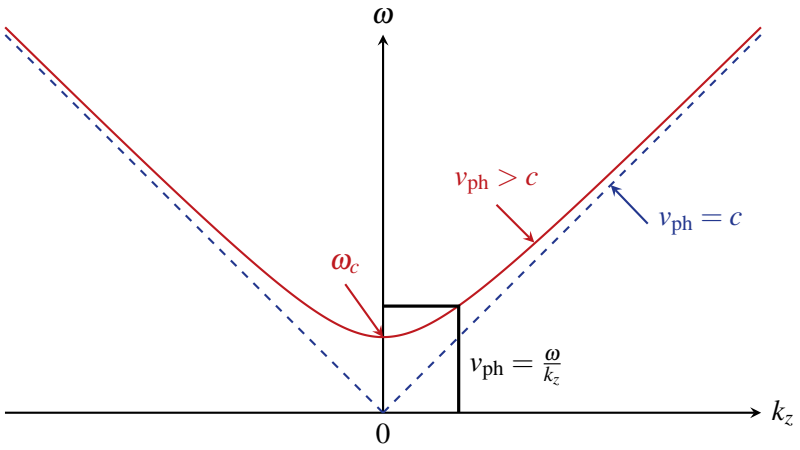


Fig. 16: Dispersion relation in a waveguide. The dotted line shows the case $v_{ph} = c$.

The slope of the dispersion relation is the group velocity

$$v_{gr} = \frac{d\omega}{dk_z}, \qquad \textbf{group velocity} \tag{99}$$

which gives the velocity with which a signal or energy is transported along the waveguide. From Fig. 16, we can conclude that:

- Each frequency has a certain phase velocity and group velocity, which means that signals with a frequency bandwidth will become deformed while travelling along a waveguide. With the help of the dispersion relation, we can easily quantify how much deformation will occur.
- The phase velocity v_{ph} is always larger than the velocity of light c, and at cut-off ($\omega = \omega_c$) it even becomes infinite ($k_z = 0$ and $v_{ph} \to \infty$).
- For acceleration, one needs synchronism between the phase velocity (the speed of the field pattern) and the velocity of the particles, which implies that acceleration in waveguides is impossible.
- Information and therefore energy travel at the group velocity, which is always slower than the speed of light.

3.4 Attenuation of waves (power loss method)

Up to this point, we have assumed perfect electrical conductors as the boundaries of our waveguides. Real waveguides and cavities have a certain resistance, and the fields therefore penetrate into the conductors, which significantly complicates the solution of the wave equation. However, we have seen in Section 2.4 that the skin depth in metals is very much smaller than the RF wavelength. This means that we can reasonably assume that the field patterns in a waveguide with ideal boundaries and in a waveguide with resistive metal boundaries will be practically identical (of course, only for good conductors such as copper or aluminium). In order to calculate the attenuation of waves, we can therefore use the fields of a waveguide with ideal electrical boundaries. From the magnetic field, we calculate the induced current in the waveguide walls, and then apply the resistance of the real material to calculate the losses and then the damping of the wave. This principle is called the power loss method and is a simplified method for calculating RF losses on the surfaces of good conductors.

We start by defining the power that is lost per unit length along the longitudinal axis of the waveguide,

$$P' = -\frac{\mathrm{d}P}{\mathrm{d}z}\,. \tag{100}$$

From

$$E, H \propto e^{-\alpha z} \quad \Rightarrow \quad P \propto e^{-2\alpha z}\,, \tag{101}$$

we immediately obtain

$$P' = -\frac{\mathrm{d}P}{\mathrm{d}z} = 2\alpha P \tag{102}$$

and thus the definition of the attenuation constant

$$\alpha = \frac{P'}{2P}\,. \qquad \textbf{attenuation constant} \tag{103}$$

In the next steps, we need to derive expressions for the power P transported through the waveguide, and the power loss per unit length P'. Using the field components of the TM_{01} mode given in Eq. (89) and the definition of the complex Poynting vector in Eq. (78), we obtain

$$P = \frac{1}{2}\int\limits_A (\mathbf{E}\times\mathbf{H}^*)\cdot\mathrm{d}\mathbf{A} = \frac{1}{2}\int\limits_0^a\int\limits_0^{2\pi} E_r H_\varphi^* r\,\mathrm{d}r\,\mathrm{d}\varphi = \frac{C^2 k_z k_\mathrm{c}^2 \pi a^2 J_1^2(k_\mathrm{c}a)}{\omega\varepsilon}\,, \tag{104}$$

where we have used

$$\int\limits_0^a J_1^2(k_\mathrm{c}r) r\,\mathrm{d}r = \frac{a^2}{2} J_1^2(k_\mathrm{c}a)\,. \tag{105}$$

In order to calculate the losses on the waveguide surface, we first need to know the surface currents that flow within the skin depth. For this purpose, we make use of Ampère's law, as shown in Fig. 17:

$$\oint\limits_c \mathbf{H}\cdot d\mathbf{l} = I = \oint\limits_c \mathbf{J}\cdot(\delta_\mathrm{s}\,\mathrm{d}\mathbf{l})\,. \qquad \textbf{Ampère's law} \tag{106}$$

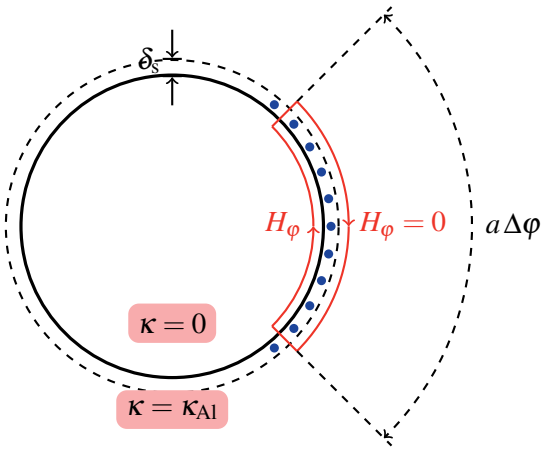


Fig. 17: Ampère's law applied to calculate the surface currents in a circular waveguide

Since the magnetic field has only an azimuthal component, we obtain

$$H_\varphi(r=a,z) = Ck_c J_1(k_c a)e^{-ik_z z} = J_z(z)\delta_s\,. \tag{107}$$

The power density (in W/m^3) in the waveguide wall is given by

$$p_v = \frac{1}{2}\mathbf{E}\cdot\mathbf{J}^* = \frac{1}{2\kappa}J_z J_z^* = \frac{\partial^3 P}{(\partial r)(r\,\partial\varphi)(\partial z)}\,, \qquad \textbf{power density} \tag{108}$$

from which we can write an expression for the power loss per unit length. Together with Eq. (107), we obtain

$$P' = \frac{\partial P}{\partial z} = \int_a^{a+\delta_s}\int_0^{2\pi} p_v r\,\mathrm{d}r\,\mathrm{d}\varphi = \frac{\pi a C^2 k_c^2 J_1^2(k_c a)}{\kappa\delta_s}\,, \qquad \textbf{power loss per unit length} \tag{109}$$

where we have used the fact that $\delta_s \ll a$ to simplify the evaluation of the integral. Now we insert Eqs. (104) and (109) into Eq. (103) and obtain an expression for the attenuation of a TM$_{01}$ mode in a circular waveguide,

$$\alpha = \frac{P'}{2P} = \frac{R_{\mathrm{surf}}}{Z_0 a\sqrt{1-(f_c/f)^2}}\,. \qquad \textbf{attenuation of TM}_{01}\textbf{ mode in circular waveguide} \tag{110}$$

In the expression above, we have used the definition of the surface resistance given in Eq. (66) and the definition of the free-space wave impedance Z_0 given in Eq. (61).

As an example, we have plotted the attenuation constant for an aluminium waveguide in Fig. 18, where we can see that for this type of waveguide:

- Large-diameter waveguides result in smaller losses, which means that a cost optimum has to be found between the cost of the waveguide, its space requirements, and the losses.
- The minimum losses occur when the operating frequency of the TM$_{01}$ mode is a factor of $\sqrt{3}$ above the cut-off frequency (try to prove this!).

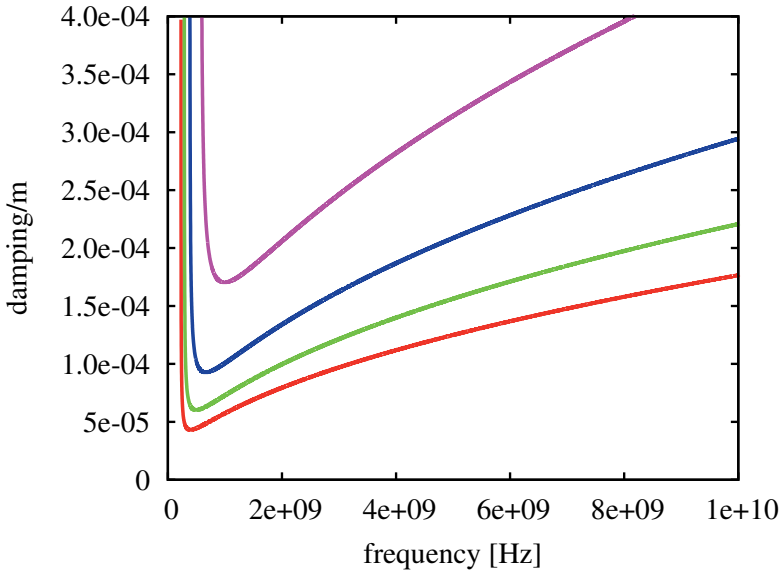


Fig. 18: Attenuation of the TM_{01} mode in a circular aluminium waveguide for several different radii: bottom to top, 0.5 m, 0.4 m, 0.3 m, 0.2 m

4 Accelerating cavities

4.1 Travelling-wave cavities

In order to accelerate particles in a 'waveguide-like' structure, the phase velocity in the structure needs to be slowed down, which can be achieved by putting some 'obstacles' into the waveguide. In Fig. 19, we see a simple example of a disc-loaded waveguide.

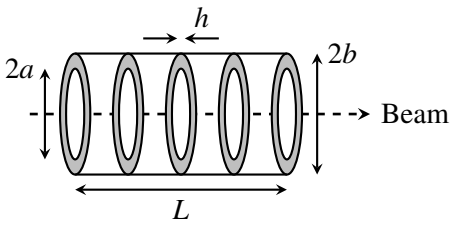


Fig. 19: Geometry of a simple travelling-wave structure

The dispersion relation for such a structure is derived in, for example, [5] as

$$\omega = \frac{2.405c}{b}\sqrt{1+\kappa(1-\cos(k_z L)e^{-\alpha h})}\,, \qquad \textbf{dispersion relation of disc-loaded circular waveguide} \tag{111}$$

where

$$\kappa = \frac{4a^3}{3\pi J_1^2(2.405)b^2 L} \ll 1 \quad \text{and} \quad \alpha \approx \frac{2.405}{a}\,. \tag{112}$$

Plotting Eq. (111) gives us the Brioullin diagram in Fig. 20, where we can see that we now obtain phase velocities that are equal to or even below the speed of light. We can also understand why the $2\pi/3$ mode is often used for acceleration in electron accelerators, because for this mode (in this example) the phase velocity is just equal to the speed of light. It should be noted that with different geometries, it is possible to operate with different modes and also at velocities $v_{\mathrm{ph}} < c$. When a structure operates in the $2\pi/3$

mode, this means that the RF phase shifts by $2\pi/3$ per cell, or, in other words, one RF period extends over three cells.

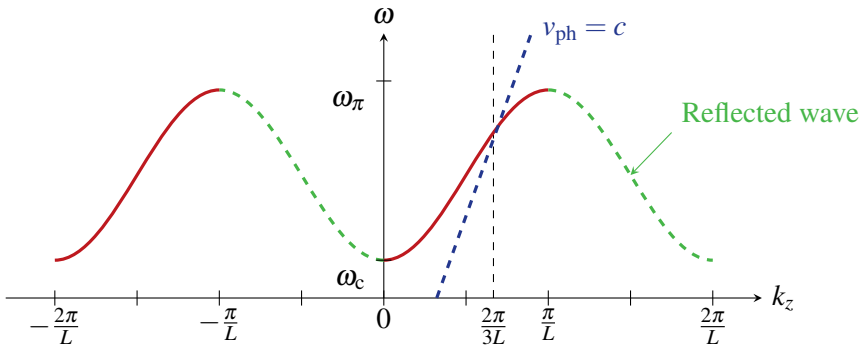


Fig. 20: Dispersion diagram for a disc-loaded travelling-wave structure. Here, the chosen operating point is $v_{ph} = c$ and $k_z = 2\pi/3L$.

By attaching an input and an output coupler to the outermost cells of the structure, we obtain a usable accelerating structure. Since the particles gain energy in every cell, the electromagnetic wave becomes increasingly damped along the structure. It is then extracted via the output coupler and dumped in an RF load. If one is interested in obtaining the maximum possible accelerating gradient in each cell, then one can counteract the decreasing fields by changing the bore radius from cell to cell. The idea is to slow down the group velocity from cell to cell and obtain a 'constant-gradient' structure, rather than a 'constant-impedance' structure where the bore radius is kept constant. Other optimizations, for example for maximum efficiency, are also possible.

4.2 Standing-wave cavities

One obtains a cylindrical standing-wave structure by simply closing both ends of a circular waveguide with electric walls. This yields multiple reflections on the end walls until a standing-wave pattern is established. Owing to the additional boundary conditions in the longitudinal direction, we obtain another 'restriction' on the existence of electromagnetic modes in the structure. Whereas a longitudinally open travelling-wave structure allows all frequencies and all cell-to-cell phase variations on the dispersion curve, now only certain 'loss-free' modes (still assuming perfectly conducting walls) with discrete frequencies and discrete phase changes can exist in a cavity. If RF power is fed in at a different frequency, then the fields excited are damped exponentially, similarly to the modes below the cut-off frequency of a waveguide.

The corresponding dispersion relation for a standing-wave cavity can again be found in textbooks (see [5] and also [6]). However, it is necessary to pay attention to whether the structure under consideration has magnetic or electric cell-to-cell coupling and what kind of end cell is assumed in the analysis. The most common form of the dispersion relation is derived from a coupled-circuit model with $N+1$ cells. Usually the model has half-cell terminations on both ends of the chain, representing the behaviour of an infinite chain of electrically coupled resonators (compare the original paper by Nagle *et al.* [7]):

$$\omega_n = \frac{\omega_0}{\sqrt{1+k\cos(n\pi/N)}}, n = 0,1,\ldots,N. \qquad \textbf{dispersion relation for half-cell-terminated standing-wave structure} \qquad (113)$$

Assuming an odd number of cells, ω_0 is the frequency of the $\pi/2$ mode and of an uncoupled single cell; k is the cell-to-cell coupling constant, and $n\pi/N$ is the phase shift from cell to cell. For $k \ll 1$, which is usually fulfilled, the coupling constant is given by

$$k = \frac{\omega_{\pi\,\mathrm{mode}} - \omega_{0\,\mathrm{mode}}}{\omega_0}. \qquad \textbf{coupling constant} \qquad (114)$$

Two characteristics of the dispersion curve are worth noting:

- The total width of the frequency band of the mode, $\omega_{\pi\,\mathrm{mode}} - \omega_{0\,\mathrm{mode}}$, is independent of the number of cells, which means that we can determine the cell-to-cell coupling constant by measuring the complete structure (but this is only true if all coupling constants are equal).
- For electric coupling, the 0 mode has the lowest frequency and the π mode has the highest. In the case of magnetic coupling, this behaviour is reversed, and one can find the corresponding dispersion curve by changing the sign before the coupling constant in Eq. (113).

In Fig. 21, we plot the dispersion curve for a seven-cell (half-cell-terminated) magnetically coupled structure according to Eq. (113).

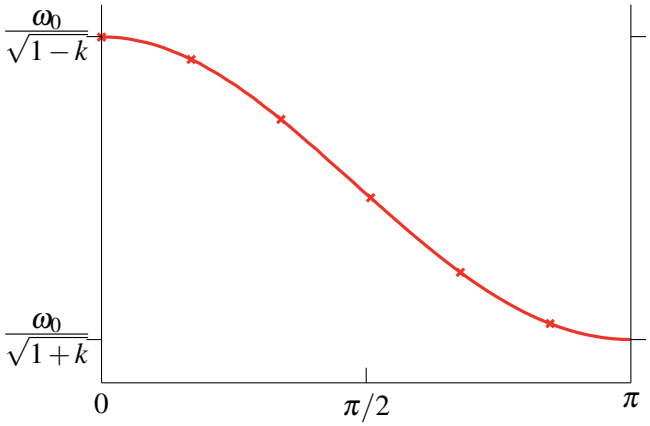


Fig. 21: Dispersion diagram for a standing-wave structure with seven magnetically coupled cells

In practice, one usually has cavities with full-cell termination, and in this case one has to detune the frequences of the end cells to obtain a flat field distribution in the cavity [8]. In this case it is possible to have a flat field distribution for either the 0 mode or the π mode but not for both at the same time, because the end cells have to be detuned by different amounts in the two cases [9].

4.3 Standing wave versus travelling wave

The principal difference between the two types of cavity is in how and how fast the cavities are filled with RF power. Travelling-wave structures are filled 'in space', which means that, basically, cell after cell is filled with power. For the following estimations, we assume a frequency in the range of hundreds of megahertz. The filling of a travelling-wave structure typically takes place with a speed of approximately 1–3% of the speed of light and results in total filling times in the submicrosecond range. Standing-wave structures, on the other hand, are filled 'in time': the electromagnetic waves are reflected at the end walls of the cavity and slowly build up a standing-wave pattern of the desired amplitude. For normal-conducting cavities, the time required for this process is typically in the range of tens of microseconds. For superconducting cavities, the filling time can easily go into the millisecond range (depending on the required field level, the accelerated current, and the cavity parameters). This means that for applications that require very short beam pulses (< 1 µs), travelling-wave structures are much more power-efficient. For longer pulses ($> n \times 10$ µs), both types of structures can be optimized to achieve similar efficiencies and costs.

Since one can have extremely short RF pulses in a travelling-wave structure, one can obtain much higher peak fields than in a standing-wave structure. This is demonstrated by the accelerating structures for CLIC [10], which have reached values of approximately 100 MV/m (limited by electrical breakdown), whereas the design gradient for the superconducting (standing-wave) cavities for the ILC [11] is just slightly above 30 MV/m (this value is generally limited by field emission and by quenches caused by the peak magnetic field).

Travelling-wave structures can, theoretically, be designed for non-relativistic particles. In existing accelerators, however, they are mostly used for relativistic particles. Low-beta acceleration is typically performed with standing-wave cavities.

Because of the lack of an obvious criterion (other than the pulse length or the particle velocity), an optimization and costing exercise has to be performed for each specific application in order to decide which structure is more efficient. Two excellent papers [12, 13] in which this exercise is performed can be used as references.

4.4 The pillbox cavity

In this chapter, we shall analyse only the simplest TM-mode cavity, the so-called pillbox cavity. A selection of cavities using other mode types is described in [14].

Resonating cavities can be represented conveniently by a lumped-element circuit consisting of an inductor (for storage of magnetic energy) and a capacitor (for storage of electric energy). Looking at Fig. 22, one can easily imagine how the lumped circuit can be transformed into a cavity.

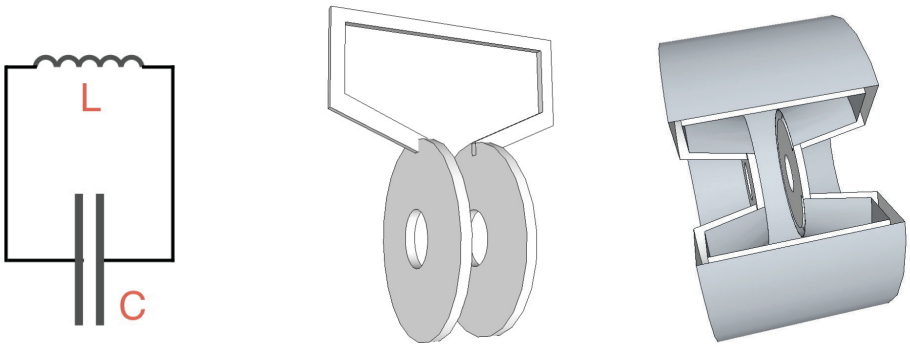


Fig. 22: Transition from a lumped resonating circuit to a resonating cavity

The pillbox cavity is nothing more than an empty cylinder with a conducting inner surface. The simplest mode in this cavity is the TM_{010} mode, which has zero full-period variations in the azimuthal direction (φ), one 'zero' of the axial field component in the radial direction (r), and zero half-period variations in the longitudinal (z) direction. We can derive the general field equations by using the vector potential for a circular waveguide given in Eq. (83) and simply superimposing two waves, one propagating in the positive z direction and one in the negative z direction:

$$A_z^{\mathrm{TM/TE}} = CJ_m(k_r r)\cos(m\varphi)\underbrace{\left(e^{-ik_z z}+e^{ik_z z}\right)}_{2\cos(k_z z)}. \qquad \begin{array}{l}\textbf{vector potential for travelling}\\ \textbf{waves in the positive and}\\ \textbf{negative } z \textbf{ directions}\end{array} \tag{115}$$

Using Eq. (79), we derive the TM field components

$$
\begin{aligned}
E_r &= \frac{i}{\omega\varepsilon}\frac{\partial H_\varphi}{\partial z} &&= i2C\frac{k_z k_r}{\omega\varepsilon}J'_m(k_r r)\cos(m\varphi)\sin(k_z z)\,,\\
E_\varphi &= -\frac{i}{\omega\varepsilon}\frac{\partial H_r}{\partial z} &&= -i2C\frac{mk_z}{\omega\varepsilon r}J_m(k_r r)\sin(m\varphi)\sin(k_z z)\,,\\
E_z &= \frac{ik_r^2}{\omega\varepsilon}A_z &&= i2C\frac{k_r^2}{\omega\varepsilon}J_m(k_r r)\cos(m\varphi)\cos(k_z z)\,,\\
H_r &= \frac{1}{r}\frac{\partial A_z}{\partial\varphi} &&= -2C\frac{m}{r}J_m(k_r r)\sin(m\varphi)\cos(k_z z)\,,\\
H_\varphi &= -\frac{\partial A_z}{\partial r} &&= -2Ck_r J'_m(k_r r)\cos(m\varphi)\cos(k_z z)\,.
\end{aligned}
\qquad \textbf{TM modes in a pillbox cavity} \quad (116)
$$

In the case of standing-wave cavities, the term 'cut-off' frequency does not really make sense, so we have replaced the symbol k_c by k_r, indicating that we have a radial dependence of the axial field component, which can also be interpreted as a radial wavenumber.

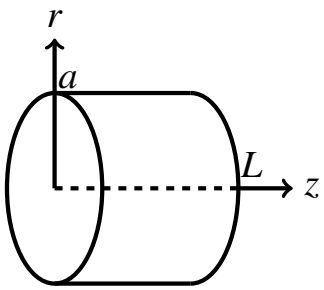


Fig. 23: Pillbox cavity

In the next step, we apply the boundary conditions for a pillbox with radius a and length L as shown in Fig. 23. We obtain

$$E_r(z=0/L),\, E_\varphi(z=0/L)=0 \quad \Rightarrow k_z=\frac{p\pi}{L}\,, \tag{117}$$

$$E_\varphi(r=a),\, E_z(r=a),\, H_r(r=a)=0 \quad \Rightarrow k_r=\frac{j_{mn}}{a}\,. \tag{118}$$

In the case of the circular waveguide, the transverse boundary condition made a discrete quantity out of k_c (which we now call k_r in the above equations), and thus defined the cut-off frequency. Now, with the second boundary in the z direction, we obtain a discrete solution for k_z also. The two boundary conditions together result in a discrete set of frequencies (the dispersion relation) for our pillbox cavity:

$$k^2=\frac{\omega^2}{c^2}=k_z^2+k_r^2 \quad \Rightarrow f_{mnp}^{\mathrm{TM}}=\frac{c}{2\pi}\sqrt{\left(\frac{j_{mn}}{a}\right)^2+\left(\frac{p\pi}{L}\right)^2}\,. \qquad \textbf{dispersion relation for TM modes in a pillbox cavity} \quad (119)$$

We note that the dispersion relation of a single-cell cavity as given above is different from the dispersion relation that can be derived for a multicell cavity, as in the case of Eq. (113). The latter is derived from a model of equivalent lumped circuits, each representing a cell resonating in the TM_{010} mode and coupled to its neighbours in order to model the behaviour of a multicell cavity, whereas Eq. (119) is directly derived from Maxwell's equations and describes a field pattern that is created by the boundary conditions of our pillbox.

The TM mode with the lowest frequency is the TM_{010} mode, with a frequency

$$f_{010}^{\mathrm{TM}} = \frac{2.405c}{2\pi a}\,, \qquad \textbf{frequency of the TM}_{010}\ \textbf{pillbox mode} \tag{120}$$

and its field components are

$$\begin{aligned} E_z &= -i2C\frac{j_{01}^2}{a^2\omega\varepsilon}J_0\left(\frac{j_{01}}{a}r\right) = E_0 J_0\left(\frac{j_{01}}{a}r\right), \\ H_\varphi &= 2C\frac{j_{01}}{a}J_1\left(\frac{j_{01}}{a}r\right) = \frac{E_0}{Z_0}J_1\left(\frac{j_{01}}{a}r\right). \end{aligned} \qquad \textbf{field components of the TM}_{010}\ \textbf{pillbox mode} \tag{121}$$

Figure 24 shows the field pattern of the TM_{010} mode, simulated by Superfish©.

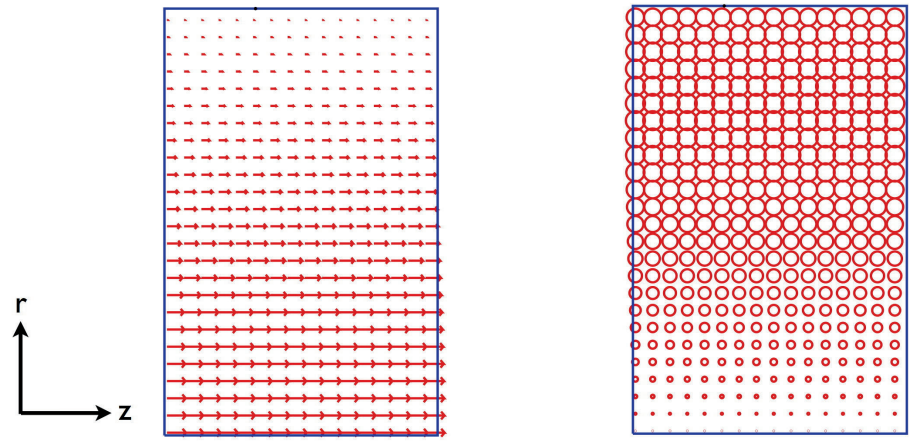


Fig. 24: Field pattern of the TM_{010} mode in a pillbox cavity

4.5 Basic cavity parameters

In order to characterize and optimize cavities, we need some commonly used figures of merit, which we shall define here in general terms and then apply to our simple pillbox cavity. In the following, we assume that we are dealing with an axially symmetric cavity resonating in the TM_{010} mode.

4.5.1 Energy gain in a cavity

For particles traversing a cavity on axis, the electric field generally has the following form:

$$E_z(r=0,z,t) = E(0,z)\cos(\omega t+\varphi)\,, \tag{122}$$

which we can use to calculate the energy gain of a particle when it traverses the cavity,

$$\begin{aligned} \Delta W &= q\int_{-L/2}^{L/2} E(0,z)\cos(\omega t+\varphi) \\ &= qV_0T\cos\varphi = qE_0TL\cos\varphi\,, \end{aligned} \qquad \textbf{energy gain in a cavity (Panofsky equation)} \tag{123}$$

where the cavity voltage is given by

$$V_0 = \int_{-L/2}^{L/2} E(0,z)\,\mathrm{d}z = E_0\,, \qquad \textbf{cavity voltage} \tag{124}$$

and the 'difficult mathematics' has been lumped into the so-called transit time factor

$$T = \frac{\int_{-L/2}^{L/2} E(0,z)\cos(\omega t(z))\,\mathrm{d}z}{\int_{-L/2}^{L/2} E(0,z)\,\mathrm{d}z} - \underbrace{\tan\phi\,\frac{\int_{-L/2}^{L/2} E(0,z)\sin(\omega t(z))\,\mathrm{d}z}{\int_{-L/2}^{L/2} E(0,z)\,\mathrm{d}z}}_{=0 \text{ if } E(0,z) \text{ is symmetric about } z=0}\,. \qquad \textbf{transit time factor} \tag{125}$$

This takes into account the fact that the RF electric field changes during the passage of the particles. It gives the ratio between the energy gained in an RF field and in a DC field and is therefore always less than 1. We note that the Panofsky equation takes account of the changing velocity of the particles when they cross the accelerating gap. This makes the integrals in the above equations difficult to evaluate. Assuming that the velocity change of the beam particles during their passage is small, however, one can say that

$$\omega t \approx \omega\frac{z}{v} = \frac{2\pi z}{\beta\lambda}\,, \tag{126}$$

which changes the expression for the transit time factor to (assuming that $E(0,z)$ is symmetric about $z=0$)

$$T = \frac{\int_{-L/2}^{L/2} E(0,z)\cos(2\pi z/\beta\lambda)\mathrm{d}z}{\int_{-L/2}^{L/2} E(0,z)\,\mathrm{d}z}\,. \qquad \textbf{transit time factor for small velocity changes} \tag{127}$$

The accelerating voltage V_{acc} is the voltage that the particle 'sees' when crossing the cavity and should not be confused with the cavity voltage V_0. We thus define

$$V_{\mathrm{acc}} = V_0 T = E_0 L T\,. \qquad \textbf{accelerating voltage} \tag{128}$$

4.5.2 Shunt impedance

The shunt impedance tells us how much voltage a cavity will provide when a certain amount of power is dissipated in the cavity walls. This is one of the parameters to be maximized in cavity design, since a large shunt impedance reduces the power consumption of an RF cavity. The general definition is

$$R_{\mathrm{s}} = \frac{V_0^2}{P_{\mathrm{d}}}\,. \qquad \textbf{shunt impedance (linac definition)} \tag{129}$$

The benefit of a high shunt impedance can easily be diminished by having a small transit time factor, because in this case the cavity voltage cannot be used efficiently to transfer energy to the beam. Therefore one usually tries to optimize both the shunt impedance and the transit time factor, which explains the

definition of the effective shunt impedance

$$R = \frac{(V_0 T)^2}{P_\mathrm{d}}. \qquad \textbf{effective shunt impedance} \tag{130}$$

When comparing multicell structures operating at different frequencies, one is interested less in the efficiency per cell (because the cell size depends on, for instance, the frequency chosen) than in the efficiency per unit length of the accelerating structure. For this reason, we define

$$Z = \frac{R_\mathrm{s}}{L} = \frac{E_0^2}{P_\mathrm{d}/L} \qquad \textbf{shunt impedance per unit length} \tag{131}$$

and

$$ZT^2 = \frac{R}{L} = \frac{(E_0 T)^2}{P_\mathrm{d}/L}. \qquad \textbf{effective shunt impedance per unit length} \tag{132}$$

4.5.3 'Linac' and 'circuit' definitions of shunt impedance

It turns out that different communities of accelerator experts use different definitions of the shunt impedance. Linac experts usually use the definitions presented above, whereas the people who deal with circular machines generally use a definition that is derived from the lumped-circuit definition of a resonator (see Section 4.7). In that definition, all shunt impedances are exactly half as large, following

$$R_\mathrm{s}^\mathrm{c} = \frac{V_0^2}{2P_\mathrm{d}}. \qquad \textbf{shunt impedance (circuit definition)} \tag{133}$$

So, before you discuss shunt impedances with anyone, make sure that you are using the same definition. In order to mark the difference clearly, we use R_s^c in this text to identify when the circuit definition is being used.

4.5.4 3 dB bandwidth and quality factor

The quality factor Q describes the bandwidth of a resonator and is defined as the ratio of the reactive power (stored energy) to the real power that is lost in the cavity walls:

$$Q = \frac{\omega}{\Delta\omega} = \frac{\omega W}{P_\mathrm{d}}. \qquad \textbf{quality factor} \tag{134}$$

If a resonator were built with ideal electrical walls (zero electrical resistance), the resonance curve would be a delta function at the resonance frequency. So, the bandwidth $\Delta\omega$ would be zero and the quality factor would be infinite. In reality, even superconducting cavities have a certain surface resistance, which is why all our cavities have a certain bandwidth and a finite quality factor. Figure 25 shows a typical resonance curve measured with a network analyser. In a measurement of this kind, two antennas penetrate the cavity. The first antenna sends an RF signal with a frequency sweep, and the second picks up the field level in the cavity. As a result, we obtain a plot of the field level versus frequency. The bandwidth is defined as the frequency width of the resonance curve, measured as the distance between the points where the field level has dropped by 50% (or -3 dB), as shown in Fig. 25.

Together with the shunt impedance, one can define another figure of merit, (R/Q), which is used to maximize the energy gain in a cavity for a given stored energy:

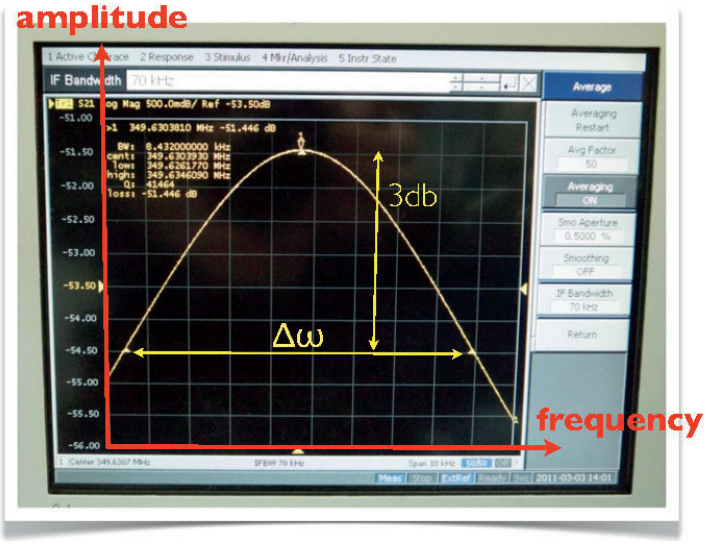


Fig. 25: Measurement of frequency, 3 dB bandwidth, and Q-factor with a network analyser

$$\left(\frac{R}{Q}\right) = \frac{(V_0 T)^2}{\omega W}. \qquad (R/Q) \tag{135}$$

(R/Q) is independent of the surface losses of the cavity and can therefore be used to qualify the geometry of an accelerating cavity.

4.5.5 Filling time of a cavity

This section is a short extract from [14], which can be consulted for more details. The dissipated power in a cavity must be equal to the rate of change of the stored energy:

$$P_\mathrm{d} = -\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{\omega_0 W}{Q_0}. \tag{136}$$

The solution of the above equation can be written as

$$W(t) = W_0 e^{-2t/\tau}, \tag{137}$$

which describes an exponential decay of the stored energy with a 'filling time constant' τ, where

$$\tau = \frac{2Q_0}{\omega_0}. \qquad \textbf{filling time constant} \tag{138}$$

If the cavity is equipped with a power coupler, we have to consider the 'loaded Q' (which will be derived later), and the filling time constant changes to

$$\tau_\mathrm{l} = \frac{2Q_\mathrm{l}}{\omega_0}. \qquad \textbf{filling time constant for a loaded cavity} \tag{139}$$

In the above definition, the electric field decays exponentially with a time constant $1/\tau$, whereas the

stored energy decays with a time constant $2/\tau$. Be aware that you can often find textbook definitions of the filling time constant where the stored energy decays with a time constant $1/\tau$.

4.6 Basic cavity parameters for a pillbox cavity

As a small exercise, in this section we calculate the cavity parameters that were defined in the previous section for a pillbox cavity of length L and radius a. Since the TM_{010} mode has no z dependence, we can simplify the expression for the transit time factor (127) to

$$T = \frac{\int_{-L/2}^{L/2} E(0,z)\cos(2\pi z/\beta\lambda)\mathrm{d}z}{\int_{-L/2}^{L/2} E(0,z)\,\mathrm{d}z} = \frac{\sin(\pi L/\beta\lambda)}{\pi L/\beta\lambda}. \qquad \textbf{transit time factor of a pillbox for small velocity changes} \tag{140}$$

In the case of relativistic particles ($\beta \approx 1$) and a cavity length $L = \lambda/2$, which is often chosen because the cavity can then be cascaded into a multicell structure, we obtain

$$T = \frac{2}{\pi} = 0.64. \qquad \textbf{transit time factor of a pillbox for relativistic particles} \tag{141}$$

With real cavities, one usually tries to increase the transit time factor by shortening the accelerating gap. This can be done by introducing nose cones on the cavity walls, as shown in Fig. 22.

We use the power loss method again to calculate the quality factor of our pillbox cavity. To evaluate Eq. (134), we need the stored energy and the power lost in the cavity walls. For the stored energy, we obtain

$$W = W_{\mathrm{el}} + W_{\mathrm{mag}} = 2W_{\mathrm{el}} = 2\int_V \frac{1}{4}\mathbf{E}\cdot\mathbf{D}^*\,\mathrm{d}V. \tag{142}$$

With

$$E_z = E_0 J_0\left(\frac{j_{01}r}{a}\right), \tag{143}$$

we obtain

$$W = \frac{\varepsilon_0}{2}\int_0^a\int_0^{2\pi}\int_{-L/2}^{L/2} E_0^2 J_0^2\left(\frac{j_{01}r}{a}\right) r\,\mathrm{d}r\,\mathrm{d}\varphi\,\mathrm{d}z = \frac{1}{2}E_0^2\varepsilon_0\pi L a^2 J_1^2(j_{01}). \tag{144}$$

To calculate the dissipated power, we integrate Eq. (108) over a volume that consists of the inner surface of the pillbox times the skin depth:

$$P_{\mathrm{d}} = \frac{\delta_{\mathrm{s}}}{2\kappa}\int_{-L/2}^{L/2} \underbrace{J_z J_z^*}_{(1/\delta_{\mathrm{s}})^2 H_\varphi^2(r=a,z)} 2\pi a\,\mathrm{d}z + \frac{\delta_{\mathrm{s}}}{\kappa}\int_0^a \underbrace{J_r J_r^*}_{(1/\delta_{\mathrm{s}})^2} H_\varphi^2(r,z=0)2\pi r\,\mathrm{d}r \tag{145}$$

$$= \frac{E_0^2\pi R_{\mathrm{surf}} a}{Z_0^2} J_1^2(j_{01})(a+L), \tag{146}$$

where we have made use of

$$H_\varphi = \frac{E_0}{Z_0} J_1\left(\frac{j_{01}r}{a}\right). \tag{147}$$

Putting everything together, we obtain

$$Q_0 = \frac{\omega W}{P_{\mathrm{d}}} = \frac{Z_0^2\omega}{2R_{\mathrm{surf}}}\frac{La}{L+a} = \frac{1}{\delta_{\mathrm{s}}}\frac{La}{L+a} \propto \sqrt{\omega}. \tag{148}$$

As we can see, the quality factor is a function of the material constants κ and μ (which are contained in ρ_s), the frequency, and the geometry of the cavity. We also note that for the same cavity shape, the quality factor increases with the frequency in proportion to $\sqrt{\omega}$.

The accelerating voltage in a pillbox cavity is given by

$$V_{\mathrm{acc}} = V_0 T = E_0 L T = E_0 L \frac{\sin(\pi L/\beta\lambda)}{\pi L/\beta\lambda}, \qquad \textbf{accelerating voltage in pillbox} \quad (149)$$

and is obviously a strong function of the transit time factor. It therefore depends on the gap length L and the speed of the particles β. Owing to their high development costs, superconducting cavities are often used over large velocity ranges without changing their cell length, and this results in a velocity-dependent acceleration efficiency. Figure 26 shows $(R/Q) \propto (V_0 T)^2$ as a function of particle velocity for a five-cell superconducting cavity whose geometric cell length corresponds to a particle speed of $\beta = 0.65$.

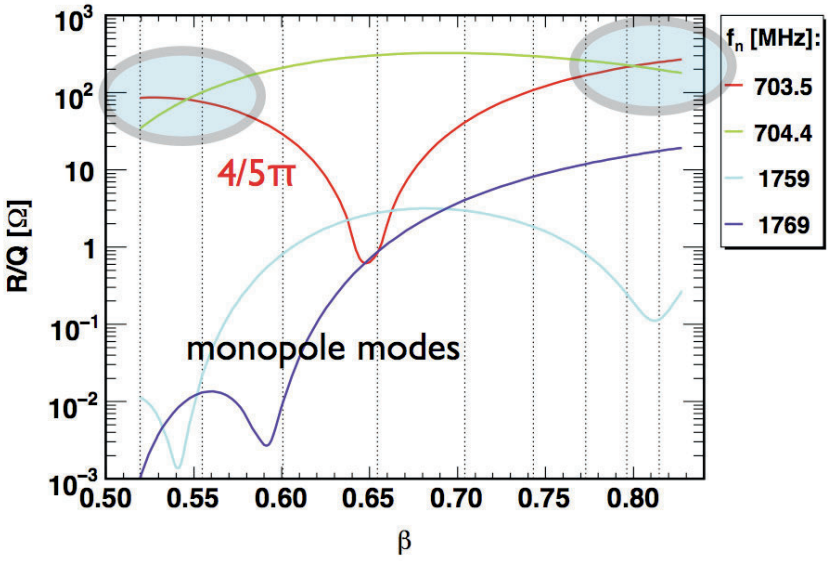


Fig. 26: Dependence of (R/Q) on particle velocity for a five-cell superconducting cavity with a geometric β of 0.65 and a frequency of 704.4 MHz. Upper curve, π mode (see also [15]).

If the cavity is used over too large a velocity range, one may find areas where the passband mode that is closest to the π mode (here, the $4/5\pi$ mode) has a higher acceleration efficiency than the accelerating mode. These areas are highlighted in Fig. 26, and should be avoided when one is designing a linac. One should also be aware that the (R/Q) of the HOMs is highly dependent on the particle velocity.

Using the expressions for the accelerating voltage $V_0 T$ (Eq. (149)) and the dissipated power P_{d} (Eq. (146)), we also obtain an analytical expression for the effective shunt impedance,

$$R = \frac{(V_0 T)^2}{P_{\mathrm{d}}} = \frac{Z_0}{\pi R_{\mathrm{surf}} J_1^2(j_{01})} \frac{\sin(\pi L/\beta\lambda)}{\pi L/\beta\lambda} \frac{L^2}{a(a+L)}. \qquad \textbf{effective shunt impedance of a pillbox} \quad (150)$$

Finally, we calculate the frequency and (R/Q) using Eqs. (120), (150), and (148):

$$f_{010}^{\mathrm{TM}} = \frac{2.405c}{2\pi a}, \qquad \textbf{pillbox frequency} \quad (151)$$

$$\left(\frac{R}{Q}\right) = \frac{2c}{\omega\pi J_1^2(j_{01})}\frac{\sin(\pi L/\beta\lambda)}{\pi L/\beta\lambda}\frac{L}{a^2}. \qquad \textbf{pillbox } (\mathbf{R/Q}) \tag{152}$$

As stated before, (R/Q) is indeed independent of any material parameters. However, it does depend on the geometry of the cavity and the transit time factor.

4.7 A cavity as a lumped circuit

In the field of RF technology, it is common practice to describe the behaviour of cavities, RF transmission lines, and couplers with equivalent lumped circuits. In this chapter, we shall present only the treatment of a cavity and a coupler, so that one can understand how to get power into a cavity. Descriptions of the transmission of RF power and the associated theory of RF transmission lines can be found in many textbooks on RF and microwave engineering. We start with the description of a cavity by a parallel *LCR* circuit as depicted in Fig. 27.

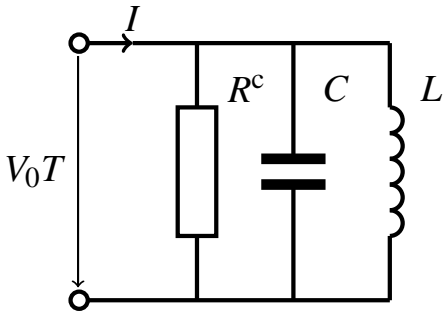


Fig. 27: Lumped-circuit equivalent of a resonant cavity

You may remember that the admittance of a parallel circuit is calculated by adding up the admittances of the individual elements, which means that we can write the cavity impedance as

$$Z^{\mathrm{c}} = \frac{1}{i\omega C + 1/i\omega L + 1/R^{\mathrm{c}}}. \qquad \textbf{lumped-circuit cavity impedance} \tag{153}$$

At resonance ($\omega = \omega_0$), the imaginary parts cancel each other and the cavity impedance becomes real, which means that

$$X = \omega_0 L = \frac{1}{\omega_0 C} = \sqrt{\frac{L}{C}}, \qquad \textbf{lumped circuit at resonance} \tag{154}$$

that the resonance frequency is given by

$$\omega_0 = \frac{1}{\sqrt{LC}}, \qquad \textbf{lumped-circuit resonance frequency} \tag{155}$$

and that the power lost in the resonator is given by

$$P_{\mathrm{d}} = \frac{1}{2}\frac{(V_0 T)^2}{R^{\mathrm{c}}}. \qquad \textbf{lumped-circuit dissipated power} \tag{156}$$

The stored energy can be written as

$$W = \frac{1}{2}C(V_0T)^2 = \frac{1}{2}\frac{(V_0T)^2}{\omega_0^2 L}, \qquad \textbf{lumped-circuit stored energy} \quad (157)$$

and from this we obtain an expression for the quality factor,

$$Q_0 = \omega_0 \frac{W}{P_\mathrm{d}} = \omega_0 C R^\mathrm{c} = \frac{R^\mathrm{c}}{\omega_0 L}. \qquad \textbf{lumped-circuit quality factor} \quad (158)$$

Our goal is to relate the lumped elements to the cavity characteristics, and for this purpose we multiply Eq. (157) by ω and, together with Eq. (154), we obtain

$$\frac{1}{\omega_0 C} = \sqrt{\frac{L}{C}} = \frac{(V_0T)^2}{2\omega_0 W} = \left(\frac{R^\mathrm{c}}{Q}\right) = \frac{1}{2}\left(\frac{R}{Q}\right). \quad (159)$$

From this, we can understand the difference between the 'circuit ohm' and the 'linac ohm', and it also provides a lumped-circuit description of a cavity, as summarized in Table 1. As we can see, three quantities are sufficient to describe a resonator. Instead of using R, L, and C, one can also use the parameters ω_0, Q_0, and (R/Q) to completely characterize an RF cavity, as in Table 2.

Table 1: Lumped-circuit elements of a cavity

Lumped circuit	Field description
R^c	$\frac{1}{2}R$
C	$\frac{2}{\omega_0(R/Q)}$
L	$\frac{1}{2\omega_0}\left(\frac{R}{Q}\right)$

Table 2: Three characteristic quantities of a cavity

Lumped circuit	Field description
$\omega_0 = \frac{1}{\sqrt{LC}}$	$\frac{2.405c}{a}$ (pillbox)
$Q_0 = \omega_0 C R^\mathrm{c} = \frac{R^\mathrm{c}}{\omega_0 L}$	$Q_0 = \frac{\omega_0 W}{P_\mathrm{d}}$
$\left(\frac{R^\mathrm{c}}{Q}\right) = \sqrt{\frac{L}{C}} = \frac{1}{2}\left(\frac{R}{Q}\right)$	$\left(\frac{R}{Q}\right) = \frac{(V_0T)^2}{\omega_0 W}$

4.8 Getting power into a cavity: couplers

In this section, we shall extend the circuit model to include the power coupler and also extend our basic equations to describe the process of coupling power into a cavity. There are two basic types of couplers

that are used in standing-wave cavities:

– *Antenna/loop couplers:* here, the coupler is usually some kind of coaxial line, with the outer conductor connected to the cavity wall and the inner conductor either penetrating into the cavity volume or connected in a loop to the inner surface of the cavity (Fig. 28).
– *Iris couplers:* here, the fields in a waveguide are coupled to the cavity fields via an opening that connects the waveguide to the cavity.

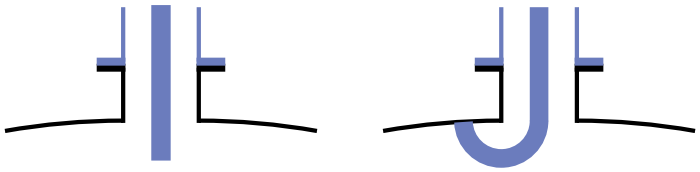


Fig. 28: Example of an antenna-type coupler (left) and a loop-type coupler (right)

When designing a coupler, one has to keep in mind the principle of reciprocity: the coupler has to produce a field pattern in the area of the coupling port that is very similar to the field pattern of the mode that will be excited in the cavity. Looking at Fig. 28, one can imagine that an antenna-type coupler would be very effective on the end walls of our pillbox, where it would couple electrically to the axial electric field lines. On the cylindrical surface of the pillbox, a loop coupler would be a better choice, with the loop oriented such that the azimuthal magnetic field penetrates the loop.

Figure 29 shows an example of a 'tuner-adjustable (waveguide) coupler' (TaCo) [16], as used for the Linac4 [17] cavities at CERN. In this case a short-circuited rectangular waveguide is coupled to a standing-wave cavity via a racetrack-shaped coupling iris. The coupling factor (more on this later) here is a function of the position of the short circuit (left side), the height of the racetrack-shaped coupling channel between the cavity and the waveguide (on the top), the size of the coupling slot, and the position of a stub tuner, which is used to fine-tune the coupling.

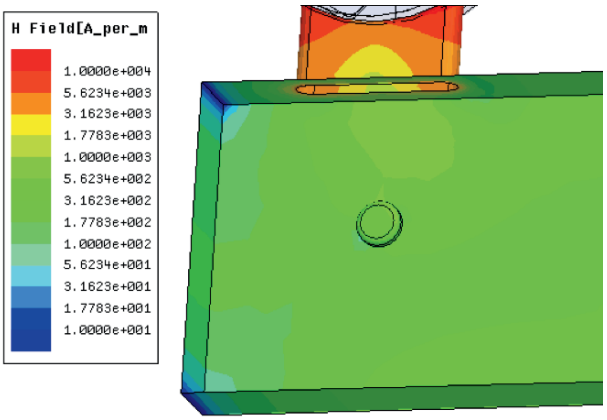


Fig. 29: Waveguide coupler connected to a Linac4 cavity

In the ideal case, the power coupler is matched to the (beam-)loaded cavity, which means that there is no reflected power returning from the cavity towards the RF power source. Here, 'matched' means that the coupler acts like an ideal transformer that transforms the impedance Z_c of the cavity into the impedance Z_0 of the attached waveguide. To keep things simple, let us assume that the RF generator is also matched to Z_0 so that we can establish a lumped-element circuit as shown in Fig. 30.

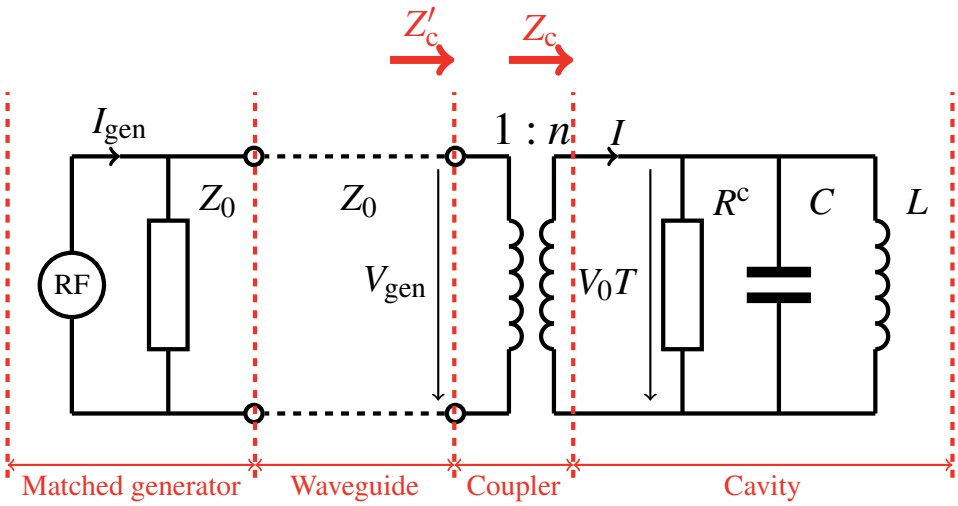


Fig. 30: Lumped-element circuit for RF power source, waveguide, power coupler, and cavity

Considering the coupler as a transformer, we can write that

$$\left.\begin{aligned} V_0T &= nV_{gen} \\ I &= \frac{I_{gen}}{n} \end{aligned}\right\} \Rightarrow Z_c = \frac{V_0T}{I} = n^2 Z'_c\,, \tag{160}$$

which means that the cavity impedance

$$Z_c = \frac{1}{i\omega C + 1/i\omega L + 1/R^c} \qquad \textbf{cavity impedance} \tag{161}$$

is transformed into

$$Z'_c = \frac{1}{i\omega n^2 C + n^2/i\omega L + n^2/R^c}\,, \qquad \textbf{cavity + coupler impedance} \tag{162}$$

which is the impedance 'seen' from the waveguide. The stored energy in the resonator, expressed in lumped-circuit values, becomes

$$W = \frac{C}{2(V_0T)^2} = n^2\frac{C}{2V_{gen}^2}\,, \qquad \textbf{stored energy} \tag{163}$$

and the dissipated power can be written as

$$P_d = \frac{(V_0T)^2}{2R^c} = n^2\frac{V_{gen}^2}{2R^c}\,. \qquad \textbf{dissipated power} \tag{164}$$

Now we can define the quality factor of the unloaded cavity with lumped-circuit elements:

$$Q_0 = \frac{\omega_0 W}{P_d} = \omega_0 R^c C\,. \qquad \textbf{unloaded } Q \tag{165}$$

When the generator is switched off, not only will the stored energy in the cavity be dissipated in the cavity walls, but a power P_{ex} will also leak out through the power coupler, where

$$P_{\mathrm{ex}} = \frac{V_{\mathrm{gen}}^2}{2Z_0}\,. \tag{166}$$

Using P_{ex}, one can define the quality factor of the external load. The external Q is thus defined as

$$Q_{\mathrm{ex}} = \frac{\omega_0 W}{P_{\mathrm{ex}}} = n^2\omega_0 Z_0 C\,. \qquad \textbf{external } Q \tag{167}$$

4.8.1 Undriven cavity

In order to understand the power balance and matching for a driven cavity with beam, we start with a simple case, assuming that the RF is switched off and that there is no beam in the cavity. The power balance is then

$$P_{\mathrm{tot}} = P_{\mathrm{d}} + P_{\mathrm{ex}}\,, \qquad \textbf{power balance of undriven cavity} \tag{168}$$

with which we can define the so-called 'loaded Q' of the ensemble of cavity and coupler by

$$\frac{1}{Q_{\mathrm{l}}} = \frac{1}{Q_{\mathrm{ex}}} + \frac{1}{Q_0}\,. \qquad \textbf{loaded } Q \tag{169}$$

The coupling between the cavity and the waveguide is described by the coupling factor β, where

$$\beta = \frac{P_{\mathrm{ex}}}{P_{\mathrm{d}}} = \frac{Q_0}{Q_{\mathrm{ex}}} = \frac{R^{\mathrm{c}}}{n^2 Z_0}\,. \qquad \textbf{coupling factor} \tag{170}$$

Optimum power transfer between the cavity (+ coupler) and the waveguide takes place when the impedance at the coupler input equals the waveguide impedance at the resonance frequency of the cavity. We know that the cavity impedance becomes real at resonance, which means that

$$Z_{\mathrm{c}} = R^{\mathrm{c}} = n^2 Z_{\mathrm{c}}' \stackrel{!}{=} n^2 Z_0 \quad \Rightarrow \quad \beta = 1\,. \tag{171}$$

It is important to keep in mind that the 'matching condition' $\beta = 1$ is only valid for a cavity without beam.

4.8.2 RF on, beam on

Once we take the beam loading into account, the power needed in the cavity increases and will yield a different value for the coupling factor β at the point of optimum power transfer. A simple way to introduce the beam is to treat it as an additional loss in the cavity, which can be added to the power dissipated in the cavity walls:

$$P_{\mathrm{db}} = P_{\mathrm{d}} + P_{\mathrm{b}}\,. \qquad \textbf{dissipated power + beam power} \tag{172}$$

As in the case without beam, maximum power transfer to the cavity is achieved when the input impedance of the coupler equals the impedance of the waveguide. This condition yields zero reflection and also implies that the power needed in the cavity, P_{bd} (for losses and beam), has to be equal to P_{ex} as defined in Eq. (166). This means that

$$\frac{P_{\mathrm{ex}}}{P_{\mathrm{db}}} = 1 = \frac{Q_{0\mathrm{b}}}{Q_{\mathrm{ex}}} \quad \Rightarrow \frac{P_{\mathrm{ex}}}{P_{\mathrm{d}}} = 1 + \frac{P_{\mathrm{b}}}{P_{\mathrm{d}}}\,, \tag{173}$$

where we have introduced a quality factor Q_{0b} for the cavity plus beam. For the matched condition, we therefore obtain a coupling factor of

$$\beta = 1 + \frac{P_b}{P_d}, \qquad \textbf{matched coupling factor with beam} \tag{174}$$

and the following quality factors:

$$Q_{ex} = Q_{0b} = \frac{\omega_0 W}{P_b + P_d} = \frac{Q_0}{1 + P_b/P_d} = \frac{Q_0}{\beta}, \qquad \textbf{external } Q \textbf{ with beam} \tag{175}$$

$$Q_l = \frac{Q_0}{1+\beta} = \frac{Q_0}{2 + P_b/P_d}. \qquad \textbf{loaded } Q \textbf{ with beam} \tag{176}$$

In the case of a superconducting cavity, one can generally assume that $P_b \gg P_d$, which means that the coupling factor for the matched condition can be written as

$$\beta = 1 + \frac{P_b}{P_d} \approx \frac{P_b}{P_d}. \qquad \textbf{matched coupling factor for SC cavity + beam} \tag{177}$$

Using

$$P_b = I_{beam} V_0 T \cos\phi_s, \tag{178}$$

we can write a simple expression for calculating the loaded and external Q values for a superconducting cavity as follows:

$$Q_l \approx Q_{ex} \approx \frac{Q_0}{P_{beam}/P_d} = \frac{V_0 T}{(R/Q) I_{beam} \cos\phi_s}. \qquad \textbf{for SC cavity } Q_{l/ex} \tag{179}$$

The results in this paragraph are summarized in Table 3.

Table 3: Definitions of Q values and coupling factors for driven and undriven cavities

	Undriven cavity	**Driven cavity**
General	$\frac{1}{Q_l} = \frac{1}{Q_{ex}} + \frac{1}{Q_0}$ $\beta = \frac{P_{ex}}{P_d} = \frac{Q_0}{Q_{ex}}$ $Q_l = \frac{Q_0}{1+\beta}$	
Matched case	$\frac{P_{ex}}{P_d} = \frac{Q_0}{Q_{ex}} = 1 \Rightarrow \beta = 1$ $Q_{ex} = Q_0 = \frac{\omega_0 W}{P_d}$ $Q_l = \frac{Q_0}{2}$	$\frac{P_{ex}}{P_{db}} = \frac{Q_{0b}}{Q_{ex}} = 1 \Rightarrow \beta = 1 + \frac{P_b}{P_d}$ $Q_{ex} = Q_{0b} = \frac{\omega_0 W}{P_d + P_b}$ $Q_l = \frac{Q_0}{2 + P_b/P_d}$

4.9 'Matching' a cavity

In the last section, it was claimed that part of an electromagnetic wave is reflected when it 'sees' a change in impedance during its propagation. In fact, the whole purpose of the power coupler was to transform the impedance of the waveguide into the impedance of the cavity. In this last section, we shall see why this is so. For this purpose, we look at a transmission line as shown in Fig. 31.

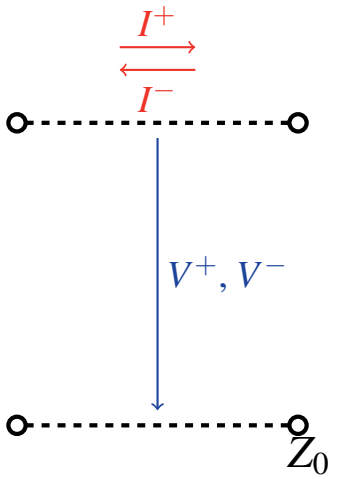


Fig. 31: Voltages and currents along a transmission line

This transmission line is representative of a waveguide, a coaxial line, or any other kind of transport geometry used to guide electromagnetic waves. Since waves can travel in the positive and negative z directions, a sign convention is introduced for the associated voltages and currents, as shown in Fig. 31, where the voltage vectors of the forward and reflected waves have the same direction and the current vectors have opposite directions.

Using the same time and location dependence as for the electric and magnetic fields in a waveguide, we can write

$$V = V_0 e^{i(kz-\omega t)} + \Gamma V_0 e^{i(-kz-\omega t)}, \tag{180}$$

$$I = \frac{V_0}{Z_0} e^{i(kz-\omega t)} - \Gamma \frac{V_0}{Z_0} e^{i(-kz-\omega t)}, \tag{181}$$

where we have introduced a reflection coefficient Γ. If we connect a cavity to an impedance Z_c' at $z = 0$, the expressions above simplify to

$$V = V_0 e^{-i\omega t}(1+\Gamma), \tag{182}$$

$$I = \frac{V_0}{Z_0} e^{-i\omega t}(1-\Gamma), \tag{183}$$

and the cavity impedance can be expressed in terms of the transmission line impedance Z_0 and the reflection coefficient Γ:

$$Z_c' = \frac{V}{I} = Z_0 \frac{1+\Gamma}{1-\Gamma}. \tag{184}$$

We can then rearrange the equation for the reflection coefficient and obtain

$$\Gamma = \frac{Z_c' - Z_0}{Z_c' + Z_0} = \frac{1-\beta}{1+\beta}. \tag{185}$$

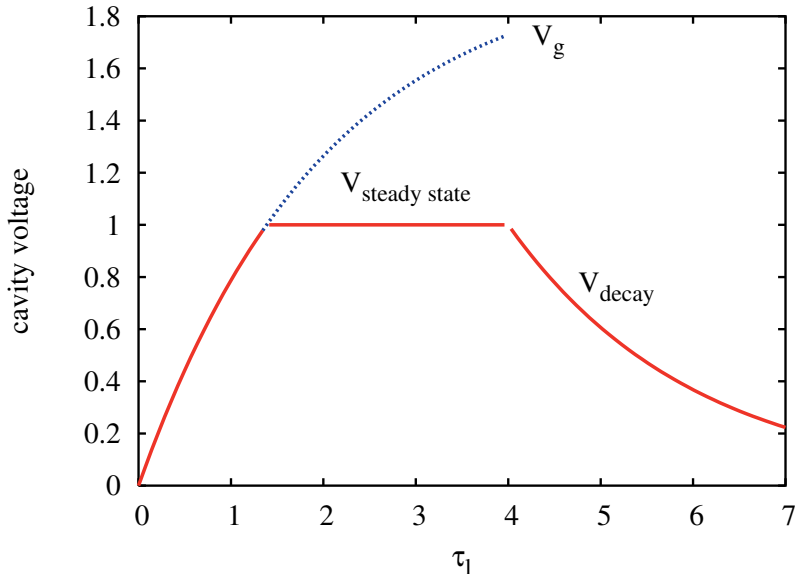


Fig. 32: Voltage profile in a pulsed superconducting cavity

From this equation, we can see that the reflection disappears only for $Z_c' = Z_0$, the 'matched condition', where the waveguide impedance equals the cavity impedance. In the case without beam, this corresponds to a coupling factor of $\beta = 1$.

In the context of matching, we therefore have to consider the following points:

- At the resonance frequency, the power coupler transforms the cavity impedance into the impedance of the waveguide.
- If the cavity is resonating off-resonance or if the coupler is mismatched, power is reflected and travels back to the RF source.
- Since the cavity impedance depends on the Q of the cavity, and since in reality most cavities have different Q values, every cavity needs a different matching.
- Beam loading increases the power needed in the cavity and changes the loaded Q and the cavity impedance. Power couplers are usually matched for the case with beam loading.
- During the start of an RF pulse (before the arrival of the beam), when the cavity is being 'filled' with RF power, the cavity is always mismatched, which means we need to make sure that the reflected power does not damage the RF source (e.g., by using a circulator between the cavity and the RF source).

The last point is especially important in the case of superconducting cavities, where the dissipated power is negligible with respect to the power taken by the beam. In this case one has, basically, full reflection of the RF wave at the beginning of the RF pulse before the cavity field increases to its nominal level. At that point the beam should enter the cavity, and from then onwards the RF generator is matched to the power needs of the RF cavity. After the RF signal is switched off, the cavity voltage decays exponentially, as shown in Fig. 32 (more details can be found in [18]).

Acknowledgements

In the preparation of this chapter, I have made extensive use of the material listed in the Bibliography below.

References

[1] R.P. Feynman, R.B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. II, *Mainly Electromagnetism and Matter* (California Institute of Technology, 1963).

[2] H. Henke, Theoretische Elektrotechnik, German script of lectures on electrodynamics at the Technical University of Berlin (1992).

[3] T. Weiland, M. Krasilnikov, R. Schuhmann, A. Skarlatos, and M. Wilke, Review of theory (I, II, III), CAS RF Engineering, Seeheim, Germany (2005).

[4] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (Dover Publications, New York, 1965).

[5] T. Wangler, *Principles of RF Linear Accelerators* (Wiley-VCH, Weinheim, 2004).

[6] M. Vretenar, Low-beta structures, CAS RF School, Ebeltoft, Denmark (2010).

[7] D.E. Nagle, E.A. Knapp, and B.C. Knapp, *Rev. Sci. Instrum.* **38** (1967) 1583.

[8] S. Schriber, Characteristics of full-cell terminated RF structures: results of analogue studies, CERN/PS/2001-067 (PP) (2001).

[9] S. Schriber, *Phys. Rev. ST Accel. Beams* **4** (2001) 122001.

[10] The Compact Linear Collider Study, http://www.cern.ch/clic-study.

[11] International Linear Collider, http://www.linearcollider.org.

[12] R.H. Miller, Comparison of standing wave and travelling wave structures, LINAC86 (1986).

[13] V.A. Moiseev, V.V. Paramonov, and K. Floettmann, Comparison of standing and travelling wave operations for positron pre-accelerator in the TESLA Linear Collider, EPAC (2000).

[14] F. Gerigk, Cavity types, CAS RF School, Ebeltoft, Denmark, arXiv:1111.4897v1 (2010).

[15] M. Schuh, F. Gerigk, J. Tuckmantel, and C.P. Welsch, Influence of higher order modes on the beam stability in the high power superconducting proton linac, *Phys. Rev. ST Accel. Beams* **14** (2011) 051001.

[16] F. Gerigk, J.M. Giguet, E. Montesinos, B. Riffaud, P. Ugena Tirado, and R. Wegner, The Linac4 power coupler, IPAC 2011, San Sebastian, Spain, CERN-ATS-2011-040 (2011).

[17] M. Vretenar *et al.*, The LINAC4 project at CERN, IPAC 2011, San Sebastian, Spain, CERN-ATS-2011-041 (2011).

[18] F. Gerigk, Formulae to calculate the power consumption of the SPL SC cavities, CERN-AB-Note-2006-011-RF (2006).

Bibliography

– T. Weiland, M. Krasilnikov, R. Schuhmann, A. Skarlatos, and M. Wilke, Review of theory (I, II, III), CAS RF Engineering, Seeheim, Germany (2005).

– T. Wangler, *Principles of RF Linear Accelerators* (Wiley-VCH, Weinheim, 2004).

– A. Wolski, Theory of electromagnetic fields, CAS RF Engineering, Ebeltoft, Denmark (2010).

– H. Henke, Theoretische Elektrotechnik, German script of lectures on electrodynamics at the Technical University of Berlin (1992)

– H. Henke, Basic concepts I and II, CAS RF Engineering, Seeheim, Germany (2005).

– K. Simonyi, *Foundations of Electrical Engineering*, Vol. 3 (Pergamon Press, New York, 1963). [Hungarian edition, *ElmÃl'leti villamossagtan Tankonyvkiado* (Budapest, 1973); German edition, *Theoretische Elektrotechnik* (VEB Deutscher Verlag der Wissenschaften, 1973).]

– H. Padamsee, J. Knobloch, and T. Hays, *RF Superconductivity for Accelerators* (Wiley, New York, 2008).

Appendices

A Cartesian coordinates (x, y, z)

A.1 Differential elements

$$\mathrm{d}\mathbf{l} = \begin{pmatrix} \mathrm{d}x \\ \mathrm{d}y \\ \mathrm{d}z \end{pmatrix}, \qquad \textbf{path element} \tag{A.1}$$

$$\mathrm{d}V = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,. \qquad \textbf{volume element} \tag{A.2}$$

A.2 Differential operators

$$\nabla\phi = \begin{pmatrix} \dfrac{\partial\phi}{\partial x} \\ \dfrac{\partial\phi}{\partial y} \\ \dfrac{\partial\phi}{\partial z} \end{pmatrix}, \qquad \textbf{gradient} \tag{A.3}$$

$$\nabla\cdot\mathbf{a} = \frac{\partial a_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial a_z}{\partial z}, \qquad \textbf{divergence} \tag{A.4}$$

$$\nabla\times\mathbf{a} = \begin{pmatrix} \dfrac{\partial a_z}{\partial y} - \dfrac{\partial a_y}{\partial z} \\ \dfrac{\partial a_x}{\partial z} - \dfrac{\partial a_z}{\partial x} \\ \dfrac{\partial a_y}{\partial x} - \dfrac{\partial a_x}{\partial y} \end{pmatrix}, \qquad \textbf{curl} \tag{A.5}$$

$$\Delta\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2}\,. \qquad \textbf{Laplace} \tag{A.6}$$

B Cylindrical coordinates (r, ϕ, z)

B.1 Transformations

$$\begin{aligned} x &= r\cos\varphi\,, \\ y &= r\sin\varphi\,, \\ z &= z\,, \end{aligned} \tag{B.1}$$

with $0 \le r \le \infty$, $0 \le \varphi \le 2\pi$.

B.2 Differential elements

$$\mathrm{d}\mathbf{l} = \begin{pmatrix} \mathrm{d}r \\ r\,\mathrm{d}\varphi \\ \mathrm{d}z \end{pmatrix}, \qquad \textbf{path element} \tag{B.2}$$

$$\mathrm{d}V = r\,\mathrm{d}r\,\mathrm{d}\varphi\,\mathrm{d}z. \qquad \textbf{volume element} \tag{B.3}$$

B.3 Differential operators

$$\nabla\phi = \begin{pmatrix} \dfrac{\partial\phi}{\partial r} \\ \dfrac{1}{r}\dfrac{\partial\phi}{\partial\varphi} \\ \dfrac{\partial\phi}{\partial z} \end{pmatrix}, \qquad \textbf{gradient} \tag{B.4}$$

$$\nabla\cdot\mathbf{a} = \frac{1}{r}\frac{\partial(ra_r)}{\partial r} + \frac{1}{r}\frac{\partial a_\varphi}{\partial\varphi} + \frac{\partial a_z}{\partial z}, \qquad \textbf{divergence} \tag{B.5}$$

$$\nabla\times\mathbf{a} = \begin{pmatrix} \dfrac{1}{r}\dfrac{\partial a_z}{\partial\varphi} - \dfrac{\partial a_\varphi}{\partial z} \\ \dfrac{\partial a_r}{\partial z} - \dfrac{\partial a_z}{\partial r} \\ \dfrac{1}{r}\left(\dfrac{\partial(ra_\varphi)}{\partial r} - \dfrac{\partial a_r}{\partial\varphi}\right) \end{pmatrix}, \qquad \textbf{curl} \tag{B.6}$$

$$\Delta\phi = \frac{\partial^2\phi}{\partial r^2} + \frac{1}{r}\frac{\partial\phi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\phi}{\partial\varphi^2} + \frac{\partial^2\phi}{\partial z^2}. \qquad \textbf{Laplace} \tag{B.7}$$

C Spherical coordinates (r, ϑ, φ)

C.1 Transformations

$$\begin{aligned} x &= r\sin\vartheta\cos\varphi, \\ y &= r\sin\vartheta\sin\varphi, \\ z &= r\cos\vartheta, \end{aligned} \tag{C.1}$$

with $0 \le r \le \infty$, $0 \le \vartheta \le \pi$, $0 \le \varphi \le 2\pi$.

C.2 Differential elements

$$\mathrm{d}\mathbf{l} = \begin{pmatrix} \mathrm{d}r \\ r\,\mathrm{d}\vartheta \\ r\sin\vartheta\,\mathrm{d}\varphi \end{pmatrix}, \qquad \textbf{path element} \tag{C.2}$$

$$\mathrm{d}V = r^2 \sin\vartheta\,\mathrm{d}r\,\mathrm{d}\vartheta\,\mathrm{d}\varphi\,. \qquad \textbf{volume element} \tag{C.3}$$

C.3 Differential operators

$$\nabla\phi = \begin{pmatrix} \dfrac{\partial\phi}{\partial r} \\ \dfrac{1}{r}\dfrac{\partial\phi}{\partial\vartheta} \\ \dfrac{1}{r\sin\vartheta}\dfrac{\partial\phi}{\partial\varphi} \end{pmatrix}, \qquad \textbf{gradient} \tag{C.4}$$

$$\nabla\cdot\mathbf{a} = \frac{1}{r^2}\frac{\partial(r^2 a_r)}{\partial r} + \frac{1}{r\sin\vartheta}\frac{\partial(a_\vartheta\sin\vartheta)}{\partial\vartheta} + \frac{1}{r\sin\vartheta}\frac{\partial a_\varphi}{\partial\varphi}, \qquad \textbf{divergence} \tag{C.5}$$

$$\nabla\times\mathbf{a} = \begin{pmatrix} \dfrac{1}{r\sin\vartheta}\left(\dfrac{\partial(a_\varphi\sin\vartheta)}{\partial\vartheta} - \dfrac{\partial a_\vartheta}{\partial\varphi}\right) \\ \dfrac{1}{r}\left(\dfrac{1}{\sin\vartheta}\dfrac{\partial a_r}{\partial\varphi} - \dfrac{\partial(r a_\varphi)}{\partial r}\right) \\ \dfrac{1}{r}\left(\dfrac{\partial(r a_\vartheta)}{\partial r} - \dfrac{\partial a_r}{\partial\vartheta}\right) \end{pmatrix}, \qquad \textbf{curl} \tag{C.6}$$

$$\Delta\phi = \frac{\partial^2\phi}{\partial r^2} + \frac{2}{r}\frac{\partial\phi}{\partial r} + \frac{1}{r^2\sin\vartheta}\frac{\partial}{\partial\vartheta}\left(\sin\vartheta\frac{\partial\phi}{\partial\vartheta}\right) + \frac{1}{r^2\sin^2\vartheta}\frac{\partial^2\phi}{\partial\varphi^2}\,. \qquad \textbf{Laplace} \tag{C.7}$$

D Useful relationships

$$\nabla\cdot(\mathbf{a}\times\mathbf{b}) = \mathbf{b}\cdot(\nabla\times\mathbf{a}) - \mathbf{a}\cdot(\nabla\times\mathbf{b})\,, \tag{D.1}$$

$$\nabla\cdot(\nabla\times\mathbf{a}) = 0\,. \tag{D.2}$$

Beam loading

Alexander Gamp
DESY, Hamburg, Germany

Abstract
We begin by giving a description of the radio-frequency generator–cavity–beam coupled system in terms of basic quantities. Taking beam loading and cavity detuning into account, expressions for the cavity impedance as seen by the generator and as seen by the beam are derived. Subsequently methods of beam-loading compensation by cavity detuning, radio-frequency feedback and feedforward are described. Examples of digital radio-frequency phase and amplitude control for the special case of superconducting cavities are also given. Finally, a dedicated phase loop for damping synchrotron oscillations is discussed.

1 Introduction

In modern particle accelerators, radio-frequency (RF) voltages in an extremely large amplitude and frequency range, from a few hundred volts to hundreds of megavolts and from some kilohertz to many gigahertz, are required for particle acceleration and storage.

The RF power, which is needed to satisfy these demands, can be generated, for example, by triodes, tetrodes, klystrons or by semiconductor devices. The continuous wave (cw) output power available from some tetrodes which were used at HERAp is 60 kW at 208 MHz and up to 800 kW for the 500 MHz klystrons for the new synchrotron light source PETRA III. The 1.3 GHz klystrons for the free-electron laser FLASH at DESY can deliver up to 10 MW RF peak power during pulses of about 1 ms length. Even higher power levels can be obtained from S- and X-band klystrons during pulse lengths on the microsecond scale.

Such RF power generators generally deliver RF voltages of only a few kilovolts because their source impedance or their output waveguide impedance is small compared with the shunt impedance of the cavities in the accelerators.

Typically, a tetrode has its highest efficiency for a load resistance of less than a kiloohm whereas the cavity shunt impedance usually is of the order of several megaohms. This is the real impedance, which the cavity represents to a generator at the resonant frequency. It must not be confused with ohmic resistances.

Optimum fixed impedance matching between generator and cavity can be easily achieved with a coupling loop in the cavity. There is, however, the complication that the transformed cavity impedance as seen by the generator depends also on the synchronous phase angle and on the beam current and is therefore not constant as we will show quantitatively. The beam current induces a voltage in the cavity, which may become even larger than that induced by the generator. Owing to the vector addition of these two voltages the generator now sees a cavity which appears to be detuned and unmatched except for the particular value of beam current for which the coupling has been optimized. The reflected power occurring at all other beam currents has to be handled.

In addition, the beam-induced cavity voltage may cause single or multibunch instabilities, since any bunch in the machine may see an important fraction of the cavity voltage induced by itself or from previous bunches. This voltage is given by the product of beam current and cavity impedance as seen by the beam. Minimizing this latter quantity is therefore essential. It is also called beam loading compensation, and some servo control mechanisms, which can be used to achieve this goal, will be discussed here.

2 The coupling between the RF generator, the cavity and the beam

For frequencies in the neighbourhood of the fundamental resonance, an RF cavity can be described [1] by an equivalent circuit consisting of an inductance L_2, a capacitor C and a shunt impedance R_S as is shown in Fig. 1. In practice, L_2 is made up by the cavity walls whereas the coupling loop L_1 is usually small as compared with the cavity dimensions.

In this example a triode with maximum efficiency for a real load impedance R_A has been taken as an RF power generator. For simplicity, we consider a short and lossless transmission line between the generator and L_1. Then there is optimum coupling between the generator and the empty (i.e. without beam) cavity for

$$N^2 = R_S / R_A = L_2 / L_1 \tag{1}$$

where, for maximum power output, R_A equals the dynamic source impedance R_I. Here N is called the transformation or step-up ratio.

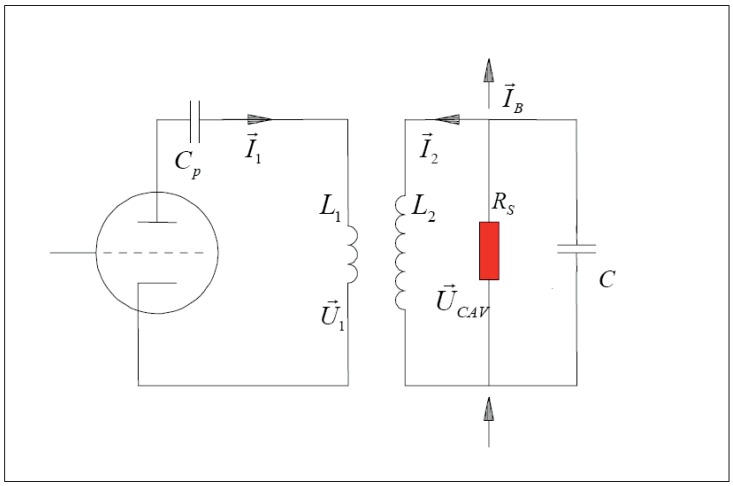


Fig. 1: Equivalent circuit of a resonant cavity near its fundamental resonance. In practice, the inductance L_2 is made up by the cavity walls whereas L_1 usually is a small coupling loop. The capacitor C denotes the equivalent cavity capacitance whereas C_p is needed only for separation of the plate dc Voltage from the rest of the circuit.

Since, in general, there may be power transmitted from the generator to the cavity and also, in the case of imperfect matching, vice versa, the voltage $\vec{U}_1$ is expressed as the sum of two voltages

$$\vec{U}_1 = \vec{U}_{forward} + \vec{U}_{reflected} \tag{2}$$

whereas the corresponding currents flow in the opposite directions, hence

$$\vec{I}_1 = \vec{I}_{forward} - \vec{I}_{reflected} \tag{3}$$

The minus sign in Eq. (3) indicates the counterflowing currents while voltages of forward and backward waves just add up.

So, in the simplest case where the beam current $\vec{I}_B = 0$ and where the generator frequency $f_{Gen} = f_{Cav}$, there is no reflected power from the cavity to the generator, and $\vec{U}_1$ and $\vec{I}_1$ are identical to the generator voltage and current, respectively. One has

$$\vec{U}_{CAV} = N\vec{U}_1 \tag{4}$$

Now we can derive an expression for the complex cavity voltage as a function of the generator and beam current and of the cavity and generator frequency.

According to Fig. 1 the cavity voltage $\vec{U}_{CAV}$ can be written as

$$\vec{U}_{CAV} = L_2\left(\dot{\vec{I}}_2 + \dot{\vec{I}}_1 / N\right) \tag{5}$$

$$\vec{I}_2 = -\left(\vec{I}_B + \vec{U}_{CAV} / R_S + C\dot{\vec{U}}_{CAV}\right) \tag{6}$$

All voltages and currents have the time dependence

$$\vec{U} = \hat{\vec{U}}e^{i\omega t} \tag{7}$$

Here $\vec{I}_B = \vec{I}_B(\omega)$ is the harmonic content at the frequency ω of the total beam current. Throughout this article we consider only a bunched beam with a bunch spacing that is small compared with the cavity filling time. In this case $\vec{I}_B(\omega)$ is quasi-sinusoidal. We also restrict the discussion to the interaction of the beam with the fundamental cavity resonance. The interaction with higher-order cavity modes can be minimized by dedicated damping antennas built into the cavity.

Inserting Eq. (6) into Eq. (5) and using

$$2\pi f_{CAV} = \omega_{CAV} = \frac{1}{\sqrt{L_2 C}} \tag{8}$$

one finds

$$\omega_{CAV}^2 \vec{U}_{CAV} = \frac{1}{C}\left[\frac{1}{N}\dot{\vec{I}}_1 - \dot{\vec{I}}_B - \frac{1}{R_S}\dot{\vec{U}}_{CAV}\right] - \ddot{\vec{U}}_{CAV} \tag{9}$$

We define

$$\Gamma = \frac{1}{2CR_S} = \frac{\omega_{CAV}}{2Q} \tag{10}$$

where the quality factor of the cavity can be expressed as 2π times the ratio of total electromagnetic energy stored in the cavity to the energy loss per cycle.

Here we would like to mention that the ratio

$$\frac{R_S}{Q} = \sqrt{\frac{L_2}{C}} \tag{11}$$

is a characteristic quantity of a cavity depending only on its geometry.

We can rewrite Eq. (9) as

$$\ddot{\vec{U}}_{CAV} + 2\Gamma\dot{\vec{U}}_{CAV} + \omega_{CAV}^2\vec{U}_{CAV} = 2\Gamma R_S\left[\frac{1}{N}\dot{\vec{I}}_1 - \dot{\vec{I}}_B\right] \tag{12}$$

This equation describes a resonant circuit excited by the current $\vec{I} = (\vec{I}_1/N - \vec{I}_B)$. The minus sign occurs because the generator-induced cavity voltage has opposite sign to the beam-induced voltage, which would decelerate the beam. It can be shown that the beam actually sees only 50 % of its own induced voltage. This is called the fundamental theorem of beam loading [2, 3].

2.1 The impedance of the generator loaded cavity as seen by the beam

To find the cavity impedance as seen by the beam we make use of Eqs. (2), (3) and (4) to express the generator current term of Eq. (12) in the form

$$\frac{1}{N}\dot{\vec{I}}_1 = \frac{1}{NR_A}\left[2\dot{\vec{U}}_{forward} - \dot{\vec{U}}_1\right] = \frac{1}{N}\left[2\dot{\vec{I}}_{forward} - \frac{\dot{\vec{U}}_{CAV}}{NR_A}\right] \tag{13}$$

The new term $\frac{\dot{\vec{U}}_{CAV}}{NR_A}$ leads to a modification of the damping term in (12)

$$\ddot{\vec{U}}_{CAV} + 2\Gamma(1+\beta)\dot{\vec{U}}_{CAV} + \omega_{CAV}^2\vec{U}_{CAV} = 2\Gamma_L R_{SL}\left[\frac{2}{N}\dot{\vec{I}}_f - \dot{\vec{I}}_B\right] \tag{14}$$

With the coupling ratio

$$\beta = R_S/(N^2R_A) \tag{15}$$

we can introduce the "loaded" damping term

$$\Gamma_L = \Gamma(1+\beta) \tag{16}$$

and consequently, in accordance with Eq. (10), the loaded cavity Q and loaded shunt impedance are

$$Q_L = Q/(1+\beta) \quad \text{and} \quad R_{SL} = R_S/(1+\beta) \tag{17}$$

In the case of perfect matching in the absence of beam, i.e. $\beta = 1$, the damping term simply doubles and Q and R_S take half their original values. This is due to the fact that the beam would see the cavity shunt impedance R_S in parallel or loaded with the transformed generator impedance $N^2 R_A = R_S$. Therefore, we find in Eq. (14) that the transformed generator current

$$\vec{I}_G = 2\vec{I}_f / N \tag{18}$$

gives rise to twice as much cavity voltage as a similar beam current would do. Here and in Eq. (15) we assume that the transformed dynamic source impedance $N^2 R_A$ is identical to the generator impedance seen by the cavity. This is strictly true only if a circulator is placed between the RF power generator and the cavity. Without a circulator it may be approximately true if the power source is a triode. Owing to its almost constant anode voltage-to-current characteristic, the impedance of a tetrode as seen from the cavity is, however, much larger than the corresponding R_A and therefore $R_{SL} \approx R_S$ in this case where a short transmission line (or of length $n\lambda/2$, n is an integer) is considered.

Following [4] we write the solution of Eq. (14) in the Fourier–Laplace representation

$$\hat{\vec{U}}_{CAV} = \frac{i\omega}{\omega_{CAV}^2 - \omega^2 + 2i\omega\Gamma_L} 2\Gamma_L R_{SL}\left[\hat{\vec{I}}_G - \hat{\vec{I}}_B\right] \tag{19}$$

For $\Delta\omega << \omega_{CAV}$ this can be approximated by

$$\hat{\vec{U}}_{CAV} \approx \frac{R_{SL}\left[\hat{\vec{I}}_G - \hat{\vec{I}}_B\right]}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}}} \tag{20}$$

where

$$\omega = \omega_{CAV} + \Delta\omega .$$

A plot of the cavity voltage modulus and its real and imaginary part as a function of $\Delta\omega$ is shown in Fig. 2. For a resonant cavity the beam-induced voltage $\hat{\vec{U}}_B$, or the beam loading, is thus given by the product of loaded shunt impedance and beam current:

$$\hat{\vec{U}}_B = -R_{SL}\hat{\vec{I}}_B \tag{21}$$

The ideal beam loading compensation would, therefore, minimize R_{SL} without increasing the generator power necessary to maintain the cavity voltage.

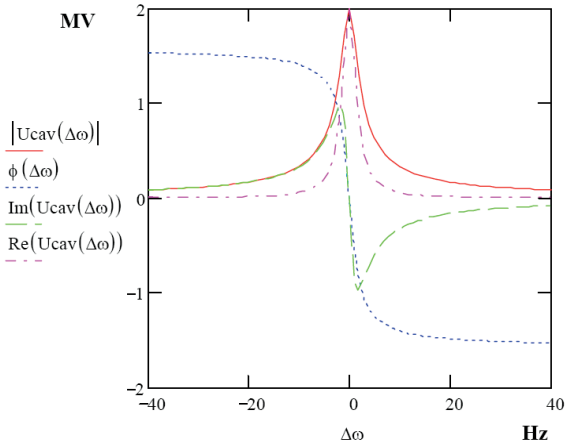


Fig. 2: Plot of the envelope of the cavity voltage modulus, its real and imaginary part calculated with (20) and of the detuning angle calculated with Eq. (43).

The beam-induced voltages are by no means negligible. For a loaded shunt impedance of, say 2.5 MΩ, and a beam current of 0.2 A, the induced voltage would be 0.5 MV. To compensate for this, a generator current of 20 A would be needed for a typical transformation ratio $N = 100$. This may lead to large values of reflected power which must be taken into consideration when designing the RF system.

2.2 The impedance of the beam loaded cavity as seen by the generator

Having just discussed the impedance, which the combined system generator and cavity represents to the beam, we would like to discuss in the following the impedance, Z, or rather admittance, $Y = 1/Z$, which the combined cavity and beam system represents to the generator.

From Eqs. (1), (5) and (6) one sees [5] that

$$Y = \frac{\vec{I}_1}{\vec{U}_1} = \frac{N^2}{R_S} + \frac{\vec{I}_B N^2}{\vec{U}_{CAV}} + \frac{N^2}{i\omega L_2}\left(1 - \frac{\omega^2}{\omega_{CAV}^2}\right) \tag{22}$$

which reduces to $Y = N^2/R_S = 1/R_A$ for a tuned cavity without beam current in the case of $\beta = 1$.

As we are now going to show, a non-vanishing real part of the quotient $\vec{I}_B/\vec{U}_{CAV}$ will necessitate a change in β to maintain optimum matching whereas the imaginary part can be compensated by detuning the cavity. To work out Re and $\mathrm{Im}(\vec{I}_B/\vec{U}_{CAV})$ we define the angle ϕ_s, as the phase angle between the synchronous particle and the zero crossing of the RF cavity voltage. The accelerating voltage is therefore given by

$$\vec{U}_{ACC} = \vec{U}_{CAV} \sin\phi_s \tag{23}$$

and the normalized cavity voltage and beam current are related by

$$\frac{\vec{I}_B}{\vec{U}_{CAV}} = \frac{\left|\vec{I}_B\right|}{\left|\vec{U}_{CAV}\right|} e^{i\left(\frac{\pi}{2} - \phi_s\right)} \tag{24}$$

Consequently,

$$\mathrm{Re}\left(\frac{\vec{I}_B}{\vec{U}_{CAV}}\right)=\left|\frac{\vec{I}_B}{\vec{U}_{CAV}}\right|\sin\phi_s \tag{25}$$

and

$$\mathrm{Im}\left(\frac{\vec{I}_B}{\vec{U}_{CAV}}\right)=\left|\frac{\vec{I}_B}{\vec{U}_{CAV}}\right|\cos\phi_s \tag{26}$$

The real part of the admittance seen by the generator then becomes

$$\mathrm{Re}(Y)=\frac{N^2}{R_S}\left(1+\frac{R_S\left|\vec{I}_B\right|}{\left|\vec{U}_{CAV}\right|}\sin\phi_s\right) \tag{27}$$

We see that the term in the parentheses describes a change in admittance caused by the beam. To maintain optimum coupling the coupling ratio β must now take the value

$$\beta=\left(1+\frac{R_S\left|\vec{I}_B\right|}{\left|\vec{U}_{CAV}\right|}\sin\phi_s\right) \tag{28}$$

This result tells us that the change in the real part of the admittance is proportional to the ratio of RF power delivered to the beam to RF power dissipated in the cavity walls. For circular electron machines, where the considerable amount of energy lost by synchrotron radiation has to be compensated for continuously by RF power, values of $\varphi_s \geq 30°$ and $\beta \geq 1.2$ are typical for high beam current and normal conducting cavities. A typical set of parameters for this case would be R_S = 6 MΩ, $\vec{U}_{CAV}=1$ MV and $\vec{I}_B(\omega)=60$ mA. This implies, of course, that for a β, which has been optimized for the maximum beam current, there will be reflected generator power for lower beam intensities. If the power source is a klystron, this can be handled by inserting a circulator in the path between generator and cavity or, in the case of a tube, by a sufficiently high plate dissipation power capability.

For superconducting cavities the situation is totally different. Here a typical set of parameters would be $R_S=10^{13}$ Ω, $\vec{U}_{CAV}=25$ MV, $\vec{I}_B(\omega)=16$ mA and $\phi_s=90^0$. Then $\beta=6401$, and for a typical unloaded $Q=10^{10}$ the loaded quantity becomes $Q_L=1.6\times10^6$. If the loaded Q is adjusted to this value, so that there is no reflection of RF power back to the cavity at the nominal beam current, it means also that there is a strong mismatch and hence almost total reflection without beam.

The complex reflection coefficient is given by

$$r(\beta,\Delta\omega)=\frac{\beta-1-\dfrac{iQ2\Delta\omega}{\omega_{cav}}}{\beta+1+\dfrac{iQ2\Delta\omega}{\omega_{cav}}} \tag{29}$$

On resonance it simplifies to

$$r(\beta)=\frac{\beta-1}{\beta+1}=\frac{Z-Z_0}{Z+Z_0}=\frac{\vec{U}_{reflected}}{\vec{U}_{forward}}=\frac{6400}{6402}\approx 1 \tag{30}$$

The voltage standing wave ratio then becomes

$$VSWR=\frac{\vec{U}_{forward}+\vec{U}_{reflected}}{\vec{U}_{forward}-\vec{U}_{reflected}}=\frac{1+r}{1-r}\approx\infty \tag{31}$$

So, for superconducting cavities beam loading is even more dramatic than it may be for normal conducting cavities, since situations where total reflection of the incident generator power occurs during significant time intervals are unavoidable.

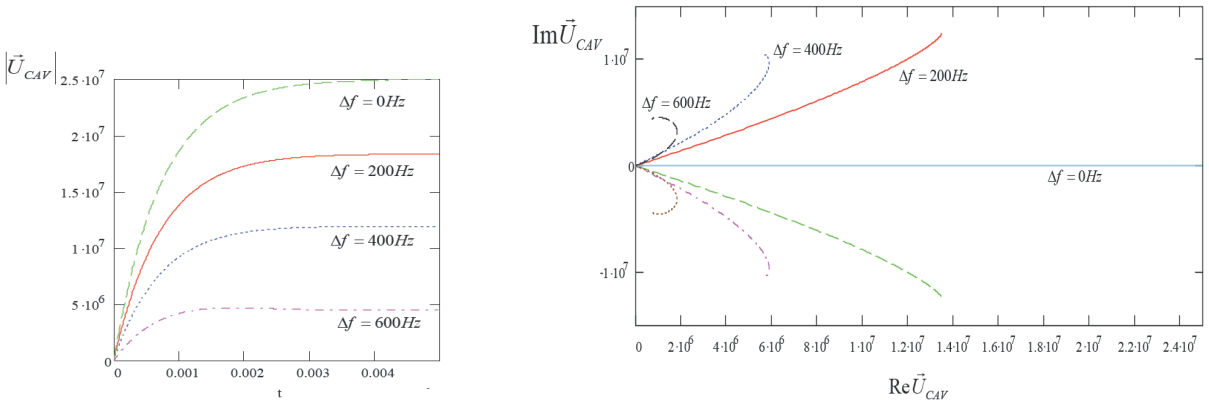
It is instructive to look at the time dependence of the envelope of the cavity voltage and of the reflected voltage. The solution of Eq. (14) yields for the envelope of the cavity voltage during filling

$$\vec{U}_{CAV}(t)\approx\frac{\hat{\vec{U}}_{CAV}(1-e^{-(1/\tau-i\Delta\omega)t})}{1+iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}}} \tag{32}$$

with the time constant

$$\tau=2Q_L/\omega_{CAV} \tag{33}$$

See Fig. 3.



(a) (b)

Fig. 3: (a) Modulus and (b) real and imaginary part of the cavity voltage calculated with Eq. (32) as a function of the detuning frequency.

On resonance Eq. (32) simplifies to

$$\vec{U}_{CAV}(t)=\hat{\vec{U}}_{CAV}(1-e^{-t/\tau}) \tag{34}$$

From the value P_{CAV} of the power transmitted into the cavity

$$P_{CAV} = P_{forward} - P_{reflected} = P_{forward}(1-r^2) = P_{forward}\frac{4\beta}{(1+\beta)^2} \tag{35}$$

the asymptotic value $\vec{U}_{CAV}(\infty) = \hat{\vec{U}}_{CAV}$ can be obtained as function of β:

$$\hat{\vec{U}}_{CAV} = \sqrt{2R_S P_{forward}\frac{4\beta}{(1+\beta)^2}} \tag{36}$$

The reflected voltage can be expressed in terms of the forward voltage and β

$$\vec{U}_{reflected} = \vec{U}_{forward}\frac{\beta-1}{\beta+1} \qquad \vec{U}_1 = \vec{U}_{forward} + \vec{U}_{reflected} = \vec{U}_{forward}\frac{2\beta}{1+\beta} \tag{37}$$

From Eqs. (37) and (2) one finds

$$\vec{U}_{reflected}(t) = \frac{1}{N}\hat{\vec{U}}_{CAV}(1-e^{-t/\tau}) - \vec{U}_{forward} = \hat{\vec{U}}_1\left[(1-e^{-t/\tau}) - \frac{1+\beta}{2\beta}\right] \tag{38}$$

See Fig. 4. For the matched case where $\beta = 1$ one sees that the reflected voltage reaches 0 asymptotically as the cavity voltage reaches the value $\vec{U}_{CAV}(\infty) = \hat{\vec{U}}_{CAV}$ given by Eq. (36). For $\beta >> 1$, however, $\vec{U}_{reflected}(t) = 0$ only at the time

$$t_{U_{refl}=0} = -\tau \ln(1 - \frac{1+\beta}{2\beta}) \tag{39}$$

At this time the cavity voltage has reached exactly the voltage for which β has been calculated with Eq. (28) for a given beam current, which is about half of the asymptotic value:

$$\vec{U}_{CAV}(t_{U_{refl}=0}) = \hat{\vec{U}}_{CAV}\frac{1+\beta}{2\beta} \approx 0.5\hat{\vec{U}}_{CAV} \tag{40}$$

This can be illustrated by taking the beam-induced voltage into account when calculating the envelope of the cavity voltage. In Eq. (41) the case where the beam is injected at $t_{U_{refl}=0}$ is considered:

$$\vec{U}_{CAV}(t) = \hat{\vec{U}}_{CAV}(1-e^{-t/\tau}) - \hat{\vec{U}}_{Beam}(1-e^{-t(>t_{U_{refl}=0})/\tau}) \tag{41}$$

This is shown in Fig. 5. The beam-induced voltage increases with the same time constant as the cavity voltage, but starting only at $t_{U_{refl}=0}$ and, in this example, with opposite sign. Therefore, the sum of the two voltages remains constant for $t > t_{U_{refl}=0}$.

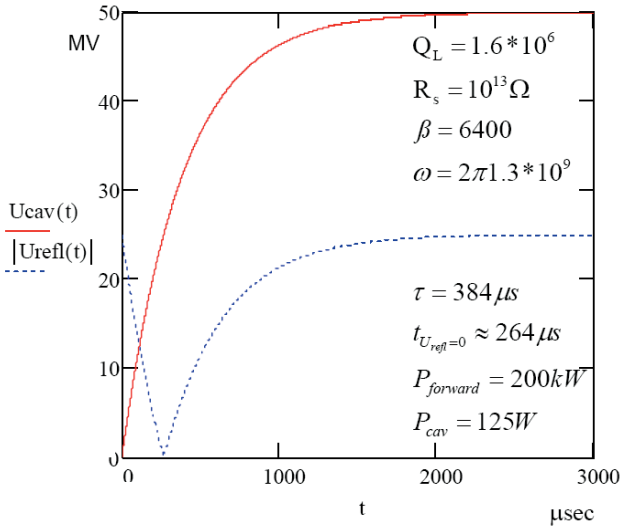


Fig. 4: Plot of the envelope of the cavity voltage and modulus of the reflected voltage calculated from Eqs. (34) and (38)

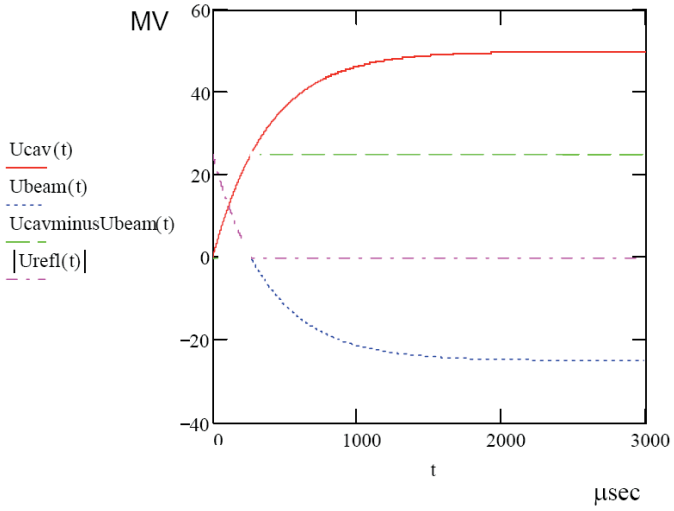


Fig. 5: Same as Fig. 4, but with the beam injected at $t_{U_{refl}=0}$. Then, for $t > t_{U_{refl}=0}$ the reflected voltage remains 0, since now there is matching with the beam, and the cavity voltage stays constant since both generator-induced and beam-induced voltages increase with the same time constant but with opposite sign. This is indicated by the dashed line calculated with (41) which coincides with the full line for $t < t_{U_{refl}=0}$.

So far we have seen that pure real beam loading, where $\hat{\tilde{U}}_B$ and $\hat{\tilde{U}}_{Gen}$ are either in phase or opposite, can be compensated for by adjustment of β and generator power.

Now we show that in contrast to this, pure reactive beam loading, where $\hat{\tilde{U}}_B$ and $\hat{\tilde{U}}_{Gen}$ are in quadrature, can be compensated for by detuning the cavity. That means that the original cavity voltage

can be restored by detuning the cavity. No additional generator power is needed in steady state, but for transient beam loading compensation significantly more power may be needed.

From the imaginary part of Eq. (22) and from Eq. (26) we find that the apparent cavity detuning caused by the beam current can be compensated for by a real cavity detuning (for example, by means of a mechanical plunger cavity tuner) of the amount

$$\frac{\omega}{\omega_{CAV}} = \sqrt{1 + \frac{R_S \left|\vec{I}_B\right|}{Q\left|\vec{U}_{CAV}\right|}\cos\phi_s} \tag{42}$$

Expanding the square root to first order we find a cavity detuning angle Ψ

$$\tan\Psi \approx \frac{R_{SL}\left|\vec{I}_B\right|}{\left|\vec{U}_{CAV}\right|}\cos\phi_s \approx 2Q_L\frac{\Delta\omega}{\omega_{CAV}} \tag{43}$$

This is essentially the ratio between the beam-induced and total cavity voltage.

To calculate the maximum amount of reflected power seen by the generator as a consequence of beam loading, we consider, for $\beta = 1$, a tuned cavity, i.e. $\omega = \omega_{CAV}$. Then, with Eqs. (25) and (26), Eq. (22) reads

$$Y = \frac{1}{R_A}\left[1 + \frac{R_S\left|\vec{I}_B\right|}{\left|\vec{U}_{CAV}\right|}\sin\phi_S + i\frac{R_S\left|\vec{I}_B\right|}{\left|\vec{U}_{CAV}\right|}\cos\phi_S\right] \tag{44}$$

Solving for $\vec{U}_{refl.}$ by means of Eqs. (2) and (3) the reflected power $P_{refl.} = \left|\hat{\vec{U}}_{refl.}\right|^2 \Big/ 2R_A$ becomes

$$P_{refl.} = R_S\hat{\vec{I}}_B^{\,2}/8 \tag{45}$$

This corresponds to half of the power given by the beam to the coupled system cavity and generator. The second half of this power is dissipated in the cavity walls. All we found is that two equal resistors in parallel dissipate equal amounts of power. As we pointed out above, this is strictly true only if a circulator is placed in between the RF power source and the cavity. Nevertheless, the amount of reflected power can be quite impressive. For an average dc beam current of, say, 0.1 A the harmonic current $\hat{\vec{I}}_B(\omega)$ may become up to twice as large. Then, taking R_S = 8 MΩ, for example, we find 40 kW of reflected power, which has to be dissipated.

For a cavity where only the reactive part of the beam loading has been compensated for by detuning according to Eq. (43), but $\beta = 1$, the reflected power is given by

$$P_{refl.} = R_S\hat{\vec{I}}_B^{\,2}\sin^2\phi_s/8 \tag{46}$$

Summarizing the results of this section we state that the beam sees the cavity shunt impedance in parallel with the transformed generator impedance. The resulting loaded impedance is reduced by the factor $1/(1+\beta)$. The optimum coupling ratio between generator and cavity depends on the amount of

energy taken by the beam out of the RF field. The coupling is usually fixed and optimized for the maximum beam current. The amount of cavity detuning necessary for optimum matching, on the other hand, depends on the ratio of beam-induced to total cavity voltage. Clearly these issues depend also on the synchronous phase angle.

3 Beam-loading compensation by detuning

In Fig. 6 a diagram of a tuner regulation circuit is shown. The phase detector measures the relative phase between the generator current and cavity voltage which depends, according to Eq. (43), on the frequency $\Delta\omega$ by which the cavity is detuned. The phase detector output signal acts on a motor which drives a plunger tuner into the cavity volume until there is resonance. An alternative tuner could be a resonant circuit loaded with ferrites. The magnetic permeability μ of the ferrites and hence the resonance frequency of the circuit can be controlled by a magnetic field. This latter method is especially useful when a large tuning range in combination with a low cavity Q is required.

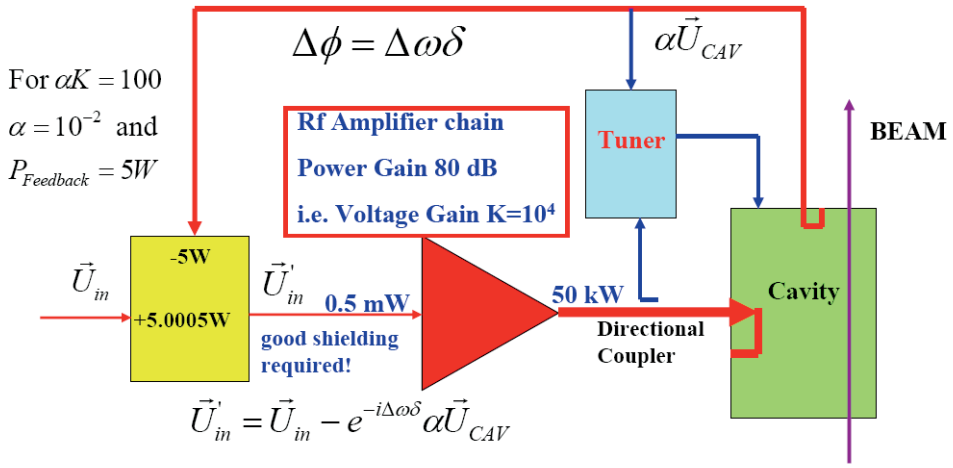


Fig. 6: Schematic of servo loops for phase and amplitude control of the HERA 208 MHz proton RF system

If proper tuner action is necessary in a large dynamic range of cavity voltages, limiters with a minimum phase shift per decibel compression have to be installed at the phase detector input. Since this phase shift is decreasing with frequency all signals should be mixed down to a sufficiently low intermediate frequency.

The signal proportional to the generator current $\vec{I}_{forw.}$ can be obtained from a directional coupler. In case the RF amplifier is so closely coupled to the cavity that no directional coupler can be installed, the relative phase between the RF amplifier input and output signal can also be used to derive a tuner signal [6].

As we have shown in the previous paragraph, stationary beam loading can be entirely compensated for by detuning the cavity, if the synchronous phase angle is small or zero. This is usually the case in proton synchrotrons during storage, where the energy loss due to the emission of synchrotron radiation is negligible. Here, the RF voltage is needed only to keep the bunch length short. Energy ramping also takes place at very small ϕ_s.

In the following, we restrict ourselves, for simplicity, to hadron machines. Consequently, $\beta = 1$, $\phi_s \approx 0$, and the generator-induced and beam-induced voltages are in quadrature.

There are, however, also in this case, several limitations to detuning as the only means of beam-loading compensation. One is known as Robinson's stability criterion [7], which we briefly explain here.

We consider a perturbation voltage $\vec{U}_{pertub.}(t) = \hat{\vec{U}}_{perturb.} e^{i\Omega t}$ such that

$$\vec{U}_{CAV}(t) = (\hat{\vec{U}}_{CAV} + \hat{\vec{U}}_{perturb.} e^{i\Omega t}) e^{i\omega_{CAV} t} \tag{47}$$

If Ω is close to Ω_S, a coherent synchrotron oscillation of all bunches with a damping constant D_S may be excited. This oscillation leads to two new frequency components $\omega \pm \Omega$ in the beam current frequency spectrum. These two components will induce additional RF voltages in the cavity. Their amplitudes are unequal, since $Z_{Cav}(\omega) \approx R_{SL}/(1 + iQ_L 2\Delta\omega/\omega_{CAV})$ and, hence, with $R(\omega) = \mathrm{Re}\, Z(\omega)$,

$$R(\omega + \Omega) \neq R(\omega - \Omega) \tag{48}$$

These two induced voltages act back on the beam current, and when the induced voltage has the same phase as and larger amplitude than the perturbation voltage the oscillation will grow and become unstable.

The stability condition can be written as

$$\frac{R(\omega + \Omega) - R(\omega - \Omega)}{\vec{U}_{CAV} \sin\phi_S} \vec{I}(\omega) < 4 \frac{D_S}{\Omega_S} \tag{49}$$

where $\omega = \omega_{beam} = h\omega_{revolution}$ and $h =$ harmonic number.

This result from Piwinski [5], which agrees with the Robinson criterion, is illustrated in Fig. 7.

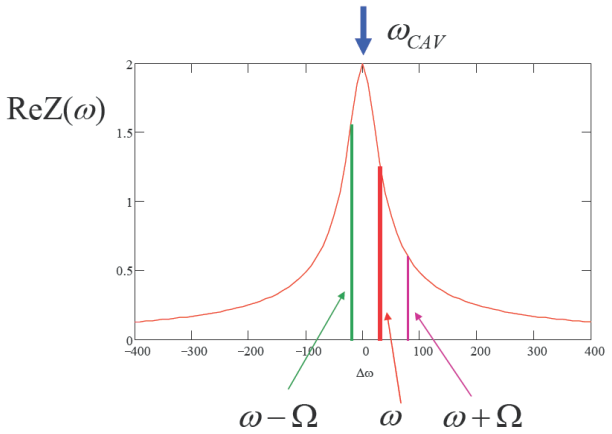


Fig. 7: Illustration of a Robinson-stable scenario since $\omega_{CAV} < \omega$ and, hence, $R(\omega + \Omega) < R(\omega - \Omega)$ (see the text)

The situation becomes more complex when there are additional resonances or cavity modes close to other revolution harmonics of the beam current $\vec{I}_B(\omega)$ which may also lead to instabilities.

Also the spectrum of the beam can become much more complicated as is schematically indicated in Fig. 8 where only the fundamental synchrotron oscillation mode is drawn.

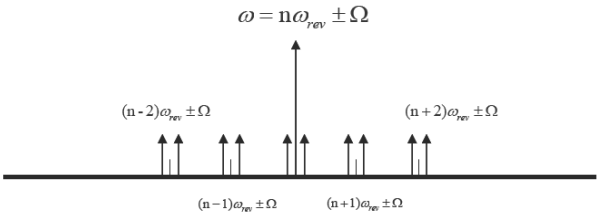


Fig. 8: Example of a beam spectrum with nearby revolution harmonics and synchrotron frequency sidebands

Damping of synchrotron oscillations can be achieved by several means. One possibility consists of an additional passive cavity with an appropriate resonance to change $R(\omega+\Omega)$ and $R(\omega-\Omega)$ such that the stability criterion (49) is fulfilled.

Another possibility is an additional acceleration voltage with slightly smaller frequency to separate the synchrotron frequencies of different bunches such that the oscillation is damped by decoherence. An active phase loop for damping synchrotron oscillations will be described in the final section.

The beam will also become unstable if the amount of detuning calculated by Eq. (43) becomes comparable to the revolution frequency of the particles in a synchrotron. The finite time of, say, a second, which is needed for the tuner to react can also create instabilities. Actually, the time scale of the cavity voltage transients, which may cause beam instabilities, is much shorter. According to Eq. (34) the cavity voltage rise after injection of a bunched beam with a current $\vec{I}_B(\omega_{CAV})$ can be approximated by

$$\vec{U}_B \approx R_{SL}\vec{I}_B\left(1-e^{-t/\tau}\right) \tag{50}$$

This voltage will add to the cavity voltage produced by the generator, and after a time $t \approx 3\tau$ the total cavity voltage becomes

$$\left|\vec{U}_{CAV}\right| \approx R_{SL}\sqrt{\left|\vec{I}_g\right|^2+\left|\vec{I}_B\right|^2} \tag{51}$$

with a phase shift given by Eq. (43).

Since, for normal conducting cavities, typical values of τ are below 100 μs and therefore much smaller than the proton synchrotron frequency in a storage ring (T_S is usually ≥ some milliseconds), these transients will, in general, excite synchrotron oscillations of the beam with the consequence of emittance blow-up and particle loss or even total beam loss. Additional compensation of transient beam loading is therefore necessary. Individual phase and amplitude loops may become unstable due to the correlation of both quantities [8, 9].

In the following section we discuss fast feedback as a possibility to overcome these problems.

4 Reduction of transient beam loading by fast feedback

The principle of a fast feedback circuit is illustrated in Fig. 6. A small fraction α of the cavity RF signal is fed back to the RF preamplifier input and combined with the generator signal. The total delay δ in the feedback path is such that both signals have opposite phase at the cavity resonance frequency. For other frequencies there is a phase shift

$$\Delta\phi = \Delta\omega\delta \quad (52)$$

Therefore, the voltage at the amplifier input is now given by

$$\vec{U}_{in}^{'} = \vec{U}_{in} - e^{-i\Delta\omega\delta}\alpha\vec{U}_{CAV} \quad (53)$$

With the voltage gain K of the amplifier we can rewrite Eq. (20) and obtain for the cavity voltage with feedback

$$\vec{U}_{CAV} \approx \frac{K\left[\vec{U}_{in} - e^{-i\Delta\omega\delta}\alpha\vec{U}_{CAV}\right] - \vec{U}_B}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}}} \quad (54)$$

or

$$\vec{U}_{CAV} \approx \frac{K\vec{U}_{in} - \vec{U}_B}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}} + e^{-i\Delta\omega\delta}\alpha K} \quad (55)$$

For $\Delta\omega = 0$ and $A_F >> 1$ this reduces to

$$\vec{U}_{CAV} \approx \frac{\vec{U}_{in}}{\alpha} - \frac{\vec{U}_B}{\alpha K} \quad (56)$$

The open-loop feedback gain A_F is defined as

$$A_F = \alpha K \quad (57)$$

One sees that there is a reduction of the beam-induced cavity voltage by the factor $1/A_F$ due to the feedback. This is equivalent to a similar reduction of the cavity shunt impedance as seen by the beam

$$Z_L \approx \frac{R_{SL}}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}}} \rightarrow \frac{R_{SL}}{1 + iQ_L 2\dfrac{\Delta\omega}{\omega_{CAV}} + A_F e^{-i\Delta\omega\delta}} \quad (58)$$

The price for this fast reduction of beam loading is the additional amount of generator current, $\vec{I}_B N$, which is needed to almost compensate for the beam current in the cavity. In terms of additional transmitter power $P^{'}$ this reads

$$P^{'} = R_S \hat{\vec{I}}_B^{\,2} / 8 \quad (59)$$

This is the power already calculated by Eq. (45). Since there is no change in cavity voltage due to P' this power will be reflected back to the generator, which has to have a sufficiently large plate dissipation power capability. Otherwise a circulator is needed. This critical situation of additional RF power consumption and reflection lasts, however, only until the tuner has reacted, and it may be minimized by pre-detuning. The generator-induced voltage is, of course, also reduced by the amount $1/A_F$, but this can be easily compensated for the low power level by increasing $\vec{U}_{in}$ by the factor $1/\alpha$ as Eq. (56) suggests. The practical implications of this will be illustrated by the following example.

Let the power gain of the amplifier be 80 dB. For a cavity power of 50 kW an input power P_{in} of 0.5 mW is thus required. This corresponds to a voltage gain of 10^4 so, for a design value of $A_F = 100$, α becomes 10^{-2}. Hence, the power, which is fed back to the amplifier input, is 5 W. To maintain the same cavity voltage as without feedback, P_{in} has to be increased from 0.5 mW to 5.0005 W. This value can, of course, be reduced by decreasing α, but then the amplifier gain has to be increased to keep A_F constant. This leads to power levels in the 100 μW range at the amplifier input. All of this is still practical, but some precautions, such as extremely good shielding and suppression of generator and cavity harmonics, have to be taken.

The maximum feedback gain, which can be obtained, is limited by the aforementioned delay time δ of a signal propagating around the loop. According to Nyquist's criterion the system will start to oscillate if the phase shift between $\vec{U}_{in}$ and $\alpha\vec{U}_{CAV}$ exceeds ≈135°. A cavity with high Q can produce a ±90° phase shift already for very small $\Delta\omega$. Therefore, once the additional phase shift given by Eq. (52) has reached $\pm\pi/4$, the loop gain must have become ≤ 1, i.e.

$$\left|A_F\left(\Delta\omega_{\max}\right)\right| \approx \frac{\alpha K}{1 + iQ_L 2\dfrac{\Delta\omega_{\max}}{\omega_{CAV}}} \leq 1 \tag{60}$$

where

$$\delta\Delta\omega_{\max} = \pm\frac{\pi}{4} \tag{61}$$

Here we assume that all other frequency-dependent phase shifts, such as those produced by the amplifiers, can be neglected. Inserting Eq. (61) into Eq. (60) we can solve for A_F:

$$A_F = \frac{Q_L}{4 f_{CAV}\delta} \tag{62}$$

This is the maximum possible feedback gain for a given δ

A fast-feedback loop of gain 100 has been realized at the HERA 208 MHz proton RF system. With a loaded cavity $Q_L \approx 27\ 000$ the maximum tolerable delay, including all amplifier stages and cables, is $\delta = 330$ ns. Therefore all RF amplifiers have been installed very close to the cavities in the HERA tunnel.

In addition, there are independent slow phase and amplitude regulation units for each cavity with still higher gain in the region of the synchrotron frequencies, i.e. below 300 Hz. Without fast feedback these units might become unstable at heavy beam loading [8, 9] since then changes in cavity voltage and phase are correlated as is shown by Eqs. (43) and (51).

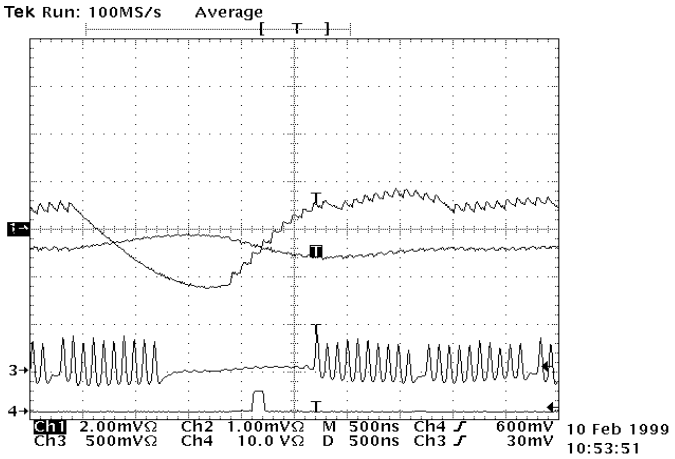


Fig. 9: Transient behaviour of the cavity voltage under the influence of fast feedback (reproduced from Ref. [10])

The effect of a fast-feedback loop is visible in Fig. 9, where the transient behaviour of the imaginary (upper curve) and real (medium curve) part of a HERA 208 MHz cavity voltage vector are displayed. The lower curve is the signal of a beam current monitor, which shows nicely the bunch structure of the beam and a 1.5 μs gap between batches of 6 × 10 bunches each. A detailed description of this measurement and of the IQ detector used is given in Ref. [10]. In this particular case the upper curve is essentially equivalent to the phase change of the cavity voltage due to transient beam loading and the middle curve corresponds to the change in amplitude.

The apparent time shift between the bunch signals and the cavity signals is due to the time of flight of the protons between the location of the cavity and the beam monitor in HERA. The transients resulting from the first two or three bunches after the gap cause step-like transients, which accumulate without significant correction. Later the fast feedback delivers a correction signal, which causes the subsequent transients to look more and more sawtooth like. From this one can estimate the time delay in the feedback loop to be of the order of 250 ns. After about 1 μs the equilibrium with beam is reached. Similarly, one observes in the left part of the picture that the feedback correction is still present during 250 ns after the last bunch before the gap has left the cavity. The equilibrium without beam is also reached after about 1 μs. Without fast feedback the time to reach the equilibrium is about 100 times larger, as one would expect for a feedback gain of 100.

To summarize this section we state that fast feedback reduces the resonant cavity impedance as seen by an external observer (usually the beam) by the factor $1/A_F$. It is important to realize that any noise originating from other sources than from the generator, especially amplitude and phase noise from the amplifiers, will be reduced by the factor $1/A_F$ because the cavity signal is directly compared with the generator signal at the amplifier input stage. Care has to be taken that no noise be created, by diode limiters or other non-linear elements, in the path where the cavity signal is fed back to the amplifier input. This noise would be added to the cavity signal by the feedback circuit. This becomes especially important for digital feedback systems, where the digital hardware (downconverters, analogue-to-digital converters, digital signal processors, etc.) is part of the feedback loop.

Amongst the great advantages of the digital technology are very easy amplitude and phase control of each channel (analogue elements are very expensive), easy application of calibration procedures and factors etc., but also cons like very high complexity.

5 Feedback and feedforward applied to superconducting cavities

So far, we have mainly considered normal conducting cavities in a proton storage ring, where the protons arrive in the cavities at the zero crossing of the RF signal, i.e. at $\varphi_s = 0°$ or a few degrees.

In the following we would like to present an example from the other extreme: superconducting cavities in a linear electron accelerator where the electrons cross the cavities near the moment of maximum RF voltage, i.e. at $\varphi_s \approx 90°$. (Note that for linear colliders usually a different definition of ϕ_s is used, namely $\varphi_s = 0°$ when the particle is on a crest. In this article we do not adopt this definition.)

In the beginning of the last decade of the last century a test facility for a TeV Energy Superconducting Linear Accelerator (TESLA) was erected at DESY. In the meantime a worldwide unique free-electron-laser user facility named FLASH, which is generating photon beams in the nanometer wavelength range for a rapidly growing user community, has emerged from this test facility. We refer to the special example of the superconducting nine-cell cavities of this accelerator, which are made of pure Nb. The operating frequency is 1.3 GHz.

The unloaded Q_0 value of these cavities is in the range $10^9 - 10^{10}$, or even higher. Hence, the bandwidth is only of the order of 1 Hz, and also the superconducting cavity shunt impedance exceeds that of normal conducting ones by many orders of magnitude. Since the particles are (almost) on a crest, only the real part of the cavity admittance as seen by the generator Eq. (27) is changed due to beam loading. This means that for beam loading compensation only a change in the coupling factor β is required and detuning plays no role for beam loading compensation in this situation. There is only perfect matching for the nominal beam current to which the cavity power input coupler has been adjusted. As we have already mentioned in Section 2.2, it takes the value $\beta = 6401$ in this case, which reflects also the fact that the ratio of the power taken away by the beam to the power dissipated in the cavity walls is much larger for superconducting cavities than for normal conducting ones. Owing to the coupling, the nominal loaded Q_L value is only 3×10^6, and the corresponding cavity bandwidth is 433 Hz. Since in this case there is a circulator with a load to protect the klystron from reflected power, the RF generator always sees a matched load.

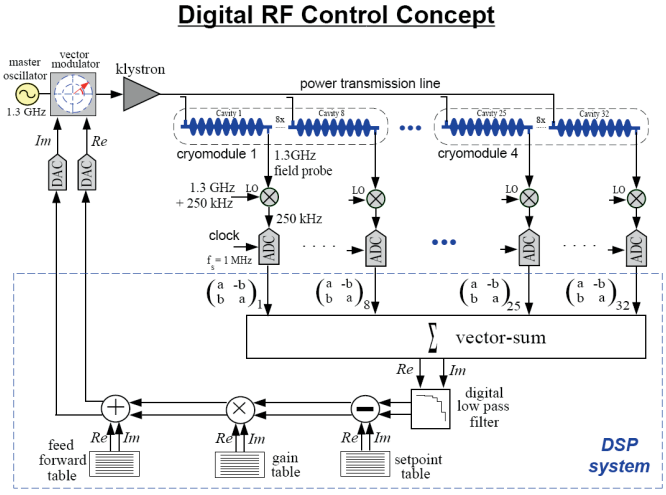


Fig. 10: Schematic of the low level RF system for control of the RF voltage of the 1.3 GHz cavities in the TESLA Test Facility (reproduced from Ref. [11])

From the circuit diagram in Fig. 10 we see that one RF generator supplies up to 32 cavities with RF power. The RF power per cavity needed to accelerate an electron beam of 8 mA to 25 MeV amounts to 200 kW, hence the minimum klystron power needed is 6.4 MW. This power is entirely carried away by the beam. In contrast to the previous example, where all of the RF power was essentially dissipated in the normal conducting cavity walls, the power needed to build up the RF cavity voltage in the superconducting cavities is only a few hundred Watts. Additional RF power is needed to account for regulation reserve, impedance mismatches etc. Therefore, high-efficiency 10 MW multibeam klystrons were developed for this project. For completeness we mention that this is pulsed power, with a pulse length of 1.5 ms and the maximum repetition rate 10 Hz. So the maximum average klystron power is 150 kW.

The RF signal seen by the beam corresponds to the vector sum of all cavity signals. Therefore, in a first step, this vector sum must be reconstructed by the low-level RF system. This is done by down conversion of the cavity field probe signals to 250 kHz intermediate frequency signals, which are sampled in time steps of 1 μs. Each set of two subsequent samples corresponds then to the real and imaginary part of the cavity voltage vector. From these signals the vector sum is generated in a computer and is compared with a table of set point values. The difference signal, which corresponds to the cavity voltage error, acts on a vector modulator at the low-level klystron input signal. In addition to this feedback a feedforward correction can be added. The advantage of feedforward is that, in principle, there is no gain limitation as in the case of feedback. If the error is known in advance, one can program a counteraction in the feedforward table. Examples for such errors could be a systematic decrease of beam current during the pulse due to some property of the electron source, or a systematic change of the cavity resonance frequency during the pulse. This effect exists indeed. The mechanical forces resulting from the strong pulsed RF field in the superconducting cavities cause a detuning of the order of a few hundred Hertz at 25 MV/m. This effect is called Lorentz force detuning.

From Eq. (62) one might infer that due to the large value of $Q_L = 3 \times 10^6$ the maximum possible feedback gain in this case could become significantly larger than for normal conducting cavities. One has to check, however, whether there are poles in the system at other frequencies, and, at least in this case, there is a fairly large loop delay of about 4 μs caused by the 12 m length of the cryogenic modules in which the cavities are placed and by the time delay in the computer. This results in a realistic maximum loop gain of 140.

The most impressive results for amplitude and phase stability recently obtained with the newly installed third harmonic RF system of the FLASH accelerator [12] are shown in Figs. 11–15. The digital RF control system used here has the same basic structure as is indicated in Fig. 10. In addition, there is a digital MIMO (multiple input multiple output) controller in the feedback path and also a learning feedforward system, which is described in detail in Ref. [13].

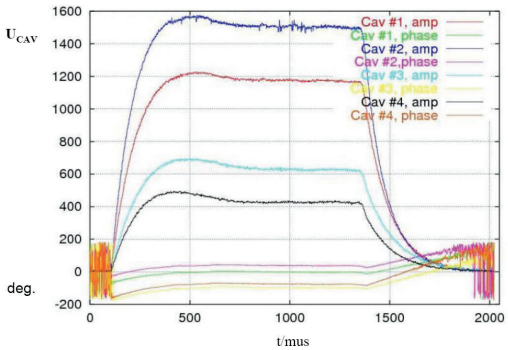


Fig. 11: Unregulated signals of RF phase (the lower four curves) and amplitude of four superconducting cavities operating at 3.9 GHz in the FLASH accelerator [14]

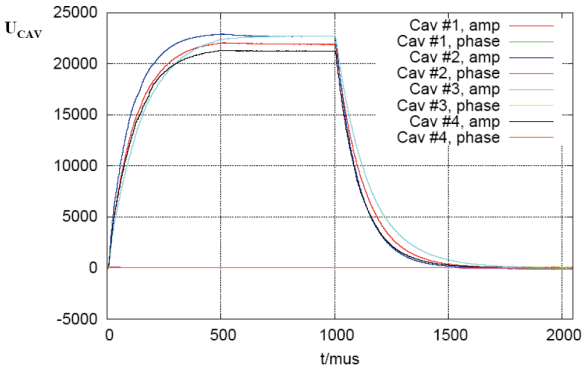


Fig. 12: Regulated signals of RF amplitude of four superconducting cavities operating at 3.9 GHz in the FLASH accelerator [14]

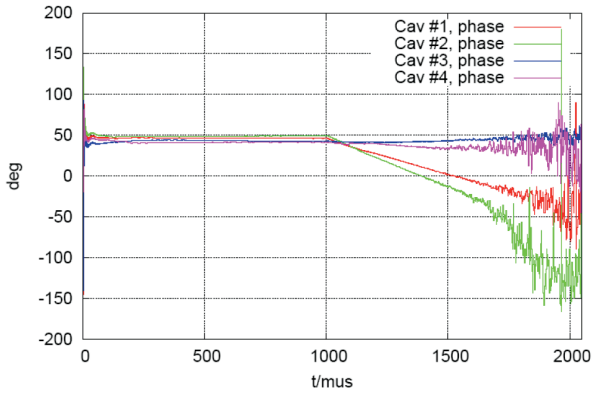


Fig. 13: Regulated signals of RF phase of four superconducting cavities operating at 3.9 GHz in the FLASH accelerator [14]

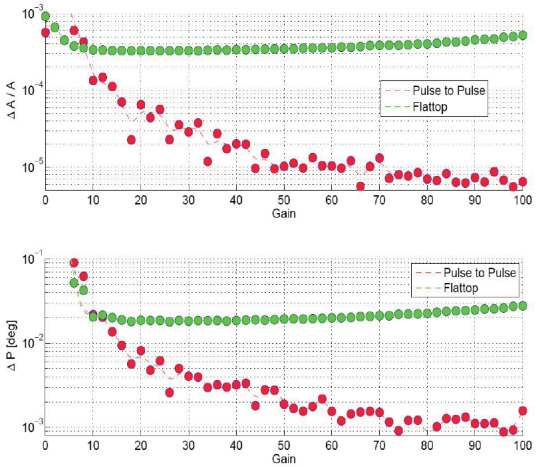
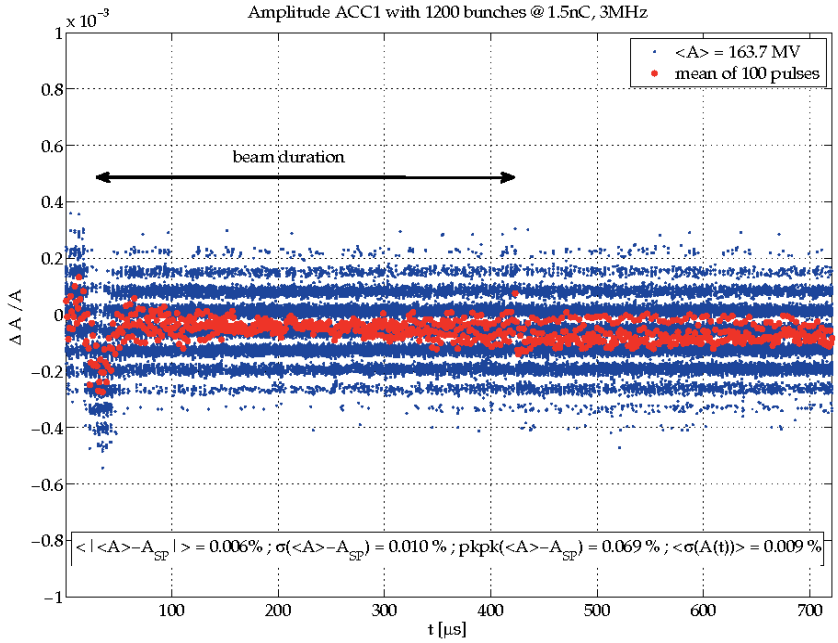
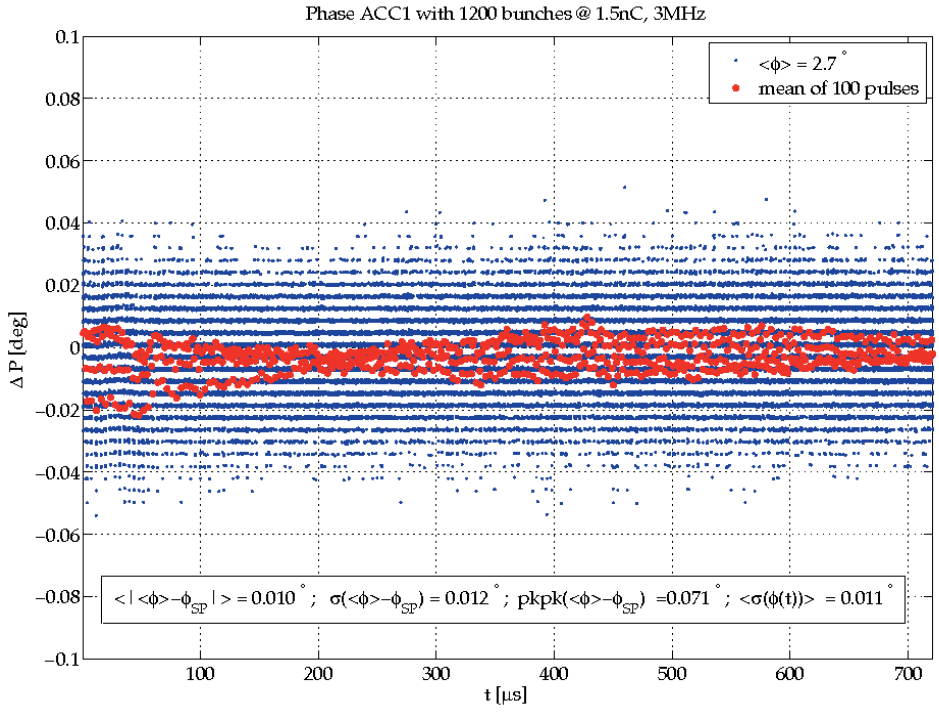


Fig. 14: Phase and amplitude root mean square stability versus feedback gain achieved by digital feedback with integrated MIMO controller and learning feedforward in 3.9 GHz cavities operating in the FLASH accelerator as a third harmonic system. The improvement in the pulse-to-pulse results is due to the effect of averaging over the measurement noise. (Reproduced from Ref. [15].)



(a)



(b)

Fig. 15: (a) Amplitude stability versus time achieved by digital feedback with integrated MIMO controller and learning feedforward in a cryogenic module containing eight nine-cell cavities operating at 1.3 GHz in the FLASH accelerator. (Reproduced from Ref. [15].) (b) Same as (a), but for phase stability.

6 Damping of synchrotron oscillations of protons in the PETRA II machine

In the preceding sections phase and amplitude control of the cavity voltage was discussed. In this last section we would like to give an example of beam control by means of a dedicated RF system for damping synchrotron oscillations of protons in the PETRA II synchrotron at DESY.

Prior to injection into HERA protons were pre-accelerated to 7.5 GeV/c and 40 GeV/c in the synchrotrons DESY III and PETRA II, respectively [16]. Timing imperfections during transfer of protons from one machine to the next one and RF noise during ramping were observed to cause synchrotron oscillations which, if not damped properly, may lead to an increase of beam emittance and to significant beam losses. Therefore, a phase loop acting on the RF phase to damp these oscillations of the proton bunches was a necessary component of the low-level RF system. The PETRA II proton RF system, which consisted of two 52 MHz cavities, each with a closely coupled RF amplifier chain and a fast-feedback loop of gain 50, was similar to that shown in Fig. 6. The block diagram of the PETRA II phase loop, on which we will concentrate now, is shown in Fig. 16.

6.1 Loop bandwidth

The maximum number of bunches was 11 in DESY III and 80 in PETRA II so that 8 DESY III cycles were needed to fill PETRA II. If synchrotron oscillations due to injection timing errors arise, all bunches of the corresponding batch are expected to oscillate coherently. Therefore, one single correction signal can damp the bunch oscillations in that batch and in total up to eight such signals were needed, one for each batch. This phase loop was a batch-to-batch rather than a bunch-to-bunch feedback. Ideally, the correction of expected errors of about 2° in the injection phase had to be switched within the 96 ns separating the last bunch of batch n from the first one of batch $n + 1$. Owing to the fast feedback of gain 50 the RF system had an effective bandwidth of about 1 MHz, it was, however, capable of performing small phase changes of the order of 1° per 100 ns, which was sufficient for damping synchrotron oscillations also in multibatch mode of operation.

6.2 The phase detector

Each bunch passage generates a signal in the inductive beam monitor also shown in Fig. 16. A passive LC filter of 8 MHz bandwidth filters out the 52 MHz component. The ringing time is comparable to the bunch spacing time as is shown in Fig. 17. Amplitude fluctuations of this signal are reduced to ±0.5 dB in a limiter of 40 dB dynamic range. So the amplitude dependence of the synchrotron phase measurement between the bunch signal and the 52 MHz RF source signal is minimized. The phase detector has a sensitivity of 10 mV per degree. Inserting a low-pass filter one can directly observe the synchrotron motion of the bunches at the phase detector output. This is shown in Fig. 18(a) for one batch of nine proton bunches circulating in PETRA II with the momentum of 7.5 GeV/c a few milliseconds after injection. The observed synchrotron period T_S = 5 ms agrees with the expected value for the actual RF voltage of 50 kV.

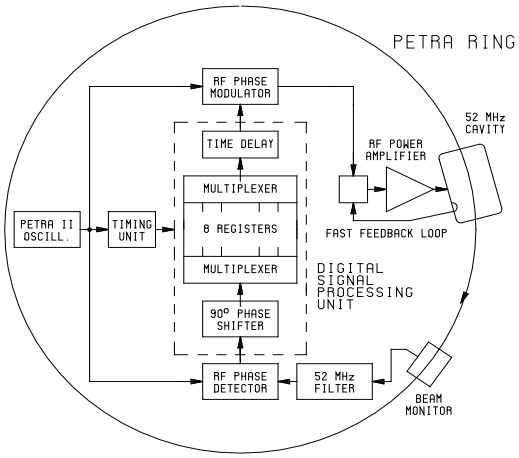


Fig. 16: Block diagram of the PETRA II phase loop. In the phase detector synchrotron oscillations of the bunches are detected by comparing the filtered 52 MHz component of the beam to the 52 MHz RF reference source. An average phase signal for each of the 8 batches of 10 bunches is phase shifted by 90° with respect to the synchrotron frequency, stored in its register and properly multiplexed to the phase modulator acting on the RF drive signal.

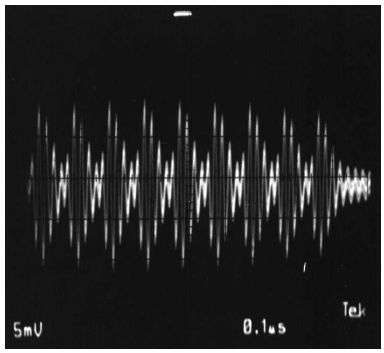
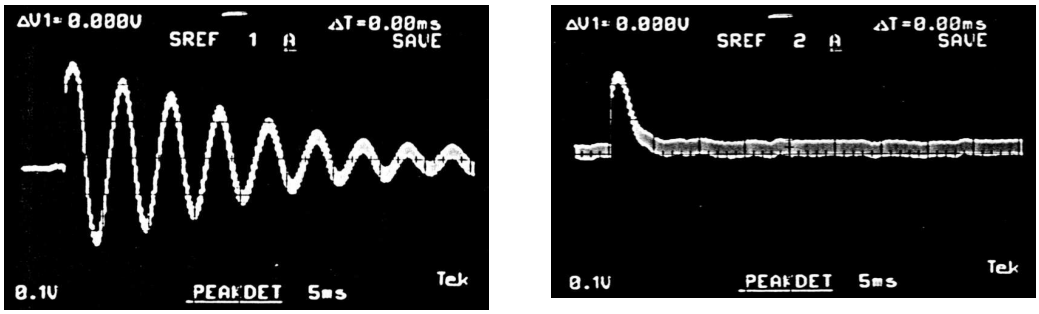


Fig. 17: Filtered signal of a batch of nine proton bunches circulating in PETRA. The bunch spacing time is 96 ns.



(a) (b)

Fig. 18: (a) The synchrotron oscillation measured at the phase detector output a few milliseconds after injection of a batch of nine proton bunches into PETRA II. It is smeared out by Landau damping after some periods. The damping loop is not active. (b) Same as (a) but with the phase loop active. The synchrotron oscillation is completely damped within half a synchrotron period of 0.5 ms.

6.3 The FIR filter as a digital phase shifter

A feedback-loop can damp the synchrotron motion if, as is indicated in Fig. 16, the synchrotron phase signal is shifted by –90° relative to the synchrotron frequency f_S , delayed properly and fed into a phase modulator acting on the 52 MHz drive signal. The necessity of the –90° phase shift relative to f_S can be seen from the equation of damped harmonic motion $\ddot{x} + a\dot{x} + bx = 0$ with the solution $x = A\sin(\omega_S t - \phi)e^{-at}$. The damping term $a\dot{x}$ is proportional to the time derivative of the solution x, i.e. a phase shift of –90°. The correction signal will coincide with the corresponding batch in the cavity if the total delay $\tau = t_f + nT_{rev}$,where t_f is the transit time from the beam monitor to the cavity, n an integer, and T_{rev} = 7.7 μs is the particle revolution time in PETRA. Since $T_S >> T_{rev}$, a delay of even more than one turn $(n > 1)$ would not be critical.

Rather than using a simple RC integrator of the differentiator network as a 90° phase shifter, which is not without problems [17], a more complex digital solution with a software controlled phase shift has been adopted. This is very attractive since during injection, acceleration and compression of the bunches the synchrotron frequency varies in the range from 200 Hz to 350 Hz. In addition, storing and multiplexing the eight correction signals for each of the eight possible batches in PETRA II can also be realized most comfortably on the digital side. The phase shifter has been built up as a three-coefficient digital FIR (finite-length impulse response) filter according to

$$g_\mu = \sum_{k=0}^{2} h_k f_{\mu-k} \tag{63}$$

with an amplitude response

$$H(\omega) = \sum_{k=0}^{2} h_k e^{-ik\omega T_S} \tag{64}$$

where f and g are input and output data, respectively. Using the coefficients $h_0 = \frac{2}{\pi}\sin\phi, h_1 = \cos\phi, h_2 = -\frac{2}{\pi}\sin\phi$ one obtains a phase shift which, in the frequency range of interest $200\ \text{Hz} \le f_S \le 359\ \text{Hz}$, deviates by less than ±0.4 from the nominal value $\phi = -\frac{\pi}{2}$ in accordance with Eqs. (63) and (64). The frequency dependence of the phase shift is mainly due to the delay in the filter which is of the order of 1 ms, i.e. two sampling periods. It can always be corrected by software, if necessary. The amplitude response is constant within a few per cent for all frequencies.

A block diagram of the filter is shown in Fig. 19. The synchrotron phase information of the eight batches is sampled at intervals T_S = 0.5 ms and passed through eight times three shift registers. The three coefficients are stored in ROM and are appropriately combined with the phase information. So, the first filter output is available after three sampling periods and is then renewed every 0.5 ms.

6.4 Performance of the phase loop

The performance of the loop is demonstrated in Fig. 19 where the phase detector output recorded by a storage scope is displayed. Complete damping of the synchrotron oscillation is achieved within less than one period. This corresponds to a damping time of less than 4 ms. If the loop is operated in the anti-damping mode, the beam is lost within some milliseconds. With the loop, losses of the proton beam in PETRA II during energy ramping could be reduced significantly.

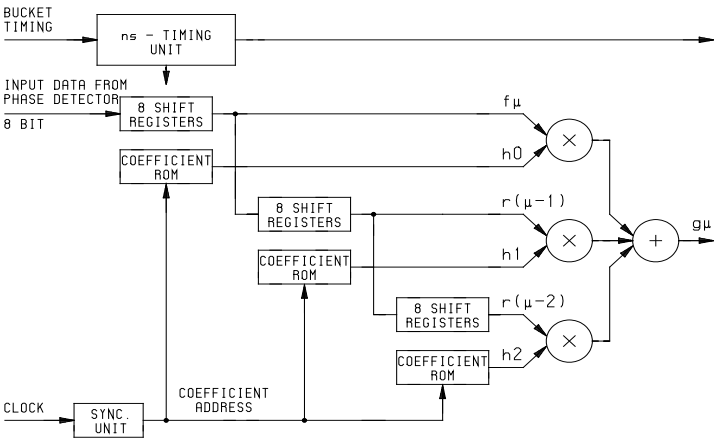


Fig. 19: Block diagram of the FIR filter. From three successive sampling periods the averaged phase signals for the eight proton batches in PETRA II are stored in shift registers and combined with the three coefficients, which are stored in ROM. The first phase-shifted output is available after three sampling periods of 0.5 ms and is renewed every sampling period.

References

[1] R.E. Collin, *Foundations for Microwave Engineering* (McGraw-Hill, New York, 1966).

[2] P.B. Wilson, CERN ISR-TH/78-23 (1978).

[3] D. Boussard, CERN SPS/86-10 (ARF) (1996).

[4] R.D. Kohaupt, Dynamik intensiver Teilchenstrahlen in Speicherringen, Lecture Notes, DESY (1987).

[5] A. Piwinski, DESY H 70/21 (1970).

[6] F. Pedersen, *IEEE Trans. Nucl. Sci.* **32** (1985) 2138.

[7] K.W. Robinson, CEA Report CEAL - 1010 (1964).

[8] D. Boussard, CERN SPS/85-31 (ARF) (1985).

[9] F. Pedersen, *IEEE Trans. Nucl. Sci.* **22** (1975) 1906.

[10] E. Vogel, Ingredients for an RF Feedforward at HERA p, DESY HERA 99-04, p 398 (1999).

[11] T. Schilcher, Thesis, DESY and Universität Hamburg, 1998.

[12] E. Vogel, *et al*., Proc. of IPAC'10, Kyoto, Japan, 2010, p. 4281.

[13] C. Schmidt, Thesis, DESY and TU Hamburg, Harburg, 2010.

[14] M. Hoffman, Private communication, DESY, 2010.

[15] C. Schmidt, Private communication, DESY, 2010.

[16] A. Gamp, W. Ebeling, W. Funk, J.R. Maidment, G.H. Rees and C.W. Planner, Proc. of the 1st European Particle Accelerator Conference, Rome, Italy, 1988.

[17] A. Gamp, Proc. of the 2nd European Particle Accelerator Conference, Nice, France, 1990.

RF transport

Stefan Choroba
DESY, Hamburg, Germany

Abstract
This paper deals with the techniques of transport of high-power radiofrequency (RF) power from a RF power source to the cavities of an accelerator. Since the theory of electromagnetic waves in waveguides and of waveguide components is very well explained in a number of excellent text books it will limit itself on special waveguide distributions and on a number of, although not complete list of, special problems which sometimes occur in RF power transportation systems.

1 Introduction

The task of a radiofrequency (RF) power transportation system in an accelerator is to transport the RF power generated by a RF power source to the cavity of an accelerator. Sometimes it is necessary to combine the power of several RF sources and very often it is necessary to transport the power to not just one cavity but a number of cavities. It is usually the aim to transport the power with high efficiency and high reliability.

Different types of transportation systems can be considered. Parallel wires or strip lines can be used to transport electromagnetic waves, but they cannot be used for high-power RF transport because of their low power capability due to radiation into the environment or electrical breakdown above a certain power level. Coaxial lines and hollow waveguides can be used to transport high-power RF. Coaxial lines are used at lower frequencies up to some 100 MHz and power of some 10 kW. Hollow waveguides typically are used for frequencies above some 100 MHz, power levels of more than some 10 kW up to several 10 MW and distances of several metres. Coaxial lines are used at these parameters only for short distances or when efficiency does not play the key role. This is due to losses in the inner conductor of the coaxial line, losses in the dielectric material or breakdown between the inner and outer conductor of the coaxial line at a high power level. There is of course no sharp line when to use coaxial lines or hollow waveguides. Although coaxial lines are in use at accelerators the next sections will concentrate on hollow waveguides since they are used in a larger number of applications.

2 Theory of electromagnetic waves in waveguides and of waveguide components

The theory of electromagnetic waves in waveguides and of waveguide components can be found in a number of excellent text books [1–4], school articles [5, 6] or school transparencies [7], of which some are listed in the references. Therefore, the theory will not be repeated here.

3 Waveguide distributions

Waveguides distributions are combinations of different waveguide components. They allow for power transport, combination and distribution of RF power. In addition they protect the RF power source from reflected power and allow for adjustment of RF parameters such as amplitude, phase or Q_{ext}.

Distributions can be complicated assemblies and sometimes different options exist to fulfil certain requirements.

The size of the waveguide is first determined by the RF frequency. It is then still possible to choose between two standard sizes. For the transport of RF power at, for example, 1.3 GHz one can consider WR650 or WR770 waveguides. The decision depends on considerations such as the maximum power to be transmitted, space availability for the installation, weight, cost or availability of waveguide components on the market. Whereas the maximum power demand might require larger size, space demands restrict us to smaller dimensions.

The type of distribution system depends on similar considerations. In addition, demands such as the required isolation between cavities (cross-talk), RF parameters to be controlled or adjusted and more requirements must be considered. It is therefore worthwhile to take into account all possible requirements as early as possible and to trade them off.

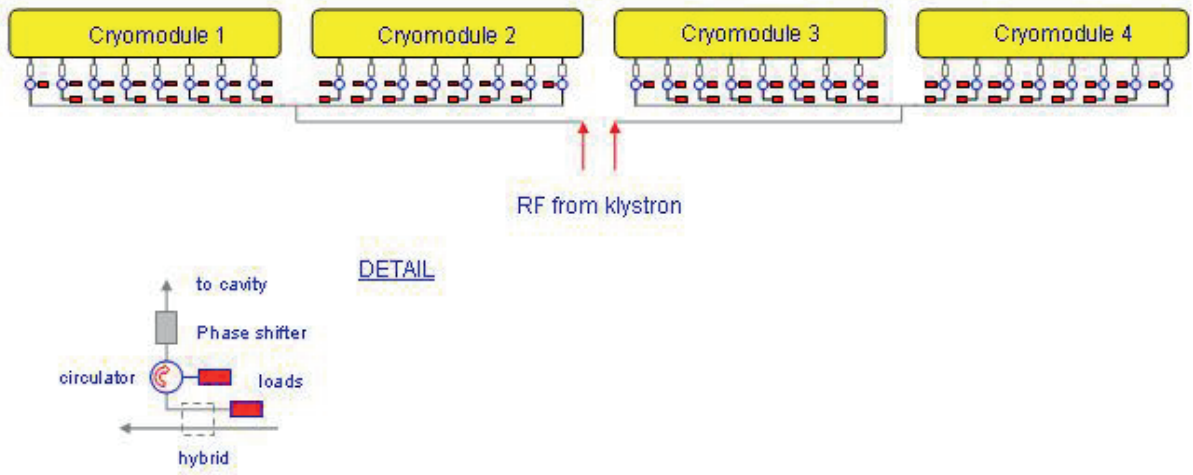


Fig. 1: Principle of a RF power transport system proposed for the TESLA linear collider

Figure 1 shows the principle of a waveguide distribution system which has been proposed for the TESLA linear collider. The RF power is generated by a 10 MW high-power klystron and extracted by two output windows towards four accelerator modules with eight superconducting cavities each. In order to achieve a gradient of 23.4 MV/m an input power of 231 kW per cavity is required. The total power produced by the RF power source must take into account losses in the waveguides and a regulation reserve. The power of each klystron waveguide arm is split again by a 3 dB hybrid. For each module a linear distribution system similar to that shown in Fig. 2 is used. Equal amounts of power are branched off for the individual cavities by hybrids with different coupling ratio. Isolators (three port circulators with loads) capable of 400 kW protect the power source from reflected power travelling back from the cavities during the filling time of the cavities or in the case of mismatch or breakdown in the cavities. Three stub tuners or phase shifters can be used for adjustment of phase and Q_{ext}.

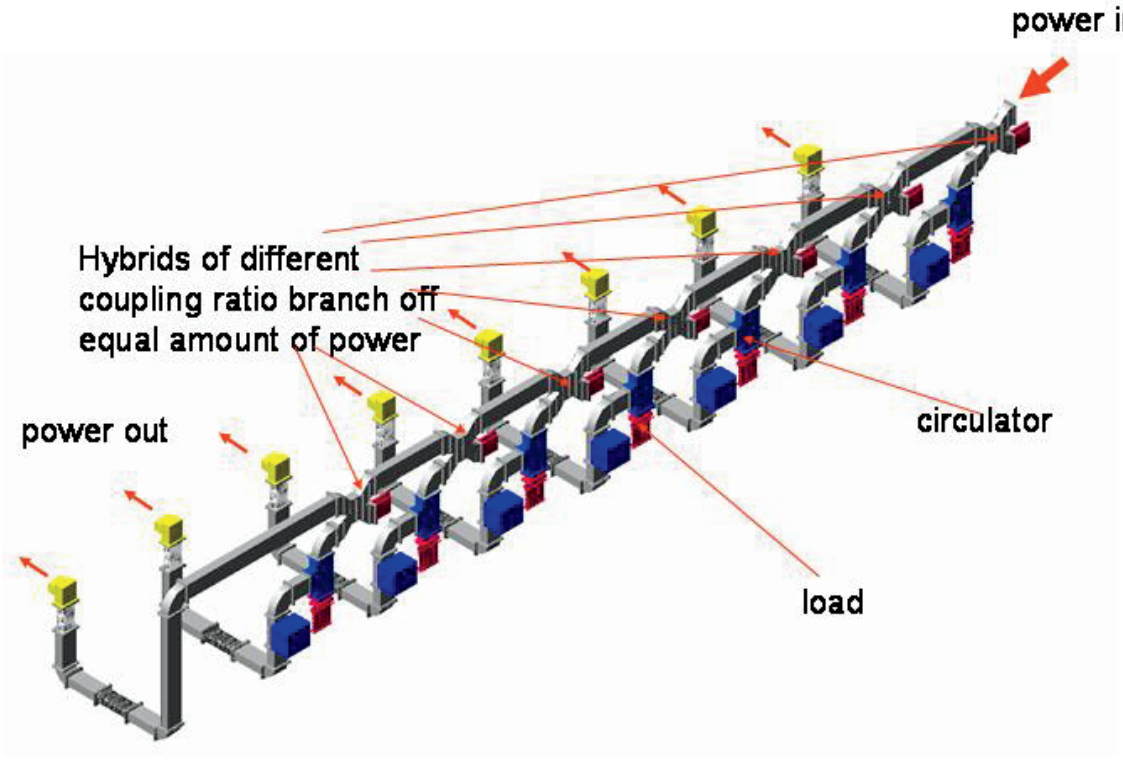


Fig. 2: Linear waveguide distribution system

Instead of a linear distribution system a tree-like system can be used (Fig. 3). The power is divided by shunt tees into several branches like the branches of a tree.

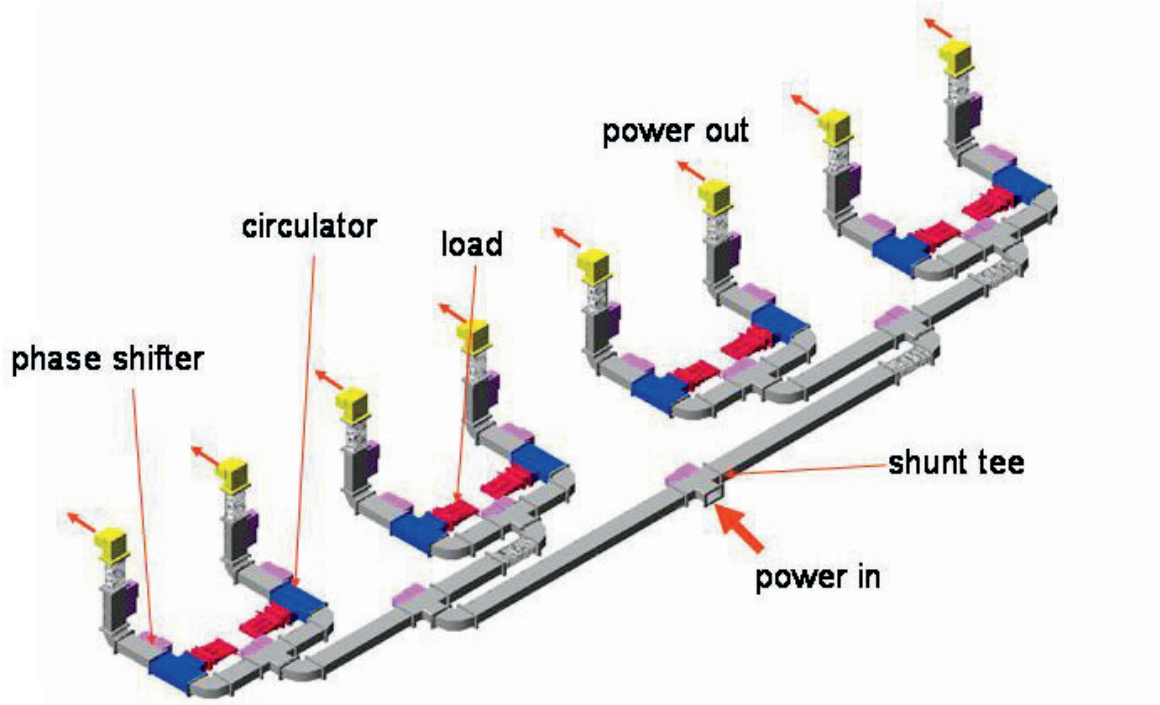


Fig. 3: Tree-like waveguide distribution system

In Fig. 4 examples of distributions meeting the requirements of power distributions for the FLASH facility at DESY are shown. The input power for the distribution can be up to 2.5 MW. Both distributions can be used to meet the requirements but because of certain advantages the first will be used in the future for the European XFEL. The latter and older system is a linear system and is used for the first accelerator modules at FLASH at DESY. The combined system proposed for the European XFEL and in use for the new RF distributions at FLASH makes use of asymmetric shunt tees of different coupling ratios. Equal amounts of power are branched off to a pair of cavities. By adjustment of tuning posts inside the tees the coupling ratios can be adjusted thus changing the branching ratio. The phase between a pair of cavities is pre-tuned by fixed phase shifters (straight waveguides with different waveguide width). The phase for the individual cavities can be adjusted by movable mechanical phase shifters after the symmetric shunt tee which splits the power for two cavities. Isolators in front of each cavity protect the power source from reflected power. This system has several advantages. Space and weight are reduced compared with the linear system. Phasing is much easier than in the purely linear system. Owing to the use of components of similar type the cost can be decreased. This system can also be pre-assembled and connected to an accelerator module before the module is installed in the accelerator tunnel, thus simplifying the entire installation procedure. An accelerator module with RF waveguide distribution can be seen in Fig. 5.

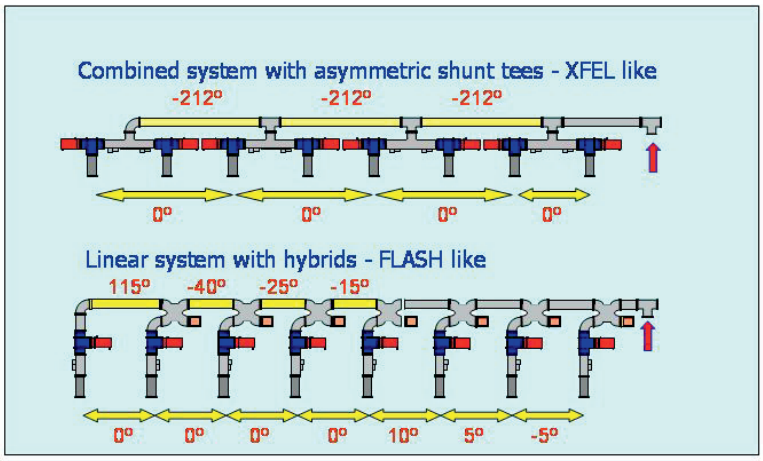


Fig. 4: Combined and linear RF waveguide distribution system

Fig. 5: Accelerator module and RF waveguide distribution

More information on the waveguide distribution systems for TESLA, FLASH and the European XFEL can be found in Refs. [8–12].

4 Limitations, problems and countermeasures

In this section some limitations, problems and possible countermeasures in RF power transportation systems are covered.

The power P_{RF}, which can be transmitted in a rectangular waveguide of size a times b in TE_{10} mode of wavelength λ is given by

$$P_{RF} = 6.63 \times 10^{-4} a[\mathrm{cm}] b[\mathrm{cm}] \left[\lambda \Big/ \lambda_g \right] E[\mathrm{V/cm}]^2$$

with

$$\lambda_g = \frac{\lambda}{\sqrt{1 - \left(\frac{\lambda}{2a} \right)^2}}$$

where E (in V/cm) is the electrical field strength of the electromagnetic wave and λ_g the guide wavelength. The maximum power is the power at the electrical breakdown limit E_{max}. In air E_{max} is 30 kV/cm, which results in a RF power of 58 MW at 1.3 GHz in a WR650 waveguide. Experience shows that this power cannot be achieved. In practice, it is 5–10 times lower. The practical power limit is lower because of a variety of different reasons, for example the smaller size of the inner waveguide dimensions (e.g. within circulators), surface effects (roughness, steps at flanges, etc., see Fig. 6), dust in waveguides, humidity of the gas inside the waveguide, reflections (VSWR) or because of higher order modes (HOMs) in TE_{nm}/TM_{nm}.

These HOMs can be generated by the power source or by non-linear effects at high power in non-reciprocal devices such as circulators. If these modes are not damped, they can be excited resonantly and reach very high field strength above the breakdown limit. In order to damp HOMs, HOM dampers can be installed. These can be complicated and specially designed devices, which couple out and damp the HOMs only, leaving the fundamental mode, which is transmitted in TE_{10}, untouched. But sometimes a quick solution must be found. This can be accomplished by inserting small antennas at the small side of the waveguides (see Fig. 7). These antennas couple to a number of HOMs in TE_{nm}/TM_{nm}. Since the field of TE_{10} is already small at the antenna position only a small amount of the power in the fundamental mode is coupled out. The antennas must be connected to loads, which must be able to handle the full amount of the power coupled out.

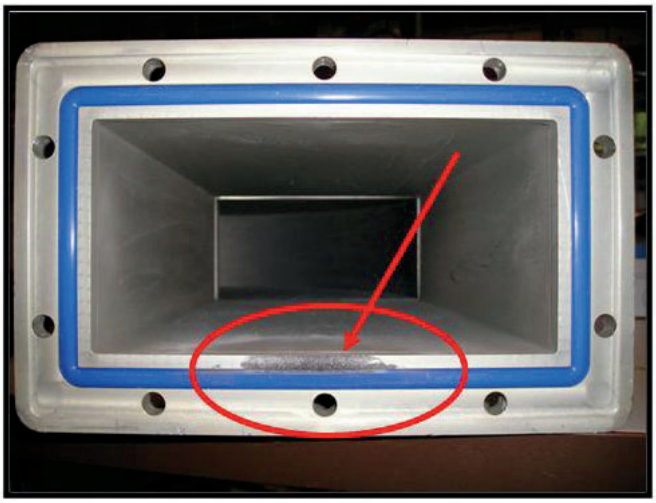


Fig. 6: Waveguide which has been damaged at the flange by breakdown

Fig. 7: Waveguide with damping antennas at the small side of the waveguide

One could increase the gas pressure inside the waveguide, which due to Paschen's law would increase the power capability, but this requires enforced and gas tight waveguides. In addition the pressure vessel rules applicable in many countries must be followed. By using SF_6 instead of air, which has E_{max} = 89 kV/cm (at 1 bar, 20°C), the power capability can also be increased. The problem with SF_6 is that although it is chemically very stable it is a green house gas and if cracked in sparks it can, together with the hydrogen of the rest amount of water, form HF, which is a very aggressive acid. HF can produce fluorides of the metal of the waveguide walls which one can sometimes find as white powder in the waveguides (see Fig. 8). Other poisonous chemicals, e.g. S_2F_{10}, are also produced. Personnel maintaining waveguide components which have been operated in SF_6 have to observe a number of safety rules and wear personal protective equipment (see Fig. 9).

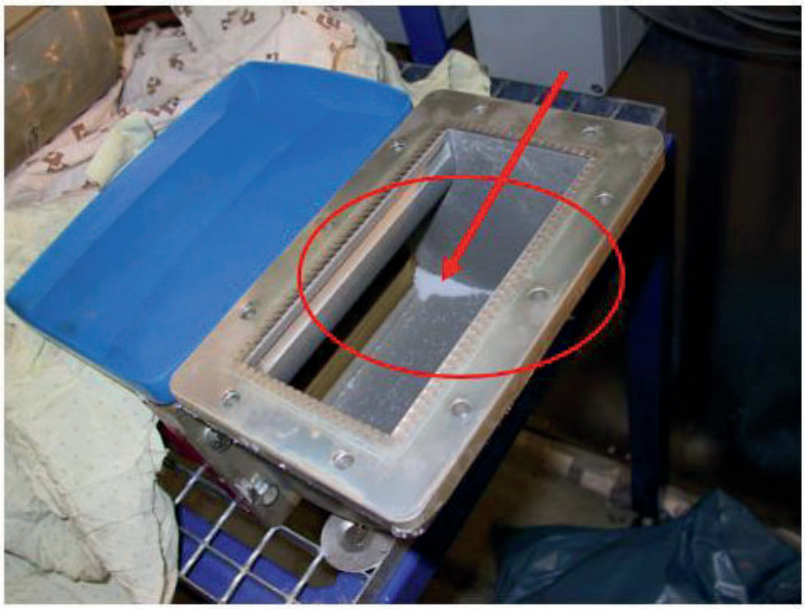


Fig. 8: Fluorides in a waveguide

Fig. 9: Staff wearing protective clothing during work with SF_6 waveguides

Acknowledgements

The author would like to thank two persons who among others supported him during the preparation of this lecture: Valery Katalev provided simulation results and pictures of the waveguide distributions; Ingo Sandvoss took many photographs.

References

[1] R.E. Collin, *Foundations for Microwave Engineering* (McGraw-Hill, New York, 1992).

[2] D.M. Pozar, *Microwave Engineering* (Wiley, New York, 2004).

[3] N. Marcuvitz, *Waveguide Handbook* (MIT Radiation Laboratory Series Vol. 10, McGraw-Hill, New York, 1951).

[4] H.J. Reich, P.F. Ordung, H.L. Krauss and J.G. Skalnik, *Microwave Theory and Techniques* (D. van Nostrand, Princeton, NJ, 1953).

[5] R.K. Cooper and R.G. Carter, High power RF transmission, Proc. CERN Accelerator School: Radio Frequency Engineering, Seeheim, Germany, 8–16 May 2000.

[6] R.K. Cooper, High power RF transmission, Proc. CERN Accelerator School: RF Engineering for Particle Accelerators, Oxford, UK, 3–10 April 1991.

[7] A. Nassiri, Microwave Physics and Techniques, USPAS, Santa Barbara, Summer 2003.

[8] V. Katalev and S. Choroba, RF power distributing waveguide system for TESLA, Proc. Russian Particle Accelerator Conf., Rupac 2002, Obninsk, Russia, 1–4 October 2002, p. 79.

[9] V. Katalev and S. Choroba, Tuning of external Q and phase for the cavities of a superconducting linear accelerator, Proc. XXII International Linear Accelerator Conf., Linac 2004, Lübeck, Germany, 16–20 August 2004, p. 724.

[10] V. Katalev and S. Choroba, Compact waveguide distribution with asymmetric shunt tees for the European XFEL, Proc. 22nd Particle Accelerator Conf., PAC07, Albuquerque, NM, 25–29 June 2007, p. 176.

[11] S. Choroba *et al*., Operation experience with the FLASH RF waveguide distribution system at DESY, Proc. XXIV Linear Accelerator Conf., LINAC08, Victoria, BC, Canada, 29 September–3 October 2008, p. 978

[12] V. Katalev and S. Choroba, Waveguide distribution for FLASH, Proc. Sixth CW and High Average Power RF Workshop, CWRF2010, Barcelona, Spain, 4–7 May 2010, http://cwrf2010.cells.es/

Superconducting versus normal conducting cavities

Holger Podlech
Institute for Applied Physics (IAP), Frankfurt am Main, Germany

Abstract
One of the most important issues of high-power hadron linacs is the choice of technology with respect to superconducting or room-temperature operation. The favour for a specific technology depends on several parameters such as the beam energy, beam current, beam power and duty factor. This contribution gives an overview of the comparison between superconducting and normal conducting cavities. This includes basic radio-frequency (RF) parameters, design criteria, limitations, required RF and plug power as well as case studies.

1 Introduction

Worldwide there is an increasing interest in a new generation of high-power proton and ion linacs. The term 'high power' refers to the product of beam energy and beam current which is the beam power. Typical applications are neutron production with low-energy deuterons, neutron production using high-energy protons, nuclear physics at rare isotope beam facilities or nuclear waste transmutation. The typical modern hadron linac consists of three major parts (Fig. 1):

1. front end (low-energy part);
2. drift tube linac (intermediate-energy part);
3. elliptical section (high-energy part).

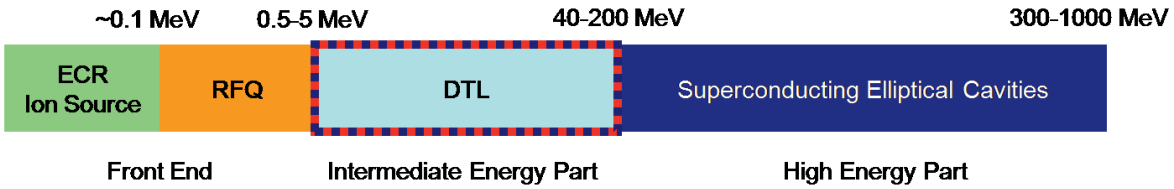


Fig. 1: Scheme of a modern high-power hadron linac

In most cases the front-end consists of an electron cyclotron resonance (ECR) ion source, the low-energy beam transport (LEBT) and a radio-frequency quadrupole (RFQ) as first accelerating RF structure. The vast majority of all RFQ structures have been realized using normal conducting technology. The high-energy section starts typically between 100 AMeV and 200 AMeV and mostly makes use of superconducting elliptical multicell cavities. Even for lower duty factors superconducting elliptical cavities are a good choice with respect to the overall required radio-frequency (RF) and plug power. In the case of the intermediate-energy part, the situation is not as clear. The choice of technology depends on the beam current, acceleration gradients and most importantly on the duty factor. Figure 1 shows schematically a modern high-power hadron linac.

The Spallation Neutron Source SNS (ORNL, USA) [1] which has been taken into operation in recent years is considered as the prototype for a new generation of high-power linacs. The driver linac provides 1 GeV protons. The duty factor is 6 % with a peak current of 38 mA resulting in an average beam current of 1.4 mA and a beam power of 1.4 MW. The front end accelerates the beam to 2.5 MeV. The intermediate-energy section consists of a classical Alvarez DTL (2.5 MeV to 87 MeV,

402.5 MHz), followed by a normal conducting CCL linac (87 MeV to 180 MeV, 805 MHz). The high-energy section consists of two groups of superconducting elliptical five-cell cavities operated at 805 MHz. The transition energy between normal conducting and superconducting cavities is 180 MeV. There are other projects such as MYRRHA [2] which will deliver a continuous wave (cw) proton beam. For this high duty factor, a much lower transition energy of 3.5 MeV has been chosen. Table 1 summarizes basic parameters of different hadron linacs which are in operation, under construction or in the design phase [1–10].

Table 1: Parameters of different modern hadron linacs

Project	Particles	Final energy (AMeV)	Transition energy (AMeV)	Duty factor (%)	Pulse beam current (mA)
SNS	p	1000	187	6	38
ESS	p	2500	50	4	50
MYRRHA	p	600	3.5	100	4
IFMIF	d	20	2.5	100	125
SPIRAL2	d, HI	20	0.75	100	5
FRIB	p-HI	200–600	0.3	100	<1
SARAF	p, d	20	1.5	100	2
LINAC4/SPL	p	5000	160	<10	40
FAIR p-Linac	p	70	70	0.04	70
GSI SHE	HI	7.5	1.4	100	1

p, protons; d, deuterons; HI, heavy ions.

Figure 2 (left) shows the transition energy as a function of the duty factor. In general, it can be observed that the transition energy is lower at higher duty factors because superconducting operation more likely becomes advantageous. There is also a clear dependence between the transition energy and the beam current (Fig. 2, right). The higher the pulse beam current the higher the transition energy typically is.

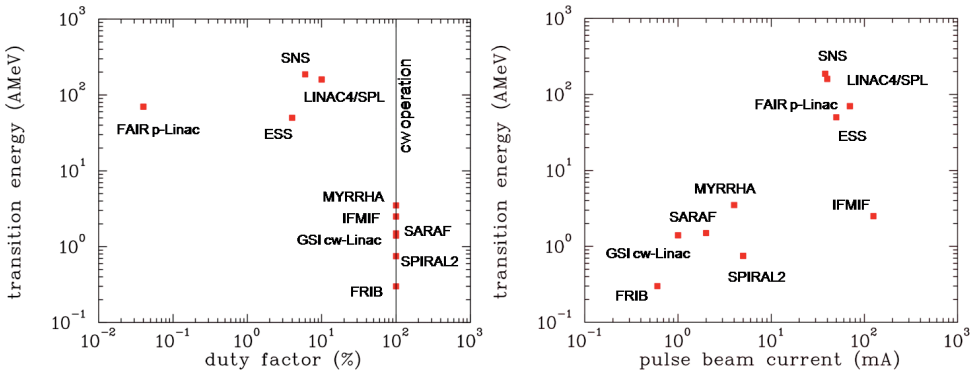


Fig. 2: Transition energy between normal conducting and superconducting technology for different duty factors (left) and for different pulse beam currents (right).

2 RF parameters

This section gives a brief description of different RF parameters. In general, these parameters can be divided into two groups:

1. parameters dependent on the surface resistance;
2. parameters independent of the surface resistance.

The surface-independent RF parameters are mostly used to compare different cavity geometries independent of the preferred technology.

2.1 Surface resistance R_s

The presence of RF fields results in a surface resistance R_s in both cases. The physical reason for this resistance and its magnitude is different, however, for normal conducting and superconducting cavities.

In the case of normal conducting cavities the skin effect leads to an expulsion of the electromagnetic fields and the corresponding current density. The decrease of the current density can be described by the following expression:

$$j(x) = j_0 e^{-x/\delta} \cos(x/\delta) \tag{1}$$

δ is the so-called "skin depth" which is the equivalent thickness of a homogeneous current density j_0. The skin depth is inversely proportional to the conductivity σ and inversely proportional to the square root of the frequency. Finally, the surface resistance can be calculated:

$$R_s = \frac{1}{\sigma\delta} = \sqrt{\frac{\pi\mu_0\mu_r f}{\sigma}} \tag{2}$$

Typical values for the surface resistance of normal conducting cavities are several milliohms.

In the case of superconducting cavities, the Meissner–Ochsenfeld effect leads to an almost complete expulsion of the electromagnetic fields from the superconductor. The current decays exponentially from the surface and reaches $1/e$ at the London penetration depth which is 39 nm for niobium. In the case of static fields, the current transport in superconductors is loss free but not for RF fields. Owing to the inertia of Cooper pairs unpaired electrons are then not completely shielded from the time-varying fields. These electrons close to the Fermi edge can be accelerated by the RF fields within the penetration depth leading to a non-vanishing surface resistance. The surface resistance can be expressed by the Bardeen–Cooper–Schrieffer (BCS) theory, for example for niobium:

$$R_s(BCS) = 2 \cdot 10^4 \frac{1}{T}\left(\frac{f}{1.5}\right)^2 e^{-\frac{17.67}{T}} \tag{3}$$

The surface resistance is highly dependent on the temperature. It decreases exponentially with lower temperatures because the number of unpaired electrons is also decreasing. On the other hand, the resistance increases quadratically with the frequency. As a consequence 2 K is chosen as the operation temperature at higher frequencies instead of 4 K. For the standard material niobium typical values for the BCS surface resistance are of the order of several nanoohms to several tens of nanoohms. The surface resistance of superconducting cavities is typically five orders of magnitude lower than for normal conducting cavities. The surface resistance of superconducting cavities is, however, higher in reality because of a frequency-independent residual resistance and because of eventually trapped magnetic flux.

2.2 Dissipated power P_c

The surface resistance results in a dissipation of energy and leads consequently to a power density

$$\rho = \frac{1}{2} R_s |H|^2 \tag{4}$$

The total power P_c can be obtained by integrating over the whole cavity surface:

$$P_c = \frac{1}{2} R_s \int_A |H|^2 \, dA \tag{5}$$

The power losses are several orders of magnitude lower for superconducting cavities because of the surface resistance dependency.

2.3 Q value

The (unloaded or intrinsic) Q value or quality factor of a cavity can be calculated by

$$Q_0 = \frac{\omega W}{P_c} \tag{6}$$

The stored energy W can be calculated either using the electric or the magnetic field. On average the energy is equally distributed in both fields. Of course the stored energy is independent of the surface resistance. It depends only on the cavity geometry and the field level:

$$W = \frac{1}{2} \mu_0 \int_V |H|^2 \, dV = \frac{1}{2} \varepsilon_0 \int_V |E|^2 \, dV \tag{7}$$

Aside from a factor 2π the Q value gives the number of RF periods until the stored energy is dissipated after the RF has been turned off. In addition, the Q value measures the full width of the resonance curve where the field amplitude reaches $1/\sqrt{2}$ of the maximum value at resonance:

$$Q_0 = \frac{f}{\Delta f} \tag{8}$$

The unloaded Q value depends inversely proportionally on the surface resistance. Typical Q values for normal conducting cavities are between 10^3 and 10^5 and between 10^7 and 10^{11} for superconducting cavities.

2.4 Shunt impedance R_a

A cavity can be described as an equivalent RCL parallel circuit. The shunt impedance R_a is the real part of the frequency-dependent complex impedance $Z(\omega)$. At resonance, the $Z(\omega)$ and R_a become identical. The shunt impedance describes the ability of the cavity to convert RF power into voltage:

$$R_a = \frac{U_a^2}{P_c} \tag{9}$$

Here U_a is the effective or accelerating voltage including the transit time factor. Sometimes the shunt impedance is normalized to the length to compare cavities with different length:

$$Z_{eff} = \frac{R_a}{L} = \frac{U_a^2}{P_c L} \tag{10}$$

2.5 R_s-independent parameters

To compare different geometries it is helpful to use parameters which are independent of the surface resistance. The so-called geometrical factor G is defined by

$$G = R_s Q_0 = \frac{R_s \omega W}{P_c} = \frac{\omega \mu_0 \int_V |H|^2 \, dV}{\int_A |H|^2 \, dA} \tag{11}$$

For a given frequency and field distribution it is the ratio between the cavity volume and cavity surface.

The shunt impedance is one of the most important parameters because it describes the cavity efficiency. This only works if the surface resistance is known. We can use the R/Q value which is independent of the surface resistance to compare different RF structures with a given surface resistance:

$$\frac{R_a}{Q_0} = \frac{U_a^2}{\omega W} = \frac{2[\int E_z \cos(\omega z/\beta c) \, dz]^2}{\varepsilon_0 \omega \int_V |E|^2 \, dV} \tag{12}$$

The R/Q value is also known as geometrical shunt impedance because it measures the ability of the cavity to focus the electric field on axis.

2.6 Example: the pillbox cavity

The pillbox cavity is a simple cylindrical cavity. The fundamental mode is the TM_{010} mode which is also used in elliptical cavities. All RF parameters of the pillbox cavity can be calculated analytically. Table 2 shows the basic RF parameters and a comparison of a normal conducting and a superconducting pillbox cavity.

Table 2: Comparison of a normal conducting and a superconducting pillbox cavity

Parameter	Normal conducting	Superconducting
Length (cm)	10	10
Radius (cm)	7.65	7.65
Frequency (MHz)	1500	1500
U_a (MV)	1	1
T (K)	300	2
R_s (Ω)	0.01	2×10^{-8}
Q_0	25 500	1.3×10^{10}
R_a (Ω)	5×10^6	2.5×10^{12}
W (J)	0.54	0.54
P_c (W)	198 000	0.4
G (Ω)	257	257
R/Q (Ω)	196	196

3 General consideration of power consumption

In this section the RF power and the grid power requirements of normal conducting and superconducting cavities are investigated. For a given RF structure the RF power is dominated by the following parameters:

1. accelerating gradient E_a;
2. beam current;
3. shunt impedance.

The required grid power is typically significantly higher than the RF power. It can be a crucial factor for operational costs especially for high duty factor accelerators. The required plug power to operate RF cavities is dominated by:

1. accelerating gradient E_a;
2. beam current;
3. shunt impedance;
4. duty factor;
5. RF amplifier efficiency;
6. efficiency of the cryogenic system (in a superconducting system);

The following considerations are limited to the cavity operation. The additional required power for magnets, heating, ventilation, etc., is not included. The given values for cavity parameters are only examples. In reality they can be higher or lower depending on the specific case and kind of RF structure.

Figure 3 shows a comparison of the required RF power (cavity loss P_c) without beam for a normal conducting and a superconducting cavity. The assumed surface resistance of the superconducting cavity is the BCS value at $T = 4$ K (12.6 nΩ).

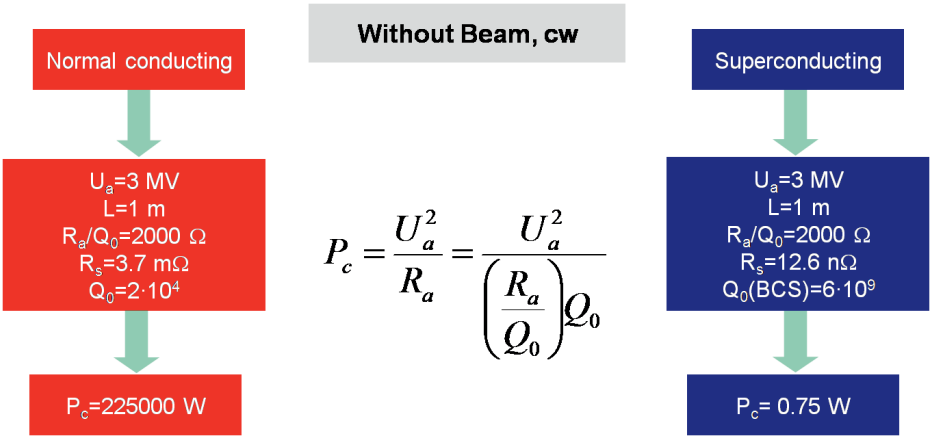


Fig. 3: Required RF power without beam for a normal conducting (left) and for a superconducting cavity (right)

In this example the difference in RF power is only due to the difference in the surface resistance. For superconducting cavities this is too optimistic because the real resistance can be significant higher. In addition to the BCS-value possible trapped magnetic flux and a frequency independent residual resistance will lead to higher power losses. The normal conducting cavity would

require 225 kW/m. In case of 100 % duty factor this corresponds to the thermal load. The present technological limit is about 100 kW/m. To reach this value the gradient has to be reduced in reality from 3 MV/m to 2 MV/m. Figure 4 shows a comparison of the grid power. For the normal conducting cavity we have to take the efficiency of the RF amplifier into account which is typically 60 %. In addition to the RF losses we have an additional heat entry into the helium bath due to static losses in the case of superconductivity. The sum of dynamic and static losses have to be removed using a cryogenic system. Owing to the very low thermodynamic efficiency η of a cryogenic system operated at 4 K given by

$$\eta = \left(\frac{4K}{300K - 4K}\right) \cdot 0.25 = 0.003$$

the grid power is much larger than the heat load in the helium bath. Finally the ratio between the grid power for a normal conducting and a superconducting cavity has been decreased from more than 4 orders of magnitude to less than a factor 100 (cw, no beam).

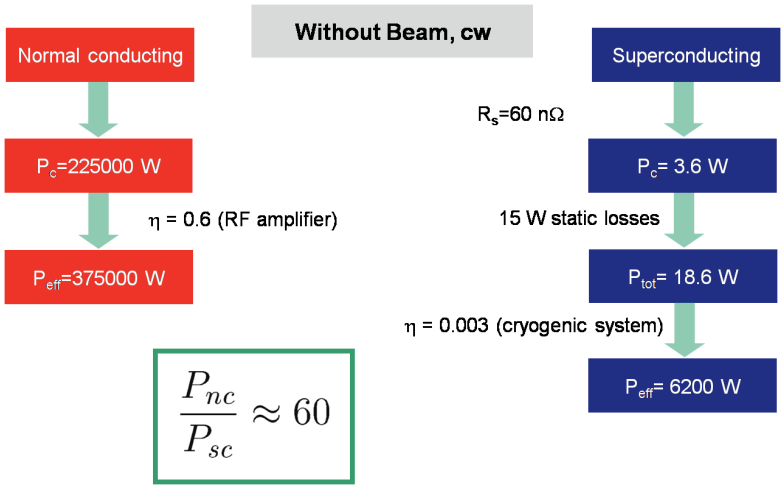


Fig. 4: Required grid power (cw operated) without beam for a normal conducting (left) and for a superconducting cavity (right)

The situation changes again in the presence of beam loading. Assuming a beam current of 20 mA we have 60 kW beam power per cavity (U_a = 3 MV). In this example the normal conducting cavity would require 475 kW and the superconducting cavity 106 kW grid power (see Fig. 5). The required grid power of superconducting cavities with heavy beam load and high duty factor is dominated by the beam power.

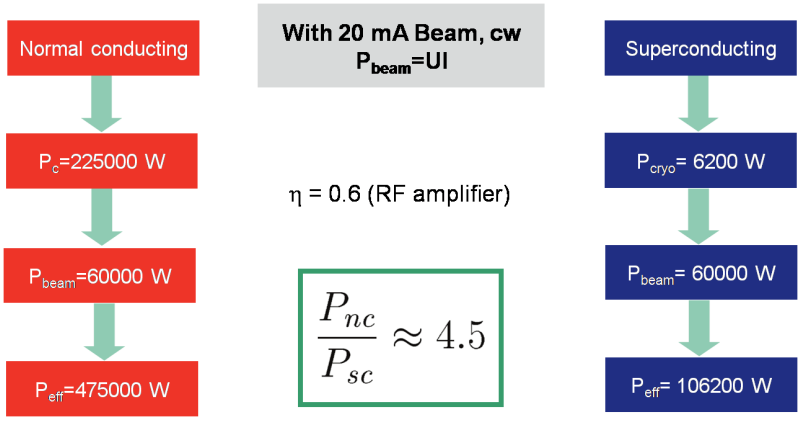


Fig. 5: Required grid power with beam (cw operated) for a normal conducting (left) and for a superconducting cavity (right)

In the case of higher beam currents and low duty factor the situation becomes more favourable for normal conducting cavities. Figure 6 shows an example with 20 mA current and a duty factor of 1 %. Owing to the static losses which are always present, the grid power is lower for the operation of normal conducting cavities.

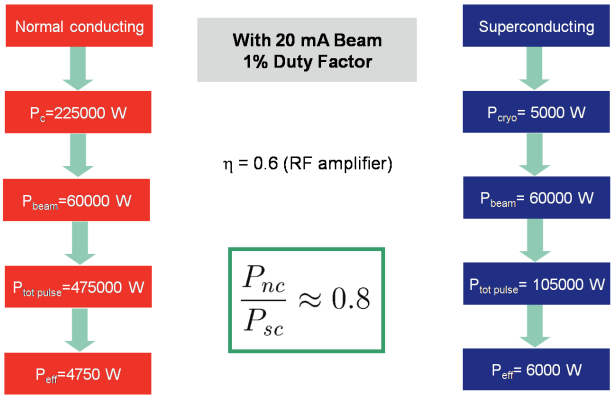


Fig. 6: Required grid power with beam and low duty factor (1 %) for a normal conducting (left) and for a superconducting cavity (right)

4 Systematic investigation of the required power

The topic of this section is a more systematic description of the required RF and grid power. The power has been calculated dependent on the gradient, duty factor, beam current, surface resistance and static losses. In each plot always one normal conducting and one superconducting cavity with typical parameters have been compared. The lengths of the cavities have been fixed to 1 m. A frequency of 325 MHz has been used. The assumed shunt impedance of the normal conducting cavity is 60 MΩ/m and the R/Q value of the superconducting cavity is 3000 Ω. Of course, shunt impedance and surface resistance change with frequency, but the general results are very similar for different cavity shapes, frequencies and the β of the cavities.

In addition, the following assumptions have been made for reasons of simplicity.

1. No additional power for over-coupling of superconducting cavities.
2. No non-Ohmic effects (field emission) are considered.
3. The operation temperature of the helium bath is 4.2 K.
4. No power for auxiliary systems.
5. No cable losses.
6. Perfect match to the beam (no reflected power).
7. Duty factor of the beam and RF is equal.

Figure 7 shows the required RF power without beam as a function of the accelerating gradient for 1 m long normal conducting and superconducting RF structures with a β of 0.2. Only Ohmic losses are considered which results in a quadratic increase of the RF power with the accelerating gradient. The horizontal line represents a power level (thermal load) of 100 kW/m which is close to the present technological limit of cooling capabilities of normal conducting RF structures. This means that for this specific case the gradient is limited to about 2.5 MV/m for cw operation. Owing to the much lower surface resistance of superconducting cavities the required power is, according to this, lower by several orders of magnitude. The different curves take various values of the residual resistance R_0 into account which has to be added to the BCS value.

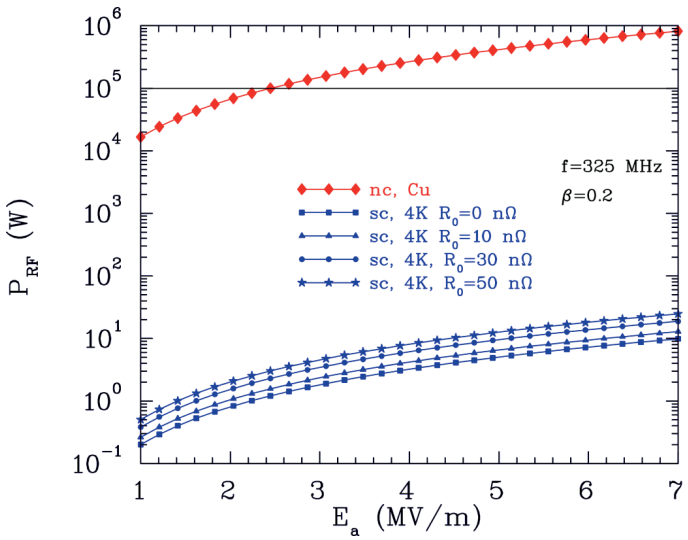


Fig. 7: RF power without beam as a function of the accelerating gradient E_a. Normal conducting cavity: f = 325 MHz, Z_{eff} = 60 MΩ/m, β = 0.2, L = 1 m. Superconducting cavity: f = 325 MHz, β = 0.2, R/Q = 3000 Ω, T = 4.2 K, G = 55 Ω.

Figure 8 shows the required plug power for a given gradient without any beam load and 100 % duty cycle. In the case of normal conducting operation an efficiency of the RF amplifier of 60 % has been assumed. The static losses for the superconducting cavities are 5 W. Owing to the low thermodynamical efficiency of the cryogenic system, the ratio between normal conducting and superconducting operation is now only between one and two orders of magnitude. At very low gradients the plug power for superconducting operation is dominated by the static losses. Beam current requires additional RF power. Figure 9 shows an example with a gradient of 4 MV/m. For each 1 mA beam current, 4 kW additional RF power has to be provided. For normal conducting cavities the power losses are still dominant up to several tens of milliamps. On the other hand, the RF power for superconducting cavities is dominated by the beam even for very small currents of less than 0.1 mA. The ratio between the total RF power of normal conducting and superconducting cavities including the beam is decreasing for higher beam currents. For very high currents such as 100 mA, the ratio decreases to a factor below two.

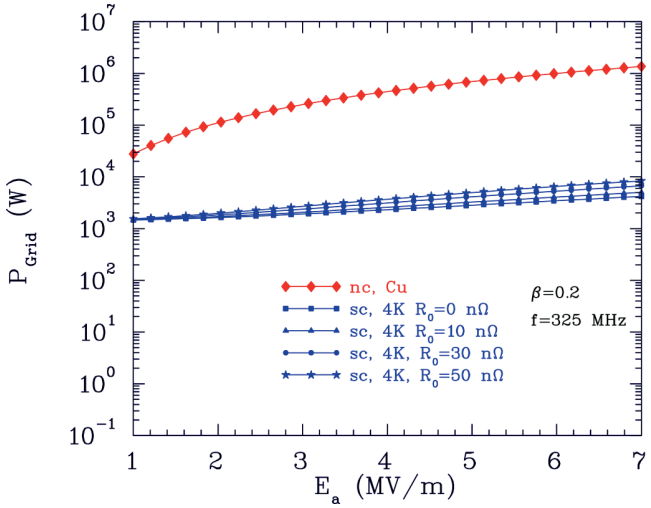


Fig. 8: Grid power without beam as a function of the accelerating gradient E_a, duty factor 100 %. Normal conducting cavity: f = 325 MHz, Z_{eff} = 60 MΩ/m, β = 0.2, L = 1 m. Superconducting cavity: f = 325 MHz, β = 0.2, R/Q = 3000 Ω, T = 4.2 K, G = 55 Ω, static losses = 5 W.

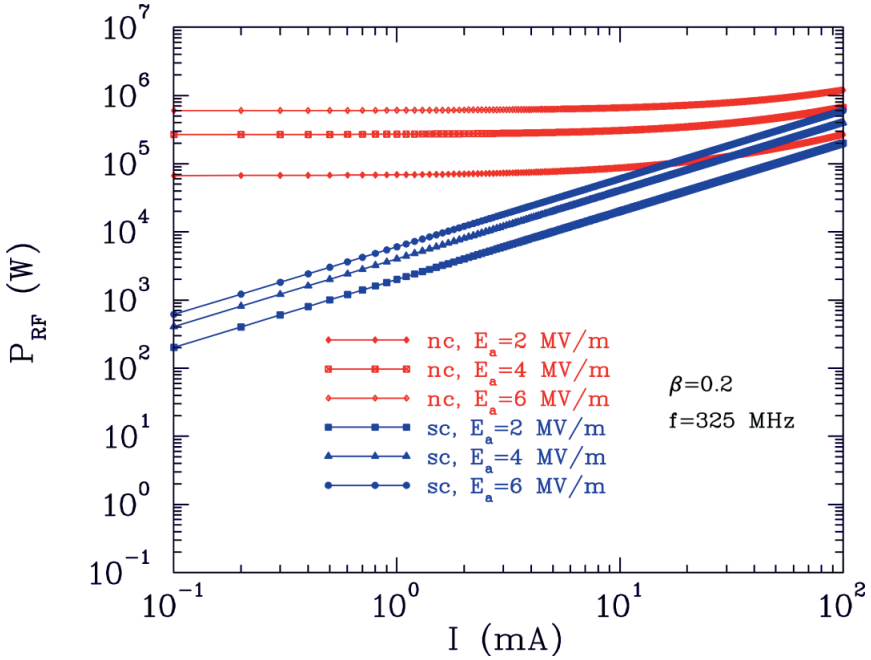


Fig. 9: RF power with beam as a function of the beam current. Normal conducting cavity: f = 325 MHz, Z_{eff} = 60 MΩ/m, β = 0.2, L = 1 m. Superconducting cavity: f = 325 MHz, β = 0.2, R/Q = 3000 Ω, T = 4.2 K, G = 55 Ω.

If superconducting cavities are operated at low duty cycles, the static losses can play an important role regarding the required grid power. Figure 10 shows the grid power as function of the beam current for different duty cycles. The higher the duty cycle the faster the grid power is dominated by the beam power in the case of superconducting cavities. The lower the duty cycle and the higher the current, the more similar is the grid power between normal conducting and superconducting cavities.

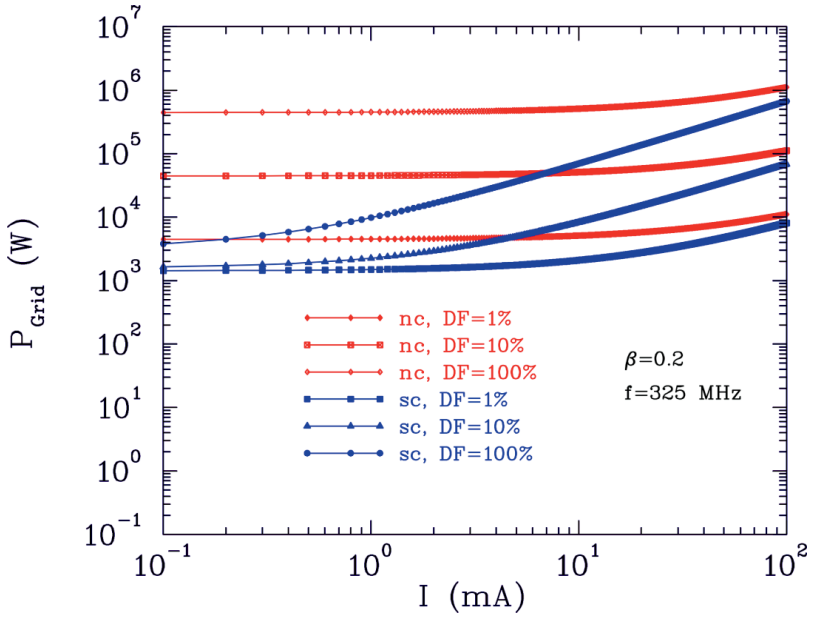


Fig. 10: Grid power with beam as a function of the beam current. Normal conducting cavity: f = 325 MHz, Z_{eff} = 60 MΩ/m, β = 0.2, L = 1 m. Superconducting cavity: f = 325 MHz, β = 0.2, R/Q = 3000 Ω, T = 4.2 K, G = 55 Ω, static losses = 5 W, R_s = 30 nΩ.

For low duty cycle machines with low beam power, the grid power can be lower for normal conducting compared with superconducting operation. Figure 11 shows the grid power without beam as a function of the duty cycle. For the superconducting cavity different values for static losses and residual resistance have been assumed. In this example ($\beta = 0.2$, $f = 325$ MHz) the break-even takes place at a duty factor of about 1 %. In particular, for lower beam energies there is often a favour for normal conducting cavities because of the existence of cavities with very high shunt impedance such as IH cavities [11].

Figure 12 shows the grid power as function of the duty factor for different beam currents ($E_a = 4$ MV/m). The power for normal conducting cavities increases linearly with the duty factor while for superconducting cavities this is only true for high beam currents and high duty factor. At lower duty factor and beam current, the power is dominated by the static losses.

Another interesting question is about the efficiency of the cavity to convert grid power into beam power (Fig. 13). As expected, the ratio of average beam power and grid power as a function of the duty cycle is constant for normal conducting cavities. It increases for higher beam current. Owing to the static losses of superconducting cavities the ratio increases for these cavities. For higher beam currents ($I > 10$ mA) the curves reach a plateau at a value which represents the efficiency of the amplifier.

Figure 14 shows the ratio of average beam power and grid power as function of the beam current. For superconducting cavities the curves reach the amplifier efficiency earlier for higher duty factor because of the higher average beam power. Normal conducting cavities typically reach only 50 % of the amplifier efficiency at reasonable beam currents.

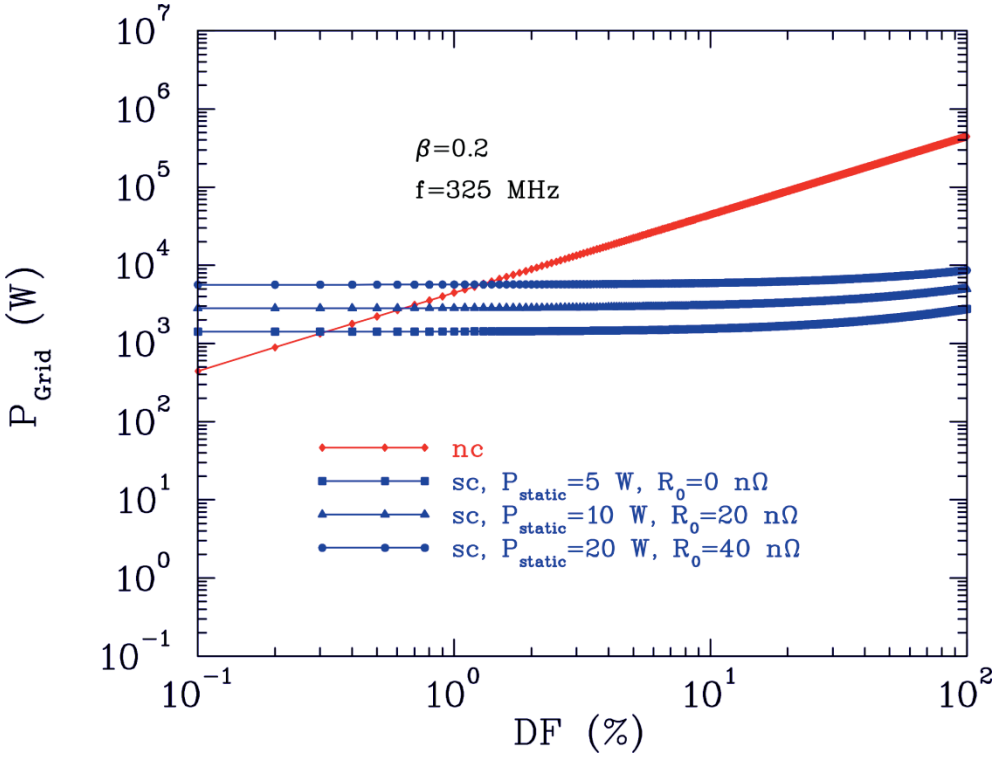


Fig. 11: Grid power without beam as a function of the duty factor, E_a = 4 MV/m. Normal conducting cavity: f = 325 MHz, Z_{eff} = 60 MΩ/m, β = 0.2, L = 1 m. Superconducting cavity: f = 325 MHz, β = 0.2, R/Q = 3000 Ω, T = 4.2 K, G = 55 Ω.

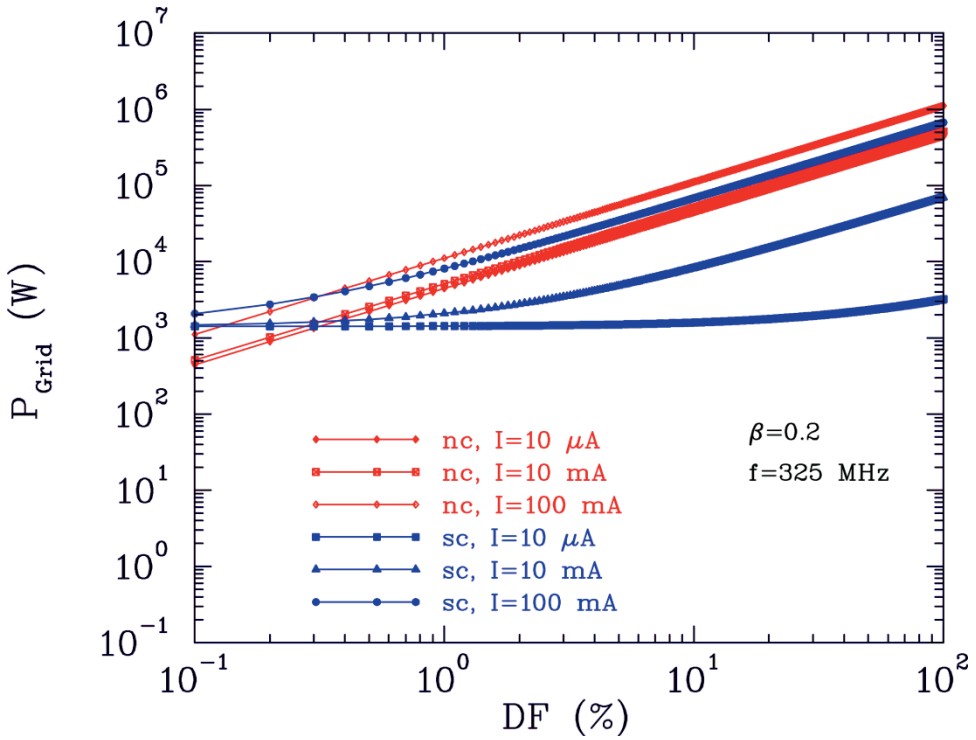


Fig. 12: Grid power with beam as a function of the duty factor, E_a = 4 MV/m. Normal conducting cavity: f = 325 MHz, Z_{eff} = 60 MΩ/m, β = 0.2, L = 1 m. Superconducting cavity: f = 325 MHz, β = 0.2, R/Q = 3000 Ω, T = 4.2 K, G = 55 Ω, static losses = 5 W, R_s = 30 nΩ.

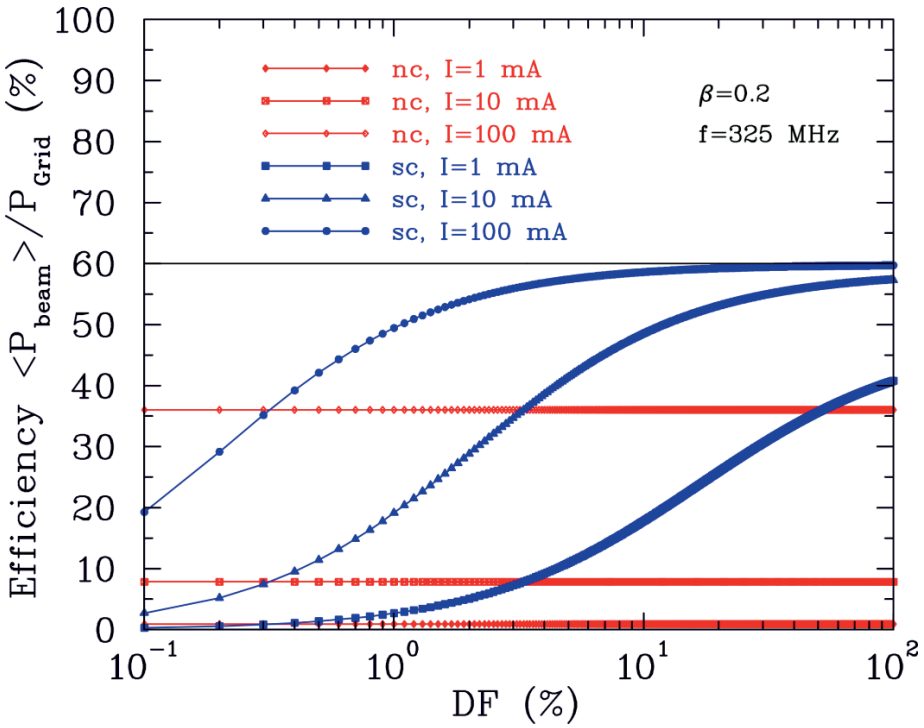


Fig. 13: Efficiency P_{beam}/P_{Grid} with beam as a function of the duty factor, E_a = 4 MV/m. Normal conducting cavity: f = 325 MHz, Z_{eff} = 60 MΩ/m, β = 0.2, L = 1 m. Superconducting cavity: f = 325 MHz, β = 0.2, R/Q = 3000 Ω, T = 4.2 K, G = 55 Ω, static losses = 5 W, R_s = 30 nΩ.

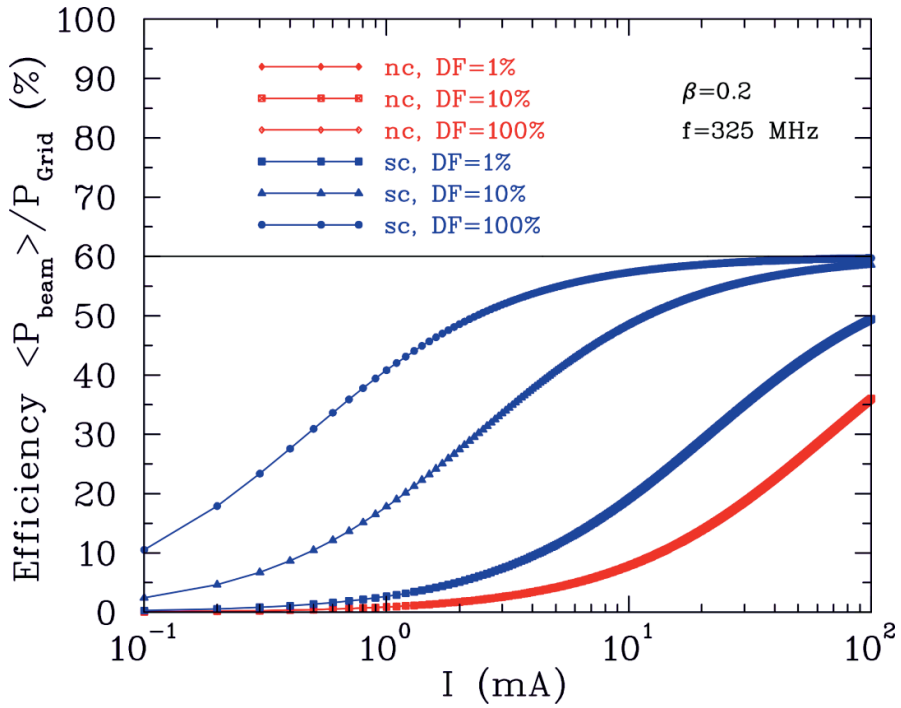


Fig. 14: Efficiency P_{beam}/P_{Grid} with beam as a function of the beam current, $E_a = 4$ MV/m. Normal conducting cavity: $f = 325$ MHz, $Z_{eff} = 60$ MΩ/m, $\beta = 0.2$, $L = 1$ m. Superconducting cavity: $f = 325$ MHz, $\beta = 0.2$, $R/Q = 3000$ Ω, $T = 4.2$ K, $G = 55$ Ω, static losses = 5 W, $R_s = 30$ nΩ.

5 Limitations of normal conducting and superconducting cavities

During testing and operating normal conducting and superconducting cavities we can observe different phenomena which are sometimes real limitations. In principle, these phenomena exist in all cavities independent of whether normal conducting or superconducting, but the severity of the problems can be very much different for normal conducting and superconducting cavities.

The most serious limitation of normal conducting cavities is the thermal load for high duty cycle operation which becomes equal to the RF losses for cw operation. Above a certain thermal load it is no longer possible to remove the heat from the cavity walls without losing performance or reliability. A typical upper limit is about 100 kW/m for normal conducting cavities, but it can be lower or higher depending on the used material (thermal conductivity), the geometry and the power density distribution.

In the case of superconducting cavities the thermal load is less important because of the low-power losses. Locally heated superconducting material has a higher surface resistance, however, resulting in even more power dissipation. This could finally lead to thermal breakdown of superconductivity (quench) above a specific field level. Sometimes there are small normal conducting inclusions within the superconducting material. In these typically sub-millimetre size defects we have significantly higher power densities. Below a power threshold, the superconducting material can conduct the additional heat while remaining superconducting. Above this threshold, the surrounding superconductor reaches the critical temperature leading to a quench. In the case of superconducting cavities there is a fundamental limitation for the maximum magnetic surface field in the cavity. Above a material specific magnetic field (200 mT for Nb) there is a breakdown of superconductivity. In reality only very few elliptical cavities have reached field levels close to this fundamental limit. Nevertheless, an important design issue for superconducting cavities is the minimization of the magnetic peak fields.

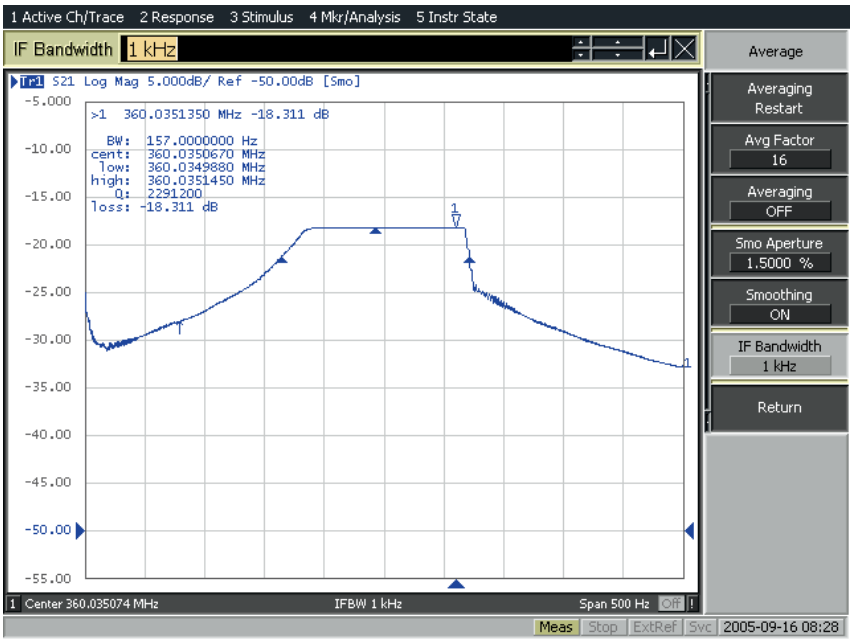


Fig. 15: Multipacting during the test of a superconducting cavity. The frequency has been swept over the resonance. Above a certain threshold (multipacting barrier) the field level remains constant.

A common problem of RF cavities can be multipacting. It is a rapid growing electron avalanche typically in the low-electric-field region. Low-energy electrons with energies of a few hundred electronvolts can hit the cavity wall and thus create secondary electrons. These electrons are accelerated and bent in the time-varying electromagnetic fields. They can hit the walls again leading to an increasing number of electrons if certain circumstances (electron energy, field distribution, frequency, field level) are present. If multipacting occurs and a so-called multipacting barrier has been reached it is not possible to increase the field level in the cavity because the stored energy and additional power is used to create the electron avalanche.

By sweeping the frequency over the resonance a flat top appears. Figure 15 shows the resonance curve with a flat top during the test of a superconducting cavity [12]. In most cases it is possible to overcome these barriers (soft barriers) with different conditioning methods. The shorter the time which is available for the avalanche creation the easier is the conditioning. This makes it clear why superconducting cavities typically suffer more from multipacting. The rise time of the fields in superconducting cavities is normally much longer than in normal conducting cavities. Stronger coupling of the cavities can decrease the rise time and shorten the conditioning time. Further information regarding multipacting can be found in Ref. [13].

The Q value is determined by the ratio of stored energy and cavity loss. Because both quantities increase quadratically with the field level, the Q value should be independent of the gradient, but most of the superconducting cavities show a significant decrease of the Q value at higher gradients. Figure 16 shows a typical measurement of the Q value as a function of the gradient [14].

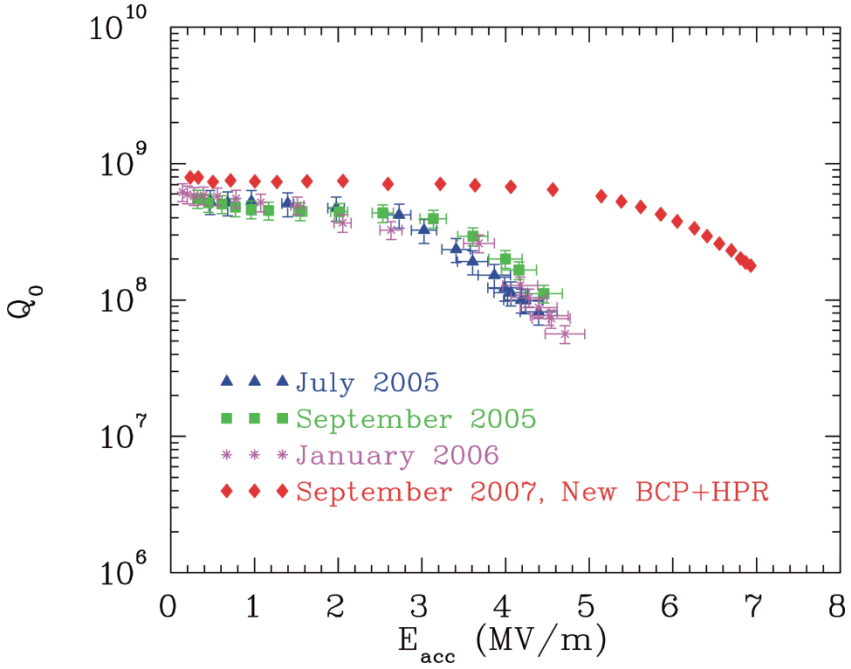


Fig. 16: Measured Q value of a superconducting cavity as function of the gradient. At higher field level the Q value decreases significantly [12].

A decreasing Q value means that the cavity power is increasing faster than the stored energy because of non-Ohmic losses. The most common reason for these additional losses is field emission. Field enhancements in the high-electric-field region can lead to emission of electrons which are then accelerated in the electric fields. Additional power is required resulting in the lower Q value. Electrons hitting the cavity are decelerated leading to the emission of X-rays. This can occur in normal conducting as well as in superconducting cavities. If the dissipated power is too high a thermal breakdown of superconductivity can happen. Typically superconducting cavities suffer by far more from field emission than normal conducting cavities because of the low dissipated power. Therefore, it is necessary to reduce the risk of field emission by special surface preparation methods (buffered chemical polishing [BCP], high-pressure rinsing [HPR]) and following assembly in a clean room. To reduce the risk of field emission one important design issue for superconducting cavities is to reduce electric peak fields and the ratio between gradient and peak field E_a/E_p.

Another possible problem is microphonics which affects mostly superconducting cavities. A cavity can oscillate mechanically in the acoustic frequency regime due to external excitations. These oscillations leading to a deformation of the cavity and finally to shifts of the RF frequency which can be much larger than the bandwidth of the cavity. There are different options to overcome this problem. We can use mechanical stiffeners to reduce the amplitude of the vibrations and to increase their frequency which makes excitation less probable. Another possibility is the use of fast-acting tuners mostly based on piezo-crystals. A stronger RF coupling (over-coupling) leads to a broader resonance curve for the price of additional RF and reflected power.

The last problem which should be mentioned is Lorentz force detuning (LFD). The electromagnetic fields inside a cavity create a pressure resulting in a field-level-dependent deformation of the cavity. This deformation leads to a detuning of the cavity ($\Delta f < 0$) which is proportional to the stored energy and the square of the field level, respectively. For cw operated superconducting cavities or normal conducting cavities in general LFD is normally not a problem, but for pulsed thin-walled superconducting cavities the detuning can be significant larger than the cavity bandwidth. In this case often fast tuners with feed-forward systems are used.

6 Case studies

In this section two linac projects are described and compared, each with normal conducting and superconducting technology. The first project is the 70 MeV proton injector for the Facility for Anti-Proton and Ion Research (FAIR) which is a typical normal conducting linac because of the high beam current and low duty cycle. The second example is the superconducting cw linac at GSI which with a duty factor of 100 % and low beam current.

6.1 NC: FAIR proton linac

The 70 MeV proton injector for the FAIR is needed to fulfil the requirements for the experimental program with respect to beam current and beam pulse structure [15]. It is the injector for the heavy ion synchrotron SIS100. It accelerates a peak proton current of up to 70 mA to a final energy of 70 MeV. The duty factor is below 0.1 %. For this project, six normal conducting CH-drift tube cavities are used [16]. The linac length is about 25 m. Owing to the very low duty cycle and the available 325 MHz, 3 MW klystrons relatively high gradients between 3 MV/m and 7 MV/m are used. Taking only the power to operate the cavities into account about 15 kW of grid power is needed. In the case of a superconducting version 14 CH cavities with a gradient of 5 MV/m would be required. This number is higher than for the normal conducting linac because it is not possible to integrate the magnetic focusing elements inside the superconducting cavities. The overall length is about the same. Assuming static losses of 10 W/m and 4 K operation we need about 100 kW of plug power to operate the cavities. Although this is much higher than for a normal conducting linac the main disadvantage of the superconducting solution would be the significant higher capital costs for the cavities, cryomodules, klystrons and cryogenic plant. Table 3 summarizes the main parameters of a normal conducting and a superconducting FAIR proton linac. Figure 17 shows the schematic layout of the normal conducting linac using CH-cavities and Fig. 18 shows the superconducting linac.

Table 3: Comparison of a normal conducting and a superconducting FAIR proton injector

Parameter	**Normal conducting**	**Superconducting**
Particles	Protons	Protons
Frequency (MHz)	325	325
Gradient (MV/m)	3–7	5
Energy (MeV)	70	70
Beam current (mA)	70	70
RF structure	nc CH	sc CH
Linac length (m)	25	25
Pulse length RF (μs)	100	200
Pulse length beam (μs)	36	36
Repetition rate (Hz)	4	4
Duty factor (%)	<0.1	<0.1
Number of cavities	6	14
P klystron (kW)	3000	500
Grid power (kW)	15	100
P_{beam}/P_{Grid} (%)	4.5	0.7

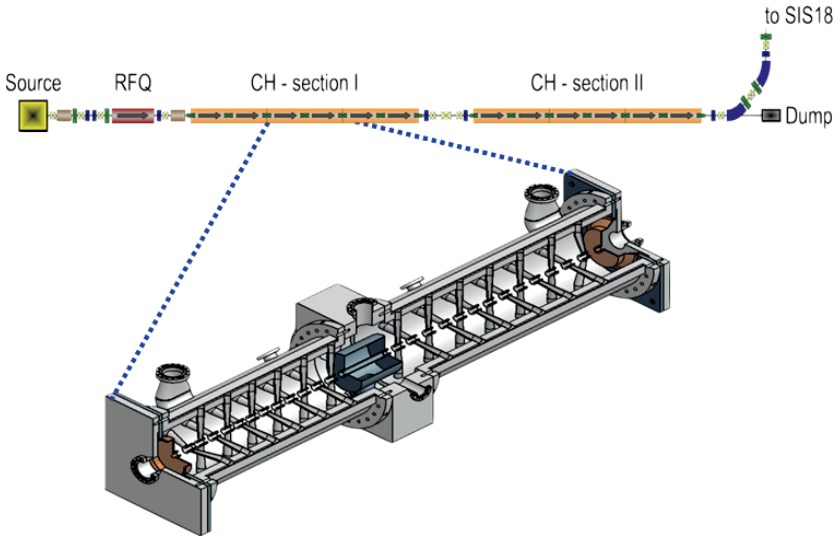


Fig. 17: Schematic layout of the normal conducting 70 MeV FAIR proton linac using CH cavities

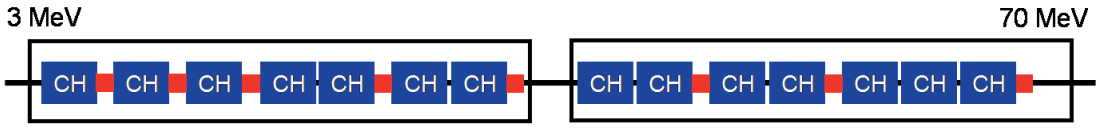


Fig. 18: In the case of a superconducting FAIR proton linac, more cavities and klystron would be required. The grid power would be significantly higher because of the low duty cycle and high beam current.

6.2 SC: GSI cw SHE Linac

The superconducting cw Super Heavy Elements (SHE) linac is a heavy ion linac operated cw [10]. It accelerates heavy ions with an A/q ratio of 6 from 1.4 AMeV to energies of up to 7.3 AMeV. Owing to the cw operation and the relatively low beam current of 100 μA, this linac is predestined to be operated in the superconducting mode. The front end is the existing High Charge State Injector at GSI [17].

The superconducting version consists of 9 CH cavities with superconducting solenoids for transverse focusing. For each cavity a 5 kW amplifier is foreseen. The total heat load at 4 K is estimated to be 500 W leading to about 200 kW of grid power. The total grid power with RF amplifiers is about 275 kW. Figure 19 shows the schematic layout of the superconducting version. The linac after the injector is about 12 m long (Fig. 19).

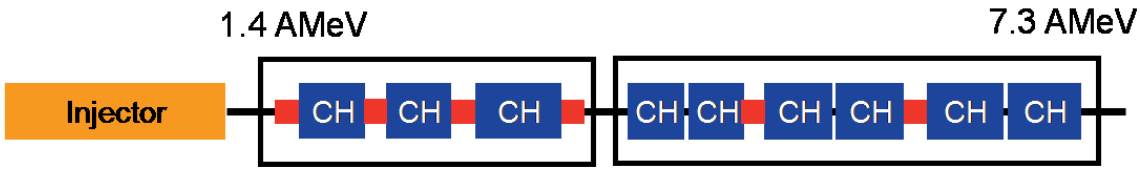


Fig. 19: Schematic layout of the superconducting cw linac at GSI. Nine CH cavities with gradients of 5 MV/m are foreseen.

In the case of a normal conducting linac we have to reduce the gradient from 5 MV/m to values between 1.5 MV/m and 2 MV/m. In total at least 14 normal conducting cavities (here IH-cavities) would be needed (Fig. 20). The required RF power for these cavities would be 900 kW resulting in a grid power of 1500 kW. This leads to 7 500 000 kW h additional energy per year. The linac length would increase by more than a factor of 2.5. Table 4 summarizes the main parameters of a normal conducting and a superconducting cw SHE linac.

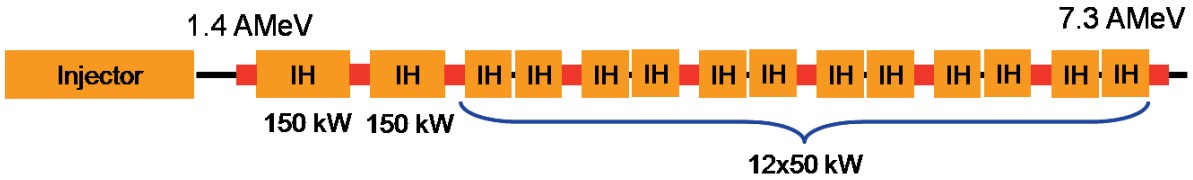


Fig. 20: In the case of a normal conducting cw linac at GSI using efficient IH cavities the required plug power would be five times the amount of a superconducting solution

Table 4: Comparison of a normal conducting and a superconducting cw SHE-linac at GSI

Parameter	**Normal conducting**	**Superconducting**
Particles	Heavy ions	Heavy ions
Frequency (MHz)	217	217
Gradient (MV/m)	1.5–2.0	5.1
Energy (AMeV)	7.3	7.3
A/q	6	6
Beam current (μA)	100	100
RF structure	nc IH	sc CH
Linac length (m)	30	12
Duty factor (%)	100	100
Number of cavities	14	9
Amplifier (kW)	50–150	5
Grid power (kW)	1500	275
P_{beam}/P_{Grid} (%)	0.24	1.13

7 Summary

Each linac project has to make the choice which is the best-suited technology for the specific application. The main issues are capital costs, operating costs (power), technical risk and reliability. During the last few decades the transition energy between normal conducting and superconducting technology decreased significantly. The main reasons for this are lower operation costs and especially the availability of suitable superconducting RF structures in the low- and medium-energy range such as quarter-wave resonators, half-wave resonators and CH cavities.

Normally the shunt impedance is higher at low energies. There are very efficient low-energy drift tube structures such as IH cavities available. Above 100 AMeV or 200 AMeV normally superconducting cavities are the best choice even for machines with lower duty cycle because there

are no real efficient normal conducting structures available. On the other side superconducting elliptical cavities can reach high gradients (E_a > 10 MV/m) resulting in a much shorter linac.

In general, normal conducting cavities are more favourable at lower energies with high beam current and low duty cycle. For superconducting cavities the opposite is valid (Fig. 21). Both technologies have their advantages and disadvantages (Fig. 22). Superconducting technology requires a cryogenic plant with associated helium distribution. The superconducting cavities are very sensitive against contaminations, pressure variations and vibrations and need in certain circumstances fast tuner systems. On the other hand, they can be operated very reliably even with cw operation.

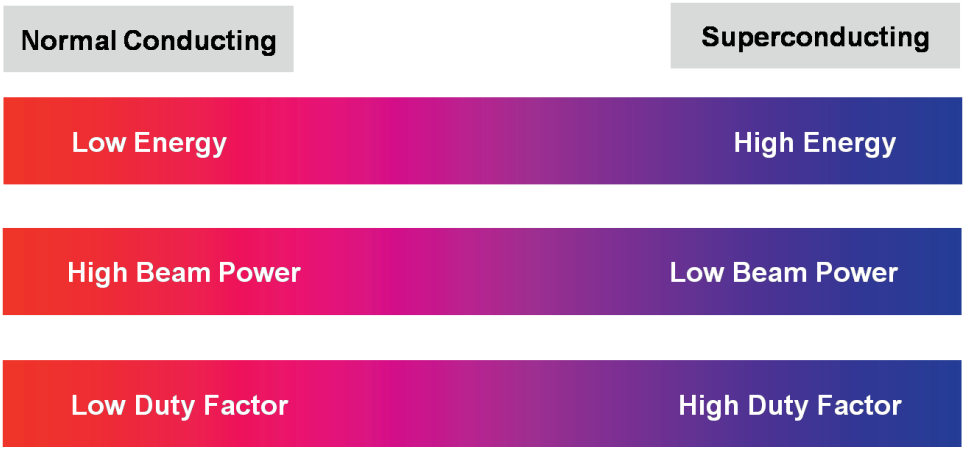


Fig. 21: Regimes of normal conducting and superconducting cavities

NC Cavities	SC Cavities
+ Less infrastructure required	+ No thermal problems
+ Simpler technology (if pulsed)	+ Stable operation
+ Simpler tuning (f and field)	+ Less RF Power
	+ Smaller amplifiers
- Higher RF power	+ Well suited for cw operation
- Expensive Amplifiers	+ Larger aperture
- Thermal problems (if high DF)	
- Lower gradients (if high DF)	- Less tolerant against beam loss
	- Cryogenic system required
	- Complicated cavity fabrication
	- Sensitive cavities

Fig. 22: Advantages and disadvantages of normal conducting and superconducting cavities

References

[1] M. White, *et al.*, Proc. of the International Linear Accelerator Conference, Gyeongju, Korea, 2002, pp. 1-5.

[2] D. Vandeplassche, *et al.*, Proc. of the International Accelerator Conference, San Sebastian, Spain, 2011, pp. 2718-2720.

[3] H. Danared, *et al.*, Proc. of the International Accelerator Conference, San Sebastian, Spain, 2011, pp. 2631-2633.

[4] A. Mosnier, Proc. of the International Linear Accelerator Conference, Victoria, Canada, 2008, pp. 1114-1118.

[5] M. Di Giacomo, International Workshop on RF Superconductivity, Cornell, NY, USA, 2002, pp. 632-636.

[6] R.C. York, Proc. of the Particle Accelerator Conference, Vancouver, Canada, 2009, pp. 70-74.

[7] A. Nagler, *et al.*, Proc. of the International Linear Accelerator Conference, Victoria, Canada, 2008, pp. 26-30.

[8] F. Gerigk, *et al.*, Proc. of the International Linear Accelerator Conference, Knoxville, TN, USA, 2006, pp. 1-3.

[9] U. Ratzinger, *et al.*, Proc. of the International Linear Accelerator Conference, Knoxville, TN, USA, 2006, pp. 1-5.

[10] S. Minaev *et al., Phys. Rev. ST Accel. Beams* **12** (2009) 120101.

[11] U. Ratzinger, *Habilitationsschrift* (University of Frankfurt, Frankfurt, 1998).

[12] H. Podlech, *et al.*, *Phys. Rev. ST Accel. Beams* **10** (2007) 080101.

[13] H. Padamssee, *et al.*, *RF Superconductivity for Accelerators*, 2nd edition (Wiley-VCH, New York, 2009).

[14] H. Podlech, *et al.*, Proc. of the International Workshop on RF Superconductivity, Beijing, China, 2007, pp. 48-54.

[15] O. Boine-Frankenheim, Proc. of the International Accelerator Conference, Kyoto, Japan, pp. 2430-2434.

[16] G. Clemente, *et al.*, *Phys. Rev. ST Accel. Beams* **14** (2011) 110101.

[17] J. Friedrich, *et al.*, Proc. of the Particle Accelerator Conference, San Francisco, CA, USA, 1991, pp. 1-3.

Beam optics and lattice design for particle accelerators

Bernhard J. Holzer
CERN, Geneva, Switzerland

Abstract
The goal of this manuscript is to give an introduction into the design of the magnet lattice and as a consequence into the transverse dynamics of the particles in a synchrotron or storage ring. Starting from the basic principles of how to design the geometry of the ring we will briefly review the transverse motion of the particles and apply this knowledge to study the layout and optimization of the principal elements, namely the lattice cells. The detailed arrangement of the accelerator magnets within the cells is explained and will be used to calculate well defined and predictable beam parameters. The more specific treatment of low beta insertions is included as well as the concept of dispersion suppressors that are an indispensable part of modern collider rings.

1 Introduction

Lattice design, in the context in which we shall describe it here, is the design and optimization of the principal elements—the lattice cells—of a (circular) accelerator, and includes the detailed arrangement of the accelerator magnets (for example, their positions in the machine and their strength) used to obtain well-defined and predictable parameters of the stored particle beam. It is therefore closely related to the theory of linear beam optics, which is treated in a number of textbooks and proceedings [1].

1.1 Geometry of the ring

Magnetic fields are used in circular accelerators to provide the bending force and to focus the particle beam. In principle, the use of electrostatic fields would be possible as well, but at high momenta (i.e., if the particle velocity is close to the speed of light), magnetic fields are much more efficient. The force acting on the particles, the Lorentz force, is given by

$$\vec{F} = q \cdot (\vec{E} + (\vec{v} \times \vec{B})).$$

Neglecting any electrostatic field, the condition for a circular orbit is given by the equality of the Lorentz force and the centrifugal force:

$$q \cdot v \cdot B = \frac{mv^2}{\rho}.$$

In a constant transverse magnetic field $\vec{B}$, a particle will see a constant deflecting force and the trajectory will be part of a circle, whose bending radius is determined by the particle momentum $p = mv$ and the external B field:

$$B \cdot \rho = p \Big/ q.$$

The term $B*\rho$ is called the beam rigidity. Inside a dipole magnet, therefore, the beam trajectory is part of a circle, and the bending angle (sketched in Fig. 1) is

$$\alpha = \frac{\int B\,\mathrm{d}s}{B\rho}\,. \tag{1}$$

For the lattice designer, the integrated B field along the design orbit of the particles (sketched roughly in Fig. 1) is the most important parameter, as it is the value that enters Eq. (1) and defines the field strength and the number of such magnets that are needed for a full circle. By requiring a bending angle of 2π for a full circle, we obtain a condition for the magnetic dipole fields in the ring. Figure 2 shows a photograph of a small storage ring [2], where only eight dipole magnets are used to define the design orbit. The magnets are powered symmetrically, and therefore each magnet corresponds to a bending angle α of the beam of exactly 45°. The field strength B in this machine is of the order of 1 T.

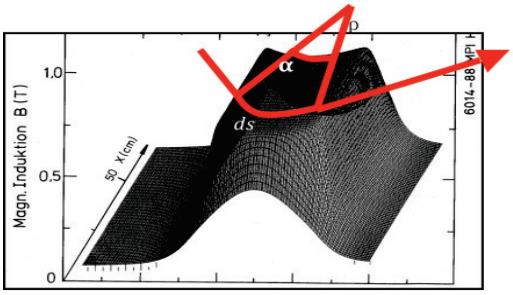


Fig. 1: Magnetic B field in a storage ring dipole and, schematically, the particle orbit

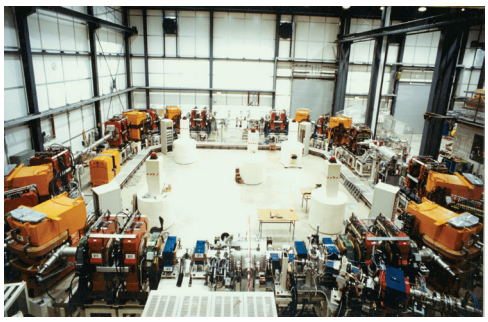


Fig. 2: The TSR heavy-ion storage ring at the Max-Planck-Institut in Heidelberg

In the case of the Large Hadron Collider LHC at CERN, for a momentum p = 7000 GeV/c, 1232 dipole magnets are needed, each having a length of 15 m and a B field of 8.3 T. As a general rule, in high-energy rings, about 66% of the circumference of the machine is used to install dipole magnets and in that way define the maximum momentum (or energy) of particles that can be stored in the machine. The remaining part of the circumference is equipped with focusing elements, RF systems for particle acceleration, diagnostics, and long straight sections for the installation of high-energy detectors.

The lattice and, correspondingly, the beam optics are therefore split into several different characteristic parts. These include arc structures, which are used to guide the particle beam and establish a regular pattern of focusing elements, leading to a regular, periodic β-function. These structures define the geometry of the ring and, as a function of the installed dipole magnets, the maximum energy of the stored particle beam. The arcs are connected by so-called insertions, long lattice sections where the optics is modified to establish the conditions needed for particle injection, to

reduce the dispersion function, or to reduce the beam dimensions in order to increase the particle collision rate (for example, where the beam in a collider ring must be prepared for particle collisions).

1.2 Equation of motion and matrix formalism

Once the geometry and specification of the arc have been determined and the layout of the bending magnets has been done, the next step is to worry about the focusing properties of the machine. In general, we have to keep more than 10^{12} particles in the machine, distributed over a number of bunches, and these particles have to be focused to keep their trajectories close to the design orbit.

As we have heard in the lecture on linear beam optics, gradient fields generated by quadrupole lenses are used to do this job. These lenses generate a magnetic field that increases linearly as a function of the distance from the magnet centre:

$$B_y = -g \cdot x, \quad B_x = -g \cdot y \ .$$

Here, x and y refer to the horizontal and vertical planes and the parameter g is called the gradient of the magnetic field. It is customary to normalize the magnetic fields to the momentum of the particles. In the case of dipole fields, we obtain from Eq. (1)

$$\alpha = \frac{\int Bds}{B\rho} = \frac{L_{eff}}{\rho},$$

where L_{eff} is the so-called effective length of the magnet. The term $1/\rho$ is the bending strength of the dipole. In the same way, the field of the quadrupole lenses is normalized to $B\rho$. The strength k is defined by

$$k = \frac{g}{B \cdot \rho},$$

and the focal length of the quadrupole is given by

$$f = \frac{1}{k\ell_q} \ .$$

The particle trajectories under the influence of the focusing properties of the quadrupole and dipole fields in the ring are described by a differential equation. This equation is derived in its full beauty in [1], so we shall just state here that it is given by the expression

$$x'' + Kx = 0 \ . \tag{2}$$

Here, x describes the horizontal coordinate of the particle with respect to the design orbit; the derivative is taken with respect to the orbit coordinate s, as usual in linear beam optics; and the parameter K combines the focusing strength k of the quadrupole and the weak-focusing term $1/\rho^2$ of the dipole field. (Note: a negative value of k means a horizontal focusing magnet.) K is given by

$$K = -k + 1/\rho^2 \ .$$

In the vertical plane, in general, the term $1/\rho^2$ is missing, as in most accelerators (but not all) the design orbit is in the horizontal plane and no vertical bending strength is present. So, in the vertical plane, we have

$$K = k \ .$$

When we are starting to design a magnet lattice, we ought to make as many simplifications as possible at the beginning of the process. Clearly, the exact solution for the particle motion has to be calculated in full detail, and if the beam optics is optimized on a linear basis, higher-order multipole fields and their effect on the beam have to be taken into account. But when we are doing the very first steps, we can make life a little bit easier and ignore terms that are small enough to be neglected.

In many cases, for example, the weak-focusing term $1/\rho^2$ can be neglected, to obtain a rough estimate that makes the formula much shorter and symmetric in the horizontal and vertical planes. Referring to the HERA proton ring as an example, the basic parameters of this machine are listed in Table 1. The weak-focusing contribution in this case, $1/\rho^2 = 2.97 \times 10^{-6}$ /m^2, is indeed much smaller than the quadrupole strength k. For initial estimates in the context of the lattices of large accelerators, this contribution can in general be neglected.

Table 1: Basic parameters of the HERA proton storage ring

Circumference C_0	6335 m
Bending radius ρ	580 m
Quadrupole gradient G	110 T/m
Particle momentum p	920 GeV/c
Weak-focusing term $1/\rho^2$	2.97×10^{-6} /m^2
Focusing strength k	3.3×10^{-3} /m^2

1.3 Single-particle trajectories

The differential equation in Eq. (2) describes the transverse motion of a particle with respect to the design orbit. This equation can be solved in a linear approximation, and the solutions for the horizontal and vertical planes are independent of each other.

If the focusing parameter K is constant, which means that we are referring to a place inside a magnet where the field is constant along the orbit, the general solution for the position and angle of the trajectory can be derived as a function of the initial conditions x_0 and x_0'. In the case of a focusing lens, we obtain

$$x(s) = x_0 * \cos(\sqrt{K} * s) + \frac{x_0'}{\sqrt{K}} * \sin(\sqrt{K} * s),$$
$$x'(s) = -x_0 * \sqrt{K} * \sin(\sqrt{K} * s) + x_0' * \cos(\sqrt{K} * s),$$

or, written in a more convenient matrix form,

$$\begin{pmatrix} x \\ x' \end{pmatrix}_s = M \cdot \begin{pmatrix} x \\ x' \end{pmatrix}_0 .$$

The matrix M depends on the properties of the magnet and, for a number of typical lattice elements, we obtain the following:

focusing quadrupole:
$$M_{QF} = \begin{pmatrix} \cos(\sqrt{K}l) & \frac{1}{\sqrt{K}}\sin(\sqrt{K}l) \\ -\sqrt{K}\sin(\sqrt{K}l) & \cos(\sqrt{K}l) \end{pmatrix}, \qquad (3a)$$

defocusing quadrupole: $$M_{QD} = \begin{pmatrix} \cosh(\sqrt{K}l) & \frac{1}{\sqrt{K}}\sinh(\sqrt{K}l) \\ \sqrt{K}\sinh(\sqrt{K}l) & \cosh(\sqrt{K}l) \end{pmatrix}, \tag{3b}$$

drift space: $$M_{\text{drift}} = \begin{pmatrix} 1 & \ell \\ 0 & 1 \end{pmatrix}. \tag{3c}$$

1.4 The Twiss parameters α, β, γ

In the case of periodic conditions in the accelerator, there is another way to describe the particle trajectories that, in many cases, is more convenient than the above-mentioned formalism, which is valid within a single element. It is important to note that in a circular accelerator, the focusing elements are necessarily periodic in the orbit coordinate s after one revolution. Furthermore, storage ring lattices have in most cases an inner periodicity: they often are constructed, at least partly, from sequences in which identical magnetic cells, the lattice cells, are repeated several times in the ring and lead to periodically repeated focusing properties.

In this case, the transfer matrix from the beginning of such a structure to the end can be expressed as a function of the periodic parameters α, β, γ, φ:

$$M(s) = \begin{pmatrix} \cos(\varphi) + \alpha_s \sin(\varphi) & \beta_s \sin(\varphi) \\ -\gamma_s \sin(\varphi) & \cos(\varphi) - \alpha_s \sin(\varphi) \end{pmatrix}. \tag{4}$$

The parameters α and γ are related to the β-function by the equations

$$\alpha(s) = -\frac{1}{2}\beta'(s) \quad \text{and} \quad \gamma(s) = \frac{1+\alpha^2(s)}{\beta(s)}.$$

The matrix is clearly a function of the position s, as the parameters α, β, γ depend on s. The variable φ is called the phase advance of the trajectory and is given by

$$\varphi = \int_s^{s+L} \frac{d\tilde{s}}{\beta(\tilde{s})}.$$

In such a periodic lattice, the relation

$$\left|\text{trace}(M)\right| < 2$$

has to be valid for stability of the equation of motion, which sets boundary conditions for the focusing properties of the lattice, as we shall see in a moment.

Given this correlation, the solution for the trajectory of a particle can be expressed as a function of the following new parameters:

$$x(s) = \sqrt{\varepsilon} \cdot \sqrt{\beta(s)} \cdot \cos(\varphi(s) - \delta),$$

$$x'(s) = \frac{-\sqrt{\varepsilon}}{\sqrt{\beta(s)}} \cdot \{\sin(\varphi(s) - \delta) + \alpha(s)\cos(\varphi(s) - \delta)\}$$

The position and angle of the transverse oscillation of a particle at a point s is given by the value of the β-function at that location, and ε and δ are constants of the particular trajectory.

As a last reminder of the linear beam optics, we state that the Twiss parameters at a position s in the lattice are defined by the focusing properties of the complete storage ring. These parameters are transformed from one point to another in the lattice by the elements of the product matrix of the corresponding magnets. Without proof, we state that if the matrix M is given by

$$M(s_1, s_2) = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix}, \tag{5}$$

the transformation rule from point s_1 to point s_2 in the lattice is given by

$$\begin{pmatrix} \beta \\ \alpha \\ \gamma \end{pmatrix}_{s_2} = \begin{pmatrix} C^2 & -2SC & S^2 \\ -CC' & SC' + S'C & -SS' \\ C'^2 & -2S'C' & S'^2 \end{pmatrix} * \begin{pmatrix} \beta \\ \alpha \\ \gamma \end{pmatrix}_{s_1} . \tag{6}$$

The terms C, S, etc. correspond to the focusing properties of the matrix. In the case of a single element, for example, they are just the expressions given in Eq. (3).

2 Lattice design

An example of a high-energy lattice and the corresponding beam optics is shown in Fig. 3, for the LHC storage ring. In general, such machines are designed on the basis of small elements, called cells, that are repeated many times in the ring. One of the most widespread lattice cells used for this purpose is the so-called FODO cell, a magnet structure consisting alternately of focusing and defocusing quadrupole lenses. Between the focusing magnet elements, the dipole magnets are installed, and any other machine elements such as orbit corrector dipoles, multipole correction coils, and diagnostics elements.

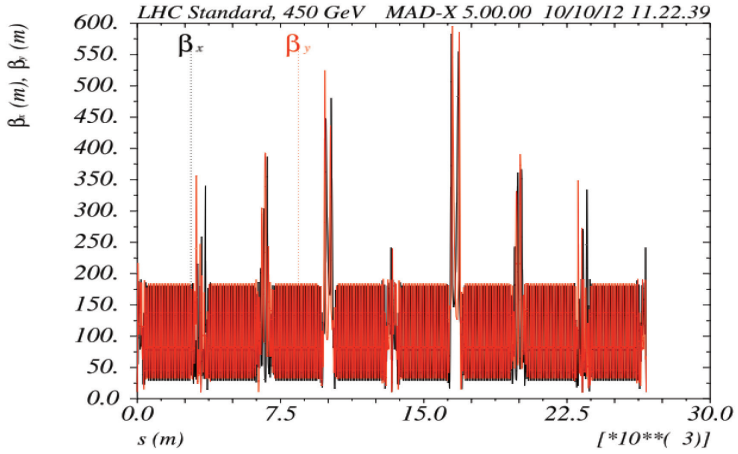


Fig. 3: Beam optics of the LHC storage ring

The optical solution for such a FODO cell is plotted in Fig. 4. The graph shows the β-function in the two transverse planes (solid line for the horizontal plane and dotted line for the vertical plane). The positions of the magnet lenses, i.e., the lattice, are shown schematically in the lower part of the plot. Owing to the symmetry of the cell, the solution for the β-function is periodic (in general, the FODO is the smallest periodic structure in a storage ring), and it reaches its maximum in the horizontal plane in the focusing lenses and its minimum in the defocusing lenses. Accordingly, the α-function is generally zero in the centre of a FODO quadrupole.

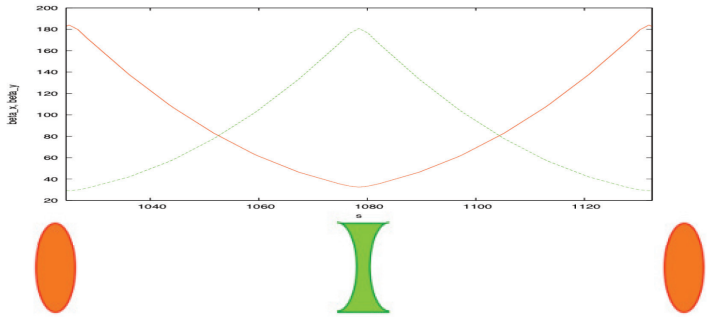


Fig. 4: A FODO cell

Table 2: Result of an optics calculation for a FODO lattice

Element	l (m)	k (1/m^2)	β_x (m)	α_x	φ_x (rad)	β_y (m)	α_y	φ_y (rad)
Start	0	–	11.611	0	0	5.295	0	0
QFH	0.25	−0.0541	11.228	1.514	0.0110	5.488	−0.78	0.0220
QD	3.251	0.0541	5.4883	−0.78	0.2196	11.23	1.514	0.2073
QFH	6.002	−0.0541	11.611	0	0.3927	5.295	0	0.3927
End	6.002	–	11.611	0	0.3927	5.295	0	0.3927

Table 2 summarizes the main parameters of the lattice magnets (quadrupole gradients, lengths, and positions) and the resulting optical properties of an example of such a single cell, calculated with an optics code. Due to symmetry reasons the calculation starts in the middle of a focusing quadrupole, named QFH in the table. Qualitatively speaking, it is already clear from the schematic drawing in Fig. 5 that the horizontal function β_x reaches its maximum value at the centre of the (horizontal) focusing quadrupoles and its minimum value at the defocusing lenses. For the vertical function β_y, a similar statement holds - vice versa - with 'maximum' and 'minimum' interchanged. The α-function in the centre of the quadrupole is indeed zero and, as $\alpha(s) = -\beta'(s)/2$, the β-function is maximum or minimum at that position.

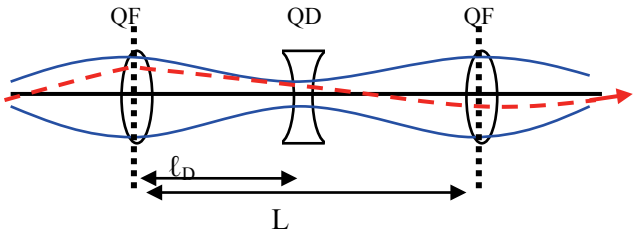


Fig. 5: Schematic drawing of a symmetric FODO cell with the beam envelope marked in blue and a single particle trajectory in red.

The phase advance of the complete machine, measured in units of 2π, is called the working point. In our case we have chosen $\varphi = 45°$, which corresponds to 0.3927 rad, as the phase advance of a single cell, and the corresponding working point is $Q_x = \int \varphi \, ds/2\pi = 0.125$. As we have chosen equal quadrupole strengths in the two planes, i.e., $k_x = -k_y$, and uniform drift spaces between the quadrupoles, the lattice is called a symmetric FODO cell. We therefore expect symmetric optical solutions in the two transverse planes.

The question now is: Can we understand what the optics code is doing? For this purpose, we refer to a single cell. In linear beam optics, the transfer matrix of a number of optical elements is given by the product of the matrices of the individual elements. In our case we obtain

$$M_{FODO} = M_{QFH} \cdot M_{Ld} \cdot M_{QD} \cdot M_{Ld} \cdot M_{QFH} . \tag{7}$$

It has to be pointed out that as we have decided to start the calculation in the centre of a quadrupole magnet, the corresponding matrix has to take this into account: the first matrix has to be that of a half quadrupole, QFH. Putting in the numbers for the length and strength $k = \pm 0.54102$ /m^2, $l_q = 0.5$ m, $l_d = 2.5$ m, where l_q and l_d refer to the length of the quadruple magnets and the drift space between them, we obtain

$$M_{\text{FODO}} = \begin{pmatrix} 0.707 & 8.206 \\ -0.061 & 0.707 \end{pmatrix}.$$

As we shall now see, this matrix describes uniquely the optical properties of the lattice and defines the beam parameters.

2.1 The most important point: stability of the motion

Taking the trace of M gives

$$\left| trace(M_{FODO}) \right| = 1.415 < 2 .$$

A lattice built out of such FODO cells would therefore give stable conditions for the particle motion. However, if new parts of the lattice are introduced, we have to go through the calculation again, as we shall see later. In addition, the matrix can be used to determine the optical parameters of the system, as described in what follows.

2.2 Phase advance per cell

Writing M as a function of α, β, γ, and the phase advance φ, we obtain for a periodic situation

$$M(s) = \begin{pmatrix} \cos(\varphi) + \alpha_s \sin(\varphi) & \beta_s \sin(\varphi) \\ -\gamma_s \sin(\varphi) & \cos(\varphi) - \alpha_s \sin(\varphi) \end{pmatrix} \tag{8}$$

and we immediately see that

$$\cos(\varphi) = \frac{1}{2} \cdot trace(M) = 0.707,$$

or $\varphi = 45°$, which corresponds to the working point of 0.125 calculated above.

2.3 Calculation of α- and β-functions

The α- and β-functions are calculated in a similar way. For β, we use the relation

$$\beta = \frac{M(1,2)}{\sin(\varphi)} = 11.611 \text{ m} ,$$

and we obtain α from the expression

$$\alpha = \frac{M(1,1) - \cos(\varphi)}{\sin(\varphi)} = 0 .$$

We can see that these analytical calculations lead to exactly the same results as the optics code used in Table 2.

To complete this first look at the optical properties of a lattice cell, I want to give a rule of thumb for the working point. Defining an average β-function for the ring, we put

$$\oint \frac{ds}{\beta(s)} \approx \frac{L}{\overline{\beta}} .$$

If we set $L = 2\pi\overline{R}$, where $\overline{R}$ is the geometric radius of the ring (which is *not* the bending radius of the dipole magnets), we can write the following for the working point Q:

$$Q = N \cdot \frac{\varphi_c}{2\pi} = \frac{1}{2\pi} \cdot \oint \frac{ds}{\beta(s)} \approx \frac{1}{2\pi} \cdot \frac{2\pi\overline{R}}{\overline{\beta}} ,$$

where N is the number of cells and φ_c denotes the phase advance per cell. So, we obtain

$$Q = \frac{\overline{R}}{\overline{\beta}} . \tag{9}$$

Therefore a rough estimate of the working point can be obtained from the ratio of the mean radius of the ring to the average β-function of the lattice.

2.4 Thin-lens approximation

As we have seen, an initial estimate of the parameters of a lattice can and should be made at the beginning of the design of a magnet lattice. If we want fast answers and require only rough estimates, we fortunately can make the task a little easier: under certain circumstances, the matrix of a focusing element can be written in the so-called thin-lens approximation.

Given for example the matrix of a focusing lens

$$M_{QF} = \begin{pmatrix} \cos(\sqrt{K}l) & \frac{1}{\sqrt{K}}\sin(\sqrt{K}l) \\ -\sqrt{K}\sin(\sqrt{K}l) & \cos(\sqrt{K}l) \end{pmatrix},$$

we can simplify the trigonometric terms if the focal length of the quadrupole magnet is much larger than the length of the lens: if

$$f = \frac{1}{k\ell_Q} >> \ell_Q ,$$

then the transfer matrix can be approximated using $kl_Q = \text{const}$, $l_Q \to 0$, and we obtain

$$M_{QF} = \begin{pmatrix} 1 & 0 \\ \frac{-1}{f} & 1 \end{pmatrix}$$

Referring to the notation used in Fig. 5, we can calculate first the transfer matrix from the centre of the focusing quadrupole to the centre of the defocusing quadrupole and obtain the matrix for half of the cell:

$$M_{\text{half cell}} = M_{\text{QD/2}} M_{\ell_D} M_{\text{QF/2}} ,$$

$$M_{\text{half cell}} = \begin{pmatrix} 1 & 0 \\ \frac{1}{\tilde{f}} & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & \ell_D \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ \frac{-1}{\tilde{f}} & 1 \end{pmatrix},$$

$$M_{\text{half cell}} = \begin{pmatrix} 1-\frac{\ell_D}{\tilde{f}} & \ell_D \\ -\frac{\ell_D}{\tilde{f}^2} & 1+\frac{\ell_D}{\tilde{f}} \end{pmatrix}. \tag{10}$$

Note that the thin-lens approximation implies that $\ell_Q \to 0$, and therefore the drift length ℓ_D between the magnets has to be equal to $L/2$. And, as we are now dealing with half quadrupoles, we must set $\tilde{f} = 2f$ for the focal length of a half quadrupole.

We obtain the second half of the cell simply by replacing $\tilde{f}$ by $-\tilde{f}$, and the matrix for the complete FODO in the thin-lens approximation is

$$M_{FODO} = \begin{pmatrix} 1+\frac{l_D}{\tilde{f}} & l_D \\ \frac{-l_D}{\tilde{f}^2} & 1-\frac{l_D}{\tilde{f}} \end{pmatrix} \cdot \begin{pmatrix} 1-\frac{l_D}{\tilde{f}} & l_D \\ \frac{-l_D}{\tilde{f}^2} & 1+\frac{l_D}{\tilde{f}} \end{pmatrix}$$

or, multiplying out,

$$M = \begin{pmatrix} 1-\frac{2\ell_D^2}{\tilde{f}^2} & 2\ell_D\left(1+\frac{\ell_D}{\tilde{f}}\right) \\ 2\left(\frac{\ell_D^2}{\tilde{f}^3}-\frac{\ell_D}{\tilde{f}^2}\right) & 1-2\frac{\ell_D^2}{\tilde{f}^2} \end{pmatrix}. \tag{11}$$

The matrix is now much easier to handle than the equivalent formulae in Eqs. (3) and (7), and the approximation is, in general, not bad.

Going briefly again through the calculation of the optics parameters, we immediately obtain from Eqs. (8) and (11)

$$\cos(\varphi) = 1-\frac{2\ell_D^2}{\tilde{f}^2}$$

and, with a little bit of trigonometric gymnastics,

$$1-2\sin^2(\varphi/2) = 1-\frac{2\ell_D^2}{\tilde{f}^2}.$$

We can simplify this expression and obtain

$$\sin(\varphi/2) = \frac{\ell_D}{\tilde{f}} = \frac{L_{cell}}{2\tilde{f}},$$

and finally

$$\sin(\varphi/2) = \frac{L_{cell}}{4f}. \tag{12}$$

In the thin-lens approximation, the phase advance of a FODO cell is given by the length of the cell $L_{cell,}$ and the focal length of the quadrupole magnets f.

For the parameters of the example given above, we obtain a phase advance per cell of $\varphi \approx 47.8°$ and, in full analogy to the calculation presented earlier, we calculate $\beta \approx 11.4$ m, which is very close to the result of the exact calculation ($\varphi = 45°$, $\beta = 11.6$ m).

2.4.1 Stability of the motion

In the thin-lens approximation, the condition for stability |trace(M) | < 2 requires that

$$\left| 2 - \frac{4\ell_D^2}{\tilde{f}^2} \right| < 2,$$

or

$$f > \frac{L_{cell}}{4}. \tag{13}$$

We have obtained the important and simple result that for stable motion, the focal length of the quadrupole lenses in the FODO has to be larger than a quarter of the length of the cell.

2.5 Scaling the optical parameters of a lattice cell

After the above discussion of stability in a lattice cell and initial estimates and calculations of the optical functions α, β, γ, and φ, we shall now concentrate a little more on a detailed analysis of a FODO concerning these parameters.

As we have seen, we can calculate the β-function that corresponds to the periodic solution — provided that we know the strength and length of the focusing elements in the cell. But can we optimize the solution somehow? In other words, for a given lattice, what would be the ideal magnet strength to obtain the smallest beam dimensions? To answer this question, we shall go back to the transfer matrix for half a FODO cell as indicated in Eq. (10), i.e., the transfer matrix from the centre of a focusing quadrupole to the centre of a defocusing quadrupole (see Fig. (4)).

From linear beam optics, we know that the transfer matrix between two points in a lattice can be expressed not only as a function of the focusing properties of the elements in that section of the ring but also, in an equivalent way, as a function of the optical parameters between the two reference points. We have used this relation already in Eq. (4) for a full turn or for one period in a periodic lattice. The general expression, in the non-periodic case, reads [1]

$$M_{1\to 2} = \begin{pmatrix} \sqrt{\frac{\beta_2}{\beta_1}}(\cos\Delta\varphi + \alpha_1 \sin\Delta\varphi) & \sqrt{\beta_2\beta_1}\sin\Delta\varphi \\ \frac{(\alpha_1 - \alpha_2)\cos\Delta\varphi - (1 + \alpha_1\alpha_2)\sin\Delta\varphi}{\sqrt{\beta_2\beta_1}} & \sqrt{\frac{\beta_1}{\beta_2}}(\cos\Delta\varphi - \alpha_2 \sin\Delta\varphi) \end{pmatrix}. \tag{14}$$

The indices refer to the starting point s_1 and the end point s_2 in the ring, and $\Delta\varphi$ is the phase advance between these points. It is evident that this matrix can be reduced to the form given in Eq. (4) if the periodic conditions $\beta_1 = \beta_2$, $\alpha_1 = \alpha_2$ are fulfilled.

We know already that β reaches its highest value in the centre of the focusing quadrupole and its lowest value in the centre of the defocusing magnet (for the vertical plane, the argument is valid 'vice versa'), and the α-functions at these positions are zero. Therefore the transfer matrix for a half cell can be written in the form

$$M = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{\check{\beta}}{\hat{\beta}}}\cos\Delta\varphi & \sqrt{\hat{\beta}\check{\beta}}\sin\Delta\varphi \\ \frac{-1}{\sqrt{\hat{\beta}\check{\beta}}}\sin\Delta\varphi & \sqrt{\frac{\hat{\beta}}{\check{\beta}}}\cos\Delta\varphi \end{pmatrix}.$$

Using this expression and putting for the matrix elements the terms that we have developed in the thin-lens approximation in Eq. (10), we obtain

$$\frac{\hat{\beta}}{\check{\beta}} = \frac{S'}{C} = \frac{1+\ell_D/\tilde{f}}{1-\ell_D/\tilde{f}} = \frac{1+\sin(\varphi/2)}{1-\sin(\varphi/2)},$$

$$\hat{\beta}\check{\beta} = \frac{-S}{C'} = \tilde{f}^2 = \frac{\ell_D^2}{\sin^2(\varphi/2)},$$

where we have set $\Delta\varphi = \varphi/2$ for the phase advance of half the FODO cell. The two expressions can be combined to calculate the two parameters $\hat{\beta}$ and $\check{\beta}$:

$$\hat{\beta} = \frac{(1+\sin(\varphi/2))L}{\sin\varphi}, \quad \check{\beta} = \frac{(1-\sin(\varphi/2))L}{\sin\varphi}. \tag{15}$$

We obtain the simple result that the maximum (and minimum) value of the β-function and therefore the maximum dimension of the beam in the cell are determined by the length L and the phase advance φ of the complete cell.

Figure 6 shows a three-dimensional picture of a proton bunch for typical conditions in the HERA storage ring. The bunch length is about 30 cm and is determined by the momentum spread and the RF potential [3]. The values of $\hat{\beta}$ and $\check{\beta}$, as determined by the cell characteristics, are typically 80 and 40 m, respectively.

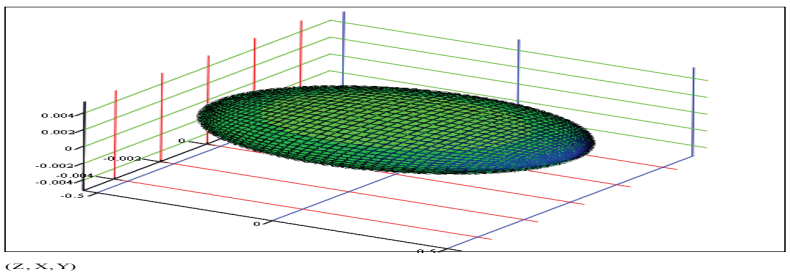


Fig. 6: Typical shape of a proton bunch in an arc of the HERA proton ring

2.5.1 Optimization of the FODO phase advance

From Eq. (15) we see that—given the length of the FODO—the maximum value of β depends only on the phase advance per cell. Therefore we may ask whether there is an optimum phase that leads to the smallest beam dimension.

If we assume a Gaussian particle distribution in the transverse plane and denote the beam emittance by ε, the transverse beam dimension σ is given by

$$\sigma = \sqrt{\varepsilon\beta}\,.$$

In a typical high-energy proton ring, ε is of the order of some 10^{-9} m·rad (e.g., for the HERA proton ring at E = 920 GeV, $\varepsilon \approx 6 \times 10^{-9}$ m·rad), and as the typical β-functions have values of about 40–100 m in an arc, the resulting beam dimension is roughly a millimetre. At the interaction point of two counter-rotating beams, even beam radii of the order of micrometres can be obtained. Figure 7 shows the result of a beam scan that was used to measure the transverse beam dimension.

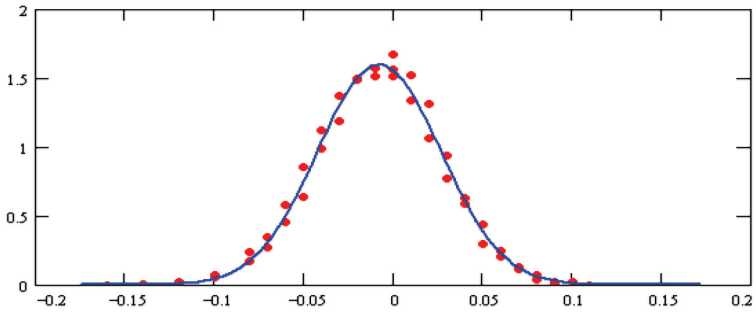


Fig. 7: Transverse beam profile of a HERA proton bunch at the interaction point. The measurement was performed by scanning the colliding beams against each other. The data points (dots) have been fitted by a Gaussian curve (line).

In general, the two emittances are equal for a proton beam, i.e., $\varepsilon_x \approx \varepsilon_y$. In this sense, a proton beam is 'round', even if the varying β-function along the lattice leads to beam dimensions in the two transverse planes that can be quite different. Optimizing the beam dimensions in the case of a proton ring therefore means searching for a minimum of the beam radius given by

$$r^2 = \varepsilon_x\beta_x + \varepsilon_y\beta_y\,,$$

and therefore optimizing the sum of the maximum and minimum β-functions

$$\hat{\beta} + \check{\beta} = \frac{(1+\sin(\varphi/2))L}{\sin\varphi} + \frac{(1-\sin(\varphi/2))L}{\sin\varphi} \tag{16}$$

at the same time. The optimum phase φ is obtained from the condition

$$\frac{\mathrm{d}}{\mathrm{d}\varphi}(\hat{\beta} + \check{\beta}) = \frac{\mathrm{d}}{\mathrm{d}\varphi}\left(\frac{2L}{\sin\varphi}\right) = 0\,,$$

which gives

$$\frac{L}{\sin^2\varphi} * \cos\varphi = 0 \quad \rightarrow \quad \varphi = 90°\,.$$

Concerning the aperture requirement of the cell, a phase advance of $\varphi = 90°$ is the best value for a proton ring. The plot in Fig. 8 shows the sum of the two β's (Eq. 16) as a function of the phase φ in the range $\varphi = 0$–$180°$. The optimization of the beam radii can be a critical issue in accelerator design.

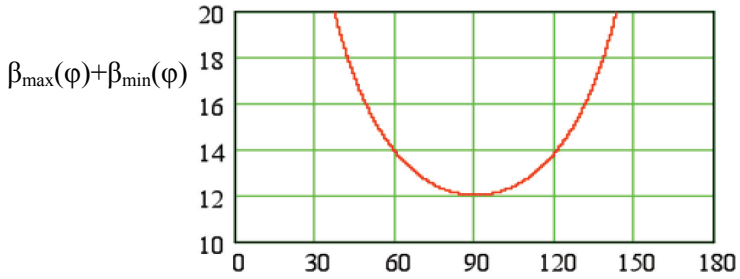


Fig. 8: Sum of the horizontal and vertical β-functions as a function on the phase advance φ

Large beam dimensions need large apertures of the quadrupole and dipole magnets in the ring. Running the machine at the highest energy can therefore lead to limitations on the focusing power, as the gradient of a quadrupole lens scales as the inverse of its squared aperture radius, i.e., $g \propto 1/r^2$, and this increases the cost of the magnet lenses. Therefore it is recommended not to tune the lattice too far away from the ideal phase advance.

Here, for completeness, I have to make a short remark about electron machines. Unlike the situation in proton rings, electron beams are flat in general: owing to the damping mechanism of synchrotron radiation [4], the vertical emittance of an electron or positron beam is only a small fraction of the horizontal emittance, such that $\varepsilon_y \approx 1$–$10\%\ \varepsilon_x$. The calculation for the optimization of the phase advance can and should be restricted to the horizontal plane only, and the condition for the smallest beam dimension is

$$\frac{d}{d\varphi}(\hat{\beta}) = \frac{d}{d\varphi}\frac{(1+\sin(\varphi/2))L}{\sin\varphi} = 0 \quad \Rightarrow \varphi = 76^o.$$

Figure 9 shows the horizontal and vertical values of β as a function of φ in this case. In an electron ring, the typical phase advance φ is in the range of 60–90°.

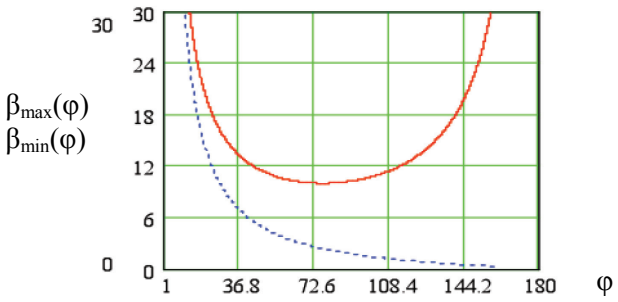


Fig. 9: Horizontal and vertical β in a FODO cell as a function of the phase advance φ

2.6 Dispersion in a FODO lattice

In our treatment of the design of a magnet lattice and our description of the optical parameters, we have restricted ourselves until now to the case where all particles have the ideal momentum. In doing so, we have followed the usual path in all books about linear beam dynamics. But, in general, the energy (or momentum) of the particles stored in a ring follows a certain distribution and deviates from the ideal momentum p_0 of the beam.

We know from linear beam optics [1] that the differential equation for the transverse motion gains an additional term if the momentum deviation is not zero: $\Delta p/p \neq 0$. We obtain an inhomogeneous equation of motion

$$x'' + K(s) \cdot x = \frac{1}{\rho} \frac{\Delta p}{p}. \tag{17}$$

The left-hand side of Eq. (17) is the same as that in the homogeneous Eq. (2), and the parameter K describes the focusing strength of the lattice element at the position s in the ring. As usual, the general solution of Eq. (17) is the sum of the complete solution x_h of the homogeneous equation and a special solution of the inhomogeneous equation, x_i:

$$x_h'' + K(s)x_h = 0,$$
$$x_i'' + K(s)x_i = \frac{1}{\rho} \cdot \frac{\Delta p}{p}.$$

The special solution x_i can be normalized to the momentum error $\Delta p/p$, and we obtain the so-called dispersion function $D(s)$:

$$x_i(s) = D(s) \cdot \frac{\Delta p}{p}. \tag{18}$$

This describes the additional amplitude of the particle oscillation due to the momentum error and is created by the $1/\rho$ term, i.e., in general, by the bending fields of the dipole magnets of our storage ring.

Starting as before from initial conditions x_0 and x_0', the general solution for the particle trajectory now reads

$$x(s) = C(s)x_0 + S(s)x_0' + D(s)\frac{\Delta p}{p}$$

or, including the expression for the angle $x'(s)$,

$$\begin{pmatrix} x \\ x' \end{pmatrix}_s = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix} \cdot \begin{pmatrix} x \\ x' \end{pmatrix}_0 + \frac{\Delta p}{p} \begin{pmatrix} D \\ D' \end{pmatrix}.$$

For convenience, in general the matrix is extended to include the second term and written

$$\begin{pmatrix} x \\ x' \\ \Delta p / p \end{pmatrix}_s = \begin{pmatrix} C & S & D \\ C' & S' & D' \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ x' \\ \Delta p / p \end{pmatrix}_0.$$

The dispersion function $D(s)$ is (obviously) defined by the focusing properties of the lattice and the bending strength of the dipole magnets $1/\rho$, and it can be shown that [1]

$$D(s) = S(s) \cdot \int_{s0}^{s} \frac{1}{\rho(\tilde{s})} C(\tilde{s})\, d\tilde{s} - C(s) \cdot \int_{s0}^{s} \frac{1}{\rho(\tilde{s})} S(\tilde{s})\, d\tilde{s}. \tag{19}$$

The variable s refers to the position where the dispersion is obtained (or measured, if you like), and the integration has to be performed over all places $\tilde{s}$ where a non-vanishing term $1/\rho$ exists (in general, in the dipole magnets of the ring).

2.6.1 Example

The 2 × 2 matrix for a drift space is given by

$$M_{\text{Drift}} = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix} = \begin{pmatrix} 1 & \ell \\ 0 & 1 \end{pmatrix}.$$

As there are no dipoles in the drift space, the $1/\rho$ term in Eq. (19) is zero and we obtain the extended 3 × 3 matrix

$$M_{\text{Drift}} = \begin{pmatrix} 1 & \ell & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

To calculate the dispersion in a FODO cell, including the 1/ρ term of the dipoles, things look quite different: We refer again to the thin-lens approximation that has already been used for the calculation of the β-functions. The matrix for a half cell has been derived above (see Eq. (10)). Again, we want to point out that in the thin-lens approximation, the length ℓ of the drift space is just half the length of the cell, as the quadrupole lenses have zero length. The matrix for a half cell is

$$M_{\text{halfcell}} = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix} = \begin{pmatrix} 1-\dfrac{\ell}{\tilde{f}} & \ell \\ \dfrac{-\ell}{\tilde{f}^2} & 1+\dfrac{\ell}{\tilde{f}} \end{pmatrix}.$$

Using this expression, we can calculate the terms D, D' of the 3 × 3 matrix:

$$D(s) = S(s) \cdot \int_{s0}^{s} \frac{1}{\rho(\tilde{s})} C(\tilde{s})\, d\tilde{s} - C(s) \cdot \int_{s0}^{s} \frac{1}{\rho(\tilde{s})} S(\tilde{s})\, d\tilde{s},$$

$$D(\ell) = \frac{\ell}{\rho}\left(\ell - \frac{\ell^2}{2\tilde{f}}\right) - \left(1 - \frac{\ell}{\tilde{f}}\right)\frac{1}{\rho}\frac{\ell^2}{2} = \frac{\ell^2}{\rho} - \frac{\ell^3}{2\tilde{f}\rho} - \frac{\ell^2}{2\rho} + \frac{\ell^3}{2\tilde{f}\rho},$$

$$D(\ell) = \frac{\ell^2}{2\rho}.$$

In an analogous way, we can derive an expression for D',

$$D'(\ell) = \frac{\ell}{\rho}\left(1 + \frac{\ell}{2\tilde{f}}\right),$$

and we obtain the complete matrix for a FODO half cell,

$$M_{\text{half cell}} = \begin{pmatrix} C & S & D \\ C' & S' & D' \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1-\frac{\ell}{\tilde{f}} & \ell & \frac{\ell^2}{2\rho} \\ \frac{-\ell}{\tilde{f}^2} & 1+\frac{\ell}{\tilde{f}} & \frac{\ell}{\rho}\left(1+\frac{\ell}{2\tilde{f}}\right) \\ 0 & 0 & 1 \end{pmatrix}.$$

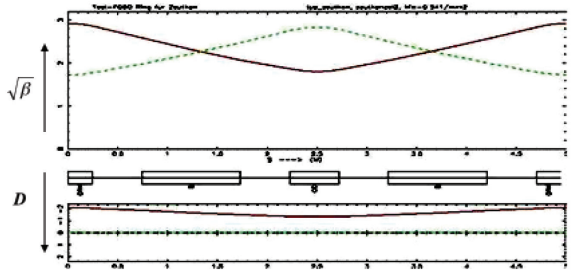


Fig. 10: β-function (top) and horizontal dispersion (bottom) in a FODO cell

Now we know that, owing to symmetry, the dispersion in a FODO lattice reaches its maximum value in the centre of a focusing quadrupole and its minimum in a defocusing quadrupole, as shown in Fig. 10, for example, where, in addition to the β-function, the dispersion is shown in the lower part of the figure. Therefore we obtain the boundary conditions for the transformation from a focusing to a defocusing lens,

$$\begin{pmatrix} \breve{D} \\ 0 \\ 1 \end{pmatrix} = M_{1/2} \begin{pmatrix} \hat{D} \\ 0 \\ 1 \end{pmatrix},$$

which can be used to calculate the dispersion at these locations:

$$\breve{D} = \hat{D}\left(1-\frac{\ell}{\tilde{f}}\right)+\frac{\ell^2}{2\rho},$$

$$0 = \frac{-\ell}{\tilde{f}^2}\hat{D} + \frac{\ell}{\rho}\left(1+\frac{\ell}{2\tilde{f}}\right).$$

Remember that we have to use the focal length of a half quadrupole,

$$\tilde{f} = 2f,$$

and that the phase advance is given by

$$\sin(\varphi/2) = \frac{L_{\text{cell}}}{2\tilde{f}}.$$

We obtain the following expressions for the maximum dispersion in the centre of a focusing quadrupole and for the minimum dispersion in the centre of a defocusing lens:

$$\hat{D} = \frac{\ell^2}{\rho}\frac{(1+(1/2)\sin(\frac{\phi_{cell}}{2}))}{sin^2(\frac{\phi_{cell}}{2})},$$

$$\breve{D} = \frac{\ell^2}{\rho}\frac{(1-(1/2)\sin(\frac{\phi_{cell}}{2}))}{sin^2(\frac{\phi_{cell}}{2})}, \tag{20}$$

It is interesting to note that the dispersion depends only on the half length ℓ of the cell, the bending strength of the dipole magnet $1/\rho$, and the phase advance φ. The dependence of D on the phase advance is shown in the plot in Fig. 11. The two values D_{max} and D_{min} decrease with increasing phase φ (which is just another way of saying 'for increasing focusing strength', as φ depends on the focusing strength of the quadrupole magnets).

To summarize these considerations, I would like to make the following remarks:

- A small dispersion needs strong focusing and therefore a large phase advance.
- There is, however, an optimum phase advance concerning the best (i.e., smallest) value of the β-function.
- Furthermore, the stability criterion limits the choice of the phase advance per cell.

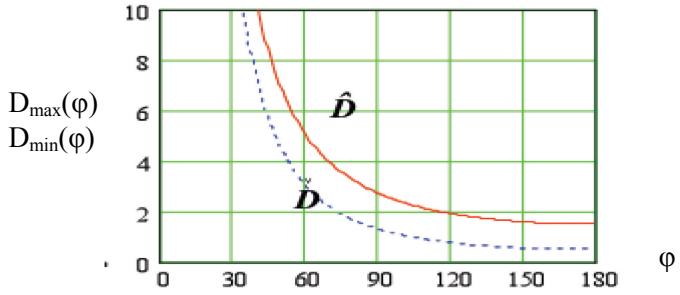


Fig. 11: Dispersion at the focusing and defocusing quadrupole lenses in a FODO as a function of the phase φ

In general, therefore, one has to find a compromise for the focusing strength in a lattice that takes into account the stability of the motion, the β-functions in both transverse planes, and the dispersion function. In a typical high-energy machine this optimization is not too difficult, as the dispersion does not have too much impact on the beam parameters (as long as it is compensated at the interaction point of the two beams).

In synchrotron light sources, however, the beam emittance is usually the parameter that has to be optimized (this means, in nearly all cases, *minimized*), and as the emittance depends on the dispersion D in an electron storage ring, the dispersion function and its optimization are of the greatest importance in these machines. In an electron ring, the horizontal beam emittance is given by the expression

$$\varepsilon_x = \frac{55}{32\sqrt{3}}\frac{\hbar}{mc}\gamma^2 \frac{\left\langle \frac{1}{R^3} H(s) \right\rangle}{J_x \left\langle \frac{1}{R^2} \right\rangle},$$

where the function $H(s)$ is defined by

$$H(s) = \gamma D^2 + 2\alpha DD' + \beta D'^2.$$

The optimization of $H(s)$ in a magnet structure is a subject of its own, and an introduction to the field of the so-called low-emittance lattices can be found, for example, in [5].

2.7 Orbit distortions in a periodic lattice

The lattice that we have designed so far consists only of a small number of basic elements: bending magnets that define the geometry of the circular accelerator and, for a given particle momentum, the

size of the machine; and quadrupole lenses that define the phase advance of the single-particle trajectories and, through this parameter, define the beam dimensions and the stability of the motion. Now it is time to fill the empty spaces in the lattice cell with some useful other components, which means we have to talk about the 'O's of the FODO.

Nobody is perfect, and this statement also holds for storage rings. In the case of a dipole magnet, an error in the bending field can be described by an additional kick θ (typically measured in mrad) on the particles,

$$\theta = \frac{ds}{\rho} = \frac{\int Bds}{p/e}$$

The beam oscillates in the corresponding plane, and the resulting amplitude of the orbit is

$$x(s) = \frac{\sqrt{\beta(s)}}{2\sin(\pi Q)} \oint \beta(\tilde{s}) \frac{1}{\rho(\tilde{s})} \cos\left(\left|\varphi(\tilde{s}) - \varphi(s)\right| - \pi Q\right) d\tilde{s}\,. \tag{21}$$

This is given by the β-function at the place of the dipole magnet $\beta(\tilde{s})$ and its bending strength $1/\rho$, and the β-function at the observation point in the lattice $\beta(s)$. For the lattice designer, this means that if a correction magnet has to be installed in the lattice cell, it should be placed at a location where β is high in the corresponding plane.

At the same time, Eq. (21) tells us that the amplitude of an orbit distortion is highest at a place where β is high, and this is the place where beam position monitors have to be located to measure the orbit distortion precisely. In practice, therefore, both the beam position monitors and the orbit correction coils are located at places in the lattice cell where the β-function in the plane considered is large, i.e., close to the corresponding quadrupole lens.

2.8 Chromaticity in a FODO cell

The chromaticity Q' describes an optical error of a quadrupole lens in an accelerator. For a given magnetic field, i.e., gradient of the quadrupole magnet, particles with smaller momentum will feel a stronger focusing force.

The chromaticity Q' relates the resulting tune shift to the relative momentum error of the particle:

$$\Delta Q = Q' \cdot \frac{\Delta p}{p}\,.$$

As it is a consequence of the focusing properties of the quadrupole magnets, it is given by the characteristics of the lattice. For small momentum errors $\Delta p/p$, the focusing parameter k can be written as

$$k(p) = \frac{g}{p/e} = g \cdot \frac{e}{p_0 + \Delta p},$$

where g denotes the gradient of the quadrupole lens and p_0 the design momentum, and the term Δp refers to the momentum error. If Δp is small, as we have assumed, we can write

$$k(p) \approx g \cdot \frac{e}{p_0}\left(1 - \frac{\Delta p}{p}\right) = k + \Delta k\,.$$

This describes a quadrupole error

$$\Delta k = -k_0 \cdot \frac{\Delta p}{p}$$

and leads to a tune shift of

$$\Delta Q = \frac{1}{4\pi} \int \Delta k \cdot \beta(s) ds,$$

or

$$\Delta Q = \frac{-1}{4\pi} \frac{\Delta p}{p} \int k_0 \cdot \beta(s) ds.$$

By definition, the chromaticity Q' of a lattice is therefore given by

$$Q' = \frac{-1}{4\pi} \int \beta(s) k(s) \, \mathrm{d}s \,. \tag{22}$$

Let us assume now that the accelerator consists of N identical FODO cells. Then, replacing $\beta(s)$ by its maximum value at the focusing quadrupoles and by its minimum value at the defocusing quadrupoles, we can approximate the integral by a sum:

$$Q' = \frac{-1}{4\pi} N \frac{\hat{\beta} - \check{\beta}}{f_Q},$$

where $f_Q = 1/(k^*\ell)$ denotes the focal length of the quadrupole magnet. We obtain

$$Q' = \frac{-1}{4\pi} N \frac{1}{f_Q} \left\{ \frac{L(1+\sin(\varphi/2) - L(1-\sin(\varphi/2)}{\sin\varphi} \right\}. \tag{23}$$

Here, we have used Eq. (15) for $\check{\beta}$ and $\hat{\beta}$. With some useful trigonometric transformations such as

$$\sin x = 2 \sin\frac{x}{2} \cos\frac{x}{2},$$

we can transform the right-hand side of Eq. (23) to obtain

$$Q' = \frac{-1}{4\pi} N \frac{1}{f_Q} \left\{ \frac{L \cdot \sin(\varphi/2)}{\sin(\varphi/2)\cos(\varphi/2)} \right\},$$

or, for one single cell (N =1),

$$Q' = \frac{-1}{4\pi} \frac{1}{f_Q} \left\{ \frac{L \cdot \tan(\varphi/2)}{\sin(\varphi/2)} \right\}.$$

Remembering the relation

$$\sin\frac{\varphi}{2} = \frac{L}{4 f_Q},$$

we obtain a surprisingly simple result for the chromaticity contribution of a FODO cell,

$$Q' = \frac{-1}{4\pi} \tan(\varphi/2) \,.$$

3 Lattice insertions

We have seen in Fig. 2 that the lattice of a typical machine for the acceleration of high-energy particles consists of two quite different types of parts: the arcs, which are built from a number of identical cells, and the straight sections that connect them and that house complicated systems such as dispersion suppressors, mini-beta insertions, and high-energy particle detectors.

3.1 Drift space

To provide some initial insight into the design of lattice insertions, I would like to start with some comments concerning a simple drift space embedded in a normal lattice structure.

What would happen to the beam parameters α, β, and γ if we were to stop focusing for a while? The transfer matrix for the Twiss parameters from the position 0 to the position s in a lattice is given by the formula

$$\begin{pmatrix}\beta\\ \alpha\\ \gamma\end{pmatrix}_s = \begin{pmatrix} C^2 & -2SC & S^2 \\ -CC' & SC'+S'C & -SS' \\ C'^2 & -S'C' & S'^2 \end{pmatrix} \cdot \begin{pmatrix}\beta\\ \alpha\\ \gamma\end{pmatrix}_0 , \tag{24}$$

where the cosine and sine functions C and S are given by the focusing properties of the lattice elements between the two points (Eqs. (3) and (7)). For a drift space of length s, this is, according to Eq. (3), as simple as

$$M = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix} = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix},$$

and the optical parameters will develop as a function of s in the following way:

$$\begin{aligned} \beta(s) &= \beta_0 - 2\alpha s + \gamma_0 s^2 , \\ \alpha(s) &= \alpha_0 - \gamma_0 s , \\ \gamma(s) &= \gamma_0 . \end{aligned} \tag{25}$$

We shall now take a closer look at these relations.

3.1.1 Location of the beam waist

From the first of these equations, we see immediately that if the drift space is long enough, even a convergent beam at the position 0 will become divergent, as the term $\gamma_0 s^2$ is always positive. This is shown schematically in Fig. 12.

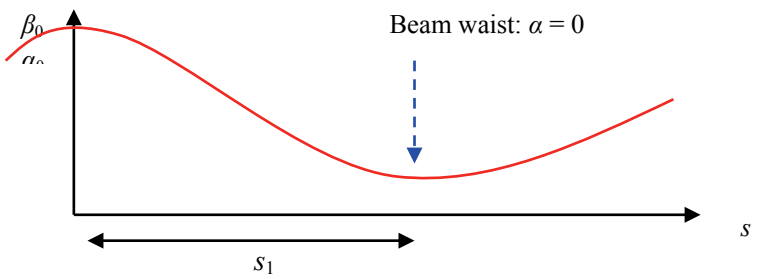


Fig. 12: Schematic drawing of a beam waist

Therefore there will be a point in the drift space where the beam dimension is smallest, in other words where the beam envelope has a waist. The position of this waist can be calculated by requiring

$$\alpha(s_1) = 0\,,$$

and the second equation of Eq. (25) then gives

$$\alpha_0 = \gamma_0 \cdot s_1,$$

or

$$s_1 = \frac{\alpha_0}{\gamma_0}\,.$$

The position of the waist is given by the ratio of the α- and γ-functions at the beginning of the drift space.

As the parameter γ is constant in a drift space and α is zero at the waist, we can directly calculate the beam size that we obtain at the waist:

$$\gamma(s_1) = \gamma_0\,, \quad \alpha(s_1) = 0\,,$$

$$\beta(s_1) = \frac{1+\alpha^2(s_1)}{\gamma(s_1)} = \frac{1}{\gamma_0}\,. \tag{26}$$

The β-function at the location of the waist is given by the inverse of the γ-function at the beginning of the drift space: a nice and simple scaling law.

3.1.2 β-function in a drift space

It is worth thinking a little bit more about the behaviour of the Twiss parameters in a drift space. Namely, the scaling of β with the length of the drift space has a large impact on the design of a lattice. At specific locations in the ring, we have to create places for beam instrumentation, for beam injection and extraction, and for the installation of particle detectors. Let us assume that we are in the centre of a drift space and that the situation is symmetric, which means that the index *0* refers to the position at the starting point, but now we want to have left–right symmetric optics with respect to it, so that $\alpha_0 = 0$. From Eq. (25), we obtain at the starting point

$$\beta(s) = \beta_0 - 2\alpha_0 s + \gamma_0 s^2,$$

and knowing already from Eq. (26) that $\alpha = 0$ at the waist, we have

$$\gamma_0 = \frac{1+\alpha_0^{\,2}}{\beta_0} = \frac{1}{\beta_0}\,.$$

Hence we obtain β as a function of the distance s from the starting point:

$$\beta(s) = \beta_0 + \frac{s^2}{\beta_0}\,. \tag{27}$$

I would like to point out two facts in this context:

- Equation (27) is a direct consequence of Liouville's theorem: the density of the phase space of the particles is constant in an accelerator. In other words, if there are only conservative forces, the beam emittance ε is constant, which leads immediately to Eq. (24) and finally Eq. (27). And, as the conservation of ε is a fundamental law, there is no trick that can be used to avoid it and no way to overcome the increase of the beam dimension in a drift space.

– The behaviour of β in a drift space has a strong impact on the design of a storage ring. As large beam dimensions have to be avoided, this means that large drift spaces are forbidden or at least very inconvenient. We shall see in the next section that this places one of the major limitations on the luminosity of colliding beams in an accelerator.

At the beam waist, we can derive another short relation that is often used for the scaling of beam parameters. The beam envelope σ is given by the β-function and the emittance of the beam by

$$\sigma(s) = \sqrt{\varepsilon \cdot \beta(s)} \,,$$

and the divergence σ' is given by

$$\sigma'(s) = \sqrt{\varepsilon \cdot \gamma(s)} \,.$$

Now, as $\gamma = (1 + \alpha^2)/\beta$, wherever $\alpha = 0$ the beam envelope has a local minimum (i.e., a waist) or maximum. At that position, the β-function is just the ratio of the beam envelope and the beam divergence:

$$\beta(s) = \frac{\sigma(s)}{\sigma'(s)} \quad \text{at a waist.}$$

If we cannot fight against Liouville's theorem, we can at least try to optimize its consequences. Equation (27) for β in a symmetric drift space can be used to find the starting value that gives the smallest beam dimension at the end of a drift space of length ℓ. Setting

$$\frac{\mathrm{d}\hat{\beta}}{\mathrm{d}\beta_0} = 1 - \frac{\ell^2}{\beta_0^2} = 0$$

gives us the value of β_0 that leads to the smallest β after a drift space of length ℓ:

$$\beta_0 = \ell \,. \tag{28}$$

For a starting value of $\beta_0 = \ell$ at the waist, the maximum beam dimension at the end of the drift space is smallest, and its value is just double the length of the drift space:

$$\hat{\beta} = 2\beta_0 = 2\ell \,.$$

3.2 Mini-beta insertions and luminosity

The discussion in the previous section has shown that the β-function in a drift space can be chosen with respect to the length ℓ to minimize the beam dimensions and, according to those dimensions, the aperture requirements for vacuum chambers and magnets. In general, the value of β is of the order of some metres and the typical length of the drift spaces in a lattice is of the same order.

However, the straight sections of a storage ring are often designed for the collision of two counter-rotating beams, and the β-functions at the collision points are therefore very small compared with their values in the arc cells. Typical values are more in the range of *centimetres* than of *metres*. Nevertheless, the same scaling law (28) holds, and the optimum length ℓ of such a drift space would be, for example, approximately 36 cm for the interaction regions of the two beams in the HERA collider. Modern high-energy detectors, in contrast, are impressive devices that consist of many large components, and they do not fit into a drift space of a few centimetres. Figure 13 shows, as an example, the ZEUS detector at the HERA collider. It is evident that a special treatment of the storage ring lattice is needed for the installation of such a huge detector.

The lattice therefore has to be modified before and after the interaction point, to establish a large drift space in which the detector for the high-energy experiment can be embedded. At the same time,

the beams have to be focused strongly to obtain very small beam dimensions in both transverse planes at the collision point or, in other words, to obtain high luminosity. Such a lattice structure is called a 'mini-beta insertion'.

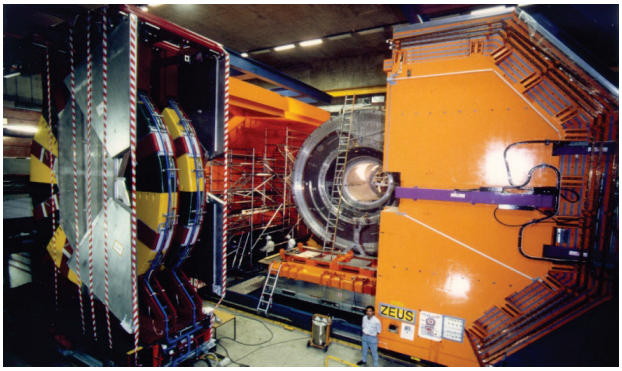


Fig. 13: Particle detector for the ZEUS collaboration at the HERA storage ring

The luminosity of a particle collider is defined by the event rate R of a specific reaction (e.g., the production of a particle in the collision of the beams):

$$R = \sigma_R \cdot L$$

The production rate of a reaction is given by its physics cross-section σ_R and a number that is the result of the lattice design: the luminosity L of the storage ring. This is determined by the beam optics at the collision point and the magnitudes of the stored beam currents [6]:

$$L = \frac{1}{4\pi e^2 f_0 b} \cdot \frac{I_1 \cdot I_2}{\sigma_x^* \cdot \sigma_y^*}$$

Here I_1 and I_2 are the values of the stored beam currents, f_0 is the revolution frequency of the machine, and b is the number of stored bunches. The quantities σ_x^* and σ_y^* in the denominator are the beam sizes in the horizontal and vertical planes at the interaction point. For a high-luminosity collider, the stored beam currents have to be high and, at the same time, the beams have to be focused at the interaction point to very small dimensions.

Figure 14 shows the typical layout of such a mini-beta insertion. It consists in general of

- a symmetric drift space that is large enough to house the particle detector and whose beam waist (where $\alpha_0 = 0$) is centred at the interaction point of the colliding beams;
- a quadrupole doublet (or triplet) on each side as close as possible;
- additional quadrupole lenses to match the Twiss parameters of the mini-beta insertion to the optical parameters of the lattice cell in the arc.

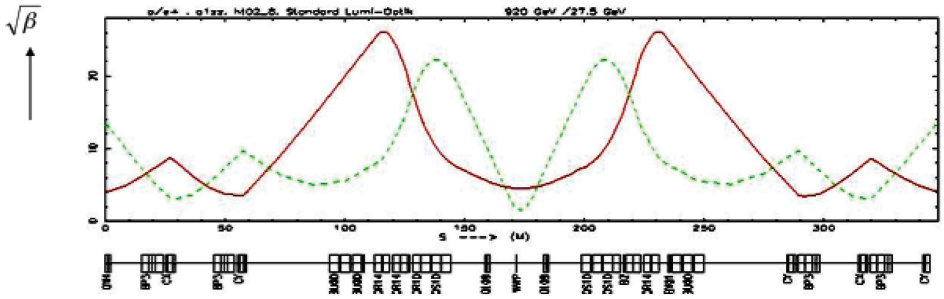


Fig. 14: Layout of a mini-beta insertion scheme

As a mini-beta scheme is always a kind of symmetric drift space, we can apply the formula that we have derived above. For $\alpha = 0$, we obtain a quadratic increase of the β-function in the drift space, and at the distance ℓ_1 of the first quadrupole lens we obtain

$$\beta(s) = \beta_0 + \frac{\ell_1^{\,2}}{\beta_0}.$$

The size of the beam at the position of the second quadrupole can be calculated in a similar way. As shown in Fig. 15, the transfer matrix of the quadrupole doublet system consists of four parts, namely two drift spaces with lengths ℓ_1 and ℓ_2, and a focusing and a defocusing quadrupole magnet. Starting at the injection point (IP), we obtain, again in the thin-lens approximation,

$$M_{\mathrm{D1}} = \begin{pmatrix} 1 & \ell_1 \\ 0 & 1 \end{pmatrix}, \qquad M_{f1} = \begin{pmatrix} 1 & 0 \\ \dfrac{1}{f_1} & 1 \end{pmatrix},$$

$$M_{\mathrm{D2}} = \begin{pmatrix} 1 & \ell_2 \\ 0 & 1 \end{pmatrix}, \qquad M_{f2} = \begin{pmatrix} 1 & 0 \\ \dfrac{-1}{f_2} & 1 \end{pmatrix}.$$

It should be noted that in general, the first lens of such a system is focusing in the vertical plane and therefore, according to the sign convention used in this school, the focal length is positive, i.e., $1/f_1 > 0$. The matrix for the complete system is

$$M = M_{QF} \cdot M_{D2} \cdot M_{QD} \cdot M_{D1}$$

$$M = \begin{pmatrix} 1 & 0 \\ -1/f_2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & \ell_2 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1/f_1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & \ell_1 \\ 0 & 1 \end{pmatrix}.$$

Multiplying out, we obtain

$$M = \begin{pmatrix} 1 + \dfrac{\ell_2}{f_1} & \ell_1 + \ell_2 + \dfrac{\ell_1 \ell_2}{f_1} \\ \dfrac{1}{f_1} - \dfrac{1}{f_2} - \dfrac{\ell_2}{f_1 f_2} & \dfrac{-\ell_1}{f_2} - \dfrac{\ell_1 \ell_2}{f_1 f_2} - \dfrac{\ell_2}{f_2} + \dfrac{\ell_1}{f_1} + 1 \end{pmatrix} = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix}.$$

Remembering the transformation of the Twiss parameters in terms of matrix elements (see Eq. (6))

$$\begin{pmatrix} \beta \\ \alpha \\ \gamma \end{pmatrix}_s = \begin{pmatrix} C^2 & -2SC & S^2 \\ -CC' & SC' + S'C & -SS' \\ C'^2 & -S'C' & S'^2 \end{pmatrix} \cdot \begin{pmatrix} \beta \\ \alpha \\ \gamma \end{pmatrix}_0,$$

we put in the terms from above and obtain

$$\beta(s) = C^2 \beta_0 - 2SC\alpha_0 + S^2 \gamma_0.$$

Here, the index 0 denotes the interaction point and s refers to the position of the second quadrupole lens. As we are starting at the IP, where $\alpha_0 = 0$ and $\gamma_0 = 1/\beta_0$, we can simplify this equation and obtain

$$\beta(s) = C^2\beta_0 + S^2/\beta_0,$$

$$\beta(s) = \beta_0 * \left(1 + \frac{\ell_2}{f_1}\right)^2 + \frac{1}{\beta_0} * \left(\ell_1 + \ell_2 + \frac{\ell_1\ell_2}{f_1}\right)^2 .$$

This formula for β at the second quadrupole lens is very useful when the gradient and aperture of a mini-beta quadrupole magnet have to be designed.

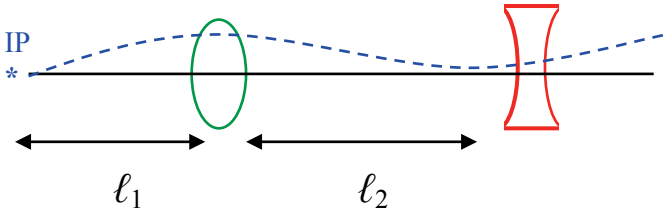


Fig. 15: Schematic layout of a mini-beta doublet

3.2.1 Phase advance in a mini-beta insertion

Unlike the situation in an arc, where the phase advance is a function of the focusing properties of the cell, in a mini-beta insertion or in any long drift space it is quasi-constant. As we know from linear beam optics, the phase advance is given by

$$\phi(s) = \int \frac{1}{\beta(s)} ds ,$$

and, inserting $\beta(s)$ from Eq. (27), we obtain

$$\phi(s) = \frac{1}{\beta_0} \int_0^{\ell_1} \frac{1}{1 + s^2/\beta_0^2} ds \ ,$$

$$\phi(s) = \arctan\frac{\ell_1}{\beta_0} \ ,$$

where ℓ_1 denotes the distance of the first focusing element from the IP, i.e., the length of the first drift space. In Fig. 16, the phase advance is plotted as a function of ℓ for a β-function of 10 cm. If the length of the drift space is large compared with the value of β at the IP, which is usually the case, the phase advance is approximately 90° on each side. In other words, the tune of the accelerator increases by half an integer in the complete drift space.

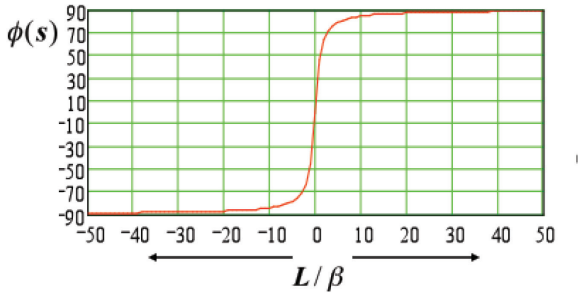


Fig. 16: Phase advance in a symmetric drift space as a function of the drift length

There are some further points that can be made concerning mini-beta sections, which will not be discussed here in detail but only mentioned briefly. As we have seen, large values of the β-function on either side of the interaction point cannot be avoided if a mini-beta section is inserted into a machine lattice. These high β values have a strong impact on the machine performance:

– According to Eq. (22), the chromaticity of a lattice is given by the strength of the focusing elements and the value of the β-function at that position:

$$Q' = \frac{-1}{4\pi}\int \beta(s)k(s)\,\mathrm{d}s \ .$$

In a mini-beta insertion, unfortunately, we have both strong quadrupoles and large beam dimensions. The contribution of such a lattice section to Q' can therefore be very large and, as it has to be corrected in the ring, it is in general a strong limitation on the luminosity in a collider ring.

– As the beam dimensions in an insertion can reach large values, the aperture of the mini-beta magnets has to be much larger than in the FODO structure of an arc. Large magnet apertures, however, limit the strength of the quadrupole. Here, a compromise has to be found between the aperture requirements, the integrated focusing strength, the spot size at the IP, and cost, as large magnets are quite expensive.

– Last but not least, the problem of field quality and adjustment has to be mentioned. Compared with the standard magnets in an arc, the lenses in a mini-beta section have to fulfil stronger requirements. A kick due to a dipole error or to an off-centre quadrupole lens leads to an orbit distortion that is proportional to the β-function at the place of the error (Eq. (21)).

As a consequence the field quality concerning higher multipole components has to be much higher and the adjustment of the mini-beta quadrupoles much more precise than for the quadrupole lenses in an arc. In general, multipole components of the order of $\Delta B/B = 10^{-4}$ with respect to the main field and alignment tolerances in the transverse plane of about a tenth of a millimetre are desired.

3.2.2 Guidelines for the design of a mini-beta insertion

– First, calculate the periodic solution for a lattice cell in the arc. This will serve to provide starting values for the insertion.

– Introduce the drift space needed for the insertion device (e.g., a particle detector).

– Put the mini-beta quadrupoles as close as possible to the IP — often, nowadays, these lenses are embedded in the detector to keep the distance s small.

– Introduce additional quadrupole lenses to match the optical parameters of the insertion to the solution for the arc cell. In general, the functions α_x, β_x, α_y, β_y and the horizontal dispersion

D_x, D'_x have to be matched. Sometimes additional quadrupoles are needed to adjust the phase advance in both planes, and in the case of HERA even the vertical dispersion D_y, D'_y needs to be corrected. So, at least eight additional magnet lenses are required.

4 Dispersion suppressors

The dispersion function $D(s)$ has already been mentioned in Section 2.6, where we have shown that it is a function of the focusing and bending properties of the lattice cell, and we calculated its size as a function of the cell parameters.

Now we have to return to this topic in the context of lattice insertions. In the interaction region of an accelerator, which means the straight section of the ring where two counter-rotating beams collide (typically designed as a mini-beta insertion), the dispersion function $D(s)$ has to vanish. A non-vanishing dispersion dilutes the luminosity of the machine and leads to additional stop bands in the working diagram of the accelerator (synchro betatron resonances) that are driven by the beam–beam interaction. Therefore, sections have to be inserted in our magnet lattice that are designed to reduce the function $D(s)$ to zero, called dispersion-suppressing schemes. In Eq. (18), we have shown that the oscillation amplitude of a particle is given by

$$x(s) = x_\beta(s) + D(s) \cdot \frac{\Delta p}{p}.$$

Here, x_β describes the solution of the homogeneous differential equation (which is valid for particles with the ideal momentum p_0), and the second term—the dispersion term—describes the additional oscillation amplitude for particles with a relative momentum error $\Delta p/p_0$.

As an example, let me present some numbers for the HERA proton storage ring. The beam size at the collision point of the two beams in the horizontal and vertical directions is determined by the mini-beta insertion, and has values $\sigma_x \approx 118$ μm and $\sigma_y \approx 32$ μm. The contribution x_D of the dispersion function to the oscillation amplitude of the particles, for a typical dispersion in the cell of $D(s) \approx 1.5$ m and a momentum distribution of the beam $\Delta p/p \approx 5 \times 10^{-4}$, is equal to 0.75 mm. Therefore a mini-beta insertion in general has to be combined with a lattice sructure to suppress the dispersion at the IP.

4.1 Dispersion suppression using additional quadrupole magnets

4.1.1 The 'straightforward way'

There are several ways to suppress the dispersion, and each of them has its advantages and disadvantages. We shall not present all of them here; instead, we shall restrict ourselves to the basic idea behind dispersion suppression. We will assume that a periodic lattice is given and that we simply want to continue the FODO structure of the arc through the straight section—but with vanishing dispersion. Given an optical solution in the arc cells, as shown for example in Fig. 17, we have to guarantee that, starting from the periodic solution for the optical parameters $\alpha(s)$, $\beta(s)$, and $D(s)$, we obtain a situation at the end of the suppressor where we have $D(s) = D'(s) = 0$ and the values for α and β are unchanged.

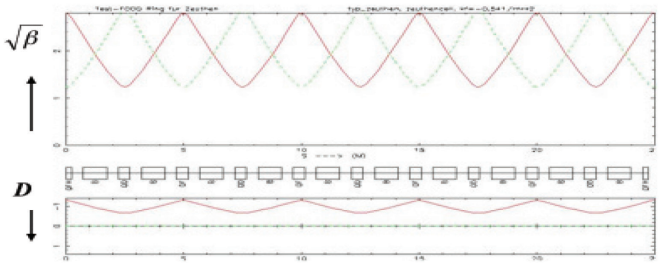


Fig. 17: Periodic FODO, including the horizontal dispersion function in the lower part of the plot

The boundary conditions

$$D(s) = D'(s) = 0,$$
$$\beta_x(s) = \beta_{x\,\mathrm{arc}}, \quad \alpha_x(s) = \alpha_{x\,\mathrm{arc}},$$
$$\beta_y(s) = \beta_{y\,\mathrm{arc}}, \quad \alpha_y(s) = \alpha_{y\,\mathrm{arc}}$$

can be fulfilled by introducing six additional quadrupole lenses, whose strengths have to be matched individually in a suitable way. This can be done by using one of the beam optics codes that are available today in every accelerator laboratory. An example is shown in Fig. 18, starting from a FODO structure with a phase advance of $\varphi \approx 61°$ per cell. The advantages of this scheme are:

- it works for an arbitrary phase advance of the arc structure;
- matching also works for different optical parameters α and β before and after the dispersion suppressor;
- the ring geometry is unchanged, as no additional dipoles are needed.

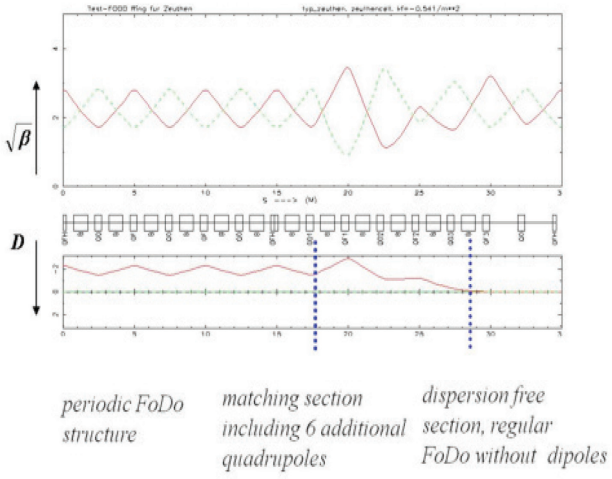


Fig. 18: Dispersion suppressor using individually powered quadrupole lenses

On the other hand, there are a number of disadvantages that have to be mentioned:

- as the strengths of the additional quadrupole magnets have to be matched individually, the scheme needs additional power supplies and quadrupole magnet types, which can be an expensive requirement;
- the required quadrupole fields are, in general, stronger than in the arc;

– the β-function reaches higher values (sometimes *really* high values), and so the aperture of the vacuum chamber and of the magnets has to be increased.

However, there are alternative ways to suppress the dispersion that do not need individually powered quadrupole lenses but instead change the strength of the dipole magnets at the end of the arc structure.

4.1.2 The 'clever way': Half-bend schemes

These dispersion-suppressing schemes consist of n additional FODO cells that are added to the periodic arc structure but where the bending strength of the dipole magnets is reduced. As before, we split the lattice into three parts: the periodic structure of the FODO cells in the arc, the lattice insertion in which the dispersion is suppressed, and a following dispersion-free section, which may be another FODO structure without bending magnets, a mini-beta insertion, or something else.

The calculation of the suppressor proceeds in several steps.

Step 1: establish the matrix for a periodic arc cell. We have already calculated the dispersion in a FODO lattice (see Eq. (20)), where we derived a formula for D in the thin-lens approximation as a function of the focusing properties of the lattice. Now we have to be a little more accurate and, instead of using the focusing strength and phase advance, we have to work with the optical parameters of the system. We know that the transfer matrix for the lattice of a storage ring can be written as a function of the optical parameters in Eq. (14):

$$M_{0\to s}=\begin{pmatrix} \sqrt{\frac{\beta_s}{\beta_0}}(\cos\phi+\alpha_0\sin\phi) & \sqrt{\beta_s\beta_0}\sin\phi \\ \frac{(\alpha_0-\alpha_s)\cos\phi-(1+\alpha_0\alpha_s)\sin\phi}{\sqrt{\beta_s\beta_0}} & \sqrt{\frac{\beta_0}{\beta_s}}(\cos\phi-\alpha_s\sin\phi) \end{pmatrix}. \tag{29}$$

The variable ϕ refers to the phase advance between the starting point *0* and the end point *s* of the transformation. This formula is valid for any starting and end points in the lattice. If, for convenience, we refer the transformation to the centre of a focusing quadrupole magnet (as we have usually done in the past), where $\alpha = 0$, and if we are interested in the solution for a complete cell, we can write the equation in a simpler form. Extending the matrix to the 3 × 3 form to include the dispersion terms (see Section 2.6) and taking the periodicity of the system into account, so that $\beta_0 = \beta_s$, we obtain

$$M_{\text{cell}}=\begin{pmatrix} C & S & D \\ C' & S' & D' \\ 0 & 0 & 1 \end{pmatrix}=\begin{pmatrix} \cos\phi_c & \beta_c\sin\phi_c & D(l) \\ \frac{-1}{\beta_c}\sin\phi_c & \cos\phi_c & D'(l) \\ 0 & 0 & 1 \end{pmatrix}. \tag{30}$$

Now ϕ_c is the phase advance for a single cell, and the index 'c' reminds us that we are talking about the periodic solution (one complete *cell*).

The dispersion elements D and D' are, as usual, given by the elements C and S according to Eq. (19):

$$D(\ell)=S(\ell)\int_0^\ell\frac{1}{\rho(\tilde{s})}C(\tilde{s})d\tilde{s}-C(\ell)\int_0^\ell\frac{1}{\rho(\tilde{s})}S(\tilde{s})d\tilde{s}\ ,$$

$$D'(\ell)=S'(\ell)\int_0^\ell\frac{1}{\rho(\tilde{s})}C(\tilde{s})d\tilde{s}-C'(\ell)\int_0^\ell\frac{1}{\rho(\tilde{s})}S(\tilde{s})d\tilde{s}.$$

The values $C(\ell)$ and $S(\ell)$ refer to the symmetry point of the cell (the centre of the quadrupole). The integrals, however, have to be taken over the dipole magnet, where $\rho \neq 0$. Assuming a constant bending radius in the dipole magnets of the arc, i.e., ρ = const (which is a good approximation in general), we can evaluate the integrals over $C(s)$ and $S(s)$ if we approximate their values by those in the centre of the dipole magnet.

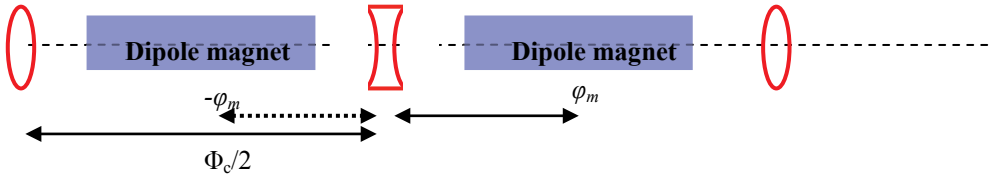


Fig. 19: Schematic view of a FODO: notation for the phase relations in the cell

Step 2: transformation of the optical functions from the centre of the quadrupole to the centre of the dipole, to calculate the functions $C(\tilde{s})$ *and* $S(\tilde{s})$. As indicated in the schematic layout in Fig. 19, we have to transform the optical functions α and β from the centre of the quadrupole lens to the centre of the dipole magnet. The formalism is given by Eq. (29), and we obtain (with $\alpha_0 = 0$)

$$C_{\mathrm{m}} = \sqrt{\frac{\beta_{\mathrm{m}}}{\beta_{\mathrm{c}}}}\cos\Delta\phi = \sqrt{\frac{\beta_{\mathrm{m}}}{\beta_{\mathrm{c}}}}\cos\left(\frac{\phi_{\mathrm{c}}}{2} \pm \varphi_{\mathrm{m}}\right),$$

$$S_{\mathrm{m}} = \sqrt{\beta_{\mathrm{m}}\beta_{\mathrm{c}}}\sin\left(\frac{\phi_{\mathrm{c}}}{2} \pm \varphi_{\mathrm{m}}\right).$$

The index 'm' tells us that we are dealing with values in the centre of the bending magnets now, and as our starting point was the centre of the focusing quadrupole, the phase advance for this transformation is half the phase advance of the cell that brings us to the defocusing lens, plus or minus the phase distance φ_{m} from that point to the centre of the dipole.

Now we can evaluate the integrals for $D(s)$ and $D'(s)$:

$$D(\ell) = \beta_c sin\phi_c \frac{L_B}{\rho}\sqrt{\frac{\beta_m}{\beta_c}}\cos\left(\frac{\phi_c}{2} \pm \varphi_m\right) - cos\phi_c \frac{L_B}{\rho}\sqrt{\beta_m\beta_c}\sin\left(\frac{\phi_c}{2} \pm \varphi_m\right), \qquad (31)$$

where L_{B} is the length of the dipole magnets. Putting $\delta = L_{\mathrm{B}}/\rho$ for the bending angle, we obtain

$$D(\ell) = \delta\sqrt{\beta_m\beta_c}\left\{sin\phi_c\left[\cos\left(\frac{\phi_c}{2} + \varphi_m\right) + \cos\left(\frac{\phi_c}{2} - \varphi_m\right)\right] - cos\phi_c\left[\sin\left(\frac{\phi_c}{2} + \varphi_m\right) + \sin\left(\frac{\phi_c}{2} - \varphi_m\right)\right]\right\}.$$

Using the trigonometric relations

$$\cos x + \cos y = 2\cdot\cos\frac{x+y}{2}\cdot\cos\frac{x-y}{2},$$

$$\sin x + \sin y = 2\cdot\sin\frac{x+y}{2}\cdot\cos\frac{x-y}{2},$$

we obtain

$$D(\ell) = \delta\sqrt{\beta_m\beta_c}\left\{sin\phi_c\, 2\cos\frac{\phi_c}{2}cos\varphi_m - cos\phi_c\, 2\sin\frac{\phi_c}{2}cos\varphi_m\right\},$$

$$D(\ell) = 2\delta\sqrt{\beta_m\beta_c}cos\varphi_m\left\{sin\phi_c\cos\frac{\phi_c}{2} - cos\phi_c\sin\frac{\phi_c}{2}\right\},$$

and with

$$\sin 2x = 2\sin x \cdot \cos x,$$
$$\cos 2x = \cos^2 x - \sin^2 x,$$

we can derive the dispersion at the centre of the quadrupole magnet in its final form:

$$D(\ell) = 2\delta\sqrt{\beta_m\beta_c}cos\varphi_m\left\{2sin\frac{\phi_c}{2}\cos^2\frac{\phi_c}{2} - \left(\cos^2\frac{\phi_c}{2} - \sin^2\frac{\phi_c}{2}\right)sin\frac{\phi_c}{2}\right\},$$
$$D(\ell) = 2\delta\sqrt{\beta_m\beta_c}cos\varphi_m sin\frac{\phi_c}{2}\left\{2\cos^2\frac{\phi_c}{2} - \cos^2\frac{\phi_c}{2} + \sin^2\frac{\phi_c}{2}\right\},$$
$$D(\ell) = 2\delta\sqrt{\beta_m\beta_c}cos\varphi_m sin\frac{\phi_c}{2}. \tag{32}$$

This is the expression for the dispersion term of the matrix in Eq. (30) at the centre of the quadrupole magnet, determined from the dipole strength $1/\rho$ and matrix elements C and S at the position of the dipole.

In full analogy, we can derive a formula for the derivative of the dispersion, $D'(s)$:

$$D'(\ell) = 2\delta\sqrt{\beta_m/\beta_c}cos\varphi_m cos\frac{\phi_c}{2}. \tag{33}$$

As we are referring to the situation in the centre of a quadrupole, the expressions for $D(s)$ and $D'(s)$ are valid for a periodic structure, namely one FODO cell. Therefore we require periodic boundary conditions for the transformation from one cell to the next:

$$\begin{pmatrix} D_c \\ D_c' \\ 1 \end{pmatrix} = M_c \cdot \begin{pmatrix} D_c \\ D_c' \\ 1 \end{pmatrix}$$

and, by symmetry,

$$D_c' = 0\,. \tag{34}$$

With these boundary conditions, the periodic dispersion in the FODO cell is determined:

$$D_c = D_c \cdot \cos\phi_c + \delta\sqrt{\beta_m\beta_c} \cdot \cos\varphi_m \cdot 2\sin\frac{\phi_c}{2},$$
$$D_c = \delta\sqrt{\beta_m\beta_c} \cdot \cos\varphi_m / \sin\frac{\phi_c}{2}. \tag{35}$$

Step 3: Calculate the dispersion in the suppressor part. In the dispersion suppressor section, starting with the value at the end of the cell, $D(s)$ is reduced to zero. Or, turning the problem around and thinking from right to left, the dispersion has to be created, starting from $D = D' = 0$. The goal is to generate the dispersion in this section in such a way that the values of the periodic arc cell are obtained.

The relation for $D(s)$ still holds in the same way:

$$D(\ell) = S(\ell)\int_0^\ell \frac{1}{\rho(\tilde{s})}C(\tilde{s})d\tilde{s} - C(\ell)\int_0^\ell \frac{1}{\rho(\tilde{s})}S(\tilde{s})d\tilde{s}.$$

But we can now take several cells into account (the number of cells in the suppressor scheme), and we have the freedom to choose a dipole strength ρ_{suppr} in this section that differs from the strength of the arc dipoles. As the dispersion is generated in n cells, the matrix for these n cells is

$$M_n = M_{\mathrm{c}}^n = \begin{pmatrix} \cos n\phi_{\mathrm{c}} & \beta_{\mathrm{c}} \sin n\phi_{\mathrm{c}} & D_n \\ \dfrac{-1}{\beta_{\mathrm{c}}} \sin n\phi_{\mathrm{c}} & \cos n\phi_{\mathrm{c}} & D_n' \\ 0 & 0 & 1 \end{pmatrix}$$

and, according to Eq. (31), the dispersion created in these n cells is given by

$$D_n = \beta_c \sin n\phi_c \cdot \delta_{\mathrm{sup}} \cdot \sum_{i=1}^{n} \cos\left(i\phi_c - \frac{1}{2}\phi_c \pm \varphi_m \right) \cdot \sqrt{\frac{\beta_m}{\beta_c}} - \cos n\phi_c \cdot \delta_{\mathrm{sup}} \cdot \sum_{i=1}^{n} \sqrt{\beta_m \beta_c} \cdot \sin\left(i\phi_c - \frac{1}{2}\phi_c \pm \varphi_m \right)$$

$$D_n = \sqrt{\beta_m \beta_c} \cdot \sin n\phi_c \cdot \delta_{\mathrm{sup}} \sum_{i=1}^{n} \cos\left((2i-1)\frac{\phi_c}{2} \pm \varphi_m \right) - \sqrt{\beta_m \beta_c} \cos n\phi_c \cdot \delta_{\mathrm{sup}} \sum_{i=1}^{n} \sin\left((2i-1)\frac{\phi_c}{2} \pm \varphi_m \right)$$

Remembering the trigonometric gymnastics shown above, we obtain

$$D_n = \delta_{\mathrm{sup}} \cdot \sqrt{\beta_m \beta_c} \cdot \sin n\phi_c \sum_{i=1}^{n} \cos\left((2i-1)\frac{\phi_c}{2} \right) \cdot 2\cos\varphi_m$$
$$-\delta_{\mathrm{sup}} \cdot \sqrt{\beta_m \beta_c} \cdot \cos n\phi_c \sum_{i=1}^{n} \sin\left((2i-1)\frac{\phi_c}{2} \right) \cdot 2\cos\varphi_m$$

$$D_n = 2\delta_{\mathrm{sup}} \cdot \sqrt{\beta_m \beta_c} \cdot \cos\varphi_m \left\{ \sum_{i=1}^{n} \cos\left((2i-1)\frac{\phi_c}{2} \right) \cdot \sin n\phi_c - \sum_{i=1}^{n} \sin\left((2i-1)\frac{\phi_c}{2} \right) \cdot \cos n\phi_c \right\}$$

$$D_n = 2\delta_{\mathrm{sup}} \sqrt{\beta_m \beta_c} \cdot \cos\varphi_m \sin(n\phi_c) \cdot \frac{\sin(n\phi_c/2) \cdot \cos(n\phi_c/2)}{\sin(\phi_c/2)}$$
$$-2\delta_{\mathrm{sup}} \sqrt{\beta_m \beta_c} \cdot \cos\varphi_m \cos(n\phi_c) \frac{\sin(n\phi_c/2) \cdot \sin(n\phi_c/2)}{\sin(\phi_c/2)}$$

$$D_n = \frac{2\delta_{\mathrm{sup}} \sqrt{\beta_m \beta_c} \cdot \cos\varphi_m}{\sin(\phi_c/2)} \left\{ 2\sin\frac{n\phi_c}{2} \cos\frac{n\phi_c}{2} \cdot \cos\frac{n\phi_c}{2} \sin\frac{n\phi_c}{2} - \left(\cos^2\frac{n\phi_c}{2} - \sin^2\frac{n\phi_c}{2} \right) \sin^2\frac{n\phi_c}{2} \right\}$$

And, finally,

$$D_n = \frac{2\delta_{\mathrm{sup}} \sqrt{\beta_m \beta_c} \cdot \cos\varphi_m}{\sin(\phi_c/2)} \sin^2\frac{n\phi_c}{2}. \tag{36}$$

This relation gives us the dispersion $D(s)$ that is created in n cells that have a phase advance of Φ_{c} per cell; δ_{sup} is the bending strength of the dipole magnets located in these n cells; and the optical functions β_{m} and β_{c} refer to the values at the centre of the dipole and of the quadrupole, respectively.

In a similar calculation, we obtain an expression for the derivative $D'(s)$ of the dispersion:

$$D'_n = \frac{2\delta_{\text{sup}}\sqrt{\beta_m / \beta_c}\cdot\cos\varphi_m}{\sin(\phi_c/2)}\sin n\phi_c . \tag{37}$$

Step 4: Determine the strength of the suppressor dipoles. The last step is to calculate the strength of the dipole magnets in the suppressor section. As the dispersion generated in this section has to be equal to that of the arc cells for the optimum match of D, we equate the expressions (34), (35) and (36), (37). For D_n, we obtain the condition

$$D_n = \frac{2\delta_{\text{sup}}\sqrt{\beta_m\beta_c}\cdot\cos\varphi_m}{\sin(\phi_c/2)}\sin^2\frac{n\phi_c}{2} = \frac{\delta_{arc}\sqrt{\beta_m\beta_c}\cdot\cos\varphi_m}{\sin(\phi_c/2)}$$

and, for D',

$$D'_n = \frac{2\delta_{\text{sup}}\sqrt{\beta_m/\beta_c}\cdot\cos\varphi_m}{\sin(\phi_c/2)}\sin n\phi_c = 0 .$$

From the latter two equations, we deduce two conditions for the dispersion matching:

$$\left.\begin{aligned} 2\delta_{\text{suppr}}\sin^2\left(\frac{n\phi_c}{2}\right) &= \delta_{\text{arc}} \\ \sin(n\phi_c) &= 0 \end{aligned}\right\} \quad \delta_{\text{suppr}} = \frac{1}{2}\delta_{\text{arc}} . \tag{38}$$

If the phase advance per cell in the arc fulfils the condition $\sin(n\phi_c) = 0$, the strength of the dipoles in the suppressor region is just half the strength of the arc dipoles. In other words, the phase has to be chosen as

$$n\phi_c = k\cdot\pi, \qquad k = 1,3,\ldots$$

There are a number of possible phase advances that fulfil this relation, but clearly not every arbitrary phase is allowed. Two possible combinations are $\phi_c = 90°$, $n = 2$ cells and $\phi_c = 60°$, $n = 3$ cells in the suppressor.

Figure 20 shows such a half-bend dispersion suppressor, starting from a FODO structure with a 60° phase advance per cell. The focusing strengths of the FODO cells before and after the suppressor are identical, with the exception that — clearly — the FODO cells on the right are 'empty', i.e., they have no bending magnets. It is evident that unlike the case for the suppressor scheme based on quadrupole lenses, the β-function is now unchanged in the suppressor region.

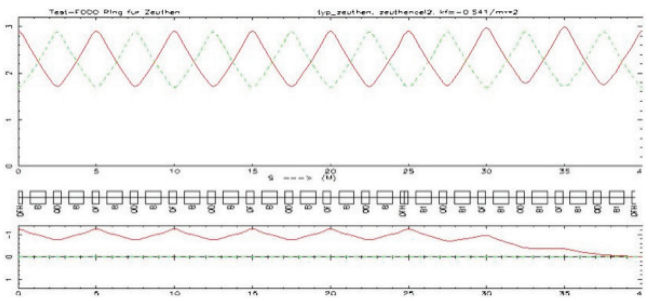


Fig. 20: Dispersion suppressor based on a half-bend scheme

Again, this scheme has advantages:

- no additional quadrupole lenses are needed, and no individual power supplies;
- the aperture requirements are just the same as those in the arc, as the β-functions are unchanged;

and disadvantages:

- it works only for certain values of the phase advance in the structure and therefore restricts the freedom of choice for the optics in the arc;
- special dipole magnets are needed (having half the strength of the arc type);
- the geometry of the ring is changed.

I should mention here, for purists only, that in these equations the phase advance of the suppressor part is equal to that of the arc structure — but this is not completely true, as the weak-focusing term $1/\rho^2$ in the arc FODO differs from the term $1/(2\rho)^2$ in the half-bend scheme. As, however, the impact of the weak focusing on the beam optics can be neglected in many practical cases, Eq. (38) is *nearly* correct.

The application of such a scheme is very elegant, but as it has a strong impact on the beam optics and geometry, it has to be embedded in the accelerator design at an early stage.

4.2 The missing-bend dispersion suppressor scheme

For completeness, I would like to present another suppressor scheme, which is also used in a number of storage rings. This consists of n cells without dipole magnets at the end of an arc, followed by m cells that are identical to the arc cells. The matching condition for this 'missing-bend scheme' with respect to the phase advance is

$$\frac{2m+n}{2}\Phi_C = (2k+1)\frac{\pi}{2},$$

and the condition for the number m of cells required is

$$\sin\frac{m\phi_c}{2} = \frac{1}{2}, \quad k = 0, 2, \ldots \quad \text{or} \quad \sin\frac{m\phi_c}{2} = \frac{-1}{2}, \quad k = 1, 3, \ldots$$

An example based on $\Phi = 60°$ and $m = n = 1$ is shown in Fig. 21. A variety of similar scenarios is feasible for different phase relations in the arc and the corresponding bending strength needed to reduce $D(s)$ [7, 8].

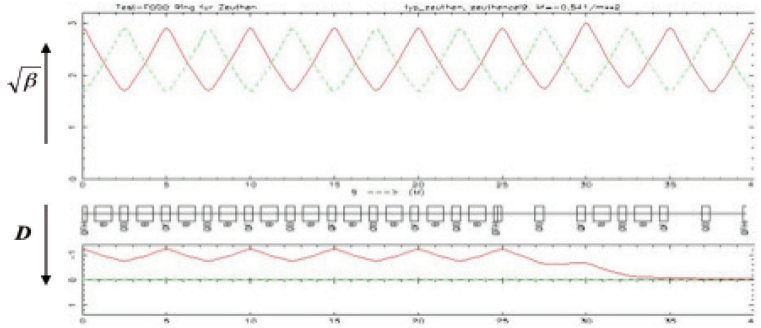


Fig. 21: Dispersion suppressor based on a missing-magnet scheme

In general, one of the above two schemes (missing-bend or half-bend suppressor) is combined with a number of individual quadrupole lenses to guarantee the flexibility of the system with respect to phase changes in the lattice and to keep the size of the β-function moderate.

References

[1] J. Rossbach and P. Schmueser, Basic course on accelerator optics, CAS Jyväskylä, 1992, CERN 94-01, CERN (1994).

[2] E. Jaeschke *et al*., The Heidelberg test storage ring for heavy ions TSR, Proc. EPAC: European Particle Accelerator Conference, Rome, Ed. S. Tazzari (World Scientific, Singapore, 1988).

[3] J. LeDuff, Longitudinal beam dynamics, CAS Jyväskylä, 1992, CERN 94-01 (1994).

[4] L. Rivkin, Electron beam dynamics with damping, contribution to this school.

[5] A. Streun, Lattices for synchrotron light sources, CAS Brunnen, 2003, CERN 2005-012 (2005).

[6] W. Herr and B. Muratori, The concept of luminosity, CAS Zeuthen, CERN 2006-002 (2006).

[7] E. Keil, Theoretical aspects of the behaviour of particle beams in accelerators and storage rings, CERN 77-13, p. 22 (1977).

[8] K. Steffen, Periodic dispersion suppressors II, DESY-HERA 83/02 (1983).

The radio-frequency quadrupole

Maurizio Vretenar
CERN, Geneva, Switzerland

Abstract
Radio-frequency quadrupole (RFQ) linear accelerators appeared on the accelerator scene in the late 1970s and have since revolutionized the domain of low-energy proton and ion acceleration. The RFQ makes the reliable production of unprecedented ion beam intensities possible within a compact radio-frequency (RF) resonator which concentrates the three main functions of the low-energy linac section: focusing, bunching and accelerating. Its sophisticated electrode structure and strict beam dynamics and RF requirements, however, impose severe constraints on the mechanical and RF layout, making the construction of RFQs particularly challenging. This lecture will introduce the main beam optics, RF and mechanical features of a RFQ emphasizing how these three aspects are interrelated and how they contribute to the final performance of the RFQ.

1 The challenges of low-energy acceleration of hadron beams

The low-energy section, between the ion source and the first drift-tube-based accelerating structure, is probably the most complicated part of any hadron linear accelerator. It is in this part of the linac that the following conditions are met:

(a) Defocusing due to space charge forces (mutually repulsive Coulomb forces between beam particles) is the highest. The space charge force acting on a single particle is inversely proportional to γ^2 (γ is the relativistic parameter here) and starts decreasing as soon as the beam becomes relativistic and the attraction between particles travelling close to the speed of light compensates for the Coulomb repulsion. The reduction will become perceptible only above few megaelectronvolts beam energy, however, leaving space charge at its maximum at energies below. To compensate for space charge, external focusing must be the highest in the low-energy range: in the usual approach, this means short focusing periods and a large number of high-gradient quadrupoles. A strong limitation to the focusing achievable at low energy, however, comes from the small dimensions of the accelerating cells. In a drift-tube structure, the distance between the centres of two quadrupoles placed inside drift tubes is $\beta\lambda$; considering that some length on the beam axis is taken by the gap and by the metal of the tubes, the space available for the quadrupole is only about $\beta\lambda/2$. At 1 MeV, $\beta = 4.6\%$ and for $\lambda \sim 1$ m the maximum length of a quadrupole is about 20 mm, nearly the same as the required aperture. The quadrupole would be dominated by fringe fields and it would be impossible to achieve on the axis a gradient sufficient to control high space charge forces.

(b) The continuous beam coming out of the source has to be bunched to be accelerated in the first radio-frequency (RF) accelerating structure. The process of bunching by means of longitudinally focusing RF forces is a critical operation: it defines the longitudinal beam emittance and can lead to the loss of a large fraction of the particles if the resulting emittance is not matched to the acceptance of the first accelerating structure.

(c) Usual low-energy accelerating structures have reduced RF efficiency and high mechanical complexity, because of the need to adapt the length of every cell to the beam velocity. Short cells

have high stray capacitances that for a given power dissipation reduce the effective voltage available for the beam, or in other terms they are particularly ineffective in concentrating on the axis the electric field required for acceleration. The result is that the accelerator cost per meter (or per megaelectronvolt acceleration) in this section tends to be the highest of all of the parts of the linac and needs to be carefully optimized.

Before the invention of the radio-frequency quadrupole (RFQs), the classical solution to cover this critical energy range was to extend as much as possible the extraction voltage from the ion source and to start the first accelerating structure, usually a drift-tube linac (DTL), from the lowest possible energy. The use of large high-voltage (HV) generators, at the limit of technology, allowed extracting from the source a beam with sufficient velocity to be injected into a DTL of relatively low frequency (to increase λ) equipped with special short quadrupoles in the first drift tubes. In these systems, however, the maximum beam current was limited by the size and aperture of the first quadrupoles and by the space charge in the transport line between the source and the DTL. Moreover, the need for a low RF frequency reduced the overall acceleration efficiency.

In the old low-energy beam transport (LEBT) lines bunching was provided by a single-gap RF cavity followed by a drift space before the DTL. The RF cavity applies a small sinusoidal modulation to the energy and velocity of the beam; after the drift, particles that were on the rising slope of the modulating voltage tend to group together, the particles that arrived first in the cavity being slower and those arriving later being faster. This will result in a higher density of particles around the particle whose energy was not changed (on the rising part of the voltage), which will be maximum at a given distance from the RF cavity. If the first accelerating gap of the DTL is placed exactly in this position, a large fraction of the beam will lie within the acceptance ("bucket") of the DTL and will be accelerated, but another fraction will be outside and will be lost in the first gaps of the DTL. A single cavity bunching section has a low transmission, of the order of 50%, and requires long drift distances where space charge can easily lead to emittance growth. To increase transmission, many bunching systems included a second harmonic cavity after the first, to linearize the overall voltage seen by the beam and extend the capture region.

As an example of low-energy section before the RFQ, Figs. 1 and 2 show the Linac2 installation at CERN as it looked between construction (in 1976) and the installation of a new RFQ replacing the original injector (in 1993). At the time of construction this was one of the proton linacs with the highest beam current in the world, 150 mA. The beam was extracted from the ion source at 750 keV, a voltage produced by a large Cockcroft–Walton generator placed in a separate HV room (Fig. 1). The 750 keV line (Fig. 2) was 5.6 m long and included four quadrupole triplets, diagnostics equipment and a double-harmonic bunching system, made of a first single-gap cavity at 202 MHz frequency followed by another at 404 MHz. The trapping efficiency of this line (ratio between current accelerated in the DTL and current extracted from the source) was as high as 80%, thanks to a careful design and to the double-harmonic bunching [1].

In spite of its sophisticated design and construction, for all operating linacs the low-energy section was not only the bottleneck in terms of beam current because of space charge limitations and of low bunching efficiency, but was also one of the main limitations in terms of reliability, the HV required for the injector being the origin of a large fraction of machine downtime.

Fig. 1: The old (1976–1993) HV installation of CERN Linac2, with the 750 kV Cockcroft–Walton in the front, the ion source electronics in its HV cabin in the background, and the sphere containing the proton source and the HV insulation to the right. The linac is in another room to the right.

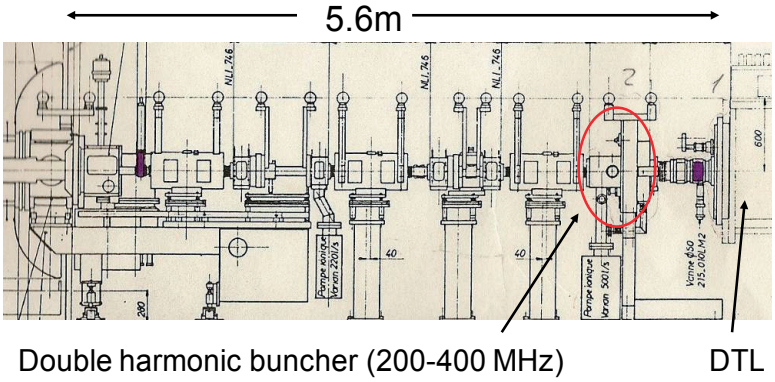


Fig. 2: Scheme of the old (1976–1993) 750 keV line of CERN Linac2, between the source of Fig. 1 and the DTL

2 The invention of the RFQ

During the 1960s and 1970s the need to build higher current proton accelerators pushed several teams, in particular in the USA and USSR, towards studying solutions to overcome the current limitations of conventional low-energy linac sections. In particular, Ilya Kapchinsky of the Institute for Theoretical and Experimental Physics (ITEP) in Moscow made a significant progress in understanding the behaviour of space charge dominated linac beams and in the frame of his studies started to develop the idea of using at low energy an electric quadrupole focusing channel excited at RF frequency as an alternative to standard electromagnetic quadrupoles. Electric quadrupole forces do not decrease for low particle velocity as the Lorentz force of a magnetic quadrupole field; if the electric field is

generated by a RF wave, a beam of particles travelling on the axis of the electric quadrupole will see an alternating gradient resulting in a net focusing force. Kapchinsky's revolutionary idea was to add to the electrodes producing the quadrupole field a longitudinal "modulation" (i.e. a sinus-like profile) which generates an additional longitudinal electric field component. By matching this longitudinal time-varying field with the velocity and phase of the particle beam it was possible to use this structure for bunching and for a moderate acceleration (more on the functioning of the RFQ will be presented in the next section). The problem of generating the quadrupole RF field was not at all trivial, and it was tackled by another Russian scientist, Vladimir Teplyakov of the Institute of High Energy Physics (IHEP) in Protvino. Together, Kapchinsky and Teplyakov published a first paper in 1969 which was the starting point for the development of the RFQ and resulted in the construction of a first experimental device in 1974 [2]. Although not classified, Kapchinsky and Teplyakov papers were published only in Russian and their work was not known in the West until 1977 when a Czech refugee brought a copy of their original paper to the linac team at Los Alamos in the USA and translated it into English. The Kapchinsky and Teplyakov device immediately looked like the long-time sought idea for generating very high currents at low energy. The Los Alamos team immediately embraced this idea and started a programme to improve it and to produce a first technological test. Los Alamos contributions consisted mainly in the development of an input radial matching section to the focusing channel and in a new resonator design that greatly reduced the non-quadrupole field components. A first proof-of-principle (POP) RFQ aimed at fusion material testing was built at Los Alamos and successfully commissioned in 1980 [3]; although it operated only for a few hours before being damaged while increasing the duty cycle, this first operational RFQ demonstrated the validity of the principle and paved the way for the successive developments. During the 1980s, the RFQ design was constantly improved and made more reliable, and RFQs started to replace the HV injectors of the main accelerator laboratories. At CERN, a first RFQ was built already in 1984, and in 1993 a second more powerful RFQ, the 200 mA RFQ2, eventually replaced the Cockcroft–Walton injector and the transport line of Linac2 [4]. Figure 3 shows the new RFQ2 (202 MHz, 1.8 m length) in front of the Linac2 DTL, with its proton source and LEBT. This new compact system occupies the same floor surface as the old 750 keV transport line of Fig. 2; the installation of the RFQ allowed decommissioning the entire HV injector of Fig. 1.

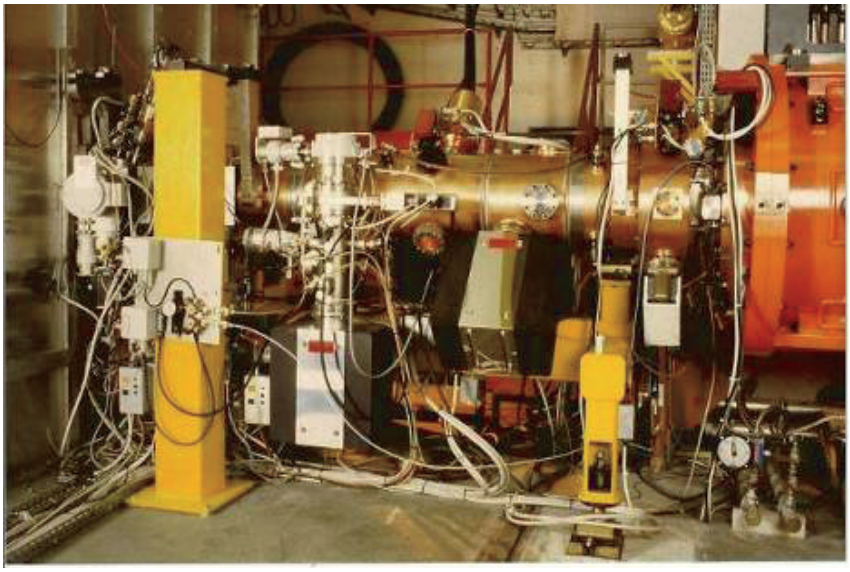


Fig. 3: The new 750 keV RFQ2 (1993) installed in front of CERN Linac2

3 The three RFQ functions

The reason why the RFQ became so popular is that it fulfils at the same time three different functions:

(i) *focusing* of the particle beam by an electric quadrupole field, particularly valuable at low energy where space charge forces are strong and conventional magnetic quadrupoles are less effective;

(ii) *adiabatic bunching* of the beam: starting from the continuous beam produced by the source it creates with minimum beam loss the bunches at the basic RF frequency that are required for acceleration in the subsequent structures;

(iii) *acceleration* of the beam from the extraction energy of the source to the minimum required for injection into the following structure.

In modern systems the ion source is followed by a short LEBT required to match transversally the beam coming from the source to the acceptance of the RFQ. Extraction from the ion source (and injection into the RFQ) is usually done at an energy of a few tens of kiloelectronvolts, achievable with small size HV installations. The RFQ follows the LEBT, and accelerates the beam up to entrance of the following structure, usually a DTL. Although a RFQ could accelerate the beam to any energy, most of the RF power delivered to the resonator goes to establishing the focusing and bunching field, with the consequence that its acceleration efficiency is very poor. For this reason, RFQs are used only in the low-energy range, up to few megaelectronvolts for protons, and their length usually reaches a maximum of a few metres. As soon as the beam is bunched and the energy is sufficiently high, it is economically convenient to pass to another type of accelerating structure. Figure 4 shows a photograph of the inside of a RFQ (CERN RFQ1, 202 MHz) and a three-dimensional view of the CERN RFQ for Linac4 (352 MHz), presently under construction.

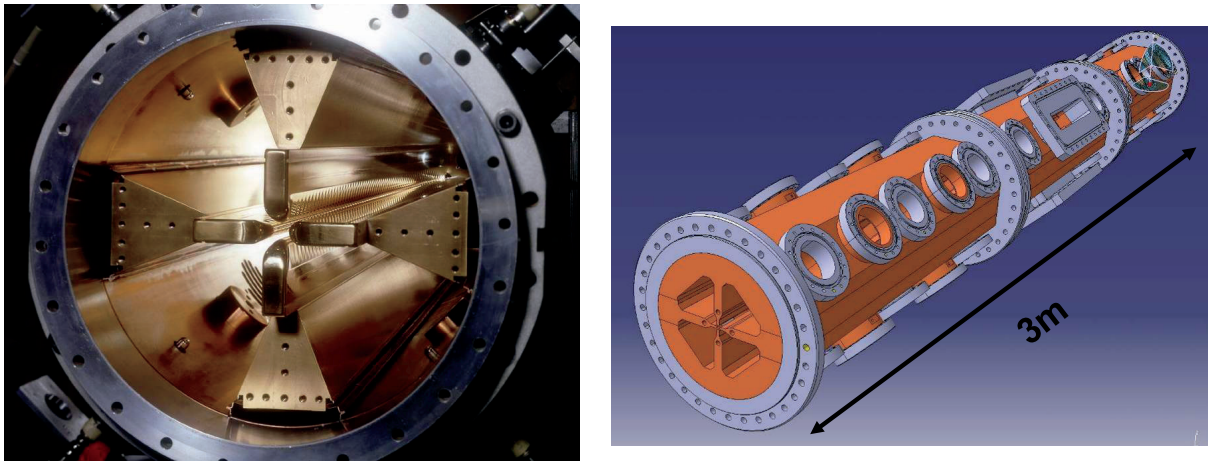


Fig. 4: The CERN RFQ1 (left) and Linac4 RFQ (right)

The generation of the quadrupole electric field requires four electrodes, visible in the left-hand side photograph of Fig. 4, which in this particular type of RFQ are called “vanes”. They are positioned inside a cylindrical tank forming a RF cavity which resonates in a mode that generates a quadrupole RF voltage between the vane tips (Fig. 5). A particle travelling through the channel formed by the four vanes will see a quadrupole electric field with polarity changing with time, at the period of the RF. Every half RF period the particle will see the polarity of the quadrupole reversed, i.e. it will see an alternating gradient focusing channel, with periodicity corresponding to the distance travelled by the particle during half RF period, i.e. $\beta\lambda/2$. The physics of this electric quadrupole channel is the same as for a magnetic focusing channel where the quadrupole gradient is replaced by the RF voltage and the space periodicity is $\beta\lambda/2$.

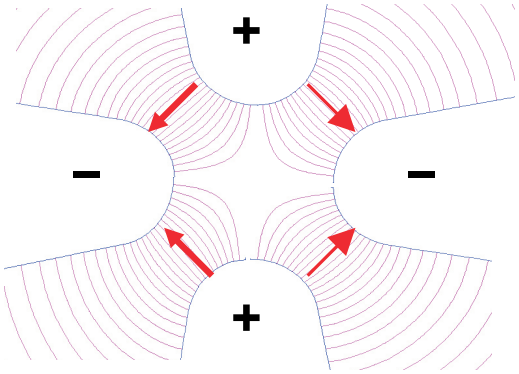


Fig. 5: Voltages and electric fields across RFQ vanes

The longitudinal focusing required for bunching and acceleration is provided by a small longitudinal modulation of the vane tips (barely visible in Fig. 4 left). On the tip of the vanes is machined a sinusoidal profile, with period $\beta\lambda$ (Fig. 6). The important point, necessary to obtain a longitudinal field component, is that on opposite vanes peaks and valleys of the modulation correspond, whereas on adjacent (at 90°) vanes peaks correspond to valleys and vice versa (Fig. 6, with adjacent vanes presented on the same plane for convenience). The arrows in the scheme for adjacent vanes of Fig. 6 represent at a given time the electric field between the two adjacent vanes which have opposite polarity (voltage difference V). On the axis, the electric field vectors can be decomposed into a transverse component, perpendicular to the direction of the beam, and in a small longitudinal component, parallel to the beam direction. The transverse component is constant along the length and represents the focusing field. The longitudinal component instead changes sign (direction) every $\beta\lambda/2$: a particle travelling with velocity β will see an accelerating field (or, in more general terms, the same RF phase) in every cell, exactly as in a standard π mode accelerating structure.

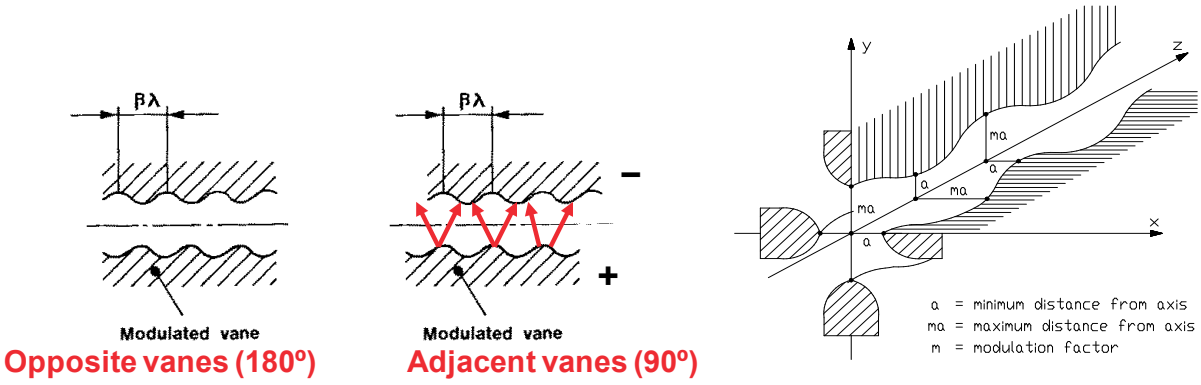


Fig. 6: RFQ vanes, field polarity and modulation parameters

As a result, from the longitudinal point of view a RFQ will be made of a large number of small accelerating cells (β being very small at the beginning of the acceleration), with the additional flexibility with respect to conventional structures that it is possible to change from cell to cell: (a) the amplitude of the modulation and therefore the intensity of the longitudinal electric field; and (b) the length of the cell and therefore the RF phase seen by the beam in its centre. It is then possible to keep the vanes flat in the initial part of a RFQ (no modulation and only focusing) and after a certain length start ramping up slowly the modulation and the longitudinal field. After the first modulated cells, the particle density will start increasing around the phase at which the RF voltage passes through zero and the bunch will be slowly formed. Over many cells, the bunching process can be carefully controlled and made "adiabatic", with the result of capturing a large fraction of the beam inside the RFQ "bucket". When the bunch is formed, the acceleration can start, and the RFQ designer can slowly

modify the cell length to bring the centre of the bunch towards the crest of the RF wave. As an example, Fig. 7 shows the evolution of the longitudinal beam emittance (energy versus phase) in 8 selected cells out of the 126 that make the CERN RFQ2 of Fig. 3 (90 keV to 750 keV): in one of the first modulated cells (top left) the continuous beam coming out of the source sees the first sinusoidal energy modulation; in the following cells (first line) the bunching proceeds until a sufficient density is achieved in the centre (second line) and the acceleration process can start. A few particles are lost in the process, corresponding to the tails visible in the fourth and fifth plots. In the last cells, a bunch is formed, ready for injection into the DTL.

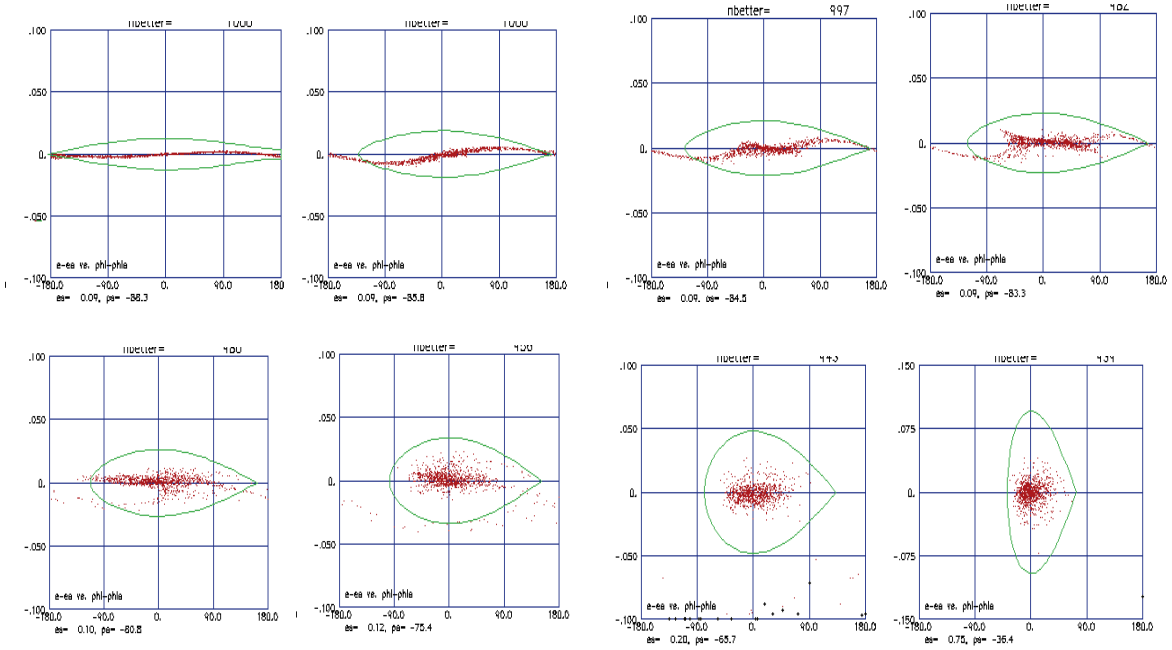


Fig. 7: Evolution of longitudinal emittance along the CERN RFQ2, in 8 representative cells out of 126

Again, it must be observed that in a RFQ only the last cells are devoted to acceleration; a RFQ is mainly a focusing and bunching device. By correctly defining the parameters of the modulation and the RF voltage, the beam dynamics designer is able to match and transport intense beams, at the same time bunching the beam with minimal particle loss. The drawback is that the beam focusing parameters are frozen forever in the beam modulation and cannot be changed during operation; the RFQ is a "one-button" machine, where only the RF voltage can be varied during operation. Its design relies completely on the beam transport codes, and it is not by coincidence that the development of the RFQs has gone in parallel with the development of the modern powerful beam simulation codes which are able to correctly treat the space charge regime.

4 A brief introduction to RFQ beam dynamics

As seen in the previous section, from the point of view of the beam a RFQ is made of a sequence of hundreds of cells with the simplified shape shown in Fig. 8. The dimensions of the region between the electrodes is small compared with the RF wavelength, thus the electric field between the vanes can be calculated in quasi-static approximation and depends only on the geometry of the electrodes. For each RFQ cell, the beam dynamics designer can use three parameters to define the action of the cell on the beam:

1. the aperture *a*, which defines the focusing strength;

2. the modulation factor m, which defines the intensity of the longitudinal field component;
3. the phase φ, which is given by the difference between the ideal modulation period ($\beta\lambda/2$) and the real one, and defines the bunching and/or accelerating action.

These parameters are specific to each cell, and can be changed, although smoothly, between one cell and the next. On top of them, the designer can act on another parameter that is common to all cells (or can be changed in more sophisticated designs, but with limited freedom), the RF voltage V.

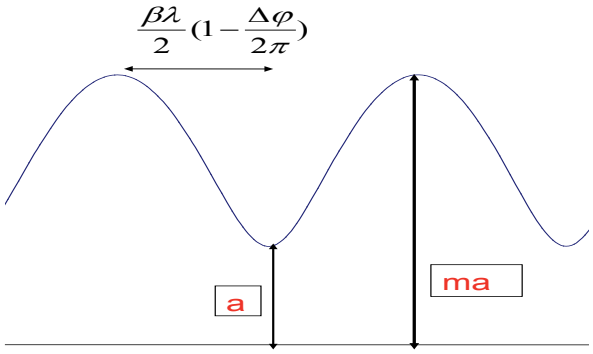


Fig. 8: Parameters of a RFQ cell

Designing a RFQ modulation consists of finding an appropriate set of $(a, m, \phi)_i$ for each cell i together with a voltage V which allows realizing in the minimum possible number of cells the following functions:

- matching of the beam out of the LEBT into the RFQ focusing channel;
- transporting the beam (transversally) with minimum emittance growth;
- bunching the beam with minimum beam loss, generating a longitudinal emittance matched to the acceptance of the following accelerator;
- accelerating the beam from the source extraction energy up to the energy required for injection into the following accelerator;
- for some more modern designs, matching the beam to the following structure using the last cells of the RFQ.

This design is usually done by computer codes. Several programs exist that from a set of input parameters and for a given voltage define the (a, m, ϕ) sequence and calculate the output beam parameters, the first and most famous being PARMTEQ, developed at Los Alamos for the POP RFQ [5]. The experience of the designer remains particularly important, however, in determining the impact of a given set of parameters on the other aspects of the RFQ, its RF and mechanical design and construction. In particular, large voltages increase the focusing but increase as well the risk of voltage breakdowns between the electrodes, too small apertures could lead to unrealistic tolerances in the electrode machining and alignment, etc. Moreover, the designer needs to devote particular care in simulating the beam evolution in presence of realistic error distributions, in particular on the positioning of the electrodes. Often the best design is not the one that gives the best performance (short RFQ, small emittance growth, small beam loss) but the one that is less sensitive to mechanical and RF errors.

Before analysing how the modulation parameters translate into beam dynamics parameters, it is important to consider how the two-dimensional treatment considered so far translates into the fully three-dimensional shape of a real electrode. In particular, we consider the vanes of a "four-vane" RFQ (Fig. 4). The original approach developed by Kapchinsky was purely analytical, at a time when powerful computer codes were not available, and allows a good insight of the RFQ field. The starting

assumption is that in the static approximation that we are allowed to use the potential must be a solution of the Laplace equation, which in cylindrical coordinates can be represented by a series of Bessel functions. The basic Kapchinsky's idea was that of all of the terms in the series of Bessel functions only two were required for a focusing and accelerating beam channel: a transverse quadrupole term and a longitudinal sinusoidal term. In mathematical form, this means that the voltage in cylindrical coordinates has to be written as the sum of the two Bessel components:

$$V(r,\vartheta,z) = A_0 r^2 \cos 2\theta + A_{10} I_0(kr) \cos kz \tag{1}$$

with $k=2\pi/\beta\lambda$. The voltage on the surface of the metallic electrodes must be constant, and this means that the three-dimensional profile of a RFQ electrode must correspond to an equipotential surface of $V(r, \theta, z)$. Such surfaces are hyperbolae in the transverse plane, presenting longitudinally the characteristic sinusoidal modulation. The electrode profile is still defined by the parameters of Fig. 7; a detailed mathematical analysis shows that the constants A_0 and A_{10} can be expressed in terms of the modulation parameters and of modified Bessel functions as

$$A_0 = \frac{V_0}{2a^2} \frac{I_0(ka) + I_0(kma)}{m^2 I_0(ka) + I_0(kma)} \qquad A_{10} = \frac{V_0}{2} \frac{m^2 - 1}{m^2 I_0(ka) + I_0(kma)} \tag{2}$$

To provide a pure quadrupole field the transverse faces of the four electrodes have to follow a hyperbolic shape; however, the electrode cannot extend indefinitely, and the hyperbola has to be truncated at a certain position. In this respect, different configurations are possible. In early RFQ designs as the RFQ1 of Fig. 7, the transverse profile followed precisely the hyperbolic shape up to a few centimetres from the vane tip, introducing only a small negligible deviation from the pure quadrupole potential. In later designs, the mechanical construction has been greatly simplified by either taking a circular cross-section for the vane tips or by even taking as electrode a circular rod instead of a vane. Such mechanical simplifications introduce multipole components that can be calculated by computer codes and whose effect on the beam can be minimized.

A complete treatment of the RFQ beam dynamics can be found in several books and reports [5–7]; here, to understand the main features of the RFQ design it is important to give the main relations that connect the RFQ design parameters (a, m, ϕ) and V with the conventional beam dynamics parameters used in a linear focusing and accelerating channel, transverse focusing coefficient B and longitudinal field E_0T:

$$B = \left(\frac{q}{m_0}\right)\left(\frac{V}{a}\right)\left(\frac{1}{f^2}\right)\frac{1}{a}\left(\frac{I_o(ka) + I_o(mka)}{m^2 I_o(ka) + I_o(mka)}\right) \tag{3}$$

$$E_0T = \frac{m^2 - 1}{m^2 I_o(ka) + I_o(mka)} \cdot V \frac{2}{\beta \cdot \lambda} \frac{\pi}{4} \tag{4}$$

An example of RFQ beam dynamics design is presented in Fig. 9; here are shown the profiles of (a, m, ϕ) along the length of the new Linac4 RFQ at CERN [8].

The Linac4 RFQ operates at 352 MHz frequency, accelerating a 70 mA beam from 45 keV up to 3 MeV energy. It is made of 303 cells for a total length of 3 m. The RF phase seen by the beam at the entrance of the RFQ is –90° (in the linac convention, counted from the crest of the wave), but soon as the bunching process starts the phase is slowly increased to reach –30° after about 120 cm. At that point the modulation factor which is very small at the beginning can be increased, ramping up the longitudinal field and starting the actual acceleration process, which takes place in the second half of the RFQ. Owing to the requirements on the length of this particular RFQ, which could not exceed 3 m for manufacturing reasons, the bunching process takes places relatively quickly and the design beam

transmission is only 95%. A theoretical transmission close to 100% is possible, but at the cost of having a longer RFQ with tight mechanical requirements.

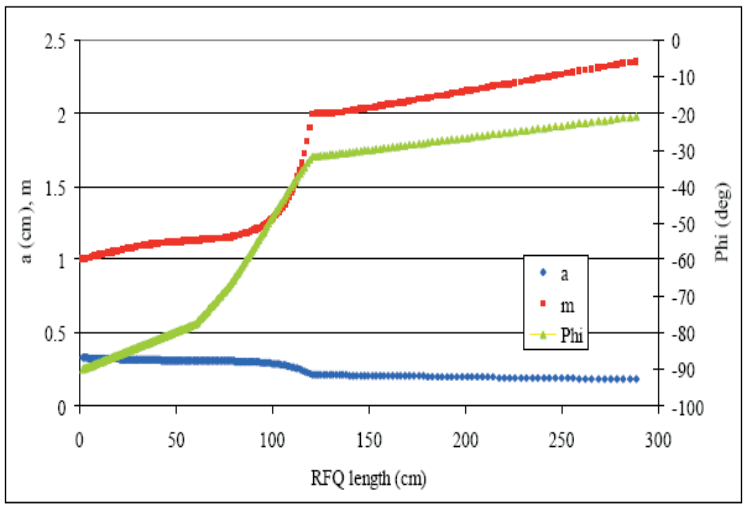


Fig. 9: Modulation parameters along the CERN Linac4 RFQ

5 The RFQ RF resonator

From the RF point of view, the problem of building a RFQ consists in creating a time-varying quadrupole-type electric field between four electrodes, keeping the voltage constant (or following a pre-defined law) along its length. To generate this field, the electrodes must be part of a RF resonator; different resonator types can be used, the most commonly used being the "four-vane" resonator, developed at Los Alamos for the POP RFQ. It can be considered as a cylindrical resonator where is excited the TE210 mode, i.e. a quadrupole mode (mode index 2 in the angular polar coordinate) with only transverse electric field components and constant fields along its length (mode index 0 longitudinally). The TE210 mode of the empty cylinder, whose electric and magnetic field symmetry is shown on the left side of Fig. 10, is transversally "loaded" by the four vanes that concentrate the electric field on the axis (Fig. 9, right). The RFQ will result in cylinder containing the four vanes, which must be connected to the cylinder all along their length.

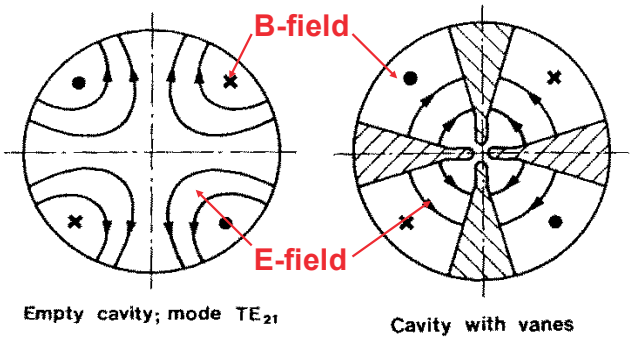


Fig. 10: Four-vane RFQ

The vanes have a twofold effect on the TE210 mode: on the one hand, they concentrate the quadrupole field on the axis, increasing the RF power efficiency of the structure and the focusing term *V/a* in Eq. (3); and, on the other hand, they increase the capacitance for this particular mode, decreasing its frequency well below that of the many other modes of the cylindrical resonator; this separation has a positive effect on the stability of the resonator. Unfortunately, the presence of the

vanes decreases in the same way as the frequency of the TE110 mode, the dipole whose field pattern is shown in Fig. 11: the RFQ resonator will present at a frequency slightly below that of the operating TE210 mode two dipole modes of TE110 type, corresponding to the two orthogonal polarizations of this mode.

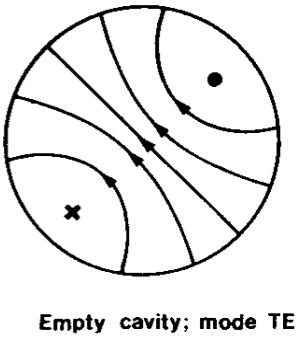


Fig. 11: Dipole modes in a cylindrical cavity

After having connected the vanes to the cylinder, it is important to terminate correctly the resonator at its two ends. Longitudinally, the voltage between the vanes must be constant, meaning that the mode of operation must be a pure TE210. The problem is that in normal conditions the TE210 is forbidden in a closed cylindrical resonator: its electric field is directed transversally to the axis, and on the end discs closing the resonator the electric field would be parallel to a metallic wall. To allow the excitation of this mode, the two end regions of the RFQ must be modified, by cutting an opening at the end of each vane (the vane "undercuts") as shown in Fig. 12. The undercuts allow the magnetic field which goes longitudinally in each quadrant to turn around the vanes and continue in the next quadrants. The ends of the vanes do not touch the covers, but leave a small gap where the turning magnetic field excites an electric field. If the resulting "end cell" is made resonant at the frequency of the TE210 mode, the electromagnetic field of the mode will see an infinitely long RFQ (i.e. will not see the presence of the end cells) and the voltage along the vanes will be constant. It should be mentioned that the correct design of the RFQ end cells requires an extensive use of three-dimensional RF simulation codes.

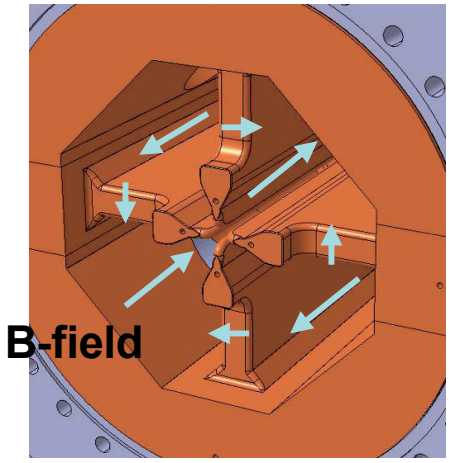


Fig. 12: End-cell of a four-vane RFQ. The arrows show the direction of the magnetic field.

The length of the RFQ, as defined by the beam optics, has an important impact on the number and distribution of high-order modes in the RFQ cavity, which will eventually determine the sensitivity of the RFQ voltage to errors in the positioning of the vanes. In a cylindrical resonator each zero-order mode, such as the TE210 and TE110, gives rise to a family of high-order modes of

increasing frequency, each one characterized by a longitudinal voltage distribution with a number of nodes (transitions through zero) equal to the order of the mode. The presence of the vanes will lower the frequency of all of the modes of the TE21*n* and TE11*n* families, bringing them in a frequency range close to the operating frequency. In particular, the distance between each high-order mode and its zero mode will be inversely proportional to $(l/\lambda)^2$, the square of the ratio between the RFQ length and the RF wavelength of the zero mode, as can be easily derived from waveguide theory. The consequence is that the longer the RFQ the lower will be the spacing between the operating mode and the higher-order modes, opening the possibility of harmful effects on the field stability of the resonator. Although the RFQ operates at fixed frequency on the TE210 mode, the presence of mechanical errors in the machining and/or positioning of the vanes will give rise to field components of the adjacent modes appearing at the operating mode frequency, whose amplitude will be proportional to the mechanical error and inversely proportional to the difference in frequency between operating and perturbing mode. The consequence is that the longer the RFQ, the more stringent will be the mechanical tolerances.

To keep the construction tolerances at a reasonable level, RFQs that are longer than about 2λ are usually equipped with special compensation devices, e.g. tuning volumes inserted inside the quadrants at different longitudinal positions allow the mechanical errors on the vanes to be compensated for by a local variation of the quadrant inductance. For RFQs that are even longer, from about 4λ, the local compensation is not sufficient and a stabilization scheme is usually implemented, under the form of a resonant or non-resonant device mounted inside the RFQ which moves the frequency of the perturbing modes away from the operating mode. As an example of a long RFQ using only compensation schemes, Fig. 13 shows the measured mode spectrum of a 425 MHz four-vane RFQ, 2.75 m long. For this RFQ, l/λ=3.9: the zero quadrupole (TE210) is surrounded by a large number of modes, with as many as three dipole modes (TE110, TE111 and TE112) at frequencies lower than the operating frequency. Each dipole mode has two polarizations, corresponding to orthogonal orientations of the electric field (see Fig. 11). These can have slightly different frequencies, each one generating its own high-order band; the notations 1-3 and 3-4 in Fig. 13 refer to the polarizations corresponding to field concentrated in pairs of opposite quadrants.

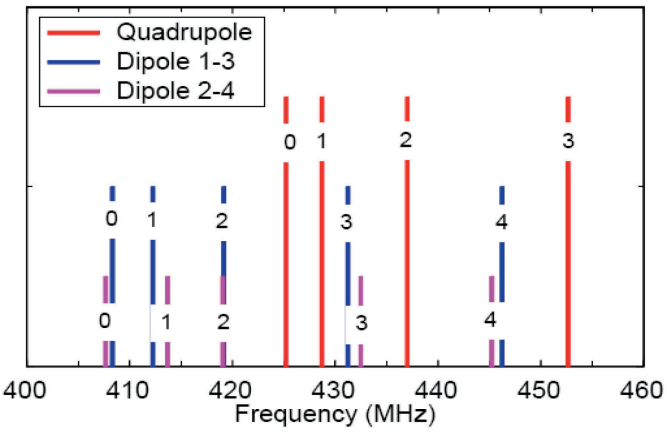


Fig. 13: Measured mode spectrum of a 425 MHz four-vane RFQ, 2.75 m long. The notations 1-3 and 3-4 refer to pairs of opposite quadrants. Excitation was in quadrant 1. Operating mode is quadrupole 0.

The high sensitivity to errors of the RFQ resonator coming from the presence of perturbing modes has to be correctly taken into account in the design, construction and tuning of the RFQ. As a first step, the length of the RFQ and the design of the end terminations have to be chosen in such a way as to avoid having dipole modes too close to the operating quadrupole mode. In the case of a RFQ

with compensation scheme, as are most of the existing RFQs, an extensive series of voltage measurements is required after construction and assembly. The measurements are then entered into an algorithm that allows the correct dimensioning of the compensation devices: this is the so-called "tuning" of the RFQ, which on top of bringing the quadrupole frequency at the required design value aims at achieving a flat (or following a predefined law) voltage along the RFQ, equal in the four quadrants. Accurate field measurements in the RFQ can be performed via "bead-pull" techniques, where a perturbing metallic bead on a plastic wire is slowly moved inside the four quadrants, to register the frequency shift which is proportional to the square of the local field.

To reduce the error sensitivity of the RFQ field, resonator designs alternative to the four-vane have been developed and are in use in many laboratories; of these, the most widely used is the so-called "four-rod" RFQ (Fig. 14) originally developed by A. Schempp at the IAP of Frankfurt University [9].

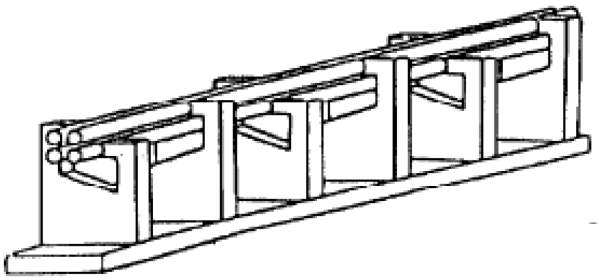


Fig. 14: Four-rod RFQ

In this device, the four electrodes are either circular rods with a modulated diameter or small rectangular bars with a modulated profile on one side; they are connected to an array of quarter-wavelength parallel plate transmission lines generating a voltage difference between the two plates (Fig. 15). Opposite pairs of electrodes are connected to the two plates of a line, resulting in a quadrupole voltage being generated between the rods. Several quarter-wavelength cells are used to cover the required RFQ length; their magnetic field couples from each cell to the next, forming a single long resonator. This "open" RFQ structure is then placed inside a tank, which forms the vacuum and RF envelope of the structure.

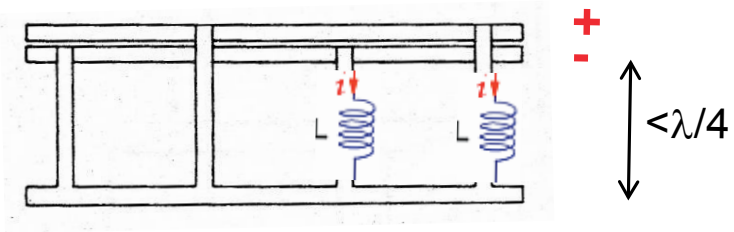


Fig. 15: The four-rod RFQ structure

RFQs of this type are free from dipole modes; however, in the standard single-support structure of Fig. 14 the transverse slope in the cut on the plate needs to be carefully defined to ensure that the voltage on two pairs of opposite rods is the same, to compensate for their different distances from the bottom plate. In other RFQ designs, such as that shown in Fig. 16, a double plate supports the electrodes, to completely eliminate dipole components. From the point of view of the longitudinal modes a four-rod RFQ is no different to a four-vane RFQ: for long RFQ structures an error compensation is required, achieved by placing short-circuiting plates or metallic volumes inside some cells to reduce the transmission line length.

The main advantages of the four-rod RFQ are the absence of dipole modes (which reduces the sensitivity to mechanical errors and simplifies the tuning), the reduced transverse dimensions as compared with the four-vane RFQ, and the simple and easy to access construction. These advantages are particularly evident for the low-frequency RFQs (up to about 100 MHz) used for heavy ions. For the higher frequencies required for protons, from about 200 MHz, the transverse dimensions of the four-rod RFQ become very small and the current and power densities reach high values in some parts of the resonator, in particular at the critical connection between the rods and the supports; cooling can be difficult, in particular for RFQs operating at high duty cycle, with the risk of excessive deformations of the rods and reduced beam transmission. For these reasons, RFQs operating at frequencies above 200 MHz or at high duty cycle are usually of the four-vane type.

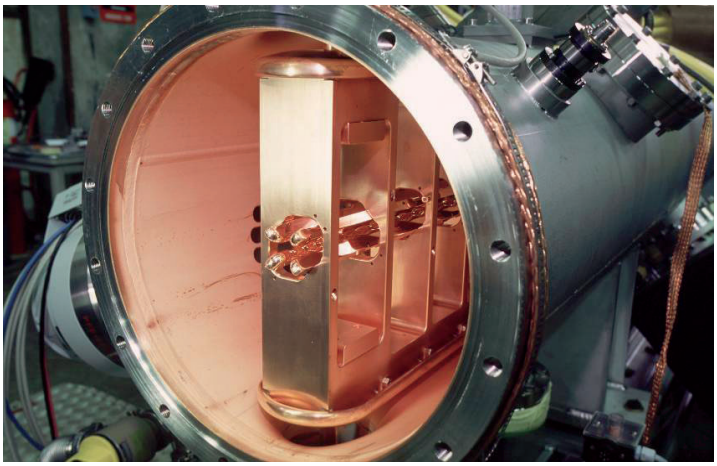


Fig. 16: The antiproton decelerating RFQ at CERN (202 MHz)

6 Mechanical construction

The mechanical design and construction of a RFQ is a challenge in itself, for two main reasons: first of all, tight tolerances in the machining and positioning of the electrodes need to be achieved and maintained during operation; and, second, because many different mechanical parts need to be joined together respecting the tolerances and providing an excellent electrical and thermal contact, to avoid excessive RF power consumption and/or overheating. On top of that, the mechanical structure has to provide sufficient access points for RF tuning and for vacuum pumping.

Usual beam dynamics tolerances in the machining and positioning of the RFQ electrodes are of the order of few tens of micrometres, a value that corresponds to about 1 % of the minimum radius of the beam channel (a in Fig. 7). For larger errors, multipoles appear in Eq. (1) resulting in a perturbation of the beam optics and in increasing beam loss in the RFQ. In four-vane RFQs the RF can generate additional dipole and higher-order mode components again proportional to the errors in the positioning of the vanes through complex RF-related algorithms. RF-related errors can be compensated for by the compensation system (tuners or other); nevertheless, the maximum permitted vane positioning errors define the size of the compensation system (number and dimension of the tuners, for example). The compromise usually adopted in this type of RFQs is to define for the RF a maximum error smaller or of the same level as the beam dynamics one, and then dimension consequently the compensation system. To give an example, Table 1 reports the permitted error budget of the CERN Linac4 RFQ, defined after a series of beam dynamics calculations in presence of random errors. The RF compensation system (tuners) is dimensioned to fully absorb the mechanical errors, leaving a residual field error of ±1 %. The electrode gap represents the distance between the vane ends at the connection between the three RFQ segments.

Table 1: Error budget of the CERN Linac4 RFQ

Linac4 RFQ Mechanical Tolerances	**Value**	**Units**
Machining error	± 20	μm
Vane modulation error	± 20	μm
Vane tilt over 1 m	± 100	μm
Vane positioning error (displacement h+v)	± 30	μm
Vane thickness error	± 10	μm
Gap between vanes (contiguous modules)	100 ± 15	μm
Section tilt over 1 m	± 30	μm
Electromagnetic field error	± 1	%

Joining the different parts is the next problem to be faced; RFQs tend to use a large variety of joining techniques, including brazing, electron-beam welding, TIG welding and simple bolting of parts using different types of gaskets and contacts. Four-rod RFQs are usually made out of parts bolted together; low- and medium-frequency four-vane RFQs are bolted or welded, and high-frequency four-vane RFQs are usually made of brazed copper elements following the scheme of Fig. 17. The RFQ is divided into longitudinal segments of about 1 m length (Fig. 4, right); the segments are composed of four copper elements brazed together, each made of a vane and of a part of the external tank to minimize the brazing surface. Cooling channels are machined inside the copper; the brazing ensures the vacuum tightness of the structure, providing at the same time a high thermal and electrical conductivity.

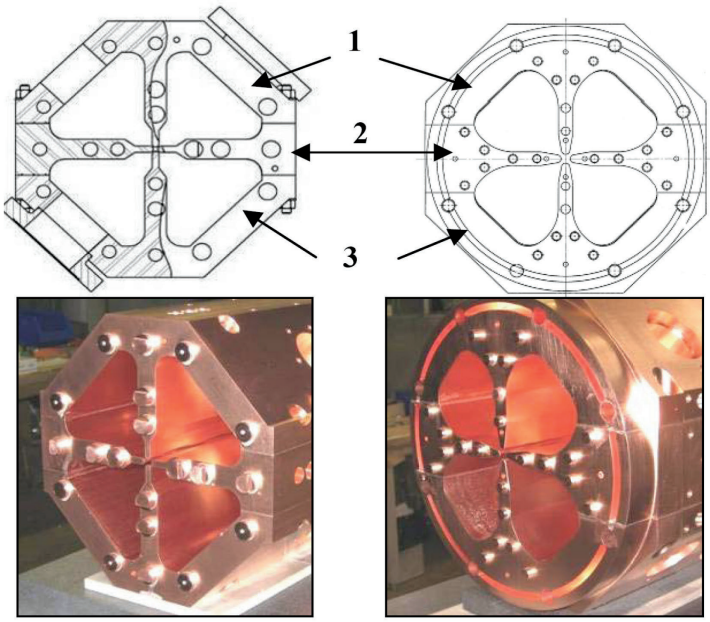


Fig. 17: Construction scheme of two European 352 MHz continuous wave RFQs: TRASCO of INFN, Italy [10] (left) and IPHI of CEA, France [11] (right)

An important requirement is that the precise alignment of the electrodes does not change when the structure is heated by the RF power dissipated on the walls and supports; this is particularly demanding for high duty cycle or continuous wave RFQs, where the power to dissipate can be of the order of 1 kW/cm. The number size and position of the water cooling channels need to be carefully dimensioned; the corresponding deformations have to be calculated and translated into beam dynamics and RF errors. In addition, a precise control of the water temperature is required, at the level of 0.1°. As an example, Fig. 18 summarizes the thermal studies performed for the design of the TRASCO RFQ of INFN. Here the deformations are translated into frequency errors in the individual RFQ quadrants.

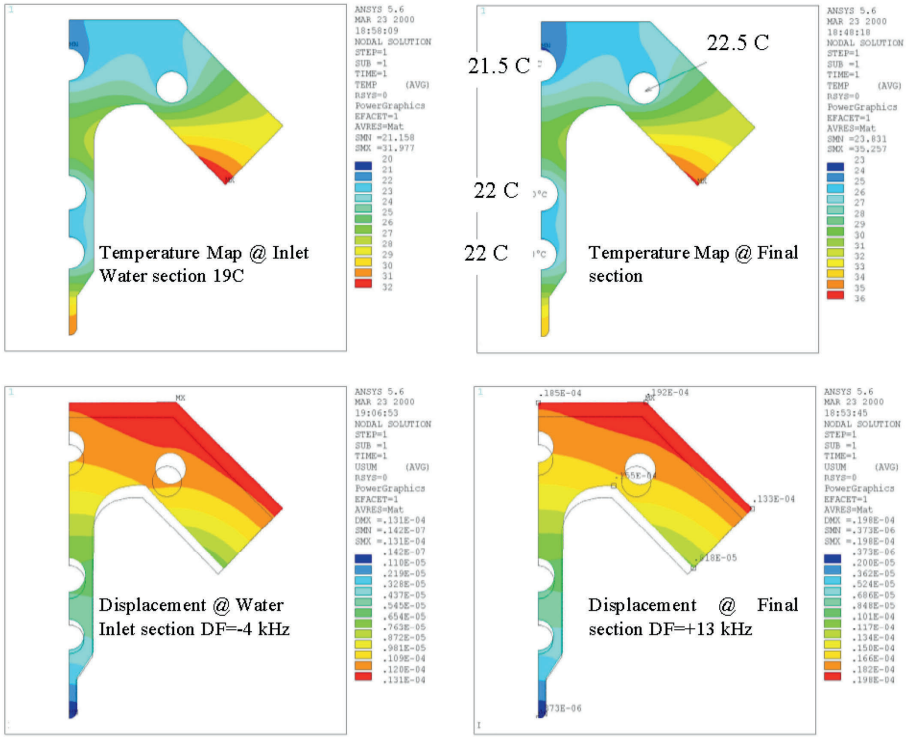


Fig. 18: Calculated temperature (top) and displacement (bottom) distributions in the TRASCO RFQ at the beginning (left) and at the end (right) of a section [10]

7 Putting it all together

In the previous sections the main aspects of RFQ theory and construction practice have been presented. What remains to be underlined now is that a RFQ is a highly multidisciplinary object, whose performance relies on a complex equilibrium between three fundamental disciplines, beam dynamics, electrodynamics and accelerator mechanics, and with important inputs from other aspects of accelerator technology such as vacuum, RF power production, survey, etc. On the one hand, mechanical errors or deformations and deviations from the ideal RF field distribution would result immediately in reductions of the RFQ beam transmission; on the other hand, an excessively demanding tolerance budget would increase the complexity of the construction leading to unnecessary challenges and sky-scraping costs. On top of that, the performance of a RFQ depends critically on its input beam parameters and therefore on the performance of the ion source and LEBT: any deviation from the design emittance or error in the input beam alignment result again in a reduction of the RFQ transmission.

The real challenge of building a RFQ is not in each single aspects of its design, but lies in putting it all together: if teamwork and good communication between the different competencies

required to build an accelerator are nowadays crucial to any project, this is even more true for RFQs where the different aspects are closely interrelated. The fact that often projects are based on international or inter-laboratory collaborations and/or rely heavily on industrial partners adds another degree of complexity that needs to be correctly managed.

If the construction of a RFQ represents a challenge, it can also be extremely rewarding. RFQs in particular for extreme parameters tend to be difficult to design and construct, but once they have been commissioned they tend to be very reliable (as far as their thermal equilibrium is not altered), operating steadily and without need for adjustment for several years: after all, they are "one-button" accelerators.

Acknowledgements

The preparation of this lecture has profited from the support and advice of A.M. Lombardi, A. Pisent, C. Rossi and J. Stovall. To all of them goes my gratitude.

References

[1] E. Boltezar, *et al*., The new CERN 50-MeV Linac, Proc. of the 1979 Linac Conference.

[2] I.M. Kapchinsky and V.A. Tepliakov, *Prib. Tekh. Eksp.* **2** (1970) 19-22.

[3] R.W. Hamm, *et al*., Proc. of the International Conference on Low Energy Ion Beams 2, University of Bath, April 1980, p. 54.

[4] E. Tanke, M. Vretenar and M. Weiss, Measurement of the CERN high intensity RFQ, Proc. of the European Particle Accelerator Conference, Berlin, 1992.

[5] K.R. Crandall, R.H. Stokes and T.P. Wangler, RF quadrupole beam dynamics design studies, Proc. of the 1979 Linear Accelerator Conference, Montauk, NY.

[6] T. Wangler, Principles of RF linear accelerators (Wiley, New York, 1998), p. 225.

[7] M. Weiss, Radio frequency quadrupole, Proc. of the 1986 CAS School, Aarhus.

[8] C. Rossi, *et al*., The radiofrequency quadrupole accelerator for the CERN Linac4, Proc. of the 2008 Linac Conference, Victoria.

[9] A. Schempp, H. Deitinghoff, M. Ferch, P. Junior and H. Klein, λ/2 RFQ for light ion acceleration, IAP Report, 1984.

[10] A. Pisent, M. Comunian, A. Palmieri, G.V. Lamanna and D. Barni, TRASCO RFQ, Proc. of Linac 2000, Monterey, CA.

[11] R. Ferdinand and P.Y. Beauvais, *AIP Conf. Proc.* **773** (2005) 84.

Linear accelerators

Maurizio Vretenar
CERN, Geneva, Switzerland

Abstract
Radio-frequency linear accelerators are used as injectors for synchrotrons and as stand-alone accelerators for the production of intense particle beams, thanks to their ability to accelerate high beam currents at high repetition rates. This lecture introduces their main features, reviewing the different types of accelerating structures used in linacs and presenting the main characteristics of linac beam dynamics. Building on these bases, the architecture of modern proton linear accelerators is presented with a particular emphasis on high-energy and high-beam-power applications.

1 Introduction, general features

A *linear accelerator* (linac) is a device where charged particles acquire energy moving on a linear path; the characteristic feature of a linac is that the particles pass only once through each of its accelerating structures.[1] In the following, we limit our analysis to radio-frequency (RF) linacs where the acceleration is provided by time-varying electric fields, leaving out of our treatment *electrostatic linacs* that are usually limited to energies outside of the scope of this lecture. Moreover, we concentrate on linacs for high-beam-power applications and therefore exclude the large and important domain of *electron linacs*.

A linac will take the continuous particle beam coming out of an ion source, bunch it a given RF frequency and then accelerate it up to the required final energy. In general, linacs are *pulsed* accelerators: the beam is generated by the source and then delivered to the users in pulses of a given length τ (between few microseconds and few milliseconds) at a given repetition frequency f (usually between 1 Hz and 100 Hz). The product of pulse length and repetition frequency is the *duty cycle* (or *beam duty cycle*, to distinguish it from the *RF duty cycle* which is always higher). A linac can as well operate continuously, producing a constant stream of particles: in this case the duty cycle is 100%, and we call it a continuous wave (CW) linac.

Together with the *kinetic energy E* of the particles coming out of the linac, its most important parameter is the beam current *I*, defined as the *average current during the beam pulse*. The current *I* is different from the average current out of the linac, which is *I* times the duty cycle; it can also be different from the *bunch current*, the average current during a RF period, in the particular case where not all bunches are populated by particles.

The *beam power P* defines the electrical power (measured in Watts) transferred to the particle beam during the acceleration process. It represents the sum of the electrical power absorbed by the beam in the different accelerating cavities that constitute the accelerator. In each cavity, the current *I* crosses a voltage *V*; the RF power going to the beam is $P = V \times I$, considering the beam current as

[1] It should be mentioned that linacs do not need to be straight; some heavy ion linacs for example incorporate 90° bends to reduce the footprint of the machine. In a similar way, some electron linac designs can include curved sections and even pass the beam two or more times through the same accelerating structures; these are the "*recirculating linacs*".

constant during the pulse at the value I and taking for V the voltage actually seen by the beam, corrected with the transit time factor of the particles in the cavity gap.[2]

Under these assumptions, the overall voltage V_{tot}, the sum of the effective voltage seen by the beam in all the RF cavities, is (numerically) equal to E, the total beam energy expressed in electronvolts, thus the total beam power produced by a linac is

$$P\,[\mathrm{W}] = E\,[\mathrm{eV}] \times I\,[\mathrm{A}] \times \text{duty cycle}$$

The beam power is important because on the one hand it represents the amount of power that the RF system has to deliver to the beam, and on the other hand it corresponds to a simple "figure of merit" for linacs dedicated to the production of secondary particles (neutrons, pions, etc.). Above a specific energy threshold, the number of secondary particles produced by the target is directly proportional to the primary beam power (i.e. to the product of beam intensity and energy).

For the production of high-power beams, linear accelerators have many advantages as compared with other types of accelerators. The repetition frequency of a linac is not limited by the rise time of the magnets as is the case for synchrotrons. It can therefore reach very high values, going up to CW operation; for example, several high-power linacs adopt as basic repetition frequency that of the mains supply, 50 Hz in Europe and 60 Hz in the US. Moreover, the fact that the beam passes only once in each section of a linac limits the effect on the beam of magnetic field errors: beam resonances that constitute the main limitation of synchrotrons in achieving high beam currents have only a minor impact on linear accelerators. Cyclotrons are often in competition with linacs for the production of low-energy CW proton and ion beams; however, their advantages in terms of compactness and cost vanish for energies above the relativistic limit where special designs are required. Linear accelerators can easily reach energies in the GeV range, although their relatively high cost per MeV of acceleration suggests the transition to a synchrotron for very high beam energies when the synchrotron repetition frequency is compatible with the beam parameter of the project. It should be observed that the optimum transition energy between a linac and a rapid cycling synchrotrons (RCS), the special type of synchrotron required for high beam power, is very difficult to define and depends on the specific beam parameters as well as on the experience of the particular laboratory.

The role of linacs with respect to cyclotrons and synchrotrons can be easily visualized considering the proton velocity curve of Fig. 1, where β = v/c is plotted as a function of kinetic energy.

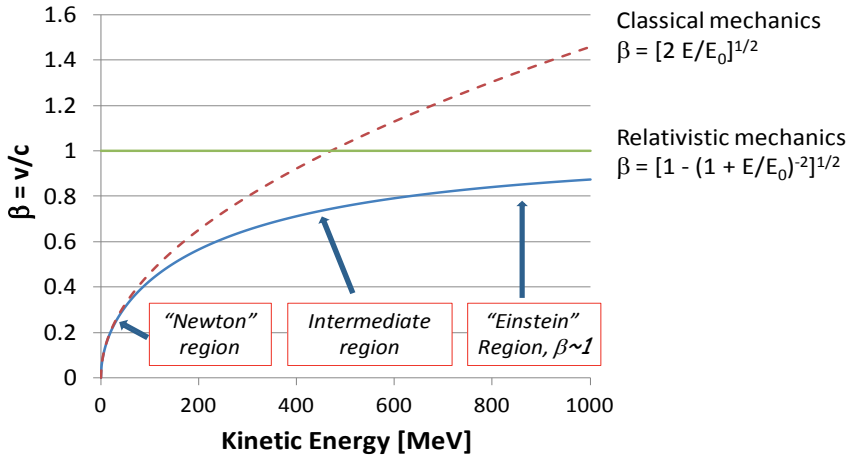


Fig. 1: Relativistic velocity as a function of kinetic energy (protons)

[2] Actually, the calculation is a bit more complicated: the voltage is sinusoidal, and only the Fourier component of the beam current at the RF frequency absorbs power following the usual formula $P = \frac{1}{2}V \times i_{RF}$; but for bunches that are short with respect to the RF period, the amplitude of the main harmonic component is $i_{RF} = 2I$, giving the well-known relation $P = V \times I$.

At the beginning of acceleration the particle velocity follows the classical square root relation, but already at proton energies of a few tens of megaelectronvolts the velocity starts to be lower than that predicted by classical mechanics; the beam is now in the domain of relativistic mechanics. From the gigaelectronvolt range, the velocity increases very slowly and starts approaching asymptotically the speed of light: the energy that our accelerator delivers to the beam goes into increasing the mass of the particles instead of increasing their velocity.

As we will see in the following, a linear accelerator is made of cells of variable length. The main feature of linacs is that by adapting the cell lengths to the increase in particle velocity one can synchronize the acceleration with the selected RF frequency. The procedure consists of defining the cell lengths in such a way that the time needed by the beam to cross a cell remains constant; the result is that the cell length will progressively increase with the energy. A linear accelerator can adapt to any beam velocity: it can cover at the same time the "Newton" range of energies, where the velocity increases with the energy, as well as the "Einstein" range, where the velocity remains approximately constant. A synchrotron instead is a typical "Einstein" machine, designed for a fixed revolution frequency and for a beam of almost constant velocity. Actually, all synchrotrons allow for a modulation of the RF frequency which permits the acceleration of beams of increasing velocity; however, the modulation range is usually small and large modulations require special and expensive RF cavities. For this reason, a linac is always used as an injector to a synchrotron, covering the energy range where the velocity variation is large while the synchrotron takes over when the variation becomes smaller. The "intermediate" region, between the Newton and Einstein regimes, can be covered by both linacs and synchrotrons: if high beam intensity and/or high beam quality are required, the high repetition frequency and the absence of resonances give an advantage to linacs, while if the objective consists in reaching a high-energy synchrotrons have an advantage because their cost per unit of acceleration is lower than that of linacs. Cyclotrons instead are special accelerators where the beam remains synchronous with the RF only in the "Newton" range. Their compactness and reduced cost give them an advantage with respect to linacs as a stand-alone machine for producing non-relativistic CW beams; to achieve higher energies cyclotrons require special technologies that largely reduce their attractiveness.

As we have mentioned, a linear accelerator is made of a sequence of accelerating gaps as in Fig. 2, where for convenience each gap is associated with an individual RF cavity.

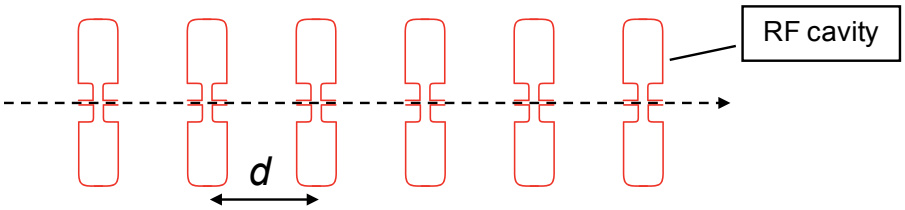


Fig. 2: Sequence of accelerating gaps/cavities

If we want to accelerate a beam of increasing velocity we can immediately observe that: (a) *the beam must be already bunched at the RF frequency when it enters the sequence*; and (b) *for a beam of increasing velocity, the distance between the cavities and their relative RF phase must be correlated.* As usual, the electric field on the gap of cavity *i* can be written as

$$E_i = E_{0i}\cos(\omega t + \varphi_i)$$

with φ_i the phase of the *i*th cavity with respect to a "reference" RF phase. To maximize acceleration, the beam has to cross the gap of each cavity at a phase φ_i on or very close to the crest of the wave ($\varphi_i = 0$); moreover, it must have a short length in time and in phase, i.e. it must be "bunched". During the time that the particles need to go from one cavity to the next the phase has changed by an amount

$\Delta\varphi = \omega\tau$ with τ the time to cross the distance d; for a particle of relativistic velocity $\beta = v/c$ the change in phase will be

$$\Delta\Phi = \omega\tau = \omega\frac{d}{\beta c} = 2\pi\frac{d}{\beta\lambda}$$

Here d is the distance between gaps and λ the RF wavelength. This means that

$$\frac{\Delta\Phi}{d} = \frac{2\pi}{\beta\lambda}$$

or, for the acceleration to take place, the *distance* and the *phase difference* between two gaps in the sequence must be correlated, their ratio being proportional to $\beta\lambda$. At every gap crossing the particle will gain some energy and its velocity will increase: in a non-relativistic regime this means that either the relative phase $\Delta\Phi$ or the distance d has to change during acceleration. In other terms, in a linear accelerator we need either to progressively increase the distance between cavities or to progressively decrease their RF phase (relative to a common reference) to keep synchronicity between the particle beam and the accelerating wave.

This requirement corresponds to two well-defined types of linear accelerators:

1. "Single-cavity" linacs (Fig. 3, top), where the *distance* between cavities is fixed, and the phase of each cavity is individually adjusted to take into account the increase in beam velocity; each cavity has to be connected to an individual RF amplifier. This scheme has the advantage of maximum flexibility, being able to accelerate different ions and/or charge states at different energies by entering a different set of phases for each ion, but has the drawback of a high cost.
2. "Coupled-cell cavity" linacs (Fig. 3, bottom), where the *phase* at each cavity/gap is fixed, and the distance changes accordingly to the beam velocity. When the RF phase is the same for each gap, several gaps can be coupled into a common RF structure and connected to a single RF power source, reducing significantly the cost of the RF system. The single gaps and resonators have to be coupled together, however, to allow fixing their relative phase (this will be the subject of the next section). This scheme is cost effective but not at all flexible: the physical distance between gaps is defined for a given energy increase, i.e. for a given particle, energy range and acceleration gradient.

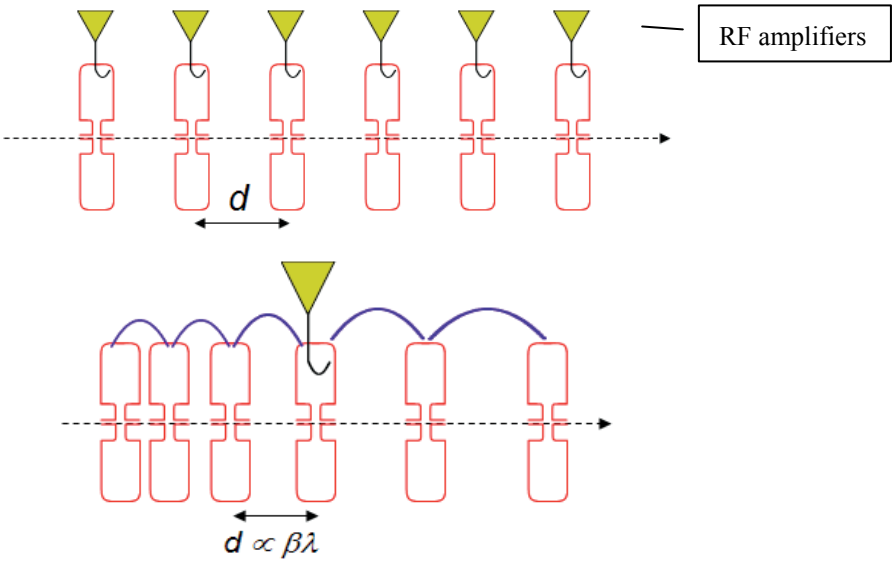


Fig. 3: Single-cavity and coupled-cell-cavity linacs

The reality of linac structures is not so antithetic, however, and most of the commonly used linacs tend to compromise between these two extreme configurations. This is particularly true at high energies: if the increase in β between two gaps is small, the phase error in the case where the gaps are equally spaced will be small and under certain circumstances acceptable. This is the case of multicell superconducting cavities such as that in Fig. 4, commonly used in the high-energy section of linacs. The structure of Fig. 4 is made of four gaps coupled together in a single resonator. At a given time, the longitudinal electric field on the axis of the cavity will have the profile shown in the figure and indicated by the arrows in the gaps: this corresponds to a constant phase difference $\Delta\Phi = 180°$ between adjacent gaps. For a particle to be accelerated, the relation between phase and gap distance must hold, with $\Delta\Phi = \pi$; solving for d, we find that the distance between gaps must be equal to $\beta\lambda/2$. This condition is respected only for one particular β, however; inside the cavity the energy and the beta are increasing, leading to a difference in the phase of the RF seen by the beam in the different cells that is called "phase slippage". For this structure to accelerate effectively, the phase slippage must be small as compared with the synchronous phase φ_i, i.e. the increase in β within the cavity must be small.

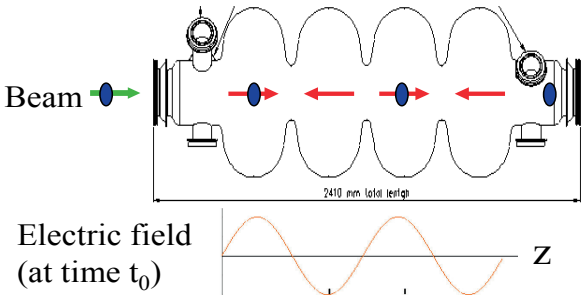


Fig. 4: Multicell superconducting cavity

Before going into the study of the different types of accelerating structures, it is important to consider that a linac is much more complex than a simple sequence of RF gaps. First of all, the sequence of gaps, usually grouped inside a RF cavity, has to be preceded by an ion source and by a bunching system[3]. Then, the gaps have to be spaced by some focusing elements, usually quadrupoles, required to keep together the particles that constitute the bunch. The RF cavity needs to be fed by a RF amplifier inserted into a feedback loop and the system requires some ancillaries to work: a vacuum system, a magnet powering system and a water cooling system to evacuate the excess power provided by the RF system. A basic block diagram of a linac system is presented in Fig. 5.

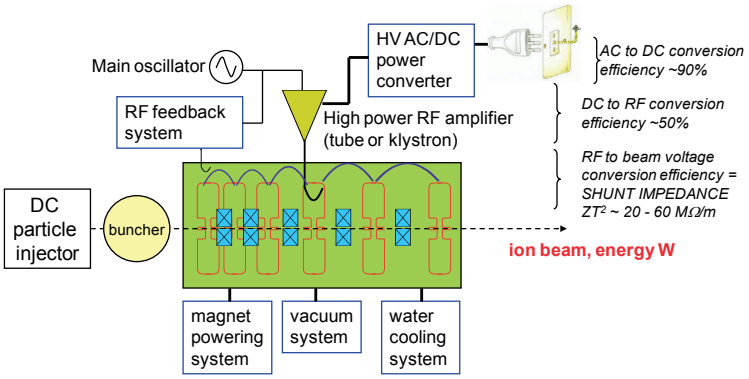


Fig. 5: Scheme of a linac accelerating system

[3] The different bunching mechanisms are described in the lecture on RFQs in these proceedings.

The representation of Fig. 5 visualizes the fact that a linac RF system transforms electrical energy taken from the grid into energy transferred to a beam of particles. The corresponding power is the beam power introduced at the beginning of this lecture. The efficiency of this transformation is usually quite low and a large fraction of the input energy is dissipated into heat released into the surrounding environment. More precisely, the energy transformation takes place in three different steps, each one characterized by a different technology and a different efficiency:

(i) The transformation of the AC power from the grid (alternate, low voltage, high current) in DC power (continuous, high voltage, low current) takes place in a power converter. Pulsed power converters are usually called "modulators"; their efficiency is usually high, of the order of 80 % to 90 %.

(ii) The following transformation of the DC power into RF power (high frequency, high voltage, low current) takes place in a RF active element: RF tube, klystron, transistor, etc.; the efficiency is the RF conversion efficiency that depends on the specific device and on its class of operation. Typical RF efficiencies are in the 50 % to 60 % range.

(iii) The final transformation of the RF power into power stored into the particle beam takes place in the gap of an accelerating cavity; the efficiency is proportional to the shunt impedance of the cavity, which represents the efficiency of the gap in converting RF power into voltage available for a beam crossing the cavity at a given velocity (see Section 6).

2 Accelerating structures for linacs

Coupled-cell cavities are the most widely used accelerating structures for linacs. To couple the elements of a chain of single-gap resonators (that from now on we will often refer to as the "cells" of our system) we need to allow some energy to flow from one cell to the next, via an aperture that permits leaking of some field (electric or magnetic) into the adjacent resonator. There will be two different types of coupling, depending on whether the opening connects regions of high magnetic field ("magnetic coupling") or regions of high electric field ("electric coupling"). The simplest magnetic coupling is obtained by opening a slot on the outer contour of the cell, whereas an electric coupling can be obtained by enlarging the beam hole until some electric field lines couple from one cell to the next. Once the cells are coupled, to find the conditions for acceleration we have to calculate the relative RF phase of the individual cells.

The simplest way to analyse the behaviour of a chain of coupled oscillators is to consider their equivalent circuits (Fig. 6) [1].

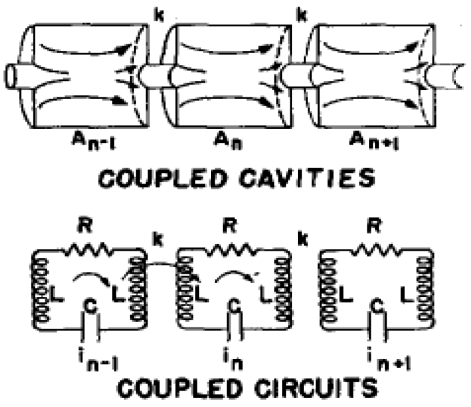


Fig. 6: From coupled cavities to coupled resonant electrical circuits (from Ref. [1])

Each coupled cavity can be represented by a standard *RLC* resonant circuit; for convenience, in Fig. 6 the inductance of each circuit is split into two separated inductances *L*. The advantage of this representation is that we can describe the (magnetic) coupling between adjacent cells as a mutual inductance *M* between two inductances *L*, which is related to a coupling factor *k* by the usual relation $M = kL$. For the series resonant circuits of Fig. 6, the behaviour of each cell is described by its circulating current I_i. The equation for the *i*th circuit can be written taking equal to zero the sum of the voltages across the different elements of the circuit (Kirchhoff's law), considering for simplicity a lossless system with $R = 0$:

$$I_i\left(2j\omega L + \frac{1}{j\omega C}\right) + j\omega kL(I_{i-1} + I_{i+1}) = 0$$

Dividing both terms of this equation by $2j\omega L$, it can be written as

$$X_i\left(1 - \frac{\omega_0^2}{\omega^2}\right) + \frac{k}{2}(X_{i-1} + X_{i+1}) = 0$$

This equation relates general excitation terms of the form $X_i = I_i / 2j\omega L$, proportional to the square root of the energy stored in the cell *i*, with the coupling factor *k* and with a standard resonance term $(1\text{-}\omega_0^2/\omega^2)$. We consider that all cells are identical, i.e. that they have the same resonance frequency $\omega_0^2 = 1/2LC$. If our system is composed of $N + 1$ cells, assuming $i = 0,1,...,N$ we can write a system of $N + 1$ equations with $N + 1$ unknowns X_i represented by the following matrix equation:

$$\begin{bmatrix} 1-\frac{\omega_0^2}{\omega^2} & \frac{k}{2} & 0 & & 0 \\ \frac{k}{2} & 1-\frac{\omega_0^2}{\omega^2} & \frac{k}{2} & \cdots & 0 \\ 0 & \frac{k}{2} & 1-\frac{\omega_0^2}{\omega^2} & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1-\frac{\omega_0^2}{\omega^2} \end{bmatrix} \begin{vmatrix} X_0 \\ X_1 \\ \cdots \\ X_N \end{vmatrix} = 0$$

or

$$MX = 0$$

The matrix *M* has elements different from zero only on three diagonals, the main one with the resonance terms and the two adjacent ones with the coupling terms. It is perfectly symmetric because we have introduced an additional simplification, closing at both ends our chain of resonators with "half cells", presenting a coupling *k* only on one side and with half the inductance and twice the capacitance of a standard cell (but the same resonant frequency). This corresponds to a physical case where the end resonators are terminated by a conducting wall passing at the centre of the gap, i.e. they are exactly one half of a standard cell. The advantage of this approach is that the matrix and the relative solutions are symmetric and lead to a simple analytical result. In the real case, the chain of resonators is terminated with full cells that need to be tuned to a slightly different frequency to symmetrise the system.

The above matrix equation represents a standard eigenvalue problem, which has solutions only for those ω giving

$$\det M = 0$$

The eigenvalue equation $\det M = 0$ is an equation of $(N + 1)$th order in ω. Its $N + 1$ solutions ω_q are the eigenvalues of the problem, which are the resonance modes of the coupled system. Whereas the individual resonators can oscillate only at the frequency ω_0, the coupled system will have $N + 1$ frequencies ω_q, the "modes", with $q = 0,1,...,N$ the index of the mode. To each mode corresponds a

solution in the form of a set of $[X_i]_q$, which is the corresponding eigenvector. It is important to observe that the number of modes is always equal to the number of cells in the system.

For the matrix M, we can find an analytical expression for the eigenvalues (mode frequencies):

$$\omega_q^2 = \frac{\omega_0^2}{1 + k\cos\frac{\pi q}{N}} \quad q = 0, \ldots, N \tag{1}$$

or, for $k << 1$, which is the operating condition for most of coupled structures commonly used in linacs where it is usually $k \sim 1$ % to 5 %:

$$\omega_q \approx \omega_0\left(1 - \frac{1}{2}k\cos\frac{\pi q}{N}\right) \qquad q = 0, \ldots, N$$

The corresponding eigenvectors (modes) are

$$X_i^{(q)} = (const)\cos\frac{\pi q i}{N} e^{j\omega_q t} \qquad q = 0, \ldots, N \tag{2}$$

The expression (1) is particularly interesting, because it indicates that each mode q is identified by a "phase":

$$\Phi_q = \frac{\pi q}{N}$$

The first mode, $q = 0$, has $\Phi = 0$ and frequency $\omega_{q=0} = \frac{\omega_0}{\sqrt{1+k}}$. The last mode, $q = N$, will have $\Phi = \pi$ and frequency $\omega_{q=N} = \frac{\omega_0}{\sqrt{1-k}}$. If we identify each mode by the value of Φ_q the first will be the "0" mode and the last the "π" mode. All other modes will have frequencies between the 0 and π mode frequencies.

For $k << 1$ the difference between π and 0 mode frequencies is

$$\Delta\omega = \omega_{q=N} - \omega_{q=0} = \omega_0\left(\frac{1}{\sqrt{1-k}} - \frac{1}{\sqrt{1+k}}\right) \approx \omega_0 k$$

i.e. the "bandwidth" of the coupled system is equal to the cell frequency times the coupling factor k.

Plotting the frequencies given by Eq. (1) as function of the phase Φ, we obtain curves such as that in Fig. 7, which corresponds to the case of five cells and five modes. This is a typical "dispersion curve", relating the frequencies of our system with their propagation constant. The permitted frequencies lie on a cosine-like curve, where the modes are represented by points equally spaced in phase. The more cells in the system, the more modes we will have on the curve, until the limit of the continuous: for an infinite number of cells, all of the modes on the curve are allowed.

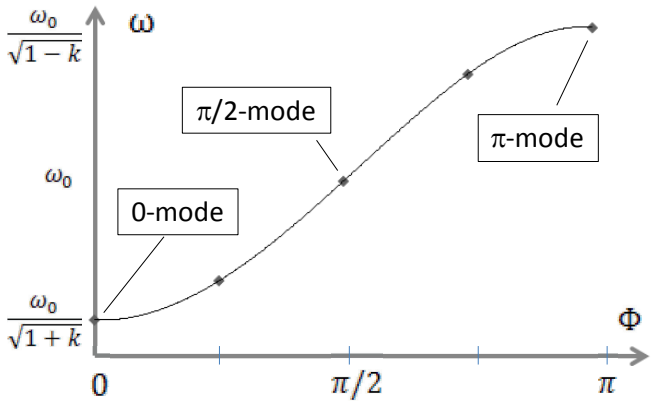


Fig. 7: Dispersion relation for a 5-cell coupled resonator chain

The field distribution in the cells is defined by the expression (2). For a given mode q, the fields will oscillate in each cell at the frequency ω_q, and *the amplitude of the oscillation will depend on the position of the cell in the chain*. The distribution of maximum field amplitudes along the chain follows a cosine-like function with argument ($\Phi_q i$), i.e. the product of the phase Φ_q times the cell number i. It is now clear that Φ_q represents the *phase difference between adjacent cells* in the coupled system. We can now draw the field distribution between the cells in the chain for the main modes, for example for a seven-cell system with $N = 6$ (Fig. 8).

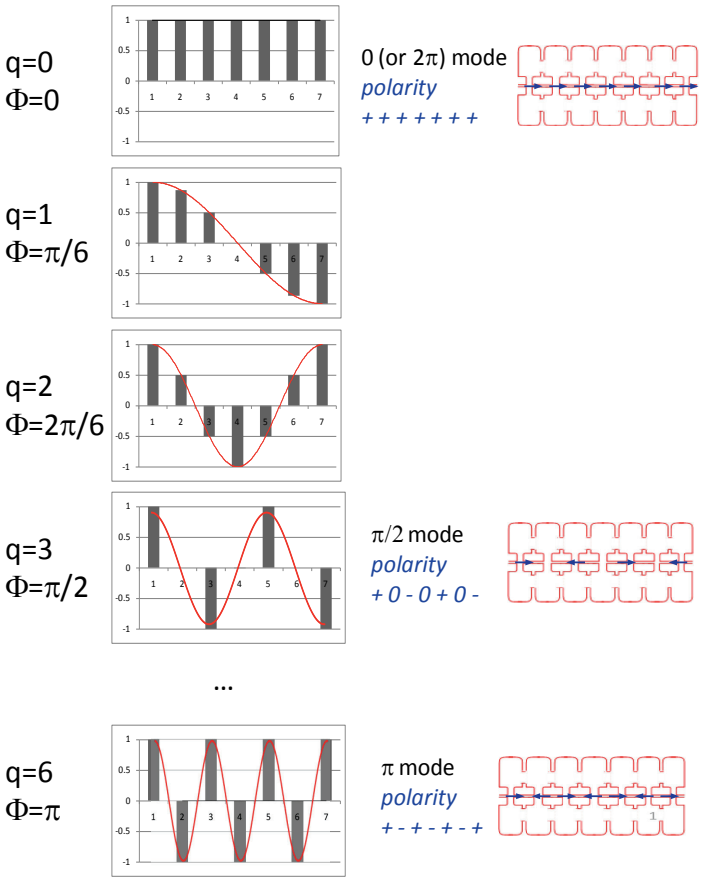


Fig. 8: Distribution of the fields in the cells of a seven-cell system and polarity of the electric field in the gaps for the modes used for particle acceleration

We must observe at this point that these are "standing-wave" modes: the plots of Fig. 8 show the field distribution at time $t = 0$. The fields are oscillating with angular frequency ω, and after $\omega t = \pi/2$ the field that we are plotting will be zero everywhere, whereas after $\omega t = \pi$ it will be maximum again, but with reversed polarity. The modes are identical to those of a vibrating string with the two ends fixed (in our case, defined by the boundary conditions).

To understand which of these modes can be used for the acceleration of particles and under what conditions, we can write the electric field for the mode q at the centre of gap i using expression (2) and then apply a simple trigonometric transformation:

$$E_i^{(q)} = E_0 \cos \Phi_q i \ \cos \omega_q t = \frac{E_0}{2}\left[\cos(\omega_q t - \Phi_q i) + \cos(\omega_q t + \Phi_q i)\right]$$

The electric field is the sum of two cosine functions. The first is the same as that we have introduced at the beginning of the lecture: for maximum acceleration, its argument must be 0 (or, more precisely, $2n\pi$) for a particle going from one cell to the next in the time τ. This gives $\omega_q \tau = \Phi_q$, leading to the synchronism condition:

$$d = \frac{\beta\lambda}{2}\frac{\Phi_q}{\pi}$$

This gives us the well-known result that the distance between the cells must be related to the beam velocity. In particular, for the 0 and π modes we obtain

$$d = \beta\lambda \qquad (0-mode, \Phi_q = 2\pi)$$

$$d = \frac{\beta\lambda}{2} \qquad (\pi-mode, \Phi_q = \pi)$$

The second cosine function instead tells us which modes can be used for acceleration: for $\omega_q\tau = \Phi_q$ it becomes equal to $\cos 2\Phi_q$ which is 1 only for $\Phi_q = 0, \pi, 2\pi, \ldots$. The conclusion is that only the modes 0 and π can be used for efficient particle acceleration. An exception is the π/2 mode, which has $\cos 2\Phi_q = 0$. This can still be used for acceleration (with d = βλ/4), but the acceleration is not very efficient, the field being present only in half of the cells. As we will see in the following, however, the π/2 mode presents the advantage of higher stability against deviation in the individual cell frequencies that justifies its use for some specific accelerating structures.

3 Zero-mode structures: the drift tube linac

The first and most important structure operating in the 0-mode is the drift tube linac (DTL), also called an Alvarez linac after the name of its inventor. It can be considered as a chain of coupled cells where the wall between cells has been completely removed to increase the coupling (Fig. 9). A high coupling offers the advantage of a large bandwidth, with sufficient spacing between the modes to avoid dangerous instabilities even when the chain is made of a large number of cells. Moreover, in the particular case of a structure operating in the 0-mode removing the cell-to-cell walls does not influence the power loss in the structure, because the RF currents flow only on the external tank and on the tubes. We must, however, keep some tubes on the axis, called "drift tubes", which hide the particles during the half RF period when the electric field on axis is decelerating. If the diameter of the drift tubes is sufficiently large, they can house focusing quadrupoles, which at low energy are required to keep the beam transversally focused. The drift tubes are suspended to the outer tank by means of supports called stems. The basic structure of a DTL is shown in Fig. 10.

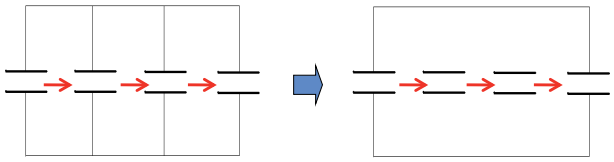


Fig. 9: From a chain or resonators operating in 0-mode to the DTL. The arrows indicate the direction of the electric field on axis.

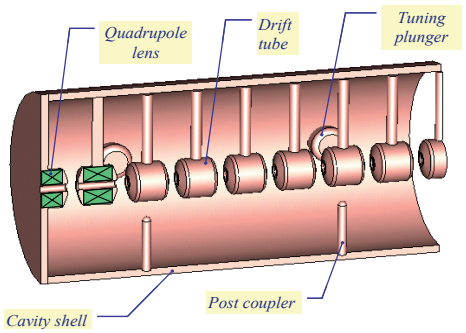


Fig. 10: The DTL structure

The DTL can be represented by a coupled circuit model similar to that in Fig. 6. Owing to the absence of the cell walls, however, the coupling mechanism is more complex, resulting in a strong electric coupling. The equivalent circuit of a DTL cell is shown in Fig. 11. The coupling factor k is in this case the ratio between the tube-to-wall and tube-to-tube capacitances C/C_0. A detailed analysis of the DTL equivalent circuit can be found in Ref. [2]. A DTL cavity is usually made of a large number of cells (up to more than 50 in a single tank), but because of the large coupling factor only the lowest modes can be observed, the others being hidden among the many different modes appearing at high frequencies.

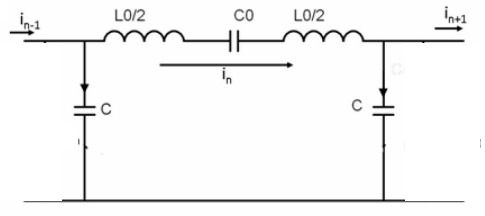


Fig. 11: Equivalent circuit of a DTL cell

The DTL is particularly suited to be used at low energies because the length of the cells (and of the drift tubes) can be easily adapted to the increasing velocity of the particle beam. We should observe that in the theoretical approach developed in Section 2, all relations depend only on the frequency and not on the inductance or capacitance of the single cells. If the capacitance and inductance from one cell to the other is changed keeping constant their product and therefore the cell frequency, all of the relations developed in Section 2 remain valid; the mode frequencies and the relative amplitudes in the cells will not change. The consequence is that the distance between gaps in a DTL can be easily adapted to the increasing beam velocity: if the length of the cells is progressively increased, keeping constant their frequency, the system will still behave in the usual way and the operating 0-mode will keep all of its properties. Tuning at the same frequency cells of different length is almost straightforward because increasing by the same proportion the cell and the drift tube lengths, the inductance will increase (longer cells) and the capacitance will decrease (larger gaps) by the same amount, and in a first approximation the variations will compensate keeping the frequency constant. Only minor adjustments to the gap lengths are required to compensate for second-order effects.

The possibility to adjust each individual cell length to the particle β together with the option of easily inserting focusing quadrupoles in the structure makes the DTL an ideal structure for the initial acceleration in a proton linac, from energies of a few megaelectronvolts to some 50 MeV to 100 MeV. As an example, Fig. 12 shows a three-dimensional open view of the CERN Linac4 DTL, which will accelerate H^- particles from 3 MeV to 50 MeV. The structure is divided into three individual 352.2 MHz resonators, for a total of 120 cells in a length of 19 m. The relativistic velocity increases from $\beta = 0.08$ to $\beta = 0.31$, and correspondingly the cell length $\beta\lambda$ increases by a factor of 3.9.

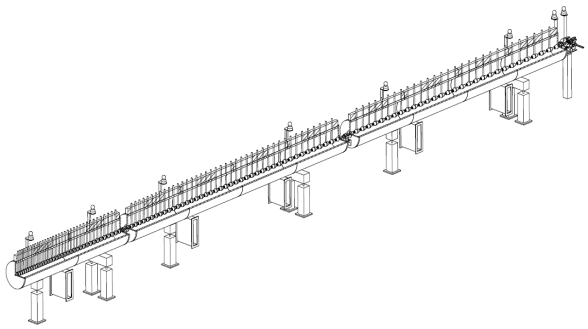


Fig. 12: Three-dimensional open view of the CERN Linac4 DTL

4 π-mode structures: PI-Mode Structure and elliptical cavities

Structures operating in the π mode are widely used, in particular in the magnetic coupled normal-conducting version and in the electric-coupled superconducting version. The coupling is provided by a slot on the external wall in the first case or by a sufficiently large opening on the axis for the latter. In both cases the cell length is kept constant inside short cavities made of a few (4 to 10, depending on the specific application) identical cells. Varying the cell length inside the cavities would complicate the design, because for π-mode structures not only the frequency but also the coupling factor depends on the cell length, and would considerably increase the construction cost. Therefore, π-mode structures are commonly used in the high-energy range of a linear accelerator for proton energies above 100 MeV, where the beam phase slippage is small.

As an example of normal-conducting π-mode structure, Fig. 13 shows the PI-Mode Structure (PIMS) that is being built at CERN for Linac4. Resonating at 352.2 MHz, it will cover the energy range between 100 MeV and 160 MeV. The PIMS cavities are made of seven cells, coupled via two slots in the connecting wall (visible on the left of Fig. 13); the pairs of slots on the two sides of a cell are rotated by 90° to minimize second-neighbour couplings that could perturb the dispersion curve. The complete PIMS section is made of 12 7-cell cavities. While the cell length inside each cavity is constant, it increases from cavity to cavity, matching the increase in β.

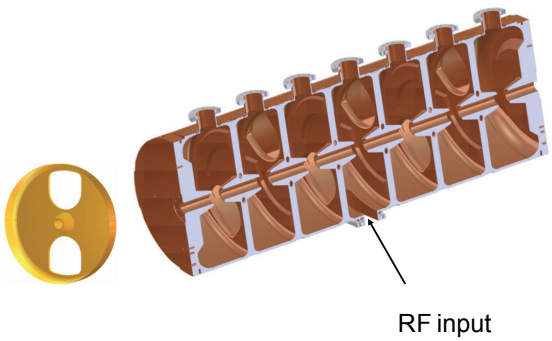


Fig. 13: The PIMS seven-cell cavity

Figure 14 shows a typical superconducting low-β cavity operating in the π mode. This particular cavity is made of five cells of identical length.

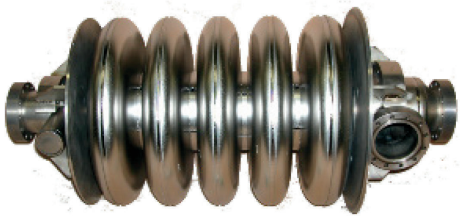


Fig. 14: A five-cell elliptical superconducting cavity

5 π/2-mode structures

To be able to operate very long chains of cells there is a particular interest in π/2-mode operation, although this mode has lower acceleration efficiency than 0 or π modes. This mode allows stable operation of long chains of coupled cells which can then be fed by single high-power RF sources, less expensive than many smaller power units.

Looking at Fig. 5, we see that the bandwidth of a coupled system is only proportional to the coupling factor and independent of the number of cells. Therefore, if we have a large number of cells and a large number of modes on the dispersion curve the modes will be very close in frequency to each other; this is particularly true for the 0 and π modes that lie in a region of the curve where the derivative is small. The modes will remain separated because their *Q* value is usually sufficiently high; however, an important consequence of having several other modes close in frequency to the operating mode is that the system becomes extremely sensitive to mechanical errors. Small deviations from the design frequency in some cells of the chain (coming, for example, from usual machining errors) would change the boundary conditions of the operating mode, forcing the system to introduce components from the adjacent modes to respect the new perturbed boundary conditions. These components are inversely proportional to the difference in frequency between operating and perturbing modes, making long structures more sensitive to errors than shorter ones.

In the π/2 mode instead, not only is the distance in frequency between the operating mode and the perturbing modes the largest, but their effect is compensated, making the chain of resonators virtually insensitive to the mechanical errors. The reason for this compensation is that the components from perturbing modes add up to the field distribution of the operating mode with a sign, positive or negative depending whether they are higher or lower in frequency than the operating mode. Observing that the modes on the two sides of the operating π/2 mode have the same field distributions in the cells that are excited (but different ones in the cells that are empty), an error in the chain of resonators will be compensated for by symmetric components of the modes higher or lower in frequency; these "perturbed" components will come with opposite sign and will cancel each other. In principle, a π/2-mode structure can be totally insensitive to errors; in practice, this requires a perfect symmetry of the perturbing modes around the operating one that is usually difficult to achieve. Reduction of the error sensitivity between a factor of 10 and a factor of 100 when going from a 0 or π mode to a π/2 mode is usually considered as satisfactory.

The best known π/2 structure is the Side-Coupled Linac (SCL) structure (Fig. 15) developed at the Los Alamos National Laboratories in the 1960s. Here the coupling is magnetic, through slots on the cell walls, and the coupling cells are moved away from the beam axis and placed symmetrically on both sides of the chain of accelerating cells. The result is that from the electromagnetic point of view, the structure operates in the π/2 mode providing stabilization of the field, whereas the beam travelling on the axis sees the typical field distribution of a π mode with maximum acceleration. Side-coupled structures are used at high energy and high RF frequency (from about 700 MHz), where high-power klystrons provide an economical way to feed a large number of cells, for which operation in 0 or π mode would be impossible because of the strong sensitivity to mechanical errors.

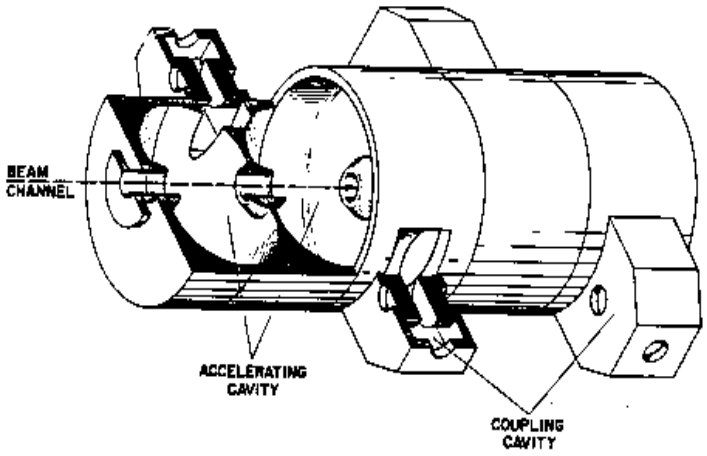


Fig. 15: The SCL structure [3]

6 Other accelerating structures: comparison of shunt impedances

We have so far considered the three main groups of coupled-cell accelerating structures, 0 mode, π mode and π/2 mode, and described the main representative for each category, the DTL, the PIMS and the SCL, respectively. Many more structures are used in linacs, however, which van be variants of these main families, can mix properties of different families or can even be based on different operating modes and a different approach. Among the most known "alternative" structures are the Separated-DTL (SDTL), a variant of the DTL without quadrupoles in the drift tubes, the Annular Coupled Structure (ACS), a π/2-mode structure with a different coupling cell geometry from the SCL, the Cell-Coupled DTL (CCDTL), a structure mixing a 0-mode with a π/2-mode operation, etc.

A particular category of linac structures is based on operation on a TE mode instead of the usual TM mode. The TE modes are called H modes in the German literature and these structures are usually referred as "H-mode" structures. A TE mode in principle has electric field only in the transverse direction and therefore cannot be used for acceleration. If drift tubes are placed in the structure connected alternatively to two sides of the resonator, however, the electric field of the TE11 mode can be forced in the longitudinal direction between the drift tubes and thus be able to provide acceleration. These are the so-called "IH structures". In a similar way, if the supports of the drift tubes are placed alternatively on the two transverse axes of the accelerating structure the TE21 mode can be forced to have a longitudinal electric field between the drift tubes: this is the "CH structure". IH and CH are very compact in terms of transverse dimensions; this is an advantage for low frequencies, but makes their construction difficult if high frequencies are required.

The choice of the most appropriate accelerating structure for a given project is very complex, being based on the comparison of many parameters. One of the most important figures of merit used for the selection of the accelerating structure is the shunt impedance, which represent the efficiency of a RF cavity in converting RF power into voltage across a gap. This is defined as

$$Z = \frac{V_0^2}{P}$$

with V_0 the peak RF voltage in a gap and P the RF power dissipated on the cavity walls to establish the voltage V_0. When the reference is to the effective voltage seen by a particle crossing the gap at velocity βc, we define the effective shunt impedance as

$$ZT^2 = \frac{(V_0 T)^2}{P}$$

with T the transit time factor of the particle crossing the gap (ratio of voltage seen by the particle during the crossing over maximum voltage available). If the structure has many gaps, we can refer to the shunt impedance per unit length, usually expressed in megaohms per metre. It must be noted that here we use the "linac" definition, considering the shunt impedance as a sort of efficiency, i.e. a ratio between useful work (the voltage available to the beam, which is proportional to the energy gained by a particle, squared for dimensional reasons) and the energy (power in this case) required to obtain it. If instead we start from the consideration that the shunt impedance is the equivalent resistance in the parallel equivalent circuit of a cavity resonator, we need to add a factor of two at the denominator of the previous relations. This is the "circuit" or "RF" definition of shunt impedance.

RF power is expensive, and the goal of every designer of normal-conducting accelerating structures is to maximize the shunt impedance, which depends on the mode used for acceleration, on the frequency and on the geometry of the structure. Other considerations come of course into play in the overall optimization; however, the shunt impedance remains one of the essential references for the structure designer.

Comparing structures in terms of shunt impedance is not easy, because the shunt impedance depends on the chosen frequency as well as on several design parameters related to the different

projects. To make the comparison as objective as possible for several types of structures, a study made in 2008 by the EU-funded Joint Research Activity HIPPI (High-Intensity Pulsed Power Injectors) derived the shunt impedance curves presented in Fig. 16, comparing eight different designs being studied in three different European Laboratories [4].

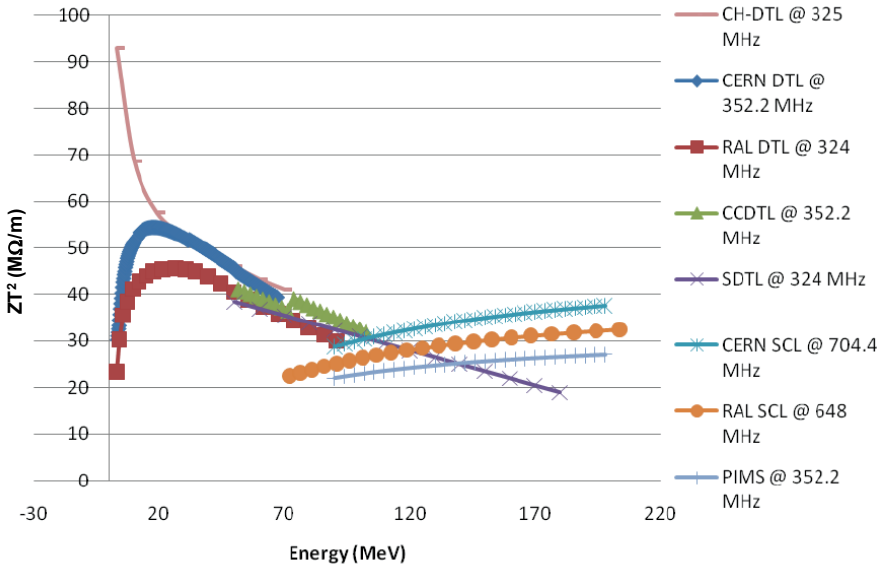


Fig. 16: Shunt impedance curves for different low-β structures

The values presented here correspond to simulations of effective shunt impedance per unit length corrected for the additional losses expected in the real case, for designs already optimized. The structures taken into consideration belong to two frequency ranges: the 324 MHz to 352 MHz range and its double harmonic, 648 MHz to 704 MHz. Higher operating frequencies have inherently higher shunt impedance; however, beam dynamics requirements in the first low-energy stages impose starting the acceleration at frequencies below about 400 MHz.

From the curves it appears that for all structures the shunt impedance has a more or less pronounced dependence on beam energy, due to the different distribution of RF currents and losses in cells of different length. Whereas 0-mode structures (DTL, but also the CCDTL in this context) have a maximum shunt impedance around 20 MeV to 30 MeV and then show a rapid decrease with energy, π-mode structures have a shunt impedance that is instead slightly increasing with energy, but starts from lower values than 0-mode structures. A natural transition point between these two types of structures would be around 100 MeV. For π-mode structures, remaining at the basic RF frequency leads to about 25 % lower shunt impedance than doubling the frequency (comparing CERN SCL and PIMS curves). Different considerations apply to H-mode structures; the CH considered in this comparison has by far the highest shunt impedance below 20 MeV. Above, its behaviour is similar to that of TM 0-mode structures.

7 Low-β superconducting structures

For superconducting structures, shunt impedance and power dissipation are not a concern, and the much lower RF power required allows using simpler and relatively inexpensive amplifiers. A separated-cavity configuration such as that shown in the bottom of Fig. 3 is therefore preferred for most superconducting linac applications at low energy, up to some 100 MeV to 150 MeV, where more operational flexibility is required and where the short cavity lengths allow having more quadrupoles

per unit length, as required by beam focusing at low energy. At higher energies, superconducting linacs use multicell π-mode cavities such as that presented in Fig. 12.

We must, however, observe that only few low-β linacs use single-gap cavities; even for superconducting structures, economic reasons suggest adopting structures with generally two or in some cases three or four gaps. The most widespread resonator used in particular for very low-beta heavy ion applications is the quarter-wavelength resonator (QWR, Fig. 17), sometimes declined in the half-wave resonator (HWR), when it is important to avoid even small dipole field components on the axis.

A resonator that has been proposed recently for several proton beam applications requiring operation at a large duty cycle, where superconductivity is an advantage, is the "spoke". In this cavity the electric field across the gaps is generated by a magnetic field turning around some supports, the spokes. Its main advantages are the compact dimensions and the relative insensitivity to mechanical vibrations. Similarly, for intense proton or deuteron beams is proposed a superconducting version of the CH resonator. Some examples of these structures are presented in Fig. 17.

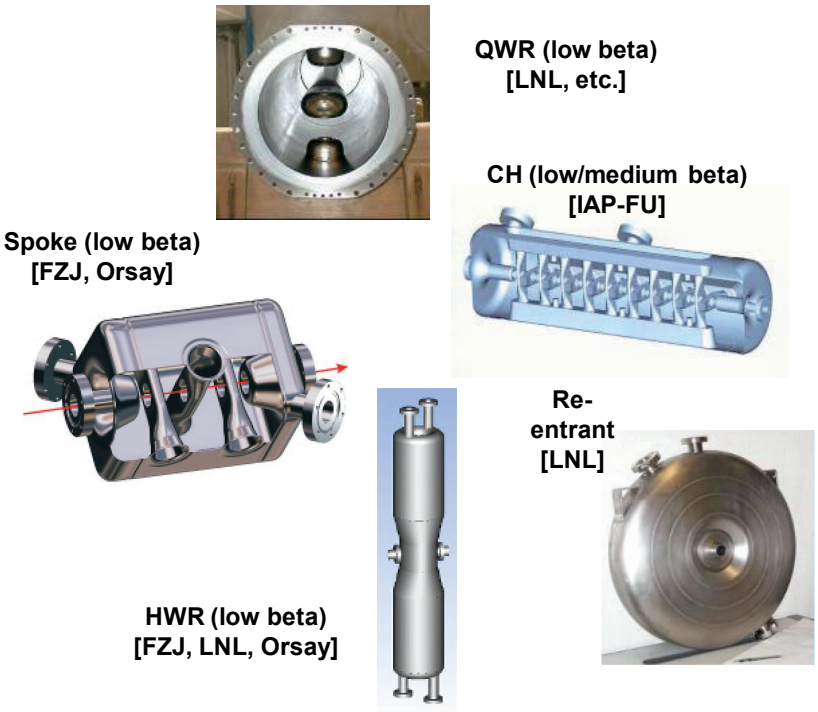


Fig. 17: Some examples of superconducting low-β structures

8 Beam dynamics in linear accelerators

8.1 Longitudinal plane

We have seen that, to achieve the maximum acceleration, bunches of particles must be synchronous with the accelerating wave. This means that they have to be injected into the linac on a well-defined phase with respect to the accelerating sinusoidal field, and then they need to maintain this phase during the acceleration process. Linac beams are usually made of a large number of particles with a given spread in phase and in energy. If the injection phase corresponds to the crest of the wave ($\varphi = 0^\circ$ in the linac definition) for maximum acceleration, particles having slightly higher or lower phases will gain less energy. They will slowly lose synchronicity until they are lost.

In linacs, the same principle of phase stability holds as in synchrotrons: if the injected beam is not centred on the crest of the wave but around a slightly lower phase, a "synchronous phase" φ_s whose typical values are between –20º and –30º, particles that are not on the central phase will oscillate around the synchronous phase during the acceleration process. The resulting longitudinal motion is confined, and the oscillation is represented by an elliptical motion of each particle in the longitudinal phase plane, i.e. the plane ($\Delta\varphi$, ΔW) of phase and energy difference with respect to the synchronous particle. The relation between the synchronous phase in an accelerating sinusoidal field and the longitudinal phase plane is presented in Fig. 18.

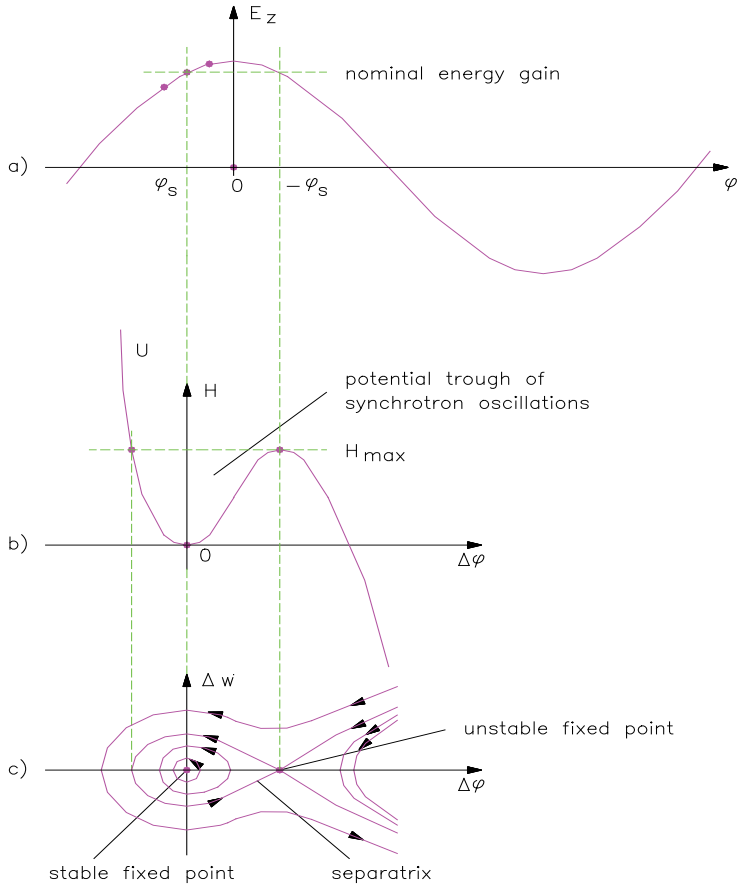


Fig. 18: Longitudinal motion of an ion beam

It is interesting to observe that the frequency of longitudinal oscillations, i.e. the number of oscillations in the longitudinal phase plane per unit time depends on the velocity of the beam. A simple approximate formula for the frequency of small oscillations ω_l can be found for example in Ref. [5]:

$$\omega_l^2 = \omega_0^2 \frac{qE_0T\sin(-\varphi)\lambda}{2\pi mc^2\beta\gamma^3}$$

Here ω_0 and λ are the RF frequency and wavelength, E_0T is the effective accelerating gradient and φ is the synchronous phase. The oscillation frequency is proportional to $1/\beta\gamma^3$: when the beam becomes relativistic, the oscillation frequency decreases rapidly. At the limit of $\beta\gamma^3 >> 1$, the oscillations will stop and the beam is practically "frozen" in phase and in energy with respect to the

synchronous particle. For example, in a proton linac $1/\beta\gamma^3$ and correspondingly ω_l can decrease by two or three orders of magnitude from the beginning of the acceleration to the high-energy section.

Another important relativistic effect for ion beams is the "phase damping", the shortening of bunch length in the longitudinal plane. This can be understood considering that, as the beam becomes more relativistic, its length in z seen by an external observer will contract due to relativity. A precise relativistic calculation shows that the phase damping is proportional to $1/(\beta\gamma)^{3/4}$:

$$\Delta\varphi = \frac{const}{(\beta\gamma)^{3/4}}$$

When a beam becomes relativistic, not only do its longitudinal oscillations slow down, but the bunch will also compact around the centre particle.

8.2 Transverse plane

Transversally in a linac, the beam will be subject to an external focusing force, provided by an array of quadrupoles or solenoids. This force has to counteract the defocusing forces that either develop inside the particle beam or come from the interaction with the accelerating field. The main defocusing contributions come from space charge forces and from RF defocusing.

8.2.1 Space charge forces

They represent the Coulomb repulsion inside the bunch between particles of the same sign. In the case of high-intensity linacs at low energy, space charge forces are one of the main design concerns. At relativistic velocity, however, the space charge repulsion starts to be compensated by the attraction due to the magnetic field generated by the beam, and finally disappears at the limit $v = c$. Space charge forces can be calculated only for very simple cases, such as that (Fig. 19) of an infinitely long cylindrical bunch with density $n(r)$ travelling at velocity v .In this case, the electric and magnetic fields active on a particle at distance r from the axis can be written as

$$E_r = \frac{e}{2\pi\varepsilon\, r}\int_0^r n(r)\, r\, dr \qquad\qquad E_r = \frac{e}{2\pi\varepsilon\, r}\int_0^r n(r)\, r\, dr$$

The resulting overall force acting on a particle in the bunch is orientated in the radial direction, and has intensity

$$F = e(E_r - vB_\varphi) = eE_r\left(1 - \frac{v^2}{c^2}\right) = eE_r(1-\beta^2) = \frac{eE_r}{\gamma^2}$$

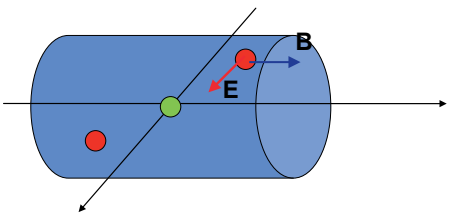


Fig. 19: Forces acting on a particle inside an infinitely long bunch

The overall space charge force is then proportional to $1/\gamma^2$ and will disappear for $\gamma \to \infty$.

8.2.2 RF defocusing forces

The RF defocusing is the transverse defocusing experienced by a particle that crosses an accelerating gap on a longitudinally focusing RF phase. We have seen in the previous section that for longitudinal

stability the beam will cross the gap when the field is increasing ($\varphi_s < 0$). Figure 20 shows a schematic configuration of the electric field in an accelerating gap. In correspondence to the entry and exit openings for the beam, the electric field has a transverse component, focusing at the entrance to the gap and defocusing at the exit, proportional to the distance from the axis. Because the field is increasing when the beam crosses the gap, the defocusing effect will be stronger than the focusing effect, and the net result will be a defocusing force proportional to the time required by the beam to cross the gap. A Lorentz transformation from the laboratory frame to the frame of the particles of the electric and magnetic field forces acting on a particle allows calculating the radial momentum impulse per period. Carrying out this calculation, one can find:

$$\Delta p_r = -\frac{\pi\, e\, E_0\, T\, L\, r \sin\varphi}{c\, \beta^2\, \gamma^2\, \lambda}$$

Again, this effect is proportional to $1/\gamma^2$, and will disappear at high beam velocity.

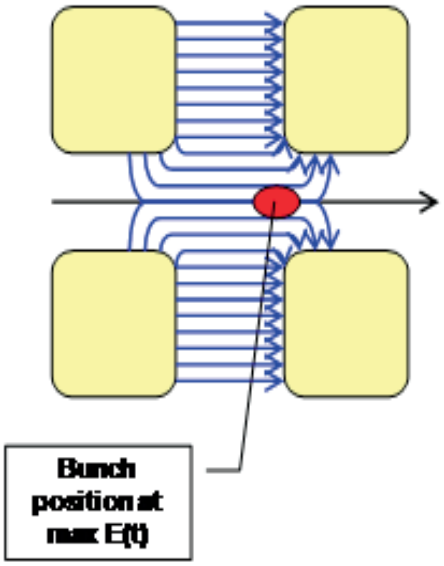


Fig. 20: Electric field line configuration around a gap and position of the bunch at maximum field

8.3 Transverse equilibrium

We have seen that in particular at low energies strong transverse defocusing forces act on the beam; to transport it with minimum particle loss we need to compensate for the defocusing forces with focusing forces, using standard alternating gradient focusing provided by quadrupoles along the beam line. As we have seen, a linac provides its acceleration with a series of accelerating structures; the standard focusing solution consists (Fig. 21) of alternating accelerating sections with focusing sections made of one quadrupole (singlet focusing), two quadrupoles (doublet focusing) or three quadrupoles (triplet focusing). In this way, we can immediately define the *focusing period* of the linac, corresponding to the length after which the structure repeats itself. It is important to observe that because the accelerating sections have to match the increasing beam velocity, the accelerating structures can have increasing lengths and therefore the basic focusing period does not necessarily have a constant length; however, in this case the travel time of the beam within a focusing period remains constant. The maximum length of the accelerating structure between the focusing elements depends on the beam energy, as we will see in the following; it goes from only one gap in the DTL to one or more structures containing many gaps at high energies.

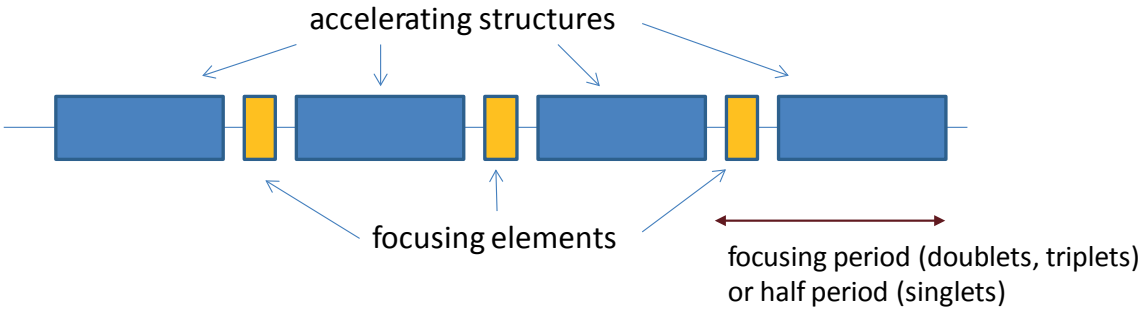


Fig. 21: Standard linac section

The beam must be transversally in equilibrium between the external focusing forces and the internal defocusing forces; the equilibrium will necessarily be dynamic, resulting in an oscillation in time and in space of the beam parameters with a frequency that will depend on the difference between focusing and defocusing forces. The oscillation can be as usual decomposed in two independent oscillations with the same frequency in the transverse planes x and y, the reference oscillating parameter being as usual the maximum beam radius. The oscillation is characterized by its frequency; instead of defining it with respect to time it is convenient to define the oscillation frequency in terms of the phase advance σ_t, the increase in the phase of the oscillation over one focusing period of the structure, which is usually constant or changes only smoothly over a linac section. Alternatively, the phase advance per unit length k_t can be used. If L is the period length of the focusing structure, then $k_t = \sigma_t/L$. Modern beam dynamics simulation codes allow us to easily calculate the phase advance for a given focusing period, accelerating structure design and quadrupole gradient; it is important to select accurately this parameter to minimize beam loss and to avoid excessive transverse emittance growth. The first basic rule is that σ_t should always be <90º, to avoid resonances that could lead to emittance growth and to beam loss; it should also be higher than some 20º, to avoid that the amplitude of the oscillations becomes too high and the beam size becomes too large.

We can find an approximate relation for the phase advance as a function of focusing and defocusing forces referred to a simple theoretical case. First of all, one has to limit the analysis to beam oscillations in a simple F0D0 quadrupole lattice (focusing–drift–defocusing-drift, corresponding to the "singlet" focusing) under smooth focusing approximation, i.e. averaging the localized effect of the focusing elements. Then, adding together the focusing and RF defocusing contributions to phase advance as derived for example in Ref. [5, Eq. (7.103)] and subtracting the space charge term as approximately calculated in the case of a uniform three-dimensional ellipsoidal bunch [5, Eq. (9.51)] we obtain for the phase advance per unit length:

$$k_t^2 = \left(\frac{\sigma_t}{N\beta\lambda}\right)^2 = \left(\frac{q\,G\,l}{2\,mc\,\beta\gamma}\right)^2 - \frac{\pi\,q\,E_0 T\sin(-\varphi)}{mc^2\lambda\,\beta^3\gamma^3} - \frac{3q\,I\,\lambda(1-f)}{8\pi\varepsilon_0\,r_0^{\,3}mc^3\beta^2\gamma^3}$$

Here $N\beta\lambda$ is the length of the focusing period in units of $\beta\lambda$. The first term on the right-hand side of the equation is the focusing component: Gl is the quadrupole integrated gradient, expressed as product of gradient G and length l of the quadrupole. The second term is the RF defocusing: E_0T is the effective accelerating gradient, λ the RF wavelength and φ the synchronous phase. For $\varphi < 0$, corresponding to longitudinal stability, $\sin(-\varphi)$ is positive and this term is negative, i.e. defocusing. The third term is the approximate space charge contribution: I is the beam current, f is an ellipsoid form factor ($0 < f < 1$) and r_0 is the average beam radius. The other parameters in the equation define the particle and medium properties (charge q, mass m, relativistic parameters β and γ, free space permittivity ε_0).

This simple equation shows, although in an approximate simplified case, how the beam evolution in a linear accelerator depends on the delicate equilibrium between external focusing and internal defocusing forces. Real cases can be solved only numerically; however, the parametric dependence given by this equation remains valid, and allows us to determine how the beam dynamics will change with the particle β. At low velocities ($\beta << 1$, $\gamma \sim 1$) the defocusing terms are dominant. To keep the beam focused with a large enough phase advance per unit length one has to increase the integrated gradient Gl and/or decrease the length of the focusing period $N\beta\lambda$, i.e. minimize the distance between focusing elements. This is for example the case of the radio-frequency quadrupole (RFQ), the structure of choice for low-energy ion beams (from $\beta \approx 0.01$ to $\beta \approx 0.1$). The RFQ provides a high focusing gradient by means of an electrostatic quadrupole field, with short cells at focusing period $\beta\lambda$. At higher energy, standard electromagnetic quadrupoles have a sufficiently high gradient and a structure alternating accelerating gaps and quadrupoles can be used. The standard structure used for protons above about 3 MeV energy is the DTL (Fig. 10), which presents a $2\beta\lambda$ focusing period when focusing and defocusing quadrupoles alternate inside the drift tubes (i.e. a F0D0 focusing). The maximum gradient achievable in the DTL quadrupoles together with the short focusing period allow keeping the phase advance in an acceptable range even for high current and high space charge beams; for example, the CERN Linac2 can accelerate a beam current up to 180 mA, the maximum achieved so far in a DTL. As an example of DTL beam dynamics design, Fig. 22 presents quadrupole gradients and the corresponding phase advance for the CERN Linac4 DTL design [6]. The corresponding oscillations of the beam envelope (in this case is plotted the maximum beam radius in x) for the given input matching conditions are shown in Fig. 23. As the phase advance decreases, the period of the oscillations becomes longer.

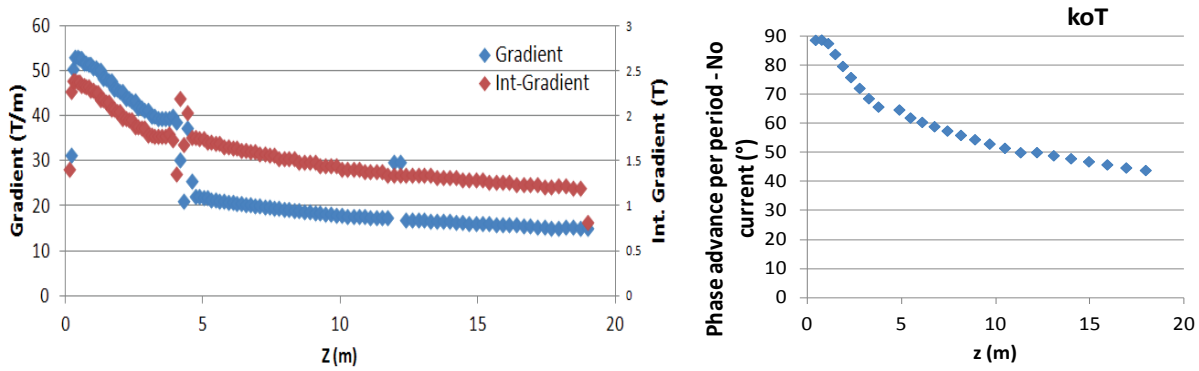


Fig. 22: Quadrupole gradients and corresponding phase advance for the Linac4 DTL design

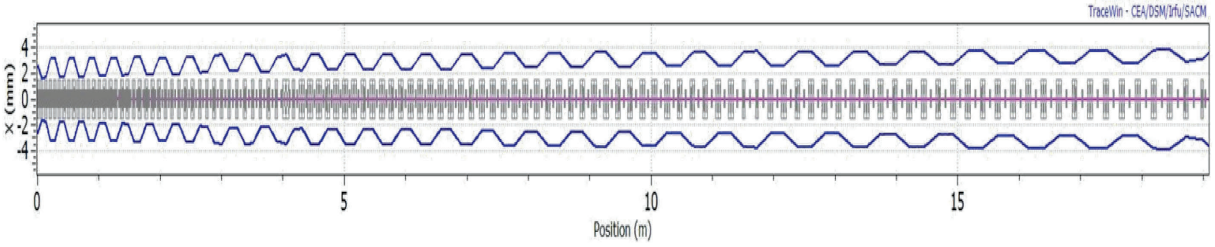


Fig. 23: Transverse root mean square beam envelope (x-plane) for the Linac4 DTL design

Going further up in energy, the defocusing terms (proportional to $1/\beta^3\gamma^3$ and $1/\beta^2\gamma^3$, respectively) decrease much faster than the focusing term (proportional to $1/\beta\gamma$). The focusing period can be increased, reducing the number of quadrupoles and simplifying the construction of the linac. Starting from energies between 50 MeV and 100 MeV modern proton linacs usually adopt multicell structures operating in π mode spaced by focusing quadrupoles; these can be normal conducting as in the

structure of Fig. 13 or superconducting as in that of Fig. 14. For example, in the CERN Linac4 design (90 keV to 160 MeV beam energy) the focusing period increases from βλ in the RFQ to 15βλ in the last π-mode accelerating structure. The corresponding beam envelope is shown in Fig. 24. Heavy ions differ from protons for the fact that usually ion currents are small and the space charge term can be neglected; immediately after the RFQ, the focusing period can be some 5βλ to 10βλ.

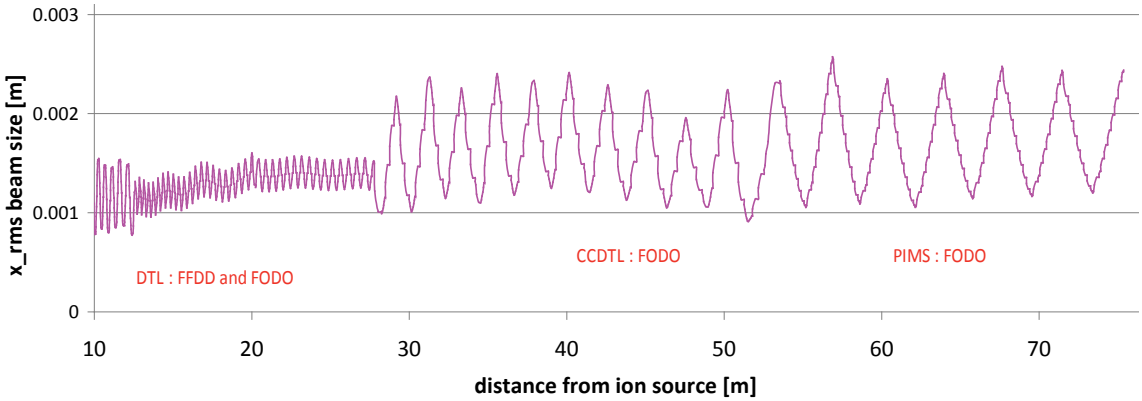


Fig. 24: Transverse root mean square beam envelope (*x*-plane) along the different Linac4 sections (3 MeV to 160 MeV)

9 Linac architecture

From the previous sections, we can derive two fundamental rules that define the basic architecture of any linear accelerator required to reach energies above a few megaelectronvolts:

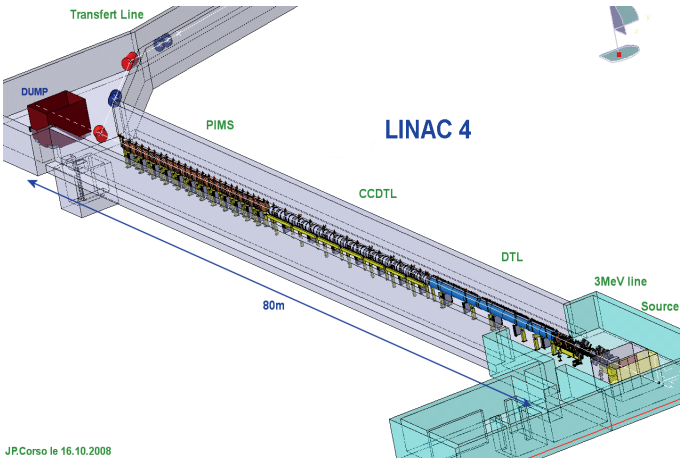
(a) at low energy the linac cells have to be different in length to follow precisely the increase in beam velocity; at higher energy instead, the accelerating structures can be made out of sequences of identical cells, thus reducing the construction costs thanks to a higher standardization and to the use of longer vacuum structures;

(b) as the beam energy increases less focusing is required because of the reduction in space charge and RF defocusing; at higher energy the focusing length can increase, reducing the number and cost of the quadrupoles.

These two rules lead to the same consequence: to keep construction and operation costs low a linac must change the type of structure and focusing scheme with the increase of energy, going from expensive structures covering the low-energy range to more economical structures and focusing layouts for the high-energy range. A linac will be thus made of different sections covering different ranges of energy. After the ion source, the first accelerating structure will be a RFQ; this is the only linac structure that can provide the strong focusing forces needed to compensate for the space charge, which is very high at low energy and constitutes the main beam current limitation in linacs. The RFQ is an expensive structure which is not particularly efficient in using the RF power; for this reason, for an energy of between 2 MeV and 3 MeV it becomes convenient to go from the RFQ to another type of structure. Most linac projects use a DTL for the following section; the DTL integrates in a single RF cavity cells of increasing length matching precisely the increase in beam velocity and quadrupoles in every cell providing a strong and uniform focusing. As an alternative to the DTL, projects required to operate at very high duty cycle (close to or at CW) and/or to provide acceleration of different ion types can find an economic interest in using short superconducting structures, e.g. with two gaps at fixed distance. Both a normal-conducting DTL and a sequence of short superconducting cavities are quite expensive, and at energies of some tens of megaelectronvolts (usually between 50 MeV and 100 MeV) it is convenient to go to structures with many identical cells interleaved with quadrupoles; whereas at intermediate energies (50 MeV to 200 MeV) special structures are required, at higher energies long

standardized modules of identical cells, often superconducting, can be used. Focusing is provided by quadrupoles placed between the modules.

As an example, Fig. 25 shows the layout of the CERN Linac4. This linac will cover the energy range between 45 keV (extraction from the ion source) and 160 MeV with four different types of accelerating structures, characterized by an increasing number of identical cells per accelerating structure and an increasing length of the focusing period.



Section	Output Energy [MeV]	Number of consecutive identical cells	Distance between quadrupoles
RFQ	3	1	-
DTL	50	1	βλ
CCDTL	100	3	4 βλ
PIMS	160	7	5 βλ

Fig. 25: Layout of Linac4 under construction at CERN

9.1 The choice of the frequency

The choice of the appropriate frequency (or sequence of frequencies) for a linac has to take into account several factors coming from mechanical, RF and beam dynamics considerations related to the different scaling with frequency of the main linac parameters. First of all, accelerator dimensions and cell length are proportional to the wavelength λ: higher frequencies will result in smaller accelerating structures, requiring less copper or steel. Machining tolerances, however, scale as well as λ: smaller structures require tighter tolerances that are more expensive to achieve. From the RF point of view, higher frequencies are also preferable, because both the shunt impedance and the maximum surface electric field scale approximately as $\sqrt{f}$.[4] In Section 8, however, we have seen that the RF defocusing scales with 1/λ, becoming excessively high at high frequencies; RF defocusing and cell length in the RFQ usually define the maximum frequency that can be used in the initial section of a linac. A summary of the dependence with the frequency for the different linac parameters is given in Table 1.

Table 1: Scaling with frequency of some basic linear accelerator parameters

RF defocusing $\sim f$
Cell length (~βλ) $\sim 1/f$
Peak electric field $\sim\sqrt{f}$
RF power efficiency (shunt impedance) $\sim\sqrt{f}$
Accelerator structure dimensions $\sim 1/f$
Machining tolerances $\sim 1/f$

[4] It should be noted that the scaling of the maximum field with the square root of the frequency is valid only approximately and for frequencies below about 10 GHz. The shunt impedance scales as the square root of the frequency in the case of a symmetric scaling of the cavity dimensions; if the beam aperture is kept constant, the dependence is smaller.

A first analysis of the frequency dependence indicates that high frequencies are economically convenient: the linac is smaller, makes use of less RF power and can reach a higher accelerating field. Limitations to the frequency come from the mechanical construction costs that depend critically on the required tolerances and on the RF defocusing in the RFQ; for these reasons, modern linac designs tend to start with a basic frequency in the RFQ and then to double it as soon as the cells become longer and the RF defocusing decreases. The availability and cost of the RF power sources is also an essential element in the choice of the frequency, and has to be considered carefully in particular when multiple frequencies are used.

All of these requirements result in some standardization in the frequencies commonly used in linacs: while in the past proton RFQs used to have frequencies around 200 MHz, nowadays frequencies in the range 325 MHz to 405 MHz are the usual standard. In the following sections, normal conducting or superconducting, a frequency jump by a factor of 2 is often applied, reaching the range 700 MHz to 800 MHz. A jump by a factor of 4 from 325 MHz allows reaching the International Linear Collider frequency of 1.3 GHz, thus making the use of protons of RF technology developed for electron linacs possible.

9.2 Superconductivity and the warm–cold transition

A constant in the architecture of modern high-energy linacs is the use of superconductivity at high energy. The advantages of superconducting accelerating structures are evident: a much smaller RF system needs to deliver only the power directly going to the beam, a large beam aperture allows for a lower beam loss (although the particles in the beam halo could be transported in the superconducting section and then lost in the following beam transport line) and finally the operating electricity costs are lower than for a normal-conducting linac, a feature particularly important given the present concerns for energy saving. The energy argument is particularly important for high duty cycle machines, where high power efficiency is an important requirement, and becomes proportionally less important for low duty machines where the beam power is a minor fraction of the power required by the machine.

In defining an optimum linac design, however, some peculiar characteristics of superconducting systems have to be considered. A superconducting system needs a large cryogenic installation requiring a significant amount of power; for a low duty linac this can be dominated by the static losses required for keeping the system at cryogenic temperature, thus greatly reducing the overall power efficiency. At low energy, the need to provide many cold–warm transitions to accommodate the large number of warm quadrupoles required because of the short focusing periods increases the cost and complexity of the installation. On top of that, the difficulty in predicting the individual gradients of superconducting cavities makes them less attractive at low energy, where as we have seen the sequence of cell lengths has to precisely follow the calculated increase in beam velocity.

The result is that while superconductivity is certainly the most attractive technology for linacs at high energy and high duty cycle, at low energy and low duty cycle normal-conducting structures remain more economical and more efficient. An exact comparison of cost and efficiency for the two technologies is particularly difficult because it depends not only on energy and duty cycle, but also on other design parameters such as repetition frequency, peak beam current and pulse length. A high repetition frequency makes a superconducting linac less efficient, more power being lost during the long pulse rise time required to feed the superconducting cavities. The maximum current during pulse plays an important role, because whereas normal-conducting linacs are more efficient operating with short pulses of high beam current, superconducting linacs prefer long pulses with less current that require a smaller RF installation. For all of these reasons, the optimum transition energy between warm and cold sections in a modern linac remains difficult to determine and requires a precise economical comparison of the two technologies for the parameters of each particular project. A special case are linacs operating in CW, where the tendency nowadays is to start the superconducting section immediately after the RFQ, usually at 3 MeV. Although the general rule remains that the higher the

duty cycle the lower the optimum transition energy, different projects have chosen different transitions; as an example, Fig. 26 plots the selected transition energy as a function of duty cycle for the most important linacs built in recent years or in the design and construction phase. The approximate trend line connects the low transition energy of CW projects with the high transition energies of low duty cycle projects and represents the "state-of-the-art" in warm–cold transition.

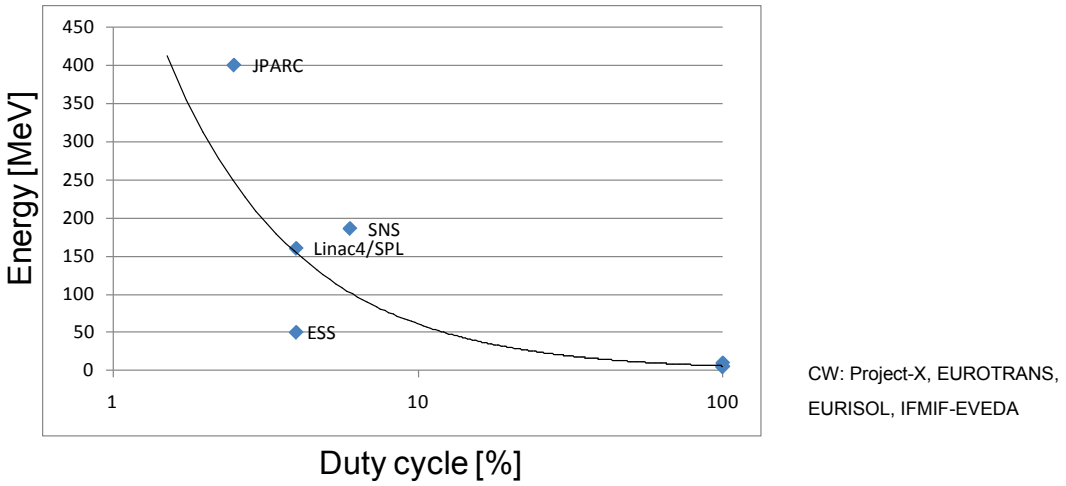


Fig. 26: Warm–cold transition for recent linac designs (built or in the construction or design phase)

References

[1] D.E. Nagle, E.A. Knapp, B.C. Knapp, *Rev. Sci. Instrum.* **38** (1967) 11.

[2] F. Grespan, G. De Michele, S. Ramberger and M. Vretenar, Circuital model for post coupler stabilization in a drift tube linac, CERN-sLHC-Project-Note-0014 (CERN, Geneva, 2010), available at: http://cdsweb.cern.ch/record/1263865/files/project-note-0014.pdf.

[3] E.A. Knapp, B.C. Knapp and J.M. Potter, *Rev. Sci. Instrum.* **39** (1968) 979–991.

[4] C. Plostinar, Comparative assessment of HIPPI normal conducting structures, CARE-Report-2008-071-HIPPI, available at: http://irfu.cea.fr/Phocea/file.php?class=std&&file=Doc/Care/care-report-08-071.pdf.

[5] T. Wangler, *Principles of RF Linear Accelerators* (Wiley, New York, 1998).

[6] A.M. Lombardi, *et al*., Beam dynamics in Linac4 at CERN, Proc. of Hadron Beam 2008, Nashville, TN, USA, available at: http://accelconf.web.cern.ch/AccelConf/HB2008/papers/wgb14.pdf.

Specific instrumentation and diagnostics for high-intensity hadron beams

Kay Wittenburg
DESY, Hamburg, Germany

Abstract
An overview of various typical instruments used for high-intensity hadron beams is given. In addition, a few important diagnostic methods are discussed which are quite special for these kinds of beams.

1 Introduction

All beam instrumentations for high-intensity hadron beams have to fulfil one important criterion: the instruments have to be as minimally invasive as possible to survive the full beam. If, for any reason, this cannot be achieved, the required diagnostics cannot be done with the full intensity of the beam and interpolations are necessary to calculate the parameters of the nominal beam. Such an interpolation might generate large error bars and therefore might not be suitable for precise beam diagnostics.

A second important feature of the instrumentation is the required dynamic range [1]. Typically the instrument has to cover signals coming from low-intensity beams during commissioning up to very-high-intensity beams after an upgrade of the accelerator (which often does not include an upgrade of the beam instrumentation). Sometimes tiny "pilot bunches" have to be diagnosed to ensure that the whole accelerator chain has been set up correctly before injecting the full beam. Also variable modes of operation, e.g. continuous wave (CW) beams, various ion types, long and short pulsed beams, have to be diagnosed. Often the beam has a large diameter, especially non-relativistic beams. Therefore, large size beam monitors with large apertures are needed.

A third important feature of high-intensity beam instruments is that some diagnostic systems have to create a beam interlock or allow the signal to protect the machine against damage from mis-steered or unmatched beams. Therefore, their high reliability and availability as well as their accurate and stable work are necessary to ensure high productivity of the accelerator.

The following sections summarize the most important instruments for sufficient beam diagnostics of high-intensity hadron beams with an emphasis on minimal invasive devices and their high dynamic range. The chapters are followed by some examples of special beam diagnostics which are important for high intense hadron beams. Instruments mainly used in electron accelerators are not mentioned here (e.g. cavity BPMs (beam position monitors) ICTs (inductive current transformer), synchrotron radiation from bending magnets, etc.). An example for the main instruments in high-intensity accelerators is given in Table 1; it summarizes the various beam diagnostic components of the J-PARC complex.

Table 1: Summary of the beam diagnostic components of the J-PARC complex; from Refs. [2–5]

LINAC: MEBT, DTL/SDTL, A0BT, L3BT	103 Beam position monitors (BPMs)
	98 Slow and fast current transformers (SCTs/FCTs)
	34 Profile monitors (wire scanners (WSs) and destructive halo monitors (beam scraper monitors (BSMs))
	125 Beam loss monitors (BLMs; scintillators and proportional chambers)
RCS	62 BPMs
	9 Current monitors (direct current transformer (DCCTs), SCTs, FCTs, wall current monitor (WCMs)), WCMs used for bunch length measurement.
	7 Secondary emission monitors (SEMs)
	2 Ionization profile monitors (IPMs), also for halo monitoring
	134 BLMs (scintillators, proportional chambers, ionization chambers)
Beam transfer lines: 3–50 BT 3 NBT	19 BPMs
	5 FCTs
	5 SEMs
	53 BLMs (Proportional and ionization chambers)
Main ring (MR)	192 BPMs
	11 Current monitors (DCCTs, FCTs, WCMs), WCMs used for bunch length measurements.
	238 BLMs (proportional and ionization chambers)
	6 Screen monitors (SEMs, luminescence screens)
	3 Profile monitors (WSs, IPMs)

2 Instruments for beam current and position measurements

Typically the electromagnetic field of the particle beam is used to determine its charge (current) and position. Its signal spectrum extends from the DC component of the beam to its radio-frequency (RF) frequency (neglecting bunch sub-structures). All electric and most magnetic signals cannot reach the region outside the conducting and non-magnetic beam chamber due to its effective shielding. Only the magnetic DC component of the beam can be detected outside the chamber while the much more useful part of the spectrum lies at higher frequencies and is therefore only accessible inside the chamber or through a "gap" in the beam pipe. A thin ceramic ring soldered at both ends of the beam pipe is required to form such a non-conducting gap. Typical beam current monitors make use of such a gap while BPMs use antennas inside the chamber together with a small ultra-high-vacuum (UHV) feedthrough to gain access to the signal at the outside of the chamber.

2.1 Resistive wall current monitor

The interruption in the beam pipe by a gap forces the image current to find a new path. By clever design of the monitor, the path and its impedance Z_{gap} are defined by the instrument designer. The voltage across the gap V_{gap} is then

$$V_{gap}(\omega) = Z_{gap}(\omega) \cdot I_{beam}(t) = Z_{gap}(\omega) \cdot (-I_{wall}(t))$$

with

$$\frac{1}{Z_{gap}} = \frac{1}{R} + \frac{1}{i\omega L} + i\omega C$$

The resistance R is formed by a resistive network across the gap, the inductance L and the capacitance C depend on geometrical and mechanical issues. Here Z_{gap} is typically of the order of a few Ohms. Many detailed design studies are necessary to achieve a flat response of the monitor over a large frequency range and to avoid dependence from the beam position:

- An UHV compatible ceramic (with relative permittivity of around 10) forming an equal spacing of the gap and a separation from the beam vacuum. The gap should be short compared with the bunch length to avoid beam shape integration. Since C depends on the gap width, it should not be too small to avoid high C and therefore a limit in the bandwidth. Typical values for C are about tens of picofarads.
- Equally spaced resistors of the same value R* around a round gap and signal summation by combining the signals from four quadrants avoiding beam position dependence.
- Well-defined bypass for image current and avoiding resonances by adding material with high μ. This increases the inductance L at low frequencies and reduces the lower cutoff frequency. Typical values for L are around 100 nH.
- Reducing other, stray currents vagabonding along the pipe by careful shielding and grounding.
- Avoiding monitor positions close to beam pipe discontinuities, since higher-order modes above cutoff can travel significant distances in the beam pipe. Even some absorbing material (e.g. ferrites) on both sides of the gap but inside the beam pipe might be useful to reduce high-frequency background [6].

Many useful design hints are given in Refs. [7, 8]. A sketch of a WCM is shown in Fig. 1. This type of monitor can have a broad frequency response from a few kilohertz up to a few gigahertz[1] with flat impedance. A frequency response variation of less than 1 dB over the full range was reached [7]. The low-frequency cutoff leads to a droop in successive signals, which has to be taken into account.

The bunch current or the number of particles in the bunch N_B can be determined by

$$N_B = \frac{\int V_{gap} \cdot dt}{Z_{gap} \cdot e \cdot K}$$

while K is a constant which takes into account various attenuation of cables, combiners and calibration constants. Note that the wall current does not contain information about the DC current component of the beam since this frequency component of the beam current penetrates the (non-magnetic) beam pipe unaffected [10]. Therefore, the baseline of the signal is shifted while the shift is proportional to the DC current. A careful baseline restoration is needed for precise bunch current measurements.

Owing to its broad frequency response the wall current monitor (WCM) is often used, in addition to bunch current determination, for measuring the longitudinal profile of the bunch, calculating its emittance and diagnosing longitudinal instabilities. Note that the ultimate bandwidth is limited by the spreading angle of the radial electrical field lines which have an opening angle of approximately $1/\gamma$. This limits the longitudinal resolution for non-relativistic beams. WCMs are also used for RF and timing feedback issues (e.g. compensating beam loading) and time-of-flight (TOF; energy) measurements since they provide very fast and large signals.

Some care has to be taken at high beam currents:
The absorption of higher-order modes (HOMs) in the magnetic material will increase its temperature. A good cooling is necessary. This is especially true for short bunches as in electron accelerators which induce strong HOMs.
High beam currents may cause saturation in the magnetic material which changes the inductance L and therefore the lower cutoff frequency. As a result the droop rate will change.

[1] A recently developed wall current monitor for CLIC reached a bandwidth of 20 GHz [9].

The power level in the monitor resistors R^* can reach some tens of Watts at $N_B = 10^{11}$ particles/bunch which may lead to high thermal load of the resistors at high repetitive signals. Many distributed resistors around the gap will help but a change in resistance with temperature might occur which will change the calibration constant of the monitor. The signal level may reach 100 V or more. Detailed calculations of the power levels for the WCM in the Large Hadron Collider (LHC) can be found in Ref. [11].

Since the signal level is quite high, a high dynamic readout can be achieved by various methods like switch able attenuators/amplifiers, logarithmic amplifiers or large bit analogue-to-digital converters (ADCs). The last two options are limited in bandwidth, so that a compromise has to be found for each specific application.

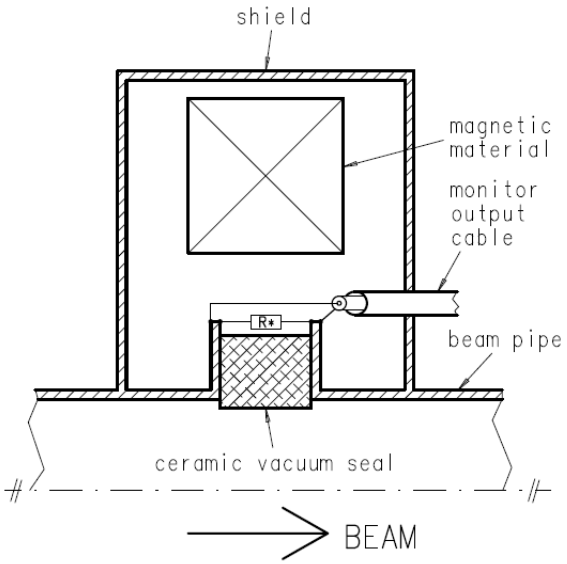


Fig. 1: Conceptual view of a WCM. The resistors R^* are distributed evenly around the gap. A magnetic material with high μ is also shown as one coaxial cable for the readout. (Courtesy of M. Siemens, DESY.)

2.2 Inductive alternating current transformers

In contrast to the capacitive (electrical) coupling of a WCM, the inductive current transformers are using the magnetic field of the beam to determine the beam current. A bunch crossing a (ceramic) gap in a beam pipe induces a magnetic flux in a high-permeability toroid around the gap, like a primary single turn winding in a classical transformer. It induces a current in a secondary winding of N_s turns and an inductance L_s. This current can be measured by the voltage V_s across a resistor R_s. By applying the classical transformer equations one receives the typical transformer response [10, 12]:

$$V_s(\omega) = \frac{i \cdot \omega \cdot \tau}{1 + i \cdot \omega \cdot \tau} \cdot R_s \cdot \frac{I_{Beam}}{N_s}$$

with $\tau = L_s/R_s$. Here τ is the time constant of the secondary winding. For $\omega >> 1/\tau$ it results in a simple proportionality of

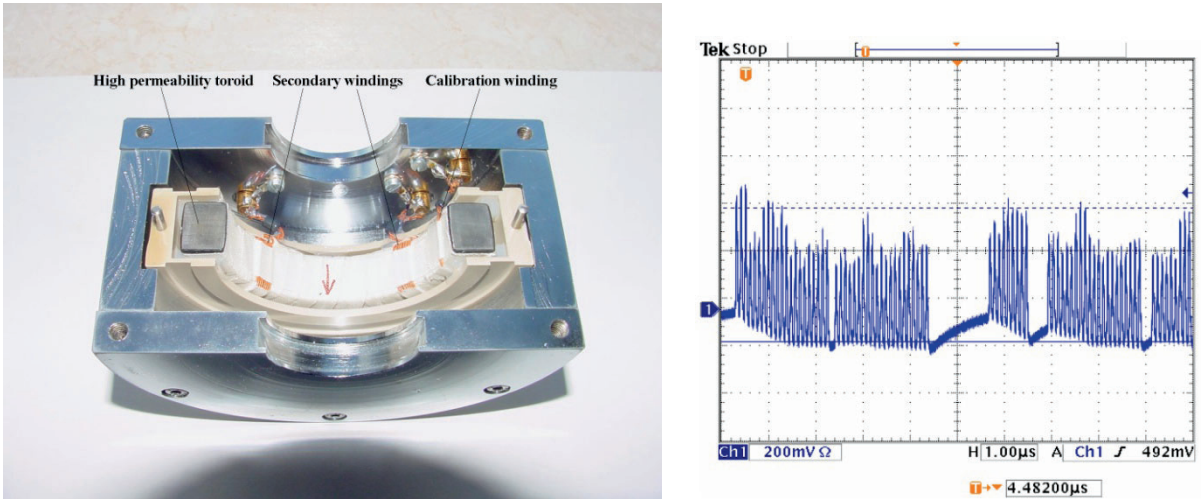
$$V_s = R_s \cdot \frac{I_{Beam}}{N_s}$$

where the measured voltage V_s is proportionally to the beam current I_{Beam} and in phase with it.

The inductance L_s depends on the permeability μ of the core material, the number of windings N_s^2 and its dimensions. Assuming typical values for L_s = 1 mH and a load resistor R_s of 50 Ω one obtains the proportionality above for frequencies ω >> 50 kHz. The high-frequency performance of such a classical current transformer is limited by stray capacitance between the windings and to ground as well as due to energy loss in the toroid material ($\sim\omega^2$). Typically the upper limit is in the some hundreds of megahertz range. Proper impedance matching and low-pass filters are essential to avoid resonances at higher frequencies in the readout loop. These limits make this simple AC transformer not suitable for the measurement of the longitudinal bunch shape but it is widely used for bunch charge/current measurements.

Since this device is a classical transformer, it cannot transmit the DC component. Therefore a certain droop rate is indispensable ($\tau_{droop} \sim L_s/R_s$); see Fig. 2. An optimization between the high-frequency response and low droop rate becomes necessary. Fast current transformers with a droop rate of <1 %/μs and an upper frequency of more than 800 MHz are commercially available [13]. Accurate DC baseline restoring is necessary, however, to avoid a measurement error in a train of successive bunches. In an accelerator/storage ring with an infinitely long bunch train an equilibrium is reached when the area below and above the zero line are equal.

An advantage of an inductive current transformer is its small dependence on the beam position. Careful magnetic shielding of the core is very important as well as a good shielding of the signal windings to avoid contamination of external noise sources. An absolute calibration of the measured value can be done by simply adding a calibration winding around the core. The response of a well-defined short calibration pulse can be used to calibrate the device. Even drifts can be compensated for by sampling of the calibration signal just before or after the passage of the beam.



(a) (b)

Fig. 2: (a) An open inductive current transformer at DESY. The magnetic core (toroid) is split into two half to allow easy mounting around a ceramic gap without opening the vacuum. (Photo by N. Wentowski, DESY.) (b) Bunch trains in HERA measured by an inductive AC current monitor. Note the droop and recovery of the baseline in the presence of signals and in the bunch gap, respectively.

High peak currents can cause magnetic core saturation which might result in non-linear behaviour. Therefore, the choice of the core material and the design of the monitor have to fit the required bunch charge range. A dynamic range of $\approx 10^3$ and a resolution of 10^{-4} of full scale can typically be reached which is quite sufficient for measuring the variation in the bunch charge. Since the voltage output is proportional to the bunch charge only the peak voltage is of interest. The

acquisition rate is the bunch repetition rate; maybe twice the rate to obtain a value between two bunches for baseline restoration. High dynamic range (12–14 bit) and high bandwidth ADCs are commercially available with sampling rates up to 100 MHz, which are in most cases sufficient for the required resolution and dynamic range. In circular machines the resolution can be improved by averaging the acquired bunch current over many turns, but taking into account the lifetime of the beam.

2.3 Direct current transformers

The integration of any alternating current transformer (ACT) monitor signal over an infinite period is always zero. Precise active baseline restoration may be used to get a DC value of the beam but due to the fact that the baseline slope and level depends on previous beam bunches (several transformer time constants), a precise measurement is difficult. Direct current transformers (DCTs) are used to measure the DC component of a bunched or unbunched beam with high precision and with a dynamic range of $>10^6$. The high dynamic range is required due to the fact that a commissioning of a circular accelerator might be done with a pilot bunch only while the design allows some orders of magnitude more bunches. Sensitivities as low as 0.5 μA exist [13] which is sufficient for low current commissioning. Obviously a DC beam current measurement does not make sense in short pulsed machines such as linacs (except CW linacs) or transport lines, but it is the only device that can measure the beam current of an unbunched or coasted beam in a circular machine.

The principle of a DCT (also called a DCCT, PCT or zero-flux current transformer) relies on a pair of identical toroids with high permeability. They are excited in an opposite direction into saturation by a common AC current (or voltage) into individual windings. Careful matching of the toroids and the exciting current is necessary. A common sense winding picks up the resulting induced current which is zero in the case of a perfect matching. A charged beam crossing the two coils drives one of the two out of saturation which leads to a modulated current in the sense of winding with a frequency of the second harmonic of the exciting frequency. This current is then proportional to the DC beam current. It can either be measured by synchronous detection or more often by a feedback loop which prevents any magnetic flux change in the cores [14, 15]. This increases the useful dynamic range to more than six decades and reduces the recovery time of the device. The bandwidth of such a DCT is limited from DC to some tens of kilohertz. A further reduction in bandwidth is often useful to reduce the low-frequency noise and to extend the resolution. If even more dynamic range is needed the only (costly) solution is then to use two DCTs with different ranges.

Some issues of DCTs are addressed in the following [16]:

- Higher harmonics in the output lead to ripple which needs to be suppressed [17].
- Temperature drifts induce a drift of the baseline. Good temperature stabilization and/or a frequent measurement of the offset in the absence of the beam are recommended.
- HOMs in the gap may induce heating; therefore, water cooling might be necessary. Care has to be taken during vacuum bake-out not to exceed a core temperature above about 60 °C to avoid damage to the core.
- A calibration winding enables an absolute determination of the beam current.
- A DCT is quite sensitive to external noise and especially to external magnetic fields. Therefore a good electrical and magnetic shielding is essential. The magnetic shield should extend along the vacuum chamber with a length of at least twice the diameter of the beam pipe and without any gaps. It should have the highest possible μ, but should not saturate. It is also necessary to shunt all external currents away from the monitor to enable it to measure the beam current only.

Application issues include the following:

- Note that the DC component of a non-relativistic beam in a circular accelerator increases with the real acceleration of the beam particles while the bunch charge remains constant.

- Beam lifetime determination in storage rings is often done by high-precision DCTs [18, 19].
- A method to determine the amount of coasting beam in a storage ring is provided by a comparison of the ACT and DCT monitors. In the absence of a coasting beam (e.g. during or immediately after acceleration) the sum of all individual bunch currents should be equal to the DC component of the beam. In fact, this can be used to calibrate the monitors. An increasing difference between the two monitors indicates an increase of coasting beam in the machine [20].

2.4 Bunch shape monitor

The longitudinal charge distribution of a highly relativistic bunch (sometimes called a "bunch shape" or often a "bunch length") can be determined by a high-bandwidth WCM (see Section 2.1). For low-β beams the electromagnetic field is not purely transversal and hence does not represent the charge distribution. Hence, a bunch shape monitor (BSM) based on secondary emission is more adequate to image the real distribution in this case. It was originally developed in Ref. [21], developed further in Ref. [22] and is now used in many proton, H^- and ion linacs. The monitor based on a coherent transformation of the temporal structure of the bunch into one of the secondary electrons and then into their spatial distribution. Figure 3 shows its principle: parts of the beam hit a metal wire target (typically tungsten) in the beam pipe (see "1" in Fig. 3). The wire emits low-energy secondary electrons of a few electronvolts. Since this process does not have a significant delay, the temporal structure of the electrons now represents that of the bunch. The electrons are accelerated radially away by a negative bias voltage ($U_{targ} \approx -10$ kV) on the wire. A fraction of the electrons passes a collimator (see "2" in Fig. 3), an electrostatic lens and a varying RF field of a deflector (see "3" in Fig. 3; $U_{RF} = A \cdot \cos(n\omega t + \Phi)$). The RF is a multiple n of the acceleration frequency of the beam and is synchronized in time. The transit time of the electron bunch should be somewhat shorter than half of the wavelength of the RF. Depending on the arrival of the electrons with respect to the phase Φ, the electrons received transversal kick so that their longitudinal distribution is transformed into a spatial distribution after some distance (see "4" in Fig. 3). At that point the distribution can be measured by various detectors, e.g.:

- phosphor screen + CCD [23];
- multichannel electron detector [24];
- scanning phase + stable slit (or stable phase + scanning slit) + collector [25].

In the special case of a H^- beam, the detached electrons originated by dissociation of the H^- ions on the target wire (of initial energy of some kiloelectronvolts) contribute to some background [26]. Energy separation by an additional spectrometer behind the second slit (see "5" reduces the background down to better than 10^{-5} of the maximum. This enables measurements of longitudinal halo with a dynamic range of 10^5 [27]; see also Fig. 4.

The resolution of a BSM is defined by some factors [28, 29]:

(1) the time uncertainty of the secondary electron emission process is far below some picoseconds;
(2) the velocity and angular spread of the secondary electrons is minimized by a high bias voltage of –10 kV on the target and a short distance to the first collimator (≈1 ps);
(3) the RF deflector might need a sufficient strength of up to some 1000 V/cm to get an image on the detector with a resolution of ≈1 ps;
(4) the slit size of the first collimator (the spot size on the detector without RF) should be small compared with the transversal dimension on the detector; this effect can be measured and subtracted;
(5) the phase stability of the RF and of the synchronization can be kept much below 1° of the RF phase;
(6) non-linear effects such as space charge and lens aberrations are assumed to be negligible.

Typical resolution achieved so far lies in the order of some 10 ps.

A BSM for measuring all three dimensions of the bunch is used in Ref. [24]. Here additional the target wire is scanned across the beam and the slit of the first collimator is scanned along the wire. The intensity distributions versus the scanning positions give the two transversal distributions of the beam. A translation of the whole BSM along the beam axis enables the measurement of a phase difference $\Delta\Phi$ between the two locations and therefore a velocity measurement of the beam particles becomes possible [30]. A translation distance of a few centimetres is sufficient to determine velocities of up to $\beta \approx 10\%$.

The energy deposition of the beam in the target implies two problems: first the creation of thermal electrons and second the melting of the target. Thermal electrons blind the whole monitor. Therefore, care has to be taken in positioning the target into the beam centre: an off-centre measurement increases the lifetime of the target. To overcome this problem the electrons created by the ionization of the residual gas are used for high-intensity beams [31]. Since the electrons have a broad spectrum of energies an electrostatic analyser is located just after the first collimator to ensure a mono-energetic beam at the RF deflector.

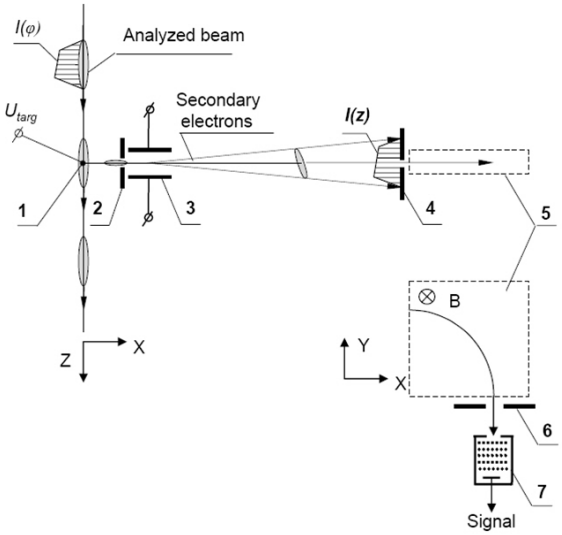


Fig. 3: Configuration of a BSM: 1, target wire; 2, input collimator; 3, RF deflector combined with an electrostatic lens; 4, output collimator or screen; 5, detector; consists here of a bending magnet; 6, collimator; 7, secondary electron multiplier (reproduced from Ref. [25])

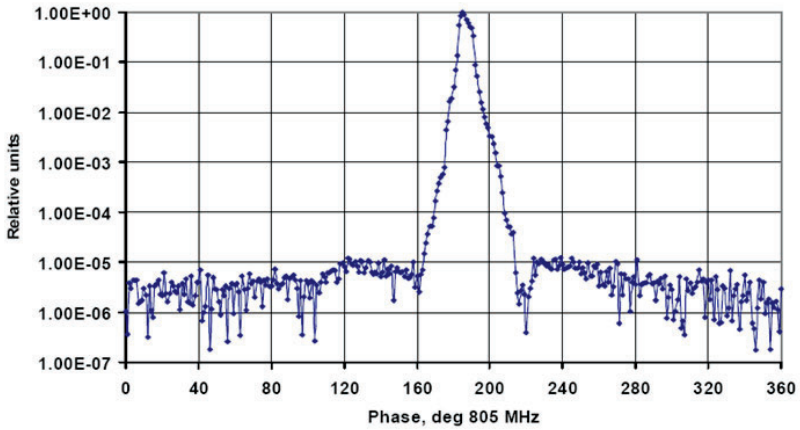


Fig. 4: Longitudinal bunch shape at SNS measured by a BSM with a high dynamic range (reproduced from Ref. [25])

2.5 Faraday cup

A Faraday cup (FC) is typically a fully destructive device which can be driven into the beam and in which the beam is completely absorbed. A full absorption of the beam enables an absolute determination of the beam charge; therefore, a FC is often used to re- and cross-calibrate the non-destructive current monitors [32]. Since the FC has to collect the whole charge of the beam, no charges must escape the cup:

- It has to be large enough to avoid any leakage of shower and multiple scattered particles. Therefore, its use is restricted to low-energy beams, otherwise its dimension became very large [32]. Hadron beam energy should be kept below about 150 MeV to stay below the π-production threshold.
- Backscattered particles have to be collected.
- Secondary and thermal electrons should not escape the FC.

The last two subjects are solved by additional negative voltage on a repeller electrode of approximately –100 V at the entrance of the cup, sometimes in combination with a magnetic field perpendicular to the incoming beam [33].

FCs (like most intercepting devices) in high-intensity beams need water cooling of the collector to take away the beam power of $P_{Beam} = E_{Beam} \times I_{Beam}$. Note that the water cooling only dissipates the average beam power. For pulsed beams the highest temperature is reached close to the penetration depth of the impact. This drives the design of the electrodes. Special shaped electrodes are necessary for intense low-energy ion beams. These ions have a very short penetration depth in materials which lead to a very dense energy deposition. Therefore a V- or saw-tooth-shaped electrode with a large inclination against the beam axis is needed to distribute the impact over a larger surface [34].

A good isolation of several gigaohms of the electrode offers a large dynamic range of beam charge measurements. Care has to be taken not to deteriorate the isolation (and the de-ionized water of the cooling) by radiation. A good vacuum is essential in the FC to avoid extra accumulated charges due to ionization for the residual gas molecules. An absolute accuracy of better than 1 % can be reached.

A high-bandwidth FC enables a measurement of the longitudinal bunch shape also for low-β beams. A careful design of the collecting electrode is necessary to achieve sufficient bandwidth. A bandwidth of some gigahertz was measured with coaxial and stripline types of electrodes in Ref. [35]. In the case of low-β beams the advanced electrical field of the bunch has to be considered. A grid in front of the electrode is required to shield the cup from this effect [36].

2.6 Beam position monitors

The fundamental principle of a BPM is to measure the transversal centre of the electromagnetic field of the beam with respect to the vacuum chamber wall. There are two ways of coupling on the electromagnetic field, by capacitive pick-ups and by inductive pick-ups. Inductive BPMs consist typically of thin loops with their open area parallel to the beam so that the magnetic field of the beam couples into the loop and induces a current (see e.g. [37]). The coupling to the beam is inductive for a thin loop and capacitive for a wide one. Nearly all modern hadron accelerators are using nowadays capacitive pick-ups which are therefore discussed in more detail following the discussions in Refs. [38, 39]. To obtain sufficient information on the beam position, the difference of the field amplitude in up–down and left–right orientation have to be measured by the pick-ups and analysed by the readout electronics.

The electromagnetic field of the beam induces an image charge on a metallic plate which is inside the vacuum chamber and insulated. For the following discussion only one plate is considered, but it is true for all four plates of a BPM. Assuming a bunched beam, the image current $I_{im}(t)$ is driven by the beam charge $Q(t)$:

$$I_{im}(t) = \frac{A}{2\pi dl} \cdot \frac{dQ(t)}{dt}$$

with $A = \pi r^2$ is the area of the electrode, d its distance to the beam centre and l its length. Here dQ/dt depends on the beam current $I_{beam}(t)$:

$$\frac{dQ(t)}{dt} = \frac{l}{\beta c}\frac{dI_{beam}(t)}{dt} = \frac{l}{\beta c} \cdot i\omega I_{beam}(t) \quad with\ I_{beam}(t) = I_0 e^{i\omega t}$$

where β is the beam velocity. To calculate the voltage drop across a resistor R, the capacity C between the plate and the grounded vacuum chamber has to be taken into account, therefore the impedance Z_{plate} of the plate is

$$Z_{plate} = \frac{R}{1 + Ri\omega C}$$

and the voltage U becomes

$$U_{im}(\omega, t) = Z_{plate}(\omega) \cdot I_{im}(t) = Z_t(\beta, \omega) \cdot I_{beam}(t)$$

while Z_t is the "transfer impedance" derived from the above:

$$Z_t(\omega) = \frac{A}{2\pi d} \cdot \frac{i\omega}{\beta c} \cdot \frac{R}{1 + Ri\omega C}$$

This impedance has a high pass characteristic which is shown in Fig. 5. Since high-frequency signals are typically transported via 50 Ω coaxial cables, a low input impedance of the first amplifier is often used. The frequency ω depends on the bunch length, which is respectively dependent on the RF frequency of the accelerator. Therefore, the coupling of a capacitive pick-up to the long bunches (e.g. in the beginning of a hadron accelerator chain) is very week. For frequencies $\omega << \omega/RC = \omega_{cut}$ the measured voltage $U_{im}(t)$ becomes

$$U_{im}(t) = \frac{R}{\beta c} \cdot \frac{A}{2\pi d} \cdot \frac{dI_{beam}(t)}{dt}$$

where $dI_{beam}(t)/dt = i\omega I_{beam}$ is the derivative of the bunch length. For $\omega >> \omega/RC = \omega_{cut}$ the voltage U_{im} follows the bunch length by

$$U_{im}(t) = \frac{1}{\beta c} \cdot \frac{1}{C} \cdot \frac{A}{2\pi d} \cdot I_{beam}(t)$$

In practice, this means that button-type pick-ups with about 100 mm^2 < A < 500 mm^2 are used only in high-energy hadron accelerators, since their coupling to the beam is strong due to the short bunch length and the high β (see Ref. [40]).[2]

For low-energy beams, large bunch length and sometimes large apertures d, the common way for a sufficient signal is to increase the size A of the pick-up and to use high-impedance amplifiers in

[2] The same argument applies to nearly all electron accelerators as well.

the readout. An example is a shoe-box type of BPM which is shown in Fig. 6. It has a large aperture but also large size electrodes which are separated diagonally with respect to the beam. Therefore, the induced voltage on both plates is proportional to the length of the beam projection on the electrodes. These types of BPMs are very linear over nearly their whole aperture [41]. Since they are used for long bunches a high-impedance readout at low frequency can be performed to obtain a useful readout voltage.

Stripline types of pick-ups are used in the case where the bunch length is shorter or about the length of the electrode. The electrode of length l forms a wide loop or transmission line between the electrode and the wall of the vacuum chamber. A signal is created by the beam on each end of the line which depends on the characteristic impedance Z_{strip} of the electrode, often $Z_{strip} = 50\ \Omega$. Depending on the termination R of the downstream port the signal there is cancelled ($R = Z_{strip}$) or appears partially ($R \neq Z_{strip}$). A complete cancellation at the downstream port happens only if the speed of the beam is equal to the speed of the signal in the transmission line which is almost true for $\beta \approx 1$. In this case a stripline is known as a "directional coupler" since the signal on one port depends on the beam direction. Such a BPM can be used to separate the beam positions of two counter-rotating beams in the same beam pipe [42]. The upstream port always includes the induced signal and the reflected inverted signal separated in time by $\Delta t = 2 \cdot l/c$ (for $\beta = 1$). The characteristic frequency of a stripline signal is defined by Δt and is $\omega_{strip} = 2 \cdot \pi \cdot 2 \cdot c/2l$. Assuming a short bunch, the Fourier transformation of the response is the transfer impedance $Z_t(\omega)$ [38]:

$$Z_t(\omega) = Z_{Strip} \cdot \frac{\alpha}{2\pi} \cdot \sin(\omega l / c) \cdot e^{i(\pi/2 - \omega l/c)}$$

where α is the azimuthal coverage angle (width of the electrode). Here $Z_t(\omega)$ shows a maximum response at a bunch repetition frequency of $\omega = \omega_{RF} = n \cdot \omega_{strip}/4$ and zero response at $\omega = \omega_{RF} = n \cdot \omega_{strip}/2$; $n = 1, 2, 3, \ldots$ (see Fig. 7). Therefore the optimum length of the stripline electrode is $l = \lambda_{RF}/4$. If, on the other hand, the bunch length exceeds the length of the electrode, partial cancellation occurs at the upstream port which reduces the signal. This is also valid for $\beta << 1$, since the field of the bunch is no longer a pure TEM wave. Note that therefore also the coupling to the four pick-up electrodes became strongly non-linear with the beam position [43, 44].

The beam position of one plane is derived from the difference of the signal of the two opposite electrodes[3] and after applying corrections to the non-linear position response of the BPM. The signals are also intensity dependent; therefore, normalization to the intensity is always necessary. There are various electronic concepts in use which are discussed in detail in Ref. [45]. Their sensitivity has to be as low as the minimum expected bunch charge to produce position readings with the required accuracy and resolution. Their dynamic range is defined by the various intensity conditions, e.g. acceleration of proton and ion beams. A dynamic range of several decades might be necessary in such a case. A bunch-by-bunch measurement of the beam position requires a broadband signal processing. In this case the dynamic range is defined by the variation in bunch charge only (plus position and β variations). An orbit measurement requires a narrowband signal processing with a maximum update rate of the revolution frequency. In this case the required dynamic range is defined by the stored or accelerated current which might also include a variable number of bunches, but due to the lower speed of the signal processing there are electronic components (e.g. ADCs) available which have a very high dynamic range.

A high bandwidth offers some advantages, e.g. a measurement of the individual bunch current by summing up all four signals from the electrodes (to be position independent). In particular, during commissioning of an accelerator this feature becomes important in detecting the positions of beam

[3] Assuming orthogonal electrode arrangement, in electron accelerators with synchrotron radiation the arrangement might be different to avoid direct irradiation of the plates.

losses. For very high bandwidth readout the bunch length can be resolved by a BPM. At even higher bandwidth a position variation along the bunch can be observed (head–tail modes) [46].

The BPM system resolves the beam position along the accelerator (= orbit). The BPMs are best located at local maxima of the β function (close to focusing quadrupoles in both planes) to obtain maximum sensitivity. This already implies two beam position measurements per betatron wavelength with a four electrode BPM. This is typically sufficient for a useful orbit distortion measurement. Often some BPMs are added at critical regions such as high dispersion, injection, extraction or collimation. In high-intensity beams the beam position excursions have to be carefully monitored and controlled. Too high excursions should cause an alarm signal to the machine protection system to avoid damage to components due to a mis-steered high-intensity beam.

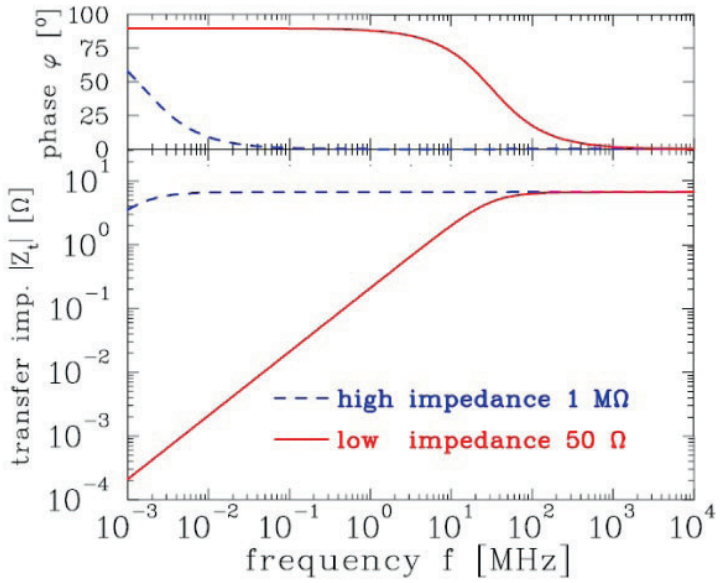


Fig. 5: Absolute value and phase of the transfer impedance for a l = 10 cm long round pick-up with a capacitance of C = 100 pF and an ion velocity of β = 50 % for high (1 MΩ) and low (50 Ω) input impedance of the amplifier. The parameters are typical for a proton/heavy ion synchrotron. (Courtesy of P. Forck [33].)

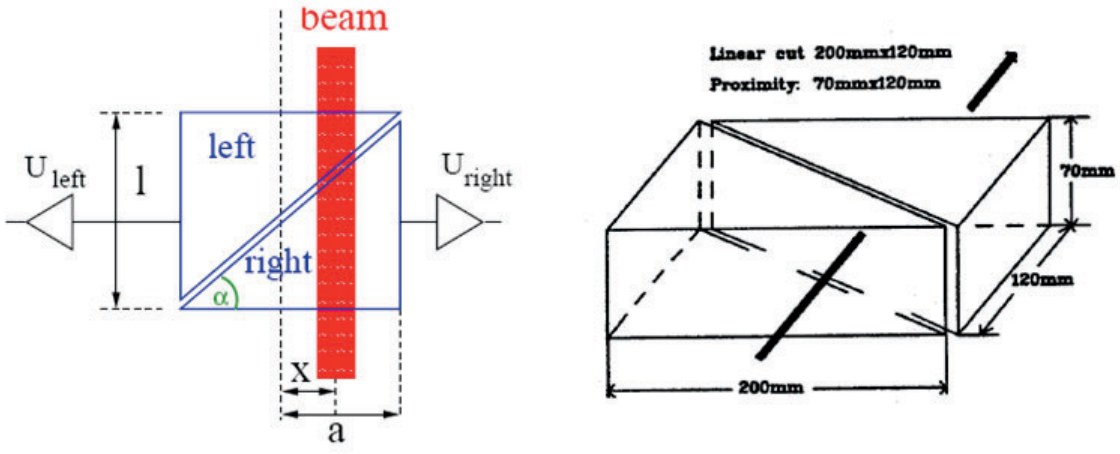


Fig. 6: Scheme of the position measurement using a shoebox BPM with linear cut and an example of an electrode arrangement for the horizontal plane (courtesy of P. Forck [33])

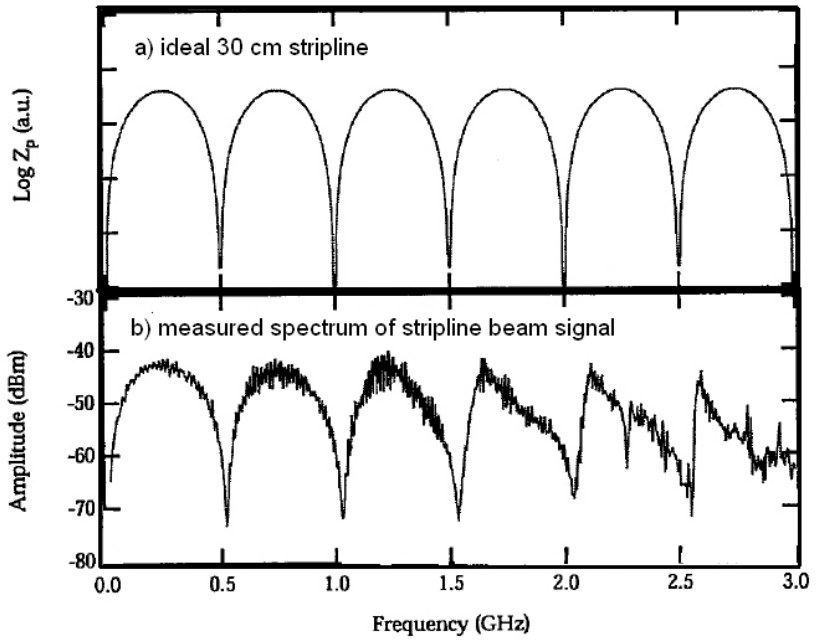


Fig. 7: Calculated transfer impedance of (a) an ideal 30 cm stripline and (b) a measurement with a spectrum analyser for a single bunch signal [47].

3 Measurement of beam emittance

The transversal emittance ε of a particle beam is used to describe the size of the beam in *x* and *y* direction as well as the angular distribution of the particles in the beam in *x*' and *y*' directions. The emittance is an ellipse in the *x*, *x*' or *y*, *y*' plane[4] and its equation can be written as

$$\varepsilon_x = \gamma(s)x^2 \cdot 2\alpha(s)xx' \cdot \beta(s)x'^2$$

where $\alpha(s)$, $\beta(s)$ and $\gamma(s)$ are the Twiss parameters of the respective plane at the position *s* and $\beta \cdot \gamma - \alpha = 1$. The emittance for Gaussian beam distributions can be expressed by the "Sigma matrix" σ:

$$\sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \varepsilon \cdot \begin{pmatrix} \beta & -\alpha \\ -\alpha & \gamma \end{pmatrix} \quad \text{with} \quad \varepsilon = \sqrt{\det \sigma} = \sqrt{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$$

In the geometrical interpretation of the emittance ellipse 39% of the particles lie in the area of $\pi \cdot \varepsilon$, therefore the typical unit of the emittance is written in π mm mrad. The root mean square (rms) emittance of any phase distribution in *x*, *x*' can be calculated by

$$\varepsilon_{rms}(x,x') = \sqrt{\langle x^2 \rangle \langle x'^2 \rangle - \langle xx' \rangle^2} \,.$$

The matrix element σ_{11} is related to the beam size $\sigma_x(s)$ by

$$\sigma_{11} = \sigma_x^2(s) = \varepsilon_x \cdot \beta_x(s) \,.$$

[4] In the following *x*, *x*' is used for the *x* or *y* phase space.

Therefore, the emittance of a particle beam can be determined by measuring the beam size σ_x, but note that the dispersion $D(s)$ of the beam also contributes to the beam size. This part has to be quadratically subtracted by

$$\varepsilon_x = \frac{1}{\beta_x(s)}\left[\sigma_x^2(s) - \left(D_x(s)\frac{\Delta p}{p}\right)^2\right]$$

where $\Delta p/p$ is the momentum spread and $\beta(s)$, $D(s)$ and $\Delta p/p$ can be determined sufficiently precisely in the case of circular accelerators by common diagnostic methods. Therefore, a measurement of the beam profile is often sufficient to determine the beam emittance. Minimally invasive instruments are needed to avoid destroying the circulating beam. These instruments are discussed in Section 4. In linear machines and transport lines the optic parameters depend on the incoming beam and the emittance has to be determined by measuring all elements of the σ-matrix or $\langle x \rangle$ and $\langle x' \rangle$. Methods and instruments for this purpose are discussed in this section. Owing to their nature of measuring the angular and spatial beam distributions at the same time, these types of instruments are fully destructive to the beam. They cannot be used in full-power high-intensity beams. They are installed in nearly all machines, however, to enable a measurement of the optic parameters at low current to tune the machine and to prove its health.

3.1 Screens and harps

Most typical devices for emittance measurement use fluorescent screens or SEM harps as detectors, therefore a brief introduction is given first. Note that both types are intercepting devices with cannot be used at full beam intensities. Also note that the energy deposition of heavy ions at very low energies is at a maximum and the penetration depth is very small. Therefore, the energy is deposited in a very small volume creating extreme hot spots in the material.

3.1.1 Screens

The observation of beam profiles on fluorescent or scintillation screens is one of the oldest and most common diagnostic techniques. Together with modern TV cameras and imaging processing this technique offers simplicity, reliability and high resolution. The resolution is limited by the grain size of the material and by optical effects, mainly due to the depth of field when viewing the screen under 45°, but the thickness of the material itself also leads to light collection in the depth which disturbs the image [48]. A thin phosphor layer (e.g. P43 or P46) with a small grain size (e.g. 5 μm) reduces this effect. Quite a number of different materials exist which are used as viewing screens in particle beams [49]. The sensitivity of the materials covers more than four orders of magnitude and decay times from nanoseconds to seconds are possible. Table 2 gives an overview of some common materials.

Table 2: Screen sensitivities for hadron beams, from Refs. [49–51]

Material	Relative sensitivity	Decay time (10%)	Maximum emission
Al_2O_3:Cr	1	> 20 ms	700 nm
CsI:Tl	≈200	900 ns	550 nm
Li Glass:Ce	≈0.05	100 ns	400 nm
Quartz SiO_2	≈0.005	1 ns to 10 ns	600 nm
P43 (Gd_2O_2S:Tb)	≈1.5	1 ms	545 nm
P46 ($Y_3Al_5O_{12}$:Ce)	≈0.2	300 ns	530 nm
ZnS:Ag	≈0.1	200 ns	450 nm

The major problem of viewing screens systems is the radiation damage of the components, the screen and the camera. Studies in Refs. [51, 52] showed that the light yield of some materials depends strongly on the integrated charge on the screen. In addition to the decrease of light it also results in broadening of the image due to the non-linear response of the screen. At high beam currents the temperature of the screen increases at the beam spot due to energy deposition in the material. Some dependence of the light yield on the temperature is reported[5] [51], also resulting in a non-linear response. Additional care has to be taken due to possible saturation of the light yield inside the material at high beam intensities [53, 54]. Unfortunately for the moment there is no clear recommendation for "the best" material for the use in high current applications and further studies are still needed.

For high-energy beams the use of optical transition radiation (OTR) became reasonable [55]. The thickness of OTR screens can be about one order of magnitude less than conventional screens. OTR is linear over a large range and it is very fast. Its use for bunch length measurements is limited due to the weak signal for hadrons. Some aging of the signal has been observed after 10^{19} protons, but the signal was still useable [56]. Many more details can be found at the Workshop of Scintillating Screen Applications in Beam Diagnostics [57].

CCD cameras suffer from radiation damage in radiation fields like most semiconductor devices. The screen itself is a source of scattering of beam particles and nuclear interactions. The resulting radiation might consist of energetic α, β, γ and neutrons which induce ionizing (electron–hole creation) and non-ionizing (e.g. displacement of atoms) processes in the semiconductors. The consequence of these processes are permanent damage of the material resulting in various effects such as increasing dark current, change of bias voltage or even complete malfunctions. CCD cameras often reply with a degradation of contrast, blind CCD pixels or gain variations. CCD cameras might become unusable already after about 10–20 Gy (see Ref. [58]). Old-fashioned vidicon cameras were somewhat more radiation resistant but they hardly exist anymore. Radiation-resistant CCD cameras are useful up to some tens of kilograys. The dynamic range of a screen + CCD station is typically limited by the dynamic range of the camera or the image processing (typically 8 bit), as long as saturation of the screen can be avoided.

3.1.2 SEM harps

SEM harps consist of stretched metallic wires orientated perpendicular (horizontal or vertical) to the beam. Each individual wire is connected to an electrical vacuum feedthrough and an amplifier. The vertical oriented harp measures the horizontal beam profile and vice versa. Beam particles hitting a wire create secondary electrons of 20 eV to 100 eV from its surface. The secondary electron yield Y is described by the Sternglass formula [59]:

$$Y = 0.002 \text{ cm MeV}^{-1} \cdot \frac{dE}{dx}$$

where dE/dx is the stopping power of the particle in the material. The secondary electron emission (SEE) efficiency ε is defined as the ratio between the number of secondary electrons and the number of traversing particles. It varies between about 300 % for low-energy protons (e.g. 40 keV) to about 2 % for minimum ionizing particles for most common metals. The SEE current in each wire is proportional to the number of beam particles hitting the wire and is linear over many orders of magnitude. An appropriate profile harp consists of defined spaced wires of defined diameter and material.[6] The spacing and the diameter of the wire defines the resolution of the instrument. Typical

[5] Note the high-energy deposition of ion beams.
[6] Note that these parameters might depend on the position within the harp, e.g. for optimum measurements of the beam core and the tail.

values of both are between 20 μm and about 2 mm. Well-suited wire materials are tungsten and titanium due to their good mechanical and thermal properties, but carbon and aluminum are also used. Using low-*Z* materials and thin foils instead of wires has the advantage of lower beam losses due to the interaction with the wire and less heating [60, 61]. Sometimes the wires are gold plated to improve the long-term stability of the SEE yield [62] or CsI plated to increase the SEE yield [63]. A negative bias voltage on the wires [62] or positive collection electrodes [64] avoids recollection of the secondary electrons by the same or another wire. The dynamic range of a SEM harp is limited by the electronic noise on the low beam current end and by thermal electron emission due to heating of the wire at the high beam current end. High dynamic range signal processing can be done by logarithmic amplifiers with a dynamic range of 10^7 (see Ref. [65]) or by selectable gain amplification [66]. A parallel readout of all wires enables a profile measurement of a single passage. Special care has to be taken for high brilliance beams to avoid too much heating of the wires. In addition to thermal electron emission, the wire itself or its support (e.g. soldering) may melt and the elongation of the wires changes with heat [67]. After a large integrated number of particles had crossed the grid some reduction, up to 50 %, in the secondary emission efficiency had been observed, especially for aluminum and gold-plated materials [68].

Only about 10% or less of the beam area is covered by the wires. Even the wires can be made of very thin strips of light material. Therefore, such a SEM harp (or a thin screen) is quite transparent for not too low-energy beams and successive harps or multiple beam passages are possible. A turn-by-turn profile measurement in a circular accelerator enables injection mismatch studies by observing beam width oscillations [69, 70].

3.1.3 Emittance measurement by slit + screen/harp and pepperpot methods

A direct measurement of the beam emittance without knowing the Twiss parameters is possible in low-energy hadron beams. Those particles which have a penetration depth of a centimetre or less can be stopped by a simple metallic plate. A small transversal (either horizontal or vertical) slit of height h_{slit} in this plate selects a beamlet which shines on a screen [71] or harp [72] monitor after a drift distance d. The width of the measured beamlet is defined by the divergence of the beam x' at the slit position x_n, on the distance d and on the resolution of the system defined by h_{slit} and the resolution of the screen/harp device. The height of the signal depends on the bunch current I_0 and on the current distribution in the bunch (the beam profile). The slit is scanned across the beam profile and for each position x_n a beamlet distribution $I(x_n, x')$ is collected. The two radii $r_{1,2}$ of the distribution give the angular spread of the beam at the position x_n by

$$x'_{1,2} = r_{1,2} / d$$

The radii of the distribution are defined by the amount of current included in the emittance ellipse. This is illustrated in Fig. 8. For each x the corresponding $x'_{1,2}$ are plotted in a contour plot where the included area A_x is then the emittance of the corresponding current level (see Fig. 9):

$$\varepsilon_x(x,x') = 1/\pi \int dx\, dx' = 1/\pi \cdot A_x$$

The rms emittance of the incident beam

$$\varepsilon_{rms}(x,x') = \sqrt{\langle x^2 \rangle \langle x'^2 \rangle - \langle xx' \rangle^2}$$

can be calculated from the measurements by a method described in Ref. [73]

$$\varepsilon_{rms}^2 \approx \frac{1}{N^2}\cdot\left[\sum_{j=1}^{p} n_j\left(x_{nj}-\bar{x}\right)^2\right]\cdot\left[\sum_{j=1}^{p}\left(n_j\cdot\left(\frac{\sigma_j}{d}\right)^2+n_j(\bar{x}'_j-\bar{x}')^2\right)\right]$$
$$-\frac{1}{N^2}\left[\sum_{j=1}^{p} n_j x_{nj}\,\bar{x}'_j - N\bar{x}\bar{x}'\right]^2$$

where N is the total number of particles crossing the slit during a scan, x_{nj} is the jth slit position, p is the number of different slit positions, n_j is the number of particles passing the though the jth slit position (proportional to the signal intensity), $\bar{x}$ is the mean position of all beamlets, $\bar{x}'$ is the mean divergence of all beamlets, $\bar{x}'_j$ is the mean divergence of the jth beamlet and σ_j is the rms divergence of the jth beamlet. The same formalism is true for the y-plane. In this case the slit has a perpendicular orientation.

The procedures described above neglect the finite resolution of the device. A small h_{slit} and small step sizes of the slit positioning (even when using more than one slit) are necessary to improve the resolution. When using a grid as detector its resolution can be improved by scanning the grid as well. A position of the device in a region with high x' is most helpful for a good resolution. In high-density beams care has to be taken that space charge in a beamlet might extend the angular spread of it on the way to the detector. Some recent studies of the influence of the slit geometry can be found in Ref. [74].

The Allison Scanner uses a FC instead of a position sensitive detector behind the slit, while the angle distribution x' is measured by an electric sweep. It is limited to small energy ion beams (<100 keV) due to the limit in sweeping voltage.

The scanning of the whole beam size in both planes needs quite a lot of time with very stable beam conditions. To overcome the scanning procedure a plate with thin holes in defined distances (pepperpot) produces a lot of beamlets in the x, y plane (see Fig. 10 and Ref. [75]). Therefore, the whole phase space can be measured with one single shot onto the pepperpot mask (see Fig. 10). This will reduce the heat load of the plate drastically with respect to slowly scanning slits.

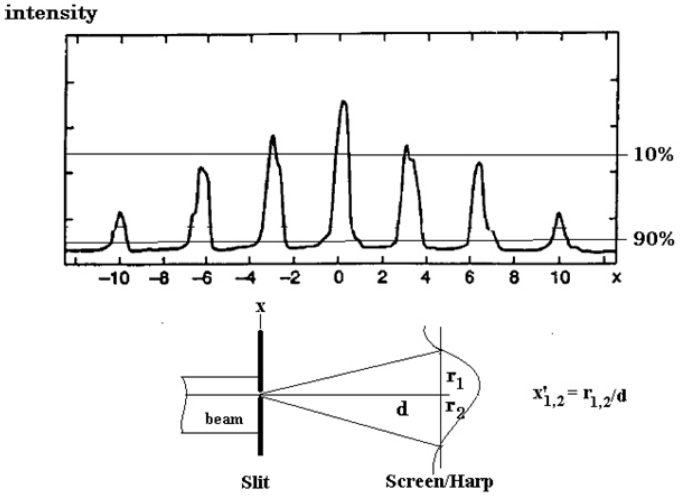


Fig. 8: Beamlet distributions at seven different x positions of the slit. The horizontal lines indicate approximately the 10% and 90% level of the beam current taken into account for the emittance calculation. Here $x'_{1,2}$ is then defined by the corresponding radius $r_{1,2}/d$.

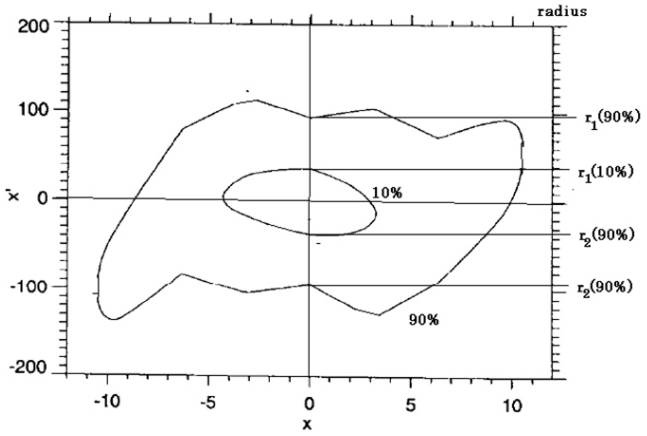


Fig. 9: Contour plot of the x, x' phase space. The measured radii at a slit position of $x = 0$ are indicated on the right-hand side.

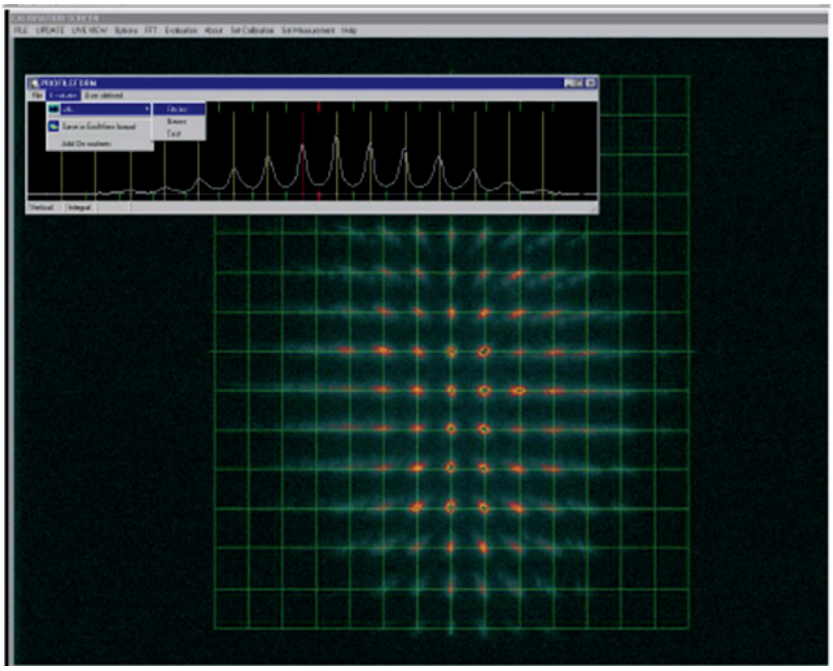


Fig. 10: Screen shot from the pepper-pot device for an Ar beam and, as an insert, the projection onto the horizontal plane (reproduced from Ref. [75])

3.2 Emittance measurements by quadrupole variation or three screens/harps methods

In a beam transport system the sigma matrix σ is transformed from one point s_0 to another s_1 by

$$\sigma(s_1) = M\sigma(s_0)M^{\mathrm{T}}$$

where M and M^{T} are the transport matrix between the two points and its transpose, respectively. In a dispersion-free beam transport all are 2 × 2 matrices and the measured width $\sigma_{measured}$ at (s_1) is then

$$\sigma^2_{measured}(s_1) = \sigma_{11}(s_1) = M^2_{11}\sigma_{11}(s_0) + 2 \cdot M_{11}M_{22}\sigma_{12}(s_0) + M^2_{12}\sigma_{22}(s_0)$$

The matrix elements M_{ij} are known by the elements of the transport system between the two points. The three unknown elements of the $\sigma_{ij}(s_0)$ matrix can now be resolved by at least three different measurements, either by three profiles at three different locations in the transport system or

by three profiles with three different transport optic settings, e.g. by variation of the focusing strength of a quadrupole. The matrix elements M_{ij} of different measurements have to differ significantly to solve the linear system, therefore the different locations have to have enough phase advances or the focal strength of the quadrupole has to be varied over a large enough range.

Typically the beam profiles are measured by thin screens or harps. This has the advantage of small and almost negligible beam blow-up due to the measurement itself so that the beam can transverse a few screens. This enables a "one shot" measurement of even a single bunch while the scanning methods need a stable beam over the scanning time.

Three measurements give a unique solution but no error estimation. Therefore, more measurements, either by more stations or by more quadrupole settings are recommended. In particular, the variation of the quadrupole settings allows a quadratic fit of the square of the measured beam size versus the quadrupole gradient together with an estimate of the errors in the measurement [76].

The standard methods descript above are valid under the assumption that:

- the dispersion along the section is zero;
- the transfer matrices are known;
- no coupling is present between the two planes; and
- no space charge or other non-linear forces are present.

In particular, a dipole in the section between the monitors creates dispersion which has to be taken into account. In this case or with initial dispersion the particle trajectory vector then includes the momentum spread $\vec{x} = (x, x', \Delta p/p)$ and the σ matrix and the transport matrix M becomes a 3 × 3 matrix. In this case at least six measurements are necessary to determine all σ matrix elements to resolve the emittance. The general case is discussed in more detail in Ref. [77].

The influence of space charge in high-intensity beams can be measured by observing the emittance evaluation in dependence on the intensity. As discussed before, care has to be taken not to saturate or even destroy the monitor with the beam. A detailed discussion of the equations of motion and of the emittance in the presence of space charge can be found in Ref. [78].

A full reconstruction of the four-dimensional phase space (x, y, x', y') by tomographic image reconstruction of the spatial beam projections (profiles) has become a useful diagnostic tool over the last few years [79]. The quadrupole scanning technique or a set of screens along the transport line with sufficient phase advance delivers the input for the computerized tomography.

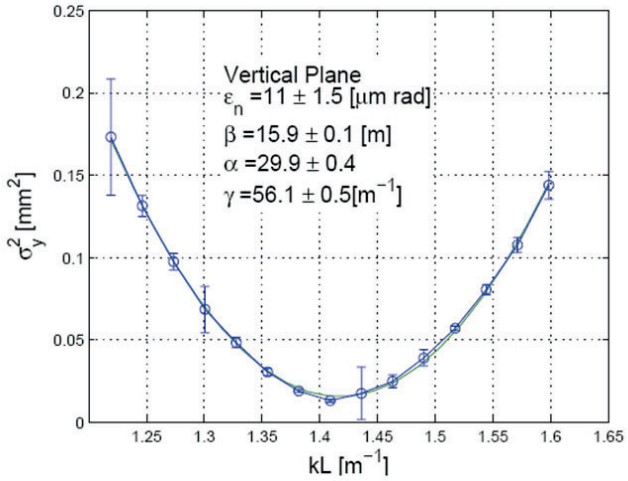


Fig. 11: Scan of the quadrupole gradient versus the measured beam width and fit to data (reproduced from Ref. [76])

3.3 Emittance measurement by tomography

Tomography is the technique of reconstructing an image from its projections; see Fig. 12. It is widely used in the medical community to observe the interior of the human body by processing multiple X-ray images taken at different angles. Beam phase space tomography reconstructs the phase space density distribution by means of one-dimensional profiles (projections) from beam profile monitors by means of a mathematical algorithm.

The main reconstruction algorithms used are [80]:

- convolution and back projection methods (FBP);
- maximum entropy (MENT) algorithm;
- maximum likelihood expectation maximization (MLEM);
- algebraic reconstruction techniques.

The "filtered back projection" or "convolution" reconstruction process is widely used because the mathematics is simple and easily programmable. For a small number of projections, however, streaking artefacts dominate the reconstructed image. The optimum algorithm depends strongly on the problem being solved. Some algorithms are better at reconstructing Gaussian distributions, whilst others are suited to detailed distributions

Some questions arise regarding the limitations of tomography technique for space charge dominated beams. The use of linear space charge forces led to inconsistent results [81].

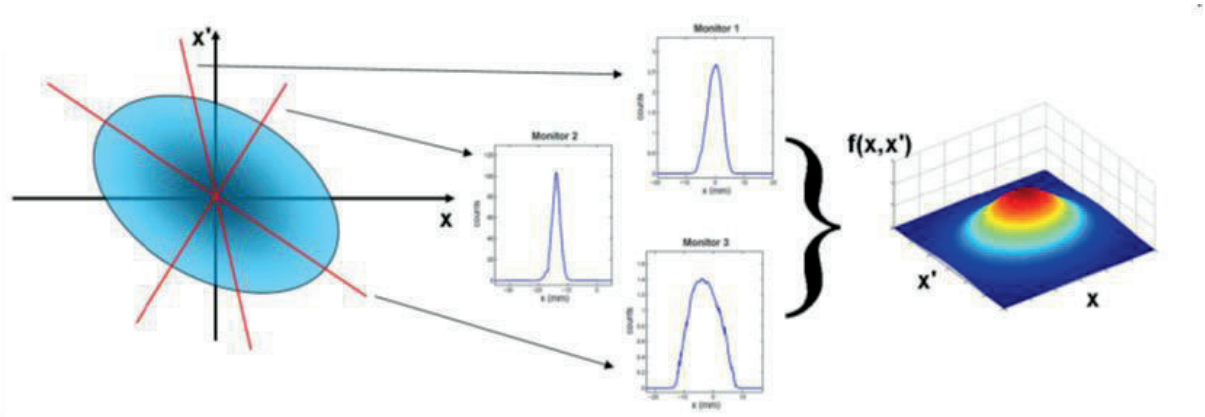


Fig. 12: The plot on the left-hand side represents the unknown beam transverse phase space distribution at the reference position $z = z_0$. Beam profile monitors acquire projections of the phase space onto the x coordinate at different locations with a certain phase advance (middle plots). These projections are related to the beam distribution at $z = z_0$ through linear transport matrices accounting for drift space and/or quadrupole magnets. In beam tomography the profile data are employed by a mathematical algorithm to reconstruct the two-dimensional beam density distribution (right plot). (Reproduced from [82].)

4 Instruments for beam profile measurements

The aim of transversal beam profile measurement is to determine the transverse shape of the beam down to about 3σ to 4σ. Further outside, halo measurement starts (see Section 6). Therefore, a dynamic range of 10^3 to 10^4 is sufficient for a single measurement. Additional constraints come from the requirement to measure the profile at different beam currents. Therefore, the profile monitor needs additional pre-gain settings to adapt to the current issue. The required spatial resolution depends on the beam size. Typical beam dimensions of hadron beams are in the millimetre range so that a resolution of 100 μm is sufficient in most cases.

The main motivation for beam profile measurements is to understand the beam dynamics in the machine and, in conjunction with that, to minimize beam losses along the accelerator (see also Section 5). There are many sources which can drive a blow-up of the beam core such as space charge, scattering, mismatch, resonances, etc., which can be observed by profile monitors. In a chain of successive accelerators profile monitors are indispensable to measure the (normalized) emittance evolvement at each step of the chain.

In contrast to the destructive emittance measurement (Section 3) the profile measurement needs to be minimally invasive for two reasons: (1) to avoid influencing the beam and (2) to avoid destroying the monitor.

4.1 Wire scanner

Wire scanners are used in many accelerators as a standard device for beam profile measurements. The device sweeps a thin wire through the beam while plotting a signal which is proportional with the number of particles interacting with the wire versus the measured position of the wire (see Fig. 13). Optical rulers can determine the position of the wire with a resolution of 1 μm to 2 μm, but only at a speed of ≤1 m s^{-1}. Higher speeds (e.g. 5 m s^{-1} [83] and up to 20 m s^{-1} [84, 85]) are required for intense and high brilliant beams in circular machines for two main reasons:

(1) Reducing the heat load of the wire due to the interaction with the beam; the heat load is inversely proportional to the speed [86].
(2) Reducing the emittance blow-up of the beam due to the wire interaction since the emittance blow-up is also inversely proportional to the speed of the wire [87].

High speed is realized by circular movement of the wire which reduces the position resolution and therefore the profile resolution to 10 μm to 100 μm. The speed of a linear wire scanner is mainly limited by the vacuum bellow stress property which limits the acceleration of the mechanical feedthrough to a few *g*. New methods for fast scanners with high resolution are under study [88].

Light materials with long radiation length are preferred to reduce the emittance blow-up and to minimize the energy deposition in the wire. On the other hand a high melting point is preferred to extend the lifetime of the wire. For that reason a thin (7 μm to 20 μm) carbon wire is often a good choice due to its high melting temperature of about 3500 °C and its excellent mechanical stability.

The main cooling processes are thermionic emission and black-body radiation. Both become important at temperatures well above 3000 °C (see Ref. [89]). This is true for the high duty cycle interaction in storage rings. At low duty cycles the heat transmission along the wire becomes dominant [90]. The calculation of the heating of the wire must include the effect of the emission of secondary particles such as delta rays and others. This reduces the amount of deposit energy in the wire by up to 70 % [86, 90], depending on the beam energy. Sublimation of the wire material takes place even before the melting temperature is reached, however, and reduces the material at each scan [89]. The heating of the wire often limits the use of wire scanners in high-intensity and high-brilliance beams to low currents only.

In linacs with low duty cycle a fast scan does not make sense since the bunch train (pulse) might be too short to allow a scan within one pulse.[7] Therefore, the wire has to crawl though the beam and the profile is acquired pulse by pulse. A few data points per 1σ beam width should be the minimum to obtain a useful profile. To avoid instantaneous overheating of the wire the charge of each pulse has to be limited, to avoid an integral overheating the repetition rate of the pulses has to be limited [90, 91]. In particular, low-energy ions will deposit huge amounts of energy even in thin wires, so that their use is very limited in such accelerators. The use of wire scanners for partially stripped ions is excluded since their interaction with the solid wire will change the charge state of the ions.

[7] Exception: Superconducting FEL Linacs might allow bunch trains of some hundred μs but with beam size of less than 100 μm

The signal from the beam–wire interaction can be detected with two different methods:

(3) Detection of scattered beam particles[8] outside the vacuum chamber. At energies above the pion threshold (>150 MeV) the beam particles mainly interact with the wire by multiple scattering and nuclear interactions. Beam and secondary particles with large scattering angles will hit the vacuum chamber and create a shower which is detected by fast loss monitors, e.g. scintillation counters. Monte Carlo studies are most helpful to find an efficient position for the detector somewhere downstream of the scanner. Note that the signal can depend on the wire position, especially when using asymmetric detector positions at large beam sizes [92]. A fast scintillation counter is able to resolve single bunches in a train or in a stored beam. While in a linac the beam profile is a composition of many (desirably similar) bunches, the profile of each individual bunch can be measured in a storage ring [93].

(4) Secondary electron emission current created by beam particles entering and leaving the conducting wire (see also Section 3). This method is often used in low-energy beams where the scattered particles cannot penetrate the vacuum pipe wall. In this low-energy regime the stopping power of the wire forces the hadron beam particle to stop in the wire, so that the signal is a composition of the stopped charge (in the case of H^-: proton and electrons) and the secondary emission coefficient. Therefore, the polarity of the signal may even change, depending on the beam energy and particle type [94, 95]. When using multiple wires on one scanner, too narrow wires may cross-talk by receiving the electrons from the other wire. If the temperature of the wire exceeds the thermionic threshold the emission of thermal electrons starts to superimpose the secondary electron emission signal. Therefore, the useable temperature range is limited by that threshold for the secondary electron emission method.

Since the signal generation during a scan is a sampling process, the beam should be quite stable during the scan. In case of linacs, the beam current and position have to be correlated with the signal for each bunch.

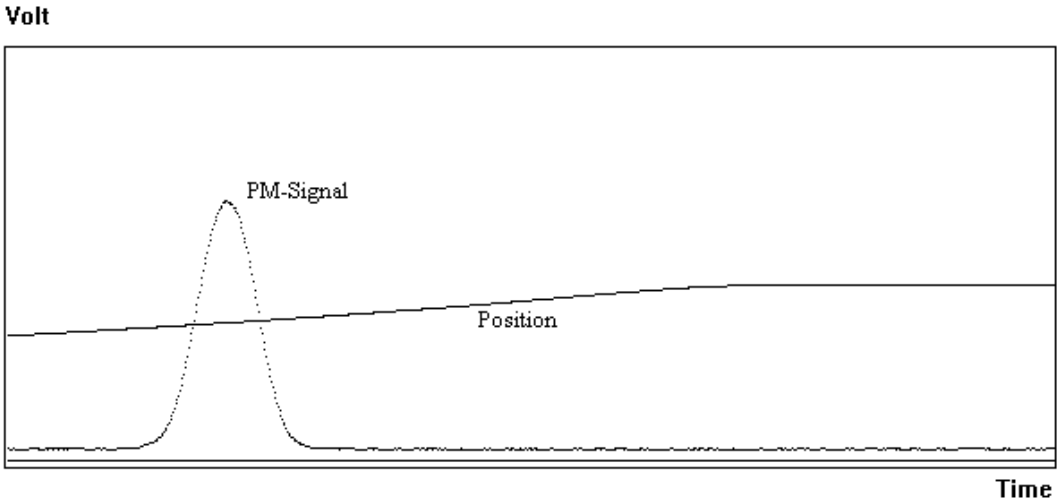


Fig. 13: Signal from a scintillation counter (PM-signal) and the measured position of the wire by a potentiometer versus the scan time. The real beam profile is a result of plotting the signal versus the position.

[8] or Bremsstrahlung in case of an electron beam

4.2 H^- laser scanner

The use of photodetachment of a H^- beam by laser photons was first used in Ref. [96] but to measure longitudinal H^- beam parameters. In Ref. [97] a well-focused laser beam was proposed to scan the intense H^- beam at the LAMPF accelerator. This nearly non-invasive method has the advantage neither to produce emittance blow-up nor intrinsic wire heating but it is applicable only for H^- beams. The cross-section for photodetachment of H^- ions is large enough (some 10^{-17} cm^2; see Ref. [98]) to neutralize a fraction of the beam-slice crossed by the laser. The number of photodetached electrons (and neutral H^0) is proportional to the beam density and the laser energy density. The cross-section has a maximum for photon wavelength around $\lambda = 1000$ nm (= 1.2 eV) so that the second electron (binding energy 13.6 eV) will not be stripped by those laser photons. The 1064 nm light from a Nd:YAG laser is very close to the optimum wavelength, but for relativistic H^- particles the Lorenz boost has to be taken into account which increases the photon energy in the rest frame of the H^- (E_{CM}) by

$$E_{CM} = \gamma\, E_{YAG}(1 - \beta \cos \Theta)$$

where Θ is the crossing angle between the laser and beam. For a H^- beam with E_{kin} = 1 GeV this reduces the cross-section to about 70 %, but the photon flux also receives a Lorenz boost in the same way keeping the photodetachment yield nearly constant $0.2 \leq E_{kin} \leq 1$ GeV. A detailed calculation of the photodetachment yield is discussed in Ref. [99].

A *Q*-switched Nd:YAG laser (up to few hundred millijoules) can be synchronized with the ion bunches. Since the laser pulse is typically much longer (some nanoseconds) than the bunches, an injection seeder is required to smooth the temporal laser pulse profile [100]. The bunch position, bunch current, laser shot-to-shot variations and drifts have to be monitored during the scan and normalized to the results. The laser focus has to be significantly smaller than the H^- beam size and its Raleigh length correspondingly larger to ensure a clean measurement. The laser is scanned across the beam by a motorized mirror system. Since a laser beam can be transported over long distances one laser can serve many scanning stations, e.g. the 9 stations along about 300 m at the SNS Linac are served by one laser [101].

Both, the liberated electrons and the remaining H^0 can be detected to measure the beam profile (see Fig. 14):

(1) The neutral H^0 reduce the bunch current. This is measured by a FCT while its amplitude is plotted versus the laser beam position.
(2) The liberated electrons are bent by a small dipole field into a FC.

Since the electron energy is only a few kiloelectronvolts a dipole field of 50–150 G is sufficient and its feedback on the H^- beam is quite small. To collect the electrons diffused by space charge a wide area FC is required, located downstream near the laser interaction point. A biased collector (≈200 V) with a repeller grid in front ensures suppression of background and secondary electron emission. A second electrode in front of the interaction point can be helpful to collect (background) electrons created by Lorentz or residual gas stripping. Beam losses near the monitor are the remaining source of background which should be avoided to get a high dynamic range of the measurement. The repeller grid is also used to measure the energy distribution of the electrons, which is a sum of their initial energy and the energy gained by the space charge of the beam [102].

The direct use of H^0 enables a direct emittance measurement using the laser as a "slit". After bending the H^- beam the neutral H^0 remain on a straight line where a screen or grid measures their distribution [103]. The laser "slit" has the advantage of avoiding the thermal problems that exist in conventional slit-grid monitors (see Section 3). To reduce the background of H^0 produced in front of the laser, Ref. [104] has the interaction with the laser in the middle of a dipole so that the laser neutralization takes place after a small bend and only those H^0 are collected on the target.

A laser pulse much shorter than the temporal current distribution of the bunch enables also a bunch length measurement. A mode locked Ti-Sapphire laser reaches a wavelength of 950 nm and pulse lengths from picoseconds down to tens of femtoseconds. For the typical some ns long H- bunches the timing requirements are somewhat relaxed. The transversal size of the laser spot and of the beam should be roughly the same and stable in their positions. The laser pulse is locked to the RF frequency and its phase is scanned across the bunch length. Measurements at SNS are reported in Ref. [105].

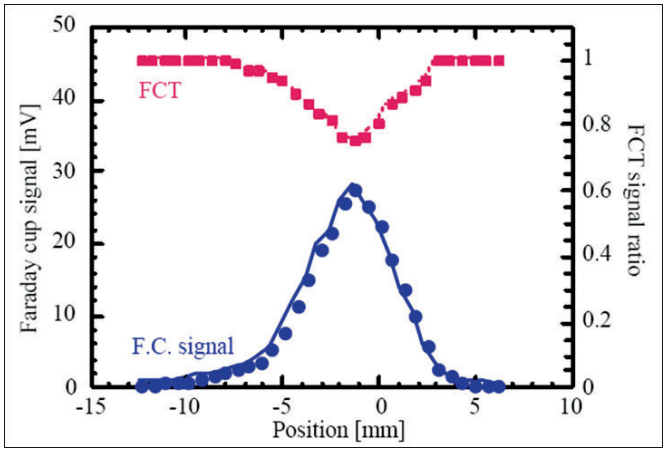


Fig. 14: FC signal and FCT signal from a laser scan (reproduced from Ref. [102])

4.3 Ionization profile monitor

Residual gas atoms or molecules are always present in the vacuum system of every accelerator. They fill the beam pipe with a homogeneous distribution, typically with a pressure of 10^{-6} mb to 10^{-9} mb. Most of the residual gas components are H_2 molecules. Assuming a mean energy of about $E_{ion} = 90$ eV needed to ionize a molecule of the residual gas (95 eV for H_2 only), the amount N of ionized particles can be derived from the Beth–Bloch Formula (dE/dx) valid for the individual pressure:

$$N = \frac{1}{E_{ion}} \cdot \frac{dE}{dx} \quad [\mathrm{cm}^{-1}\ \mathrm{particle}^{-1}]$$

More accurate measurements of the ionization cross-sections were performed in Ref. [106]. The simple model from the formula above results in $N \approx 40$ ion–electron pairs per centimetre at a H_2 pressure of 10^{-9} mb and a passage of 10^{13} minimum ionizing particles.

The IPM separates the resulting electron ion pairs by the use of an extraction field E_{ext} perpendicular to the beam axis. Typical values for E_{ext} lie between 50 V mm^{-1} and 300 V mm^{-1}, depending on the gap between the electrodes, practical power supply and space charge distortion (see below) considerations.

Field-forming electrodes and their careful design guarantee a highly parallel field so that the electrons or ions are projected onto the readout plane (see Fig. 15). Extended electrodes [107] and/or coated walls are useful for cleaning the environment from secondary electrons and ions not generated in the extraction volume. Most existing IPMs are now using one or two micro-channel plates (MCPs) inside the beam vacuum to amplify the resulting current. Just after the MCP the amplified current is collected by a phosphor screen or multi-anode strips. Early examples of these system can be found in Ref. [107] (phosphor screen) and Ref. [108] (strips); the first IPM was described in Ref. [109] but without using a MCP. An internal amplification can also be achieved by using a gas curtain [110] or a gas bump [111] in the monitor. The gain of a single MCP reaches up to 10^3 while a double stack

(chevron) reaches up to 10^6. Since the distance and diameter of the microchannels are of the order of 10 μm, a MCP does not distort the projected beam profile significantly as long as the width is larger than some 100 μm. The secondary electrons at the output of the MCP are accelerated by a second electrical field onto a phosphor screen or a position-sensitive anode configuration.

The phosphor screen is viewed by a standard CCD video camera which provides sufficient resolution and sensitivity in most cases [112]; see Fig. 16. A carefully designed optic is important, however, to achieve a good resolution [113]. The video frame rate limits the bandwidth of this readout to 50 (60) Hz which is sufficient for storage rings, but not for cycling synchrotrons or linacs. A fast readout can be achieved by position-sensitive (silicon) photomultipliers or photodiode arrays [114, 115]. The use of a fast decaying phosphor type (e.g. P47) is then required.

A very fast readout can be achieved by using anode strips as a collector of the MCP electrons. The separation of the strips defines the spatial resolution. A pitch of down to 250 μm is possible providing sufficient resolution. Connecting each strip via a charge-sensitive amplifier to a fast ADC enables turn-by-turn [116] and even bunch-by-bunch [117] resolution. The use of a resistive plate or wedge-and-stripe anodes [118, 119] can reduce the number of channels and vacuum feedthroughs. Since this is based on the detection of single particles, it disables the very fast readout opportunity. On the other hand it can improve the resolution up to the limit of the MCP by applying high statistics.

IPMs in high-intensity accelerators suffer from the high space charge of the (bunched) beam, exceeding the extraction field E_{ext}. The space charge disturbs the exact projection of the beam profile by bending the trajectories of the ions and electrons [116]. In particular, the light electrons get such a large kick that a profile measurement becomes impossible. Applying a magnetic field parallel to the extraction field forces the electrons to spiral around the magnetic lines with the cyclotron radius $r_c = m_e v_\perp / eB$. The radius depends only on the initial transverse kinetic energy defined by the kinematics of the ionization process and is below 50 eV for 90 % of the electrons. A magnetic field of about 0.1 T is then sufficient to keep the radius (and therefore the monitor resolution) below about 0.1 mm (see Ref. [120]). Different design with permanent magnets [121] and electromagnets have been realized [117], but note that some electrons may reach much higher kinetic energies which lead to tails of the distribution produced by intrinsic effects and not by the beam halo. The extraction field E_{ext} and the magnetic field B have to be compensated by opposite fields close to the monitor to minimize any influence on the beam.

By changing the polarity of the extraction field E_{ext}, it is also possible to collect the ions on the MCP. The heavy ions are not so strongly affected by the space charge and the distortion can be analysed and subtracted. A precise correction includes also terms from the collision impact and from the thermal movement of the residual gas molecules. The collision impact on the ions is quite small but the thermal velocity spread contributes already with a profile broadening of approximately 200 μm to 300 μm, depending on E_{ext} and the monitor geometry. The broadening of the measured profile due to space charge is reasonable at lower bunch densities but can exceed by far the real beam width at high bunch density. Under those conditions a correction might become useless. Since the broadening depends on all bunch parameters, detailed calculations and simulations are required to estimate the error contributions in detail [122–125].

A well-known problem of MCPs, phosphor screens and anodes is their aging with the amplified charge. This leads to a reduced gain just in the centre of the beam distribution and broadens the measured profile. Therefore, a continuous monitoring of the gain distribution is required. Various methods are in use or discussed, such as an α source [118], an electron generator plate [126], a tungsten filament emitter [127], an ultraviolet lamp [128] or a motorized 90° flip of the MCP [129]. A simple way to prove the aging is to steer the beam to an unused part of the sensitive area.

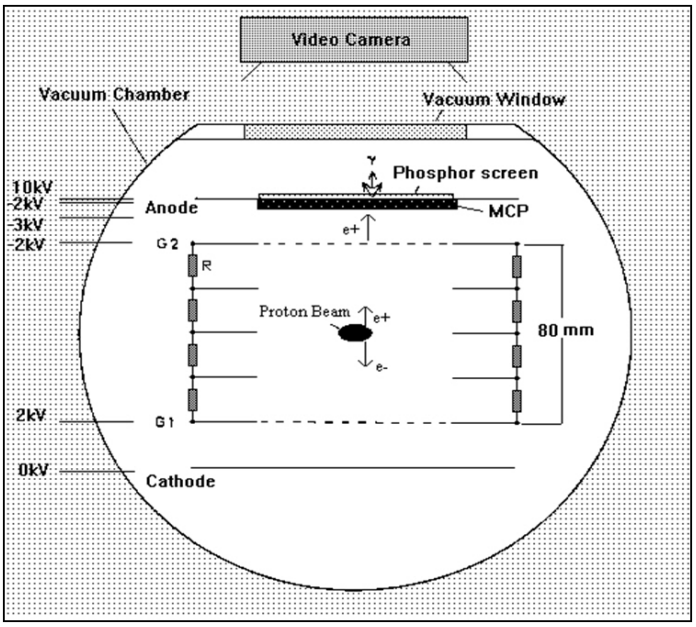


Fig. 15: Sketch of an IPM (in the (y, z)-plane) with MCP and phosphor screen. The extraction field E_{ext} is applied between G1 and G2 (grids). A resistive network R and the field shaping electrodes provide a homogeneous field distribution. The voltage configuration shown is an example to collect residual gas ions on the MCP.

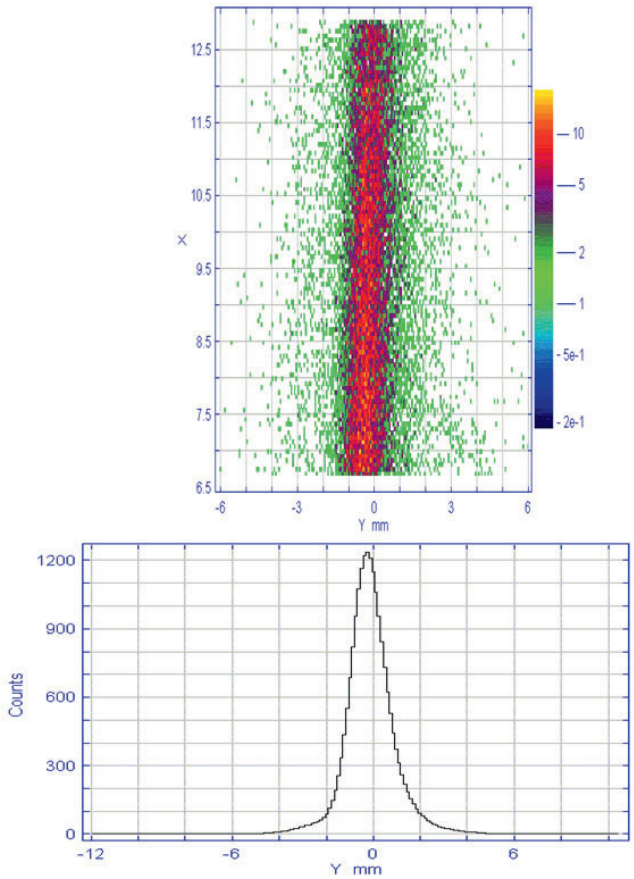


Fig. 16: Image of the beam on the phosphor screen (upper part) and its projection on the y-plane of the beam (lower part) (reproduced from [118])

4.4 Gas scintillation

The molecules of the residual gas are not only ionized, but are also excited by the beam particles. During their de-excitation they emit light in the visible range (depending on the gas type). Assuming a homogeneous distribution of the gas atoms in the vacuum, the focused light distribution represents the beam profile. The cross-section for gas scintillation is much smaller than for gas ionization; only a small percentage of the ionization loss is converted into detectable visible light [130]. This still does not include the fact that the light is emitted into the full solid angle while the detector only covers a fraction of it. Also not included is the finite sensitivity of a photocathode of a position-sensitive detector. Owing to this small efficiency typically a pressure bump is needed to increase the signal to a sufficient level. Nitrogen (N_2) is often used for local gas bumps since the vacuum system can efficiently pump this gas. The dependence of the cross-section on the ionization loss (Bethe–Bloch formula) has been proven over a wide energy range with proton and ion beams, as well as its linear dependence on the N_2 gas pressure [131, 132], but note that a partially stripped ion beam cannot handle significant added gas pressure.

If the excited residual gas atoms are still neutral atoms, they not affected by the space charge of the beam and this method is quite suitable for high-intensity beams and bunches, but the excitation cross-section is highest for the ionized state N_2^+ (see Ref. [133]) with a half lifetime of the de-excitation of about 60 ns (see Refs. [131, 134]). The movement of the excited ions during this time depends mainly on the space charge of the bunch. Simulations for LHC have shown drifts of some hundred micrometres, creating long tails in the measured beam profile [135]. Using Xe as a working gas with a lifetime of 6 ns (Xe^+) can improve this situation [136] but a much lower cross-section has to be taken into account. Other contributions which further broaden the measured profile of the order of some tens of micrometres include [135, 137]:

- thermal movement of the ions;
- momentum exchange during ionization;
- finite impact parameter;
- secondary electrons create additional excitations far away from the beam.

There is some experience with other gases used for a gas scintillation monitor. Xenon and other noble gases were studied in Refs. [131,136, 138], with a first result that He is excluded due to large tails in the beam profile.

High-sensitivity light detectors with position resolution are necessary for this type of monitor, even with a pressure bump. High gain image intensifiers, intensified CCD cameras or position-sensitive photomultiplier tubes (PMTs) are typically used. With practical gas bumps signal integration over many turns is still required to get a useful signal so that turn-by-turn or single shot profile measurements are excluded. All devices need good protection against radiation from adjacent beam losses to reduce background signals as well as radiation damage of the detector.

5 Beam loss measurements

A serious problem for high-current and high-brilliance accelerators is the high power density of the beam. A misaligned beam is able to destroy the beam pipe or collimators and may break the vacuum. This fact makes the BLM system one of the primary diagnostic tools for beam tuning and equipment protection in these machines. In addition to the task of machine protection the BLM system has more major goals:

- It should limit the losses to a level which ensures hands-on-maintenance of accelerator components during shutdown and it should limit the radiation outside the accelerator shielding. The hands-on limit has been found approximately between 0.1 W m^{-1} and 1 W m^{-1} (see Refs. [139, 140]). A value of 1 W m^{-1} corresponds to 1 GeV·nA m^{-1}; note that the limit of losses shrinks with beam energy.

- Ground water activation and radiation damage to components may put additional constraints on tolerable beam losses [139].
- Detecting the physical locations of a beam loss within a certain resolution in space. Often the resolution is limited by the spacing of the individual BLMs.
- Determination of the fraction of the lost particles relative to the beam, within a certain time interval.
- The system should be sensitive enough to enable machine fine tuning and machine studies with the help of BLM signals; sometimes even at low beam intensity to avoid high losses and/or during machine commissioning and at various energies during acceleration. This includes the comparison of the detected loss with computer models (Monte Carlo and beam tracking programs) and the analysis of the behaviour.

Therefore, one of the main issues of a BLM system is its very high dynamic range. It has to deal with two different types of losses; the regular losses which are unavoidable but suitable for beam diagnostics and the uncontrolled losses which generates additional radiation and risks [141].

Uncontrolled losses may occur with a fast transient, therefore the reaction time of the BLM system has to be matched to the transient time. In linacs even a bunch-by-bunch loss measurement is required while in (superconducting) storage rings about 0.1 ms to 1 ms are sufficient [142–144]. An integration of the signal over the required period is compared with a predefined threshold to generate alarm signals in case it exceeds the threshold. The threshold of tolerable beam losses depends on the specific requirements of the adjacent accelerator, e.g. quench limits, heating, radiation, residual activation, background, etc. Dangerous conditions are defined by the acceptable energy deposition of the lost particles and its adjacent shower in sensitive materials of the accelerator environment. Monte Carlo simulations are most helpful in calculate the thresholds for each specific BLM location as well as to calibrate the response of the BLM in terms of lost particles [145, 146].

Regular losses might occur continuously during operational running and correspond to the lifetime/transport efficiency of the beam in the accelerator. The lowest possible loss rate is defined by the theoretical beam lifetime limitation due to various effects, like residual gas scattering, diffusion, space charge, etc.; controlled losses due to scraping, beam extraction and injection (stripping foil), collision, etc. also fall into this category. These losses should be localized on the collimator system or on other known and properly designed aperture limits. At these locations, the measurement of losses can also be used for machine diagnostic purposes (in addition to their protection task), e.g. for optimizations of injection, lifetime, beam transport, background conditions and residual activation as well as for tail and tune scans, for measurement of diffusion processes and much more. For details see Refs. [141, 147].

BLMs should be localized in areas with higher probability of beam losses, e.g. collimators, high dispersion regions, high-β amplitudes. Different types of BLMs are used sometimes at the same location to extend the dynamic range of the system: sensitive BLMs to measure small losses and more insensitive ones to cover the high loss rates [142, 148]; or to cover different time scales: scintillator-based BLMs for nanosecond response times and ionization chambers for microsecond response times. In particular, at beam energies below the pion threshold (< 150 MeV) the (additional) use of neutron-sensitive BLMs is useful since the charged particles hardly escape the vacuum chamber [149].

Many factors are important for the design of a proper BLM system. In particular, for high-intensity beams a common aspect is the required large dynamic range, but also the radiation resistance, saturation characteristics and more. A summary of important considerations when selecting a BLM design are listed in Ref. [150]. A detailed discussion of various types of BLMs can be found in Ref. [147]. The following discussion about BLMs will concentrate on the aspect of their dynamic range.

5.1 Ionization chambers

Short ionization chambers are used as BLMs in many accelerators [148, 151–154]. An ionization chamber in its simplest form consists of two parallel metallic electrodes (anode and cathode) separated by a gap of width *D* and an applied bias voltage of some hundreds of Volts. The gap is filled with gas (air, argon, xenon[9]) of density ρ. The gas-filled volume between the electrodes defines the sensitive volume of the chamber. Ionizing radiation creates electron–ion pairs in this sensitive volume. The electrons can escape an immediate recombination if the electric field between the electrodes is larger than the Coulomb field in the vicinity of the parent ion. If all charges are collected the signal does not depend on the applied voltage (ionization region). The flatness of the plateau of the ionization region depends on the collection efficiency of the electrons or ions on the electrodes. In particular, at high radiation levels electrons on their way to the anode may be captured by positive ions produced close to their trajectory (by other incoming particles) and do not contribute to the charge collection. Therefore, a high voltage and a small gap *D* are preferred to achieve a high dynamic range as well as to achieve a faster response time of the ionization chamber [152]. Electron collection times of less than 1 μs are achieved, even in large chambers, by an appropriate arrangement of the electrodes [151].

The dynamic range of an ionization chamber is defined by its upper and lower current signal. The upper limit is given by the non-linearity due to the recombination rate at high dose; the typical chamber current in such a case is a few hundred microamperes. The lower limit is given by the dark current between the two electrodes. A very careful design of the chamber is necessary to very low dark currents in the order of few picoamperes. This gives a dynamic range of up to 10^8. Such a high dynamic range needs some special signal processing. Solutions such as variable gain amplifiers [155], logarithmic amplifiers [156], high ADC resolution [153] and current-to-frequency conversion [157] are applied.

Ionization chambers can be built from radiation-resistant materials such as ceramic, glass and metal with no radiation and time aging. Special care has to be taken for the feedthroughs and to the preamplifiers. Up to more than 10^8 rad can be tolerated by a careful design focused on radiation hardness. Air-filled ionization chambers require virtually no maintenance, leakage in N_2 filled chambers is not critical, but sealed Ar-filled chambers also give very few reasons for maintenance.

An enhanced sensitivity is provided by using the internal gas amplification of an ionization chamber in the proportional regime. In Ref. [158] an internal gas amplification of 6×10^4 at 2 kV, a dynamic range of 10^3 and a fast rise time of 100 ns were reported.

A “short” ionization chamber covers only a small part of an accelerator; therefore, a large number need to be installed to detect all losses. To overcome this problem a long, gas-filled coaxial cable has been used as an ionization chamber. Position sensitivity is achieved by reading out at one end the time delay between the direct pulse and the reflected pulse from the other end. The time resolution is about 50 ns (~15 m) for 6 km long cables, for shorter chambers about 5 ns (~1.5 m) was achieved [159]. This principle of longitudinal resolution works for one-shot (turn) accelerators (and transport lines) with a bunch train much shorter than the length of the cable. For particles travelling significantly slower than the signal in the cable (≈0.92c) the resolution of multiple hits in the cable becomes difficult. In this case and for circular and multibunch machines it is necessary to split the cable. Each segment has to be read out separately, with spatial resolution equal to the length of the unit. This is done in linacs [160–162] and in some rings and transport lines [163–165]. Since the chamber geometry is not optimized for high dynamic range, their linear range is limited to about 10^3 to 10^4 depending on the gas contents. Long ionization chambers made of commercial cables are simple to use, cheap and they have a uniform sensitivity. The isolation is not very radiation resistant, nevertheless these cables were used in SLAC for more than 20 years without serious problems.

[9] Electronegative gases (O_2, H_2O, CO_2, SF_6, etc.) capture electrons before reaching the electrode. Noble gases have negative electron affinities (Ar, He, Ne) which reduces recombination.

5.2 PIN diodes

One can treat a PIN diode with its intrinsic depletion layer as a "solid-state ionization chamber". The required energy to create an electron–hole pair in a semiconductor is about 10 times smaller than creating an electron–ion pair in gas. Also the density of a semiconductor is about three orders of magnitude larger than for gas (at 1 atm). Therefore, the signal created by radiation is much higher per unit path length than in a gas volume, but the active volume is much smaller: a sensitive area of 1 cm^2 and a depletion layer of d = 100 μm to 200 μm is already one of the largest PIN diodes which are commercially available. At about U = 30 V to 40 V the width d reaches its maximum. The transit time and the rise time of the signal of the order of a few nanoseconds due to the small gap d, a high electric field $E = U/d$ and a capacitance C = 10 pF to 100 pF. A dark current of a few nanoamperes is typical (due to the finite resistance between the two electrodes (p^+ and n^+) of the diode) and limits its dynamic range in this "photocurrent" mode. Modest radiation damage at 10^6 rad [166] leads to an increase of the dark current while most of the other parameters remain unchanged. Note that not too strong magnetic fields do not influence the charge collection in PIN diodes. Therefore, they can work as radiation monitors in stray fields of magnets, e.g. in high-energy experiments [167, 168].

PIN diodes are not very sensitive to γ-radiation but highly efficient to charged particles due to their thin active volumes P. The (hadronic) shower created by beam losses includes a large number of charged particles. The HERA BLM system consists of two PIN photodiodes mounted close together (face to face) and readout in pulsed mode and in a coincidence circuit [169]. Thus, charged particles crossing through the diodes give a coincidence signal, while γ radiation which interacts in only one diode (already with small efficiency) does not [170]. In this way the background of γ radiation (e.g. synchrotron radiation) and internal noise (dark counts) can be suppressed very efficiently. In contrast to the analogue charge detection of most other BLM systems, coincidences are counted while the count rate is proportional to the loss rate as long as the number of overlapping coincidences is small. Counting of charged particles crossing both diodes has a few implications:

- Both channels need a discriminator to suppress dark counts due to noise. Since the signal of one minimum ionizing particle (MIP) is still weak, the threshold cuts also some of the MIP signals which reduce the efficiency. The efficiency for a coincident detection of MIPs was found to be about ε_{count} = 30 % to 35 % per MIP including the readout electronic characteristic [171, 172].
- The dark count rate in coincidence mode is very small, typically <0.01 Hz
- The counter cannot distinguish between one or more MIPs crossing both diodes at the same time. The shortest signal length is defined by the response time of the diodes, but in practice it is defined by the readout electronics. An efficient counting type of BLM should have a signal length shorter than the bunch distance, so that the maximum measured loss rate is the bunch repetition rate of the accelerator. Saturation effects occur even before the maximum rate but they can be corrected by applying Poisson statistics [173].
- The dynamic range lies between the dark count rate of <0.01 Hz and the maximum rate (e.g. 10.4 MHz for HERA) and might reach 10^9.

A PIN diode BLM system has been successfully operated between 1992 and 2007 in HERA without significant problems or radiation damage.

5.3 Secondary emission monitors

Gaseous ionization chambers have the disadvantage that their charge collection is slower than the bunch distance in most accelerators. Counting mode devices have to integrate the counts over a lot of bunches to get a statistically relevant signal. In some cases a bunch resolved fast signal is required, e.g. for fast machine protection [174]. A simple, robust and fast BLM is a secondary emission chamber. Secondary electrons are emitted from a surface due to the impact of charged particles with an efficiency of a few percent [175]. Secondary electron emission (SEE) is a very fast effect, but its very low sensitivity makes secondary electron emission useable only in high radiation fields, with the

additional advantage that it consists of nothing more than a few layers of metal. Therefore, it is a very radiation-resistant monitor. The monitor has to be evacuated to avoid contamination of the signal due to gas ionization. Since the efficiency of gas ionization is much higher, a gas pressure of better than 10^{-4} mb should be achieved to get <1% signal from ionization. In particular, in high-radiation fields gas ionization will lead to non-linearities while secondary electron emission is a very linear process over a wide range of intensities [175, 176]. Unavoidable ionization at the feedthroughs and connectors limits the linearity at the lower end of the signal; the upper end is not seriously studied [151]. A dynamic range of $>>10^5$ is expectable [174].

A SEE multiplier extends the use of SEE BLMs to small radiation intensities. As long ago as 1971 aluminum cathode electron multipliers (ACEMs) have been used for beam loss measurements [177]. This device is a PMT where the photocathode is replaced by a simple aluminum cathode. The SEE electrons are guided to dynodes where they are amplified; amplifications up to 10^6 are possible. An example for recent use of ACEMs can be found in Refs. [178, 179].

5.4 Scintillation detectors

SEE-based BLMs are very fast but still have a moderate sensitivity. An equivalent speed (a few nanoseconds) but much higher sensitivity can be achieved with scintillation counters: a combination of a scintillating material and a PMT. Large area plastic (organic) and liquid scintillators are available. In particular, plastic scintillators can be modulated in nearly all shapes and sizes while inorganic scintillators are expensive and limited in size. Descriptions of details of the scintillation process can be found in Ref. [180] and in various text books, e.g. Refs. [181, 182]. Large scintillators can be useful to enhance the solid angle of beam loss detection if the resulting radiation is not distributed uniformly. This is often true if the BLM is located very close to the beam pipe where the radiation is peaked into a solid angle and at low beam energies. Typically a thin layer of scintillator (0.3 cm to 3 cm) is sufficient to ensure sensitive loss detection, even at very limited space conditions [183].

Note that the light transmission through the scintillator (and the light guide) changes due to radiation damage. This depends strongly on the scintillator and light guide material, but for organic scintillators a typical value can be assumed: the transmission decreases to $1/e$ of its original value after about 0.01 MGy to 1 MGy (1 Mrad to 100 Mrad) collected dose. Liquid scintillators are somewhat radiation harder and have about the same sensitivity [184]. Inorganic scintillators such as BGO or CsJ(Tl) have about a factor of 10–50 higher sensitivity but their radiation resistance is poor and large size crystals are very expensive.

The gain of the same type of photomultipliers (PMT_{gain}) varies within a factor of 10. Therefore a careful inter-calibration of the BLM sensitivities is necessary by adjusting the high voltage (HV). The drift of the gain is a well-known behaviour of PMTs. A stabilized HV source and continuous monitoring of the photomultiplier gain over the run period are necessary to keep the calibration error small. The adjustable gain of the PMT increases the dynamic range of this type of BLM. At high gain the noise of a PMT is still quite low but non-linearities appear at low gain and high losses in the PMT; the space charge of the signal cloud cannot be compensated any more by the low voltage between the dynodes. A dynamic range of 10^8 was measured at LEDA [185].

A special BLM uses Cherenkov light created in the glass tube of the PMT which is then detected directly [186]. It is a quite radiation-tolerant system; however, the darkening of the PMT glass has to be compensated for by increasing the PMT gain. Such a system is not sensitive enough to measure “small normal” losses but it is used to control and limit strong and dangerous losses.

Cherenkov light created in long optical fibres is used to determine the longitudinal position of beam losses. The fast response of the Cherenkov signal is detected with photomultipliers at the end of the irradiated fibres. A time measurement provides the position measurement along the fibre while the integrated light amplitude gives the amount of losses. A longitudinal position resolution of

20 cm (= 1 ns at $v = 0.66c$) is possible. High-purity quartz fibres (Suprasil) withstand 30×10^9 rad and generate no scintillation. Scintillating fibres are about 1000 times more sensitive but are not very radiation hard [187]. Examples for Cherenkov fibre-based BLM systems can be found in Refs. [188, 189].

So far all detectors are sensitive to "local" losses that occur within proximity of the detector. Hadron beam losses are typically connected with higher neutron flux, while neutrons can travel quite a long distance along the accelerator. Therefore, neutron detectors (NDs) are good at detecting losses occurring metres away from the detector itself. This makes NDs hard to interpret but more reliable for **Machine Protection System** (MPS) purposes. Solely relying on "normal" BLMs can lead to hiding of losses because a machine tuning process sometime moves the loss to a place where it is not seen by the "normal" BLM. An example is the SNS ND with a PMT + scintillator inside an X-ray shielding (lead) and surrounded by a polyethylene neutron moderator [190]. It is used in addition to ionization chambers and scintillator PMTs.

5.5 Summary

Different types of beam losses together with some examples have been shown. Beam loss monitoring techniques for measuring losses along an entire accelerator have been discussed with a focus on the sensitivity of the various types.

The most common BLM is a short ionization chamber. Whether a simple air-filled chamber is adequate or an argon- or nitrogen-filled chamber with superior higher dynamic range must be used depends on the conditions of the particular accelerator. Ionization chambers can be built very radiation resistant.

Long ionization chambers using a single coaxial cable work well for one-shot accelerators or transport lines. To achieve spatial resolution of losses along an entire accelerator either the first two or the third of the following conditions must be fulfilled: (1) the machine must be much longer than the bunch train; (2) the particles must be relativistic; (3) the long chamber has to be split into short parts which are readout individually.

PIN diodes with thick depletion layers can be used as "solid-state" ionization chambers. They have a high sensitivity but they exist only in small sizes. The combination of two PIN photodiodes in a coincidence counting mode results in a detector with very large dynamic range and extremely effective rejection of noise. A limitation is the inability to distinguish overlapping counts, so that the response is linear only for losses which are less than one count per coincidence interval.

A very fast and sensitive BLM system is a PMT in combination with a scintillator. Owing to the adjustable gain the dynamic range can be large, but the calibration of each device must be adjusted and monitored over time.

Long optical fibres can be used as in long ionization chambers with the same limitations in the bunch repetition rate. Cherenkov-based fibres are much more radiation hard but much less sensitive to losses than scintillating fibres.

Table 3 summarizes the different BLM types used in various high-intensity hadron accelerators.

Table 3: BLM types used at some high-intensity hadron accelerators

Scintillator		
LEDA	CsI scintillator PMT based	[185]
ISIS	Plastic scintillator (BC408)	[183]
J-PARC RCS, MR, LINAC	GSO scintillator	[148]
SNS Ring	Scintillator PMTs	[190]
SNS Linac	PMTs with a neutron converter	[190]
PSR	Liquid scintillator with PMT (old)	[191]
CSNS	Scintillator PMTs	[192]
Ionization chambers		
LEDA	160 cm^3 N_2 ion chamber	[193]
ISIS	Long Ar ionization tubes (3 m to 4 m)	[183]
SNS Ring	113 cm^3 Ar ion chambers	[152]
SNS Linac	113 cm^3 Ar ion chambers	[190]
PSI	Air ionization chambers	[194]
PEFP		
J-PARC RCS, MR, LINAC	Ar+CO_2 proportional counters (80 cm) and coaxial cable ion chambers, air filled (4 m to 5 m)	[148]
PSR	ion chambers filled with 160 cm^3 of N_2 gas	[191]
LANSCE	180 cm^3 N_2 ion chamber	[195]
CSNS	110 cm^3 Ar ion chamber	[196]
AGS	Ar-filled long coaxial ion chambers	[163, 197]
NuMI	Ar-filled Ion glass tubes	[156]
SPS, CNGS	Air-filled ion chambers (1 litre)	[192]
APT	Same as LEDA	
Tevatron, MI, Booster	Ar-filled Ion glass tubes, 190 cm^3	[198, 199]
CERN LHC	N_2-filled ion chambers 1.5 litre	[142]
Rhic	Ar-filled Ion glass tubes	[197]
SEM chambers		
LHC	SEM chambers	[175]
PIN diodes		
HERA	PIN diodes in counting mode	[145]
Tevatron	PIN diodes in counting mode	[200]
Rhic	PIN diodes in counting mode	[201]

We now give some examples for beam diagnostics for high-intensity hadron beams.

6 Transversal beam halo measurements

Particles which are expelled from the beam core form a halo around the beam. This halo can cause harmful beam losses, especially at higher beam energies. It contributes to activation of the environment and to background in the experiments. There are numerous sources of halo formation, in linear and circular accelerators, which are summarized in Ref. [202].

In the summary of the HALO'03 workshop [203] is written: "…it became clear that even at this workshop (HALO'03) a general definition of 'Beam Halo' could not be given, because of the very different requirements in different machines, and because of the differing perspectives of

instrumentation specialists and accelerator physicists… From the diagnostics point of view, one thing is certainly clear – by definition halo is low density and therefore difficult to measure…". A quantification of the halo requires a more or less simultaneous measurement of the core and the halo of the beam. Therefore, halo measurements require very high dynamic range instruments and methods as well as very sensitive devices to measure the few particles in the halo. The difference between "halo" and "tail" can be defined as tails are deviants from the expected beam profile of the order of a few percent or per mille while halos are much less than this.

A measurement of the halo should result in a quantification of the halo; therefore, it is important to have a definition of the halo in at least one-dimensional spatial projection since this is relatively easy to obtain by a beam profile/halo monitor. For a complete understanding of the halo it might be necessary to extend the one-dimensional work to the whole phase space, in the measurement (location of the monitors) as well as in the theoretical work. This leads finally to the kinematic invariants imposed by Hamilton's equations [204]. Such a consideration is mainly used in simulations [205, 206].

In any case, the separation between the halo and the main core of the beam is not well defined. This leads to uncertainties to define a good description of the halo content of a beam. Typically beam halo is defined as an increased population of the outer part of the beam relative to the expected distribution which describes the core. Three different methods are commonly in use to characterize beam halo:

- kurtosis [204, 206, 207];
- Gaussian area ratio [208];
- ratio of beam core to offset [209].

An important feature of such quantifiers is that they are model independent and rely only on the characteristics of the beam distribution itself. Note that a measurement always contains instrumental effects. To define the halo contents in such a theoretical way one has to exclude these effects in advance.

The following sections concentrate on the instruments which are able to measure the beam halo and its evolution directly. Since the definition of halo is something like "$< 10^{-4}$ of the beam core", some usual beam profile monitors might have intrinsic limitations to get the required dynamic range. For example, ionization-beam profile monitors (IPM), luminescence beam profile monitor and laser-based monitors are not (yet) sufficient for very high dynamic range halo measurements [210].

6.1 Beam halo measurements with wire scanners

Wire scanners are widely used for halo measurements with huge dynamic range and high sensitivity. This instrument provides a direct halo measurement by analysing the signal amplitude directly or in combination with particle counting. A combination of a wire and a scraper can be used to improve the sensitivity. Typically the signal is read out by the secondary electron emission current of the wire (low beam energy) or by scintillators measuring the scattered particles (high beam energy). The problems of wire scanners are well known, e.g. emittance blow-up and wire heating (see Section 4).

The direct beam profile and halo measurement is done by correlation of the signal with the position of the wire with a high dynamic range readout. A dynamic range of 10^5 was achieved by linear amplification and the use of a 16-bit ADC as well as using a logarithmic amplifier which allows a standard 12-bit ADC [211]; see Fig. 17. The use of the secondary electron emission signal in low-energy accelerators has the advantage of avoiding an intrinsic error of measuring asymmetric tails by the asymmetric location of external detectors and/or large beam offsets [212]. This effect vanishes at higher energies and smaller beams. Therefore, scintillation counters outside the vacuum chamber can be used to measure the amount of scattered and shower particles created at the wire. Such scintillators are also sensitive to background due to, e.g., residual gas scattering, bremsstrahlung and other sources

of beam losses. A telescope counter using a coincidence technique can reduce this background dramatically as well as dark counts (noise) from the counters itself so that a dynamic range of 10^8 can be achieved [213–216]. The lower limit is defined by the remaining background rate.

In low-energy accelerators and/or at low bunch repetition rates as in a linac the counting method might not be very useful. In addition, a secondary electron emission current readout of a thin wire in the beam halo does not deliver enough current for a reliable measurement. Therefore, the wire size has to be increased even to a solid scraper to increase the achievable signal [217]. Their halo scanner consists of a 33 μm carbon fiber and two halo scrapers consisting of two graphite plates. Special care has to be taken that the beam does not induce too much heating of the scraper. Like in the counting method, the wire scanner and two scraper data sets must be joined to plot the complete beam distribution for each axis [218].

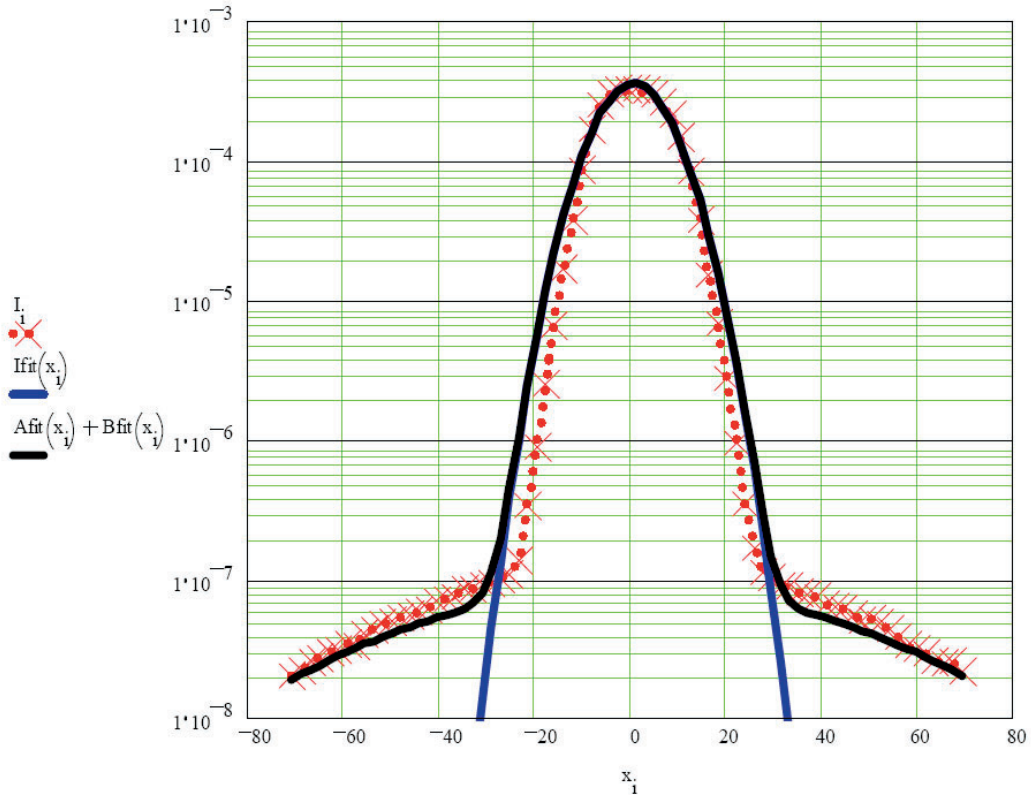


Fig. 17: A normal function shown in solid blue has been fit to the data (red crosses). A sum of two normal functions is shown in solid black. The *x*-axis is scaled as scanner position in millimetres and the *y*-axis is log-ampere input current in Amperes [211].

6.2 Beam halo measurements with interceptors/scrapers

Halo collimators are designed to remove the halo of the beam, but halo measurements can also be performed by moving one jaw of a collimator closer to the beam in steps. Either the beam current or the signal from adjacent BLMs can be recorded for each jaw position. The derivate of the signal gives the halo distribution. Very high sensitivity can be achieved by using BLMs close to the collimator jaws. The signal of the BLMs is proportional to the inverse lifetime of the beam which gives loss rates directly in terms of equivalent lifetimes. By moving the collimators closer until significant lifetime reductions were observed, the lifetimes calculated from beam currents can be used to calibrate the BLMs. Since this scraping method is a slow process it is very important to normalize each data point to the measured beam, to the measured beam size of the beam core and to the beam position [219].

Note that in high-energy and/or high-intensity accelerators/storage rings a complete scan of the whole beam is impossible since the jaws are typically not designed to withstand the full beam intensity [220, 221]. Therefore, a calibration of the halo contents (relative to the beam core) is often not possible or contains large errors, but relative changes of the halo can be detected at a very low level and far outside the beam core, e.g. ground motion frequencies and diffusion parameters [222–224].

Note also that in a synchrotron one jaw of a collimator will always scrape both sides of the beam distribution due to the β oscillation of the beam particles. Therefore, one will always measure a symmetric halo distribution.

Instead of a collimator with BLM readout other sensitive detectors can be moved into the halo to generate directly a signal from the halo particles. Various techniques are reported using, e.g., ionization chambers [225], scintillation fibres [226], vibrating wire scanners [227] or secondary electron emission foil [228]. All of these devices have the same strong limitation in determining the halo relative to the beam core, but relative changes of the halo can be observed with high sensitivity and resolution.

6.3 Optical halo monitors

For hadron beams optical methods are barely used since electromagnetic light generation (e.g. by synchrotron radiation, optical transition radiation) by hadrons is suppressed due to their high mass. Therefore, it is discussed here only very briefly.

The previously discussed methods to measure the halo distribution are relatively slow. Scanning of the halo typically needs seconds to minutes. One needs a stable beam and precise correlations with the beam size and position are mandatory. Optical methods have to give enough light to measure the core of even one single bunch at one passage. The light generation of these effects is linear over a huge dynamic range.[10] The dynamic range of the light detector (e.g. CCD cameras) can be improved by special optical systems:

- CID camera system with a dynamic range of $>10^8$ (see Ref. [229]);
- micro-mirror array [230].

Most optical applications suffer from diffraction limits which create diffraction fringes of 10^{-2} to 10^{-3} of the peak intensity which makes halo observations of lower than 10^{-3} impossible. A coronagraph with a so-called 'Lyot stop' [231, 232] removes this fringes and a background level of 6×10^{-7} was observed. More details can be found in Ref. [210].

7 Longitudinal beam halo measurements

The meaning of "longitudinal halo" can be divided into three different classes of different interests:

- *Beam in the abort gap*. High-intensity and superconducting hadron storage rings need a gap in the bunch train to have enough time for loading the dump kickers to ensure a clean beam dump. In the case of a beam dump any particles in the gap will be lost around the ring risking a quench.
- *Coasting beam*. Experiments in colliders need very clear background conditions and precise time structures of the bunch crossings. Particles outside the main bunches may contribute to background as well as to undefined timing of the trigger counters in the experiment.
- *Neighbour bunches or bunch purity*. In time resolved experiments on synchrotron light sources a clear signal from one bunch without contributions from the adjacent (neighbour) buckets is desired. The neighbour bunches have to be determined on level of better than 10^{-6}. This topic is mainly related to synchrotron light sources and will not be discussed here.

[10] The light generation in scintillation and phosphor screens suffers from non-linearities [233, 234] and therefore might not be applicable for huge dynamic range measurements.

7.1 Beam in gap

Stringent particle loss constraints in high current accelerators and in superconducting machines require a clean beam gap. Extraction of the beam (to the experiments or to the dump) is done by kickers with limited rise times, typically a few microseconds. This time is known as the abort gap where no particles should be stored. Any beam in this gap (bunched or coasting beam) will spray around the machine if the dump kicker is fired causing some problems:

- quenches (superconducting magnets);
- activation;
- spikes in experiments;
- equipment damage.

Reasons for beam in the abort gap can be:

- injection errors (timing);
- space-charge pushing particles out of the RF bucket;
- debunching;
- diffusion;
- RF noise/glitches;
- other technical problems.

Therefore, a continuous determination of the amount of beam in the gap is necessary to either clean the gap[11] or dump the whole beam before major problems arise. In high-energy storage rings like the LHC, Tevatron or HERA the presence of particles in the gap can be detected by the synchrotron radiation they emit, using the synchrotron radiation profile monitor port. Note that in principle any other fast process, e.g. beam-induced gas scintillation or secondary electron emission or BLM signals (e.g. at halo scrapers) can serve as a signal source, which are not limited to very high beam energy [233–235]. A fast and gateable detector which is synchronized by the revolution frequency is most useful to avoid saturation due to the signal of the main bunches. Optical methods have the advantage of existing detectors which are fast and sensitive enough to measure even a small amount of beam in the gap. A gated MCP PMT is able to measure both components of the beam in the gap, the bunched (AC) and the unbunched (DC) components while an intensified gated CCD or CID camera integrates over many turns and measures the DC component only [236]. Typical gate rise times of about 1 ns are sufficient for this application. Often the display of the analogue signal of the MCP PMT versus the gate time is sufficient but the dynamic range is limited to about 10^3 due to the noise of the PMT. When using the gating technique one has to take into consideration the maximum duty cycle of the instrument. A typical maximum duty cycle of 1 % (e.g. Hamamatsu R5916U-50 MCP PMT) might not allow a complete gate over the whole gap at every turn. Therefore, the gate repetition rate has to slow down or a shorter gate has to be scanned across the gap [237]. The dynamic range and the signal-to-noise ratio can be increased by applying the time-correlated single photon counting method. First results with MCP PMTs and fast avalanche photo diodes are reported in Refs. [238, 239].

7.2 Coasting beam

The coasting beam is the part of the beam which is not captured by the RF system; its energy is not being replenished by the RF system. Even in high-energy hadron storage rings uncaptured protons lose only a few electronvolts per turn so that they can be stored for many minutes up to hours. Uncaptured beam slowly spirals inward and is lost on the tightest aperture in the ring. RF noise, glitches or intra-beam scattering can cause diffusion out of the RF buckets leading to coasting beam [240, 241]. The total uncaptured beam intensity is a product of the rate at which particles leak out of the buckets and

[11] Abort gap cleaning by, e.g., fast kickers, resonant excitation, electron lens, etc. are not discussed here.

the time required for them to be lost. This kind of beam loss causes additional activation of the collimators as well as additional background in the experiments. In particular, collider experiments might suffer from this background; therefore, they are interested in measuring the amount of coasting beam. Very sensitive methods are needed to measure small fractions of coasting beam in an appropriate time. Therefore, the experiments themselves use their sensitive detectors and fast trigger equipment which have a very large detection efficiency as well as very small dead times. Detailed measurements of coasting beam are reported from HERA-B (HERA) and CDF (Tevatron) [242–244]. Both detectors use as the signal source the beam loss in the detector while HERA-B even increased the loss rate using its internal wire target. The time structure of the losses is measured by fast counters and TDCs (HERA-B) or by integrating counts versus a sliding time interval (CDF).

Note that the signal source comes from the far transverse halo of the beam. Its time structure might differ strongly from the time structure of the beam core [242], especially because uncaptured beam slowly spirals inward. Therefore, a total determination of the amount of coasting beam will have a large uncertainty. As soon as the amount of coasting beam is large enough an absolute determination can be done by comparing the AC and DC beam current monitors readings. The DC monitor measures all circulating particles while the AC monitor is sensitive only to the bunched beam component (Fig. 18). The calibration of both monitors to each other can be done just after finishing the acceleration where no coasting beam had survived.

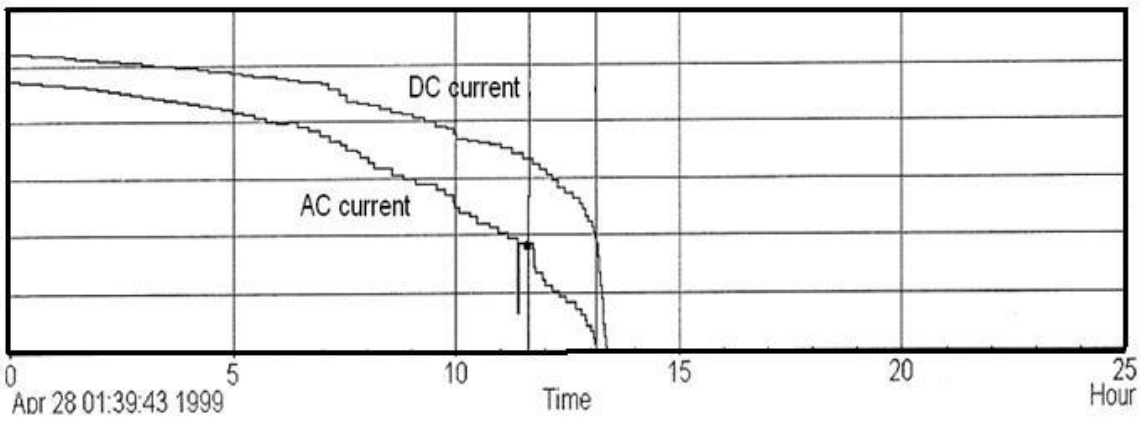


Fig. 18: DC beam current (includes coasting beam) and AC beam current (sum of all bunches) in HERA during an unusual store with a large amount of increasing coasting beam

8 Diagnostics for electron clouds

A charged beam can generate low-energy electrons by various, often unavoidable effects such as synchrotron radiation, residual gas ionization or stray particles. These electrons can strike the beam pipe wall and can create multiple electrons leading to multipactoring. Repetitive bunch crossings can lead to a quasi-stationary electron cloud (EC). A charged beam might interact destructively with this cloud resulting in beam instabilities and particle losses. Since the electrons are able to desorb gas from the wall, the first hint of an EC is typically an increase in the vacuum pressure in that section. The increase and observation of the vacuum is quite a slow process [245] and not very suitable for detailed analysis of ECs. Some more suitable instruments for EC diagnostics are discussed in the following sections.

8.1 Shielded BPMs

In front of the electrode of a button-type BPM, a grounded grid shields the electrode against the wake fields of the bunch. While the electrode is positive biased against ground it collects all low-energy electrons in its vicinity. A variable DC bias voltage enables electrons to be attracted or repelled depending on their energy. Such a relatively simple device is able to obtain time resolved information on the EC density (e.g. build up and decay) but an estimate of the EC line density λ is also possible:

$$\lambda = I_e /(e \cdot f_b \cdot tr \cdot A_e / A_{ch})$$

where $I_e = U_e/Z_e$ is the measured current on the electrode (Z_e is the impedance of the electrode), f_b is the bunch frequency, tr is the transparency of the grid, A_e is the area of the electrode and A_{ch} is the inner area of the chamber.

In Refs. [245, 246] the resonant build-up of an EC is clearly diagnosed with such a shielded button-type BPM. Figure 19 shows a sketch of the design of a shielded pickup in the CESR-TA ring.

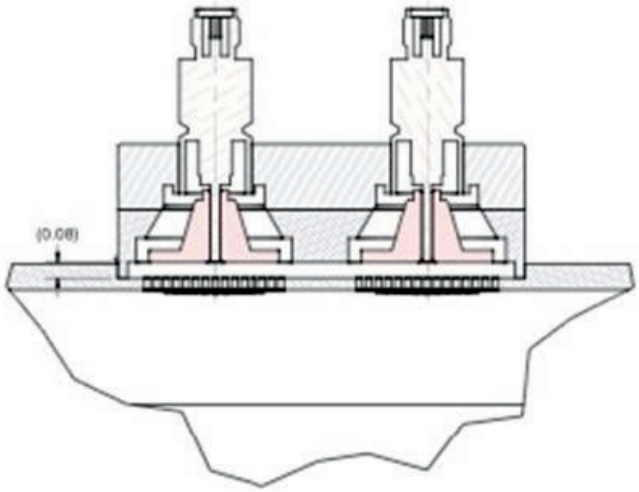


Fig. 19: Sketch of the CESR-TA shielded pick-up buttons [246]

8.2 Retarded field analyser

A retarded field analyser (RFA) is based on the same principle as the shielded BPM but it has a second retarding grid between the shield and the electrode [247]. The second grid is biased at a retarding potential (E_r) such that only electrons with kinetic energies greater than this are transmitted to the electrode (collector). The collector has a low secondary emission yield and is biased by a positive voltage. To amplify weak signals a MCP or channeltron can be used but usually the signal of ECs are sufficient for electronic amplification. The advantages of a dedicated RFA with respect to a shielded BPM are:

- increased surface area;
- higher sensitivity;
- better energy separation (see Fig. 20).

Examples of the extensive use of RFAs can be found in Refs. [248, 249].

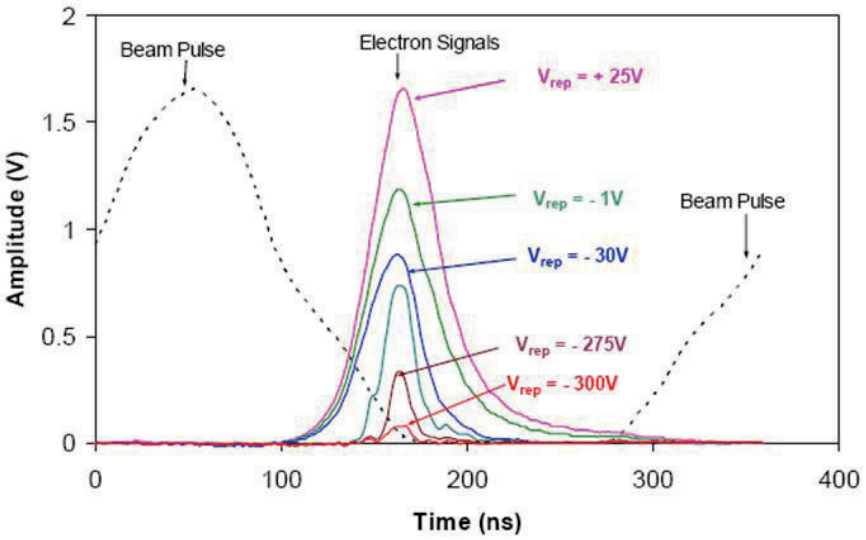


Fig. 20: RFA signals with different retarding voltage in time reference to the circulating beam pulse at PSR [248]. This experiment shows the low-energy distribution of the electrons of the cloud.

8.3 EC diagnostic with microwaves

The discussed EC monitors are suitable for localized measurements only. If the EC is created not in the vicinity it will not be detected by the monitor. In contrast, microwave transmission measurements are sensitive to the average EC density over a long section of the accelerator.

If a microwave of frequency ω is transmitted through electron plasma (EC) of length L it will undergo a phase shift Δϕ:

$$\Delta\varphi \approx \frac{L \cdot \omega_p^2}{2c \cdot \sqrt{\omega^2 - \omega_c^2}} \quad \text{with} \quad \omega_p = \sqrt{\frac{N_e e^2}{\varepsilon_0 m_e}} \text{ the plasma frequency}$$

where N_e is the electron density (typically 10^{11} m^{-3} to 10^{12} m^{-3} in ECs), e is the electron charge, ε_0 is the vacuum permeability, m_e is the electron mass and ω_c is the cutoff frequency of the beam pipe [250].

Therefore, the phase shift depends only on the electron density N_e while all other parameters are constant and relatively well known.

The setup of a transmission measurement is shown in Fig. 21(a). For exciting a TE wave into the beam pipe a BPM can be used which has to be optimized for TE mode emission by using 180° hybrids and combiners. Splitting the power between pairs of opposite buttons, or striplines, lowers the power on a single electrode and improves the coupling to the TE mode electric field [251]; see Fig. 21(b). Note that the reversed power from the beam signal may disturb the signal generation. Hybrids are also used on the receiving BPM to suppress common beam position signals. The carrier frequency ω is chosen by measuring the optimum of the transmission function (obviously above ω_c). With a constant EC a phase shift could hardly be detected. Therefore, one has to change the EC density during the measurement, typically by having a long enough gap between bunch trains to remove the EC. This gap creates than a phase modulation at the receiver (e.g. spectrum analyser) which appears as side bands to the carrier frequency in a distance of the revolution frequency. Its amplitude relative to the carrier is proportional to the phase shift Δϕ.

DeSantis, at ECLOUD'10, stated "Although having a simple formulation, the practical application of the TE wave method is not straight forward". Many problems might hinder the analysis [251–253]:

- The coupling efficiency of BPMs is small above cutoff, impedance is not well matched (by design).
- Non-linearities by reflections in generator and receiver add sidebands to the spectrum.
- The strong beam harmonics superimpose the weak EC sidebands.
- AM modulation by resonant coupling to e^- trapped in the magnetic field (near the cyclotron frequency) add sidebands.
- Owing to reflections of the carrier, L can be underestimated.
- L is not always the distance between the BPMs, the cloud might be shorter.
- Ensure that cleaning gap is larger than the decay time of cloud. Take into account EC rise and fall times.
- How precise the cutoff frequency ω_c is known?

A local transmission measurement (below cutoff) can be done with the setup of Fig. 21(c). First tests were done in Ref. [253] but a complete understanding of the physics of this method is still under discussion.

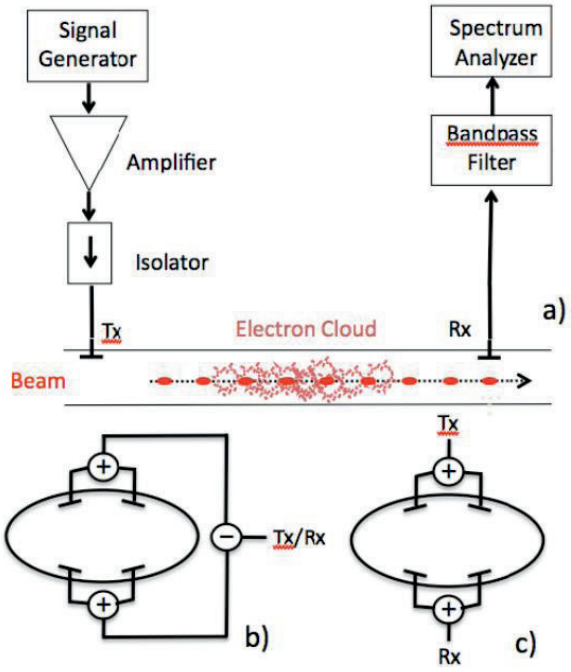


Fig. 21: (a) Microwave transmission setup (Tx, transmitter; Rx, receiver), (b) BPM arrangement for transmission measurement (⊕, splitters, Θ, 180° hybrid), (c) BPM arrangement for "local transmission" measurement (reproduced from [253])

9 Injection mismatch

As a rule, proton/ion accelerators need their full aperture at injection, thus avoiding mismatch allows a beam of larger normalized emittance ε_n and containing more protons. In proton/ion ring accelerators any type of injection mismatch will lead to an emittance blow-up. Off-axis injection will lead to orbit oscillations. These oscillations can be detected easily by turn-by-turn BPMs in the ring (before Landau damping occurs). The orbit mismatch can be corrected by a proper setup of the steering magnets, kickers and septa. Any mismatch of the optical parameters α, β, γ (space charge), however, will also lead to an emittance blow-up (and beam losses) and is not detectable by BPMs.

Figure 22(a) shows the phase ellipse at a certain location in a circular accelerator. The ellipse is defined by the optics of the accelerator with the emittance ε and the optical parameters β (beta function), $\gamma = (1 + \alpha)/\beta$ and the slope of the beta function $\alpha = -\beta'/2$. Figure 22(b)–(d) show the process of filamentation after some turns.

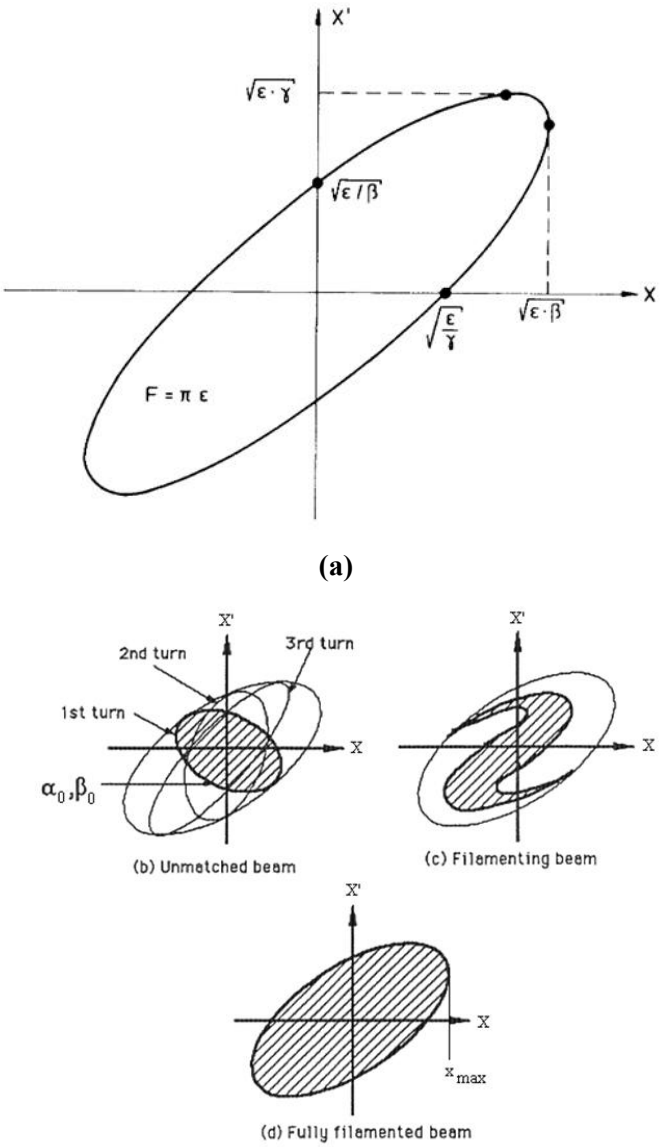


Fig. 22: (a) A phase space ellipse of a circular accelerator, defined by α, β, γ, ε. (b) Filamentation of an unmatched beam. (Reproduced from Ref. [254].)

Assuming a beam is injected into the circular machine, defined by β_0 and α_0 (and, therefore, γ_0) with a given emittance ε_0. For each turn i in the machine the three optical parameters will be transformed by

$$\begin{pmatrix} \beta_{i+1} \\ \alpha_{i+1} \\ \gamma_{i+1} \end{pmatrix} = \begin{pmatrix} C^2 & -2SC & S^2 \\ -CC' & SC' + S'C & -SS' \\ C'^2 & -2S'C' & S'^2 \end{pmatrix} \cdot \begin{pmatrix} \beta_i \\ \alpha_i \\ \gamma_i \end{pmatrix} \quad (\text{starting with } i = 0)$$

where C and S are the elements of the Twiss matrix ($\mu = 2 \cdot \pi \cdot q$, where q is tune):

$$\begin{pmatrix} C & S \\ C' & S' \end{pmatrix} = \begin{pmatrix} \cos\mu + \alpha_0 \cdot \sin\mu & \beta_0 \cdot \sin\mu \\ -\gamma_0 \cdot \sin\mu & \cos\mu - \alpha_0 \cdot \sin\mu \end{pmatrix}$$

and $\gamma = (1 + \alpha^2)/\beta$

Without any mismatch, the three parameters will be constant while a mismatch will result in an oscillation of the parameters at twice the betatron tune [255, 257]. A mismatch of, e.g., the betatron phase space will result in transverse shape oscillations, at least for some 10 turns, before the different phases of the protons lead to a filamentation of the beam. A measurement of width (or shape) oscillations at injection is a very efficient method to detect an optical mismatch that increases the emittance in the circular accelerator. A measurement of the turn-by-turn shape oscillation is possible with a fast (turn-by-turn) readout of:

(1) thin screen (OTR, phosphor); see Ref. [258] for details;
(2) secondary electron emission grids [259];
(3) IPM [260];
(4) quadrupole (QP) pickup [261];
(5) synchrotron radiation (SR) monitor (electrons) [262].

The effect of the monitors on the beam include the following:

- Screen/grid: emittance blow-up and losses.
- IPM: very small, a sufficient signal at each turn needs a pressure bump leading to emittance blow-up and losses.
- QP pickup: none but very difficult to suppress the dipole mode.
- SR monitor: none, but no light from protons at low energy.

9.1 Blow-up

A screen/grid or IPM pressure bump will give an additional constant increase of the emittance, but it can easily be separated from the oscillation observation. The protons receive a mean kick at each traverse through a screen resulting in an additional angle θ:

$$\theta = \frac{0.014}{p \cdot \beta} \cdot Z \cdot \sqrt{\frac{d}{l_{rad}}} \left[1 + \frac{1}{9} \log_{10} \left(\frac{d}{l_{rad}} \right) \right] \quad \text{in radians}$$

where p is the momentum in GeV/c and $Z = 1$ the charge number of the proton, $\beta = v/c$ the velocity, d the thickness of the foil and l_{rad} the radiation length of the material of the foil. This formula describes the Gaussian approximation of the mean scattering angle of the protons after one traverse. The change of the emittance δε for every turn can be calculated by

$$\delta\varepsilon_{rms} = \sqrt{2 \cdot \pi} \cdot \theta^2 \cdot \beta$$

which adds quadratic to the 1σ emittance of the previous turn. The emittance blow-up due to a thin foil is much too large at low energies. A harp of thin wires produces less emittance blow-up. Assuming a harp of 20 μm titanium wires at a separation of 1 mm, the blow-up can be calculated as in a 0.2 μm foil. The secondary electron emission current created in the wires can be read out by fast ADCs turn by turn. Such a readout schema is applied in the PS-Booster at CERN [258].

9.2 Losses

The relative proton losses per turn dN/N_0 in the foil (thickness d) is given by the nuclear interaction length L_{nuc}:

$$\frac{dN}{N_0} = \frac{d}{L_{nuc}} \qquad \text{with} \qquad L_{nuc} = \frac{A}{\rho \cdot N_A \cdot \sigma_{nuc}}$$

here L_{nuc} depends on the total nuclear cross-section of the nuclear interaction σ_{nuc}, the density ρ of the foil and the Avogadro constant $N_A = 6.0225 \times 10^{23}$ mol^{-1}. The nuclear cross-section σ_{nuc} depends on the proton momentum and on the material of the foil and is shown for different materials in Table 4 between a momentum of $0.3 < p < 40$ GeV/c.

Table 4: Nuclear total cross-sections, interaction length and particle losses

Material A [g/mol] ρ [g/cm³]	Momentum [GeV/c]	σ_{nuc} [mb]	L_{nuc} [cm]	Relative loss/turn $dN/N_0 \times 100$ [%] with $d = 10$ μm
Carbon	0.3	280	31.5	3×10^{-3}
12.01	7.5	360	24.5	4×10^{-3}
2.26	40	330	22.5	4.4×10^{-3}
Aluminum	0.3	550	30.2	3.3×10^{-3}
26.98	7.5	700	38.4	2.6×10^{-3}
2.70	40	640	35.1	2.8×10^{-3}
Copper	0.3	950	12.4	8.1×10^{-3}
63.546	7.5	1350	17.6	5.7×10^{-3}
8.96	40	1260	16.4	6.1×10^{-3}

The loss rate is negligibly small even at the injection energies of proton machines and will not influence the mismatch measurement.

9.3 Some notes on the readout

The optical readout of screens/IPMs is slow. A turn-by-turn observation needs a 100 kHz (3 km) data collection of the whole image. Line sensors with a larger pixel size (for better sensitivity) nowadays have a readout frequency of >15 MHz/pixel. Assuming 128 pixels will give a maximum readout frequency of 117 kHz for a one-dimensional image.

A secondary electron emission signal as well as the QP pickup signal can be picked up with very high frequencies, even bunch by bunch (100 MHz) and is therefore preferred for smaller ring diameters with a higher revolution frequency.

10 Beam energy

10.1 Beam energy determination using spectrometers

The most common method for determining the momentum/energy of a particle is a spectrometer. This includes any circular accelerator where the main dipole field and the closed orbit, resp. the central frequency define the particle energy [263] while spectrometer magnets making use of this effect are widely used in hadron Linacs. Relative energy resolutions are of the order of 10^{-4} [264]. Spectrometers measure the particle momentum by precisely determining the angle of deflection Θ in a dipole magnetic field B:

$$\Theta \propto \frac{1}{p} \int B ds$$

A very good determination of the magnetic field (10^{-5} or better) and the beam position at the entrance and exit of the spectrometer magnet is essential for a precise measurement. A position-

sensitive detector at the end of the spectrometer arm enables a precise momentum and momentum spread measurement. A collimator in front of the spectrometer magnet and a detector position at a low β and high dispersion value improves the precision of the measurement [265].

10.2 Beam energy determination using TOF

The resulting profile at a spectrometer detector is a mixture of the transverse and longitudinal beam parameters. An independent measurement can be performed for non-relativistic energies using the TOF method.

Two or more fast beam pick-ups are installed in a straight section with a typical distance L of several meters while L has to be known exactly, with a typical precision of about 1 mm. Each kind of fast pick-up can be used as a signal generator; their well-known signal properties will define the start and end of the measured time t, e.g. maximum or half height of a unipolar signal, zero crossing of a bipolar signal (more precise). When picking up the same bunch its velocity β is simply given by $t = L/\beta c$, but the value has to be corrected for signal propagation delays along the cables [266]:

$$t_j = \frac{l_{cab,j}}{v_{cab}} + \frac{L_j}{\beta \cdot c}$$

where v_{cab} is the cable phase velocity, $l_{cab,j}$ is the length of the cable of each station j and $L_1 = 0$. Modern digital processing allows an I/Q method in a FPGA which results in better precision of the TOF measurement [267] as well as a possible comparison of the bunch with the cavity phase [268].

10.3 Energy measurement with other methods

The use of Rutherford scattering to extract the beam energy is limited to low energies only. In Ref. [269], a 0.3 mg/cm^2 thick gold foil was inserted into the beam periphery and the scattered protons were detected by two 500 μm thick silicon particle detectors. The detectors were placed at a distance of approximately 30 cm from the target, at angles of 45° and 100° with respect to the incident beam direction. Careful positioning of the foil in the beam halo is necessary to avoid saturation of the detector. The (full absorbing) detector measures the energy spectrum of the scattered particles with a strong peak at the beam energy. A fast detector (e.g. diamond) enables also a bunch length method with this technique [270].

It is possible to measure the energy of a laser- or gas-stripped electron of a H^- beam. Beam electrons have the same velocity as the beam and therefore an energy of 1/1836 of the beam protons. A 200 MeV H^- beam yields 109 keV electrons. The beam energy spectrum can then be determined by measuring the electron charge versus repelled voltage on a FC [271].

In Ref. [272] a longitudinal movement of a Feschenko-type monitor was proposed while the bunch shape functions are measured along a phase axis φ. Measuring Δφ and d one can find the beam velocity β.

11 Machine protection systems

For this quite large topic I would like to refer to the comprehensive report of R. Schmidt on “Machine Protection” at CAS 2008 in Dourdan, France. Most of the pictures of the “Little Shop of Horrors” from the talk can be found in the recent ICFA Advanced Beam Dynamics Workshops “High Intensity High Brightness Hadron Beams”.

12 Tune and chromaticity

During the presentations within this CAS a question about tune and chromaticity measurement was asked. The answer was given in the diagnostic talk “on the fly” and was not prepared as a special topic of high-intensity diagnostics in this report. Therefore, I would like to refer to the 5th workshop in the framework of CARE-N3-HHH-ABI , Novel Methods for Accelerator Beam Instrumentation, "Schottky, Tune and Chromaticity Diagnostic (with real time feedback)", 11–13 December 2007 in Hotel Prieuré, 74404 Chamonix Mont-Blanc, France, for tutorials and details about this diagnostic as well as for Schottky diagnostics [273].

References

[1] J.D. Gilpatrick, *AIP Conf. Proc.* **737** (2004) 365–371.

[2] S. Lee, *et al.,* The beam diagnostics system in the J-PARC linac, Proc. LINAC 2004, Lübeck, Germany.

[3] T. Toyama, *et al*., Beam diagnostics at the first beam commissioning of the J-PARC MR, Proc. PAC09, Vancouver, BC, Canada.

[4] H. Hotchi, *et al*., *Phys. Rev. STAB.* **12** (2009) 040402.

[5] S. Lee, *et al.,* Design study of a nondestructive beam profile and halo monitor based on residual gas ionization for the J-PARC RCS, Proc. 14th Symposium on Accelerator Science and Technology, Tsukuba, Japan, November 2003.

[6] B. Fellenz and J. Crisp, An improved resistive wall monitor, Proc. of the Beam Instrumentation Workshop (BIW98), Stanford, CA, 1998, pp. 446–445.

[7] R.C. Webber, Longitudinal emittance an introduction to the concept and survey of measurement techniques including design of a wall current monitor, Proc. of the Beam Instrumentation Workshop (BIW 1989), Upton, NY, 1989.

[8] R.C. Webber, Tutorial on beam current monitoring, FERMILAB-CONF-00-119, June 2000, and Proc. 9th Beam Instrumentation Workshop (BIW 2000), Cambridge, MA, 8–11 May 2000.

[9] A. D'Elia, R. Fandos and L. Soby, High bandwidth wall current monitor for CTF3, Proc. 11th Biennial European Particle Accelerator Conference (EPAC'08), Genoa, Italy, 23–27 June 2008.

[10] R.C. Webber, Charged particle beam current monitoring tutorial, FERMILAB-CONF-94-333, October 1994 and Proc. 6th Beam Instrumentation Workshop (BIW 94), Vancouver, BC, Canada, 2–6 October 1994.

[11] T. Bohl and J.F. Malo, The APWL wideband wall current monitor, CERN-BE-2009-006, February 2009.

[12] R.C. Webber, A tutorial on non-intercepting electromagnetic monitors for charged particle beams, FERMILAB-PUB-07-394-APC, July 2007 and Proc. AccApp'07, 8th International Topical Meeting on Nuclear Applications and Utilization of Accelerators, 30 July–2 August 2007, Pocatello, Idaho.

[13] Bergoz Instrumentation, http://www.bergoz.com

[14] K.B. Unser, *AIP Conf. Proc.* **252** (1992) 266-275

[15] K.B. Unser, *IEEE Trans. Nucl. Sci.* **16** (1969) 934–938.

[16] P. Odier, DCCT technology review, CARE-Conf-04-023-HHH, Lyon, France, 1–2 December 2004 and A. Peters, H. Schmickler and K. Wittenburg, eds., Proc. CARE-HHH-ABI Workshop on DC Current Transformers and Lifetime Calculations.

[17] S. Hiramatsu and M. Arinaga, Wideband beam DCCTs with parallel feedback circuits, Tsukuba KEK, KEK-Preprint-99-117.

[18] K. Knaack and M. Lomperski, Lifetime calculations at DESY: improving the reaction time of the measurements in the presence of low frequency DCCT noise, Proc. CARE-HHH-ABI Workshop on DC Current Transformers and Lifetime Calculations, CARE-CONF-2004-023-HHH, Lyon, France, 1–2 December 2004. Ed. A. Peters, H. Schmickler and K. Wittenburg

[19] A.J. Burns, *et al*., Real time monitoring of LEP beam currents and lifetimes, Proc. 4th European Particle Accelerator Conference (EPAC94), London, 27 June–1 July 1994.

[20] R. Neumann, Comparison: ACCT–DCCT, Proc. CARE-HHH-ABI Workshop on DC Current Transformers and Lifetime Calculations, CARE-CONF-2004-023-HHH, Lyon, France, 1–2 December 2004. Ed. A. Peters, H. Schmickler and K. Wittenburg

[21] R. Witkover, *Nucl. Instrum. Meth.* **137** (1976) 203–211.

[22] A.V. Feschenko and P.N. Ostroumov, Bunch shape monitor and its application for an ion linac tuning, Proc. 1986 Linac Conf., Stanford, CA, 2–6 June 1986.

[23] N.Y. Vinogradov, *et al*., Bunch shape measurement of CW heavy-ion beam, Proc. LINAC2002, Gyeongju, Korea, 2002.

[24] S. K. Esin, *et al*., A three dimensional bunch shape monitor for the CERN Proton Linac, Proc. 18th International Linear Accelerator Conference (Linac96), Geneva, Switzerland, 26–30 August 1996.

[25] A.V. Feschenko, *et al*., The first results of bunch shape measurements in SNS Linac, Proc. LINAC 2004, Lübeck, Germany.

[26] A.V. Feschenko, *et al*., Bunch shape monitors for the DESY H- Linac, Proc. 1997 Particle Accelerator Conference, 12–16 May 1997, vol. 2, pp. 2078–2080.

[27] A.V. Feschenko, *et al*., Longitudinal beam parameters study in the SNS Linac, Proc. PAC07, Albuquerque, NM, 2007.

[28] E.S. McCrory, Use of an INR-style bunch-length detector in the Fermilab Linac, Proc. 16th International LINAC Conference, Ottawa, ON, Canada, 23–28 August 1992.

[29] A.V. Feschenko and V.A. Moiseev, Space charge effects in bunch shape monitors, Proc. XX International Linac Conference, The Monterey Conference Center, Monterey, CA, 21–25 August 2000.

[30] Yu.V. Bylinsky, *et al*., Bunch length and velocity detector and its application in the CERN Heavy Ion Linac, Proc. 4th European Particle Accelerator Conference (EPAC94), London, 27 June–1 July 1994.

[31] P. Forck and C. Dorn, Measurements with a novel non-intercepting bunch shape monitor at the high current GSI-LINAC, Proc. DIPAC 2005, Lyon, France.

[32] T. Suwada, *et al*., Recalibration of a wall-current monitor using a faraday cup for the KEKB injector linac, Proc. 1999 Particle Accelerator Conference, New York, 1999.

[33] P. Forck, Lecture notes on beam instrumentation and diagnostics, Proc. Joint University Accelerator School, January–March 2009.

[34] R.W. Muller and P. Strehl, *Nucl. Instrum. Meth.* **415** (1998) 305–309.

[35] M. Ferianis, *et al*., Characterisation of fast Faraday cups at the ELETTRA Linac, Proc. DIPAC 2003, Mainz, Germany, 2003.

[36] W.R. Rawnsley, *et al*., *AIP Conf. Proc.* **546** (2000) 547–554.

[37] S. Battisti, *et al*., Magnetic beam position monitors for the LEP Pre-Injector, Proc. CERN-PS-87-37, March 1987 and Proc. 12th IEEE Particle Accelerator Conference (PAC), Washington, DC, 16–19 March 1987.

[38] G.R. Lambertson, Electromagnetic detectors, LBL-26075-mc (microfiche), 1989, Presented at 3rd Joint U.S.-CERN School on Particle Accelerators, Capri, Italy, 20–26 October 1988.

[39] R.E. Shafer, Beam position monitoring, Proc. Accelerator Instrumentation, Upton NY, 23-26 October 1989. Eds. E. R. Beadle, V. J. Castillo (AIP Conf. Proc. 212, 1990), p. 26.

[40] O.R. Jones, LHC beam instrumentation, Proc. Particle Accelerator Conference (PAC 07), Albuquerque Convention Centre, Albuquerque, NM, USA, 25–29 June 2007.

[41] P. Kowina, Optimisation of "shoe-box type" beam position monitors using the finite element methods, Proc. DIPAC 2005, Lyon, France.

[42] R.E. Shafer, *IEEE Trans. Nucl. Sci.* **32** (1985) 1933–1937.

[43] Y.-F. Ruan, *et al*., *Chin. Phys. C* **33** (2009) 240–244.

[44] R.E. Shafer, Beam position monitor sensitivity for low-β beams, LA-UR--93-3724, Proc. Beam Instrumentation Workshop, Santa Fe, NM, 20–23 October 1993.

[45] G. Vismara, Signal processing for beam position monitors, CERN-SL-2000-056-BI, Proc. 9th Beam Instrumentation Workshop, Cambridge, MA, 8–11 May 2000.

[46] O.R. Jones and H. Schmickler, The measurement of *Q*' and *Q*'' in the CERN-SPS by head–tail phase shift analysis, CERN-SL-2001-020-BI, Proc. PAC2001, Chicago, IL, June 2001.

[47] J.M. Byrd, Bunched beam signals in the time and frequency domain, Proc. Joint US–CERN–Japan–Russia School on Particle Accelerators, Beam measurement, Montreux and CERN, Switzerland, 11–20 May 1998 (World Scientific, Singapore, 1999).

[48] B.K. Scheidt, Upgrade of the ESRF fluorescent screen monitors, Proc. DIPAC 2003, Mainz, Germany, 2003.

[49] R. Jung, *et al*., Single pass optical profile monitoring, Report No. CERN-AB-2003-064, Proc. DIPAC 2003, Mainz, Germany, 2003.

[50] A. Peters, *et al*., 2D-characterization of ion beams using viewing screens, Proc. 8th European Particle Accelerator Conference (EPAC'02), Cité de la Science et de l'Industrie, La Villette-Paris, 3–7 June 2002.

[51] P. Forck, *et al*., Scintillation screen investigations for high current ion beams at GSI Linac, Proc. Beam Instrumentation Workshop (BIW08), Granlibakken Conference Center and Lodge in Tahoe City, CA, USA, 4–8 May 2008.

[52] C. Bal, Scintillating screens study for LEIR / LHC heavy ion beams, Report CERN-AB-2005-067, Proc. DIPAC 2005, Lyon, France (CERN, Geneva, 2005).

[53] G. Kube and W. Lauth, Investigation of the light yield of luminescent screens for high energy and high brilliant electron beams, Proc. DIPAC09, Basel, Switzerland, 2009.

[54] A. Murkov, *et al.*, Limitations on the resolution of YAG:CE beam profile monitor for high brightness electron beam, Proc. 2nd ICFA Advanced Accelerator Workshop on the Physics of High Brightness Beams, San Jose, CA, 15–18 November 2009.

[55] J. Bosser, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **238** (1985) 45–52.

[56] V. Scarpine, OTR imaging of intense 120 GeV protons in the NuMI beamline at FNAL, Proc. Particle Accelerator Conference (PAC 07), Albuquerque Convention Center, Albuquerque, NM, 25–29 June 2007.

[57] Workshop on Scintillating Screen Applications @ GSI, GSI Helmholtz Centre for Heavy Ion Research Darmstadt, Germany, 14–15 February 2011. http://www-bd.gsi.de/ssabd/proceedings.htm

[58] S. Hutchins, *et al*., Radiation tests on solid state cameras for instrumentation, CERN-AB-2005-061, Proc. DIPAC 2005, Lyon, France.

[59] E.J. Sternglass, *Phys. Rev.* **108** (1957) 1 and J.E. Borovsky and D.M. Suszeynsky, *Phys. Rev. A* **43** (1991) 1416.

[60] Z. Pavlovic, Studies of beam heating of proton beam profile monitor SEMs, FERMILAB-TM-2312-AD, May 2005.

[61] L. Badano, *et al*., Segmented foil SEM grids for high-intensity, *Proc. DIPAC 2007*, Venice, Italy, 2007.

[62] J. Camas, *et al*., Screens versus SEM grids for single pass measurements in SPS, LEP and LHC, CERN SL/95-62 (BI).

[63] J. Camas, *et al*., High sensitivity beam intensity and profile monitors for the SPS extracted beams, Proc. 1993 Particle Accelerator Conference, Washington, DC, 17–20 May 1993.

[64] D.A.G. Neet, *IEEE Trans. Nucl. Sci.* **16** (1969). 914-918

[65] R. Dölling, *et al*., Profile measurement of scanning proton beam for LISOR using carbon fibre harps, Proc. 10th Beam Instrumentation Workshop Brookhaven National Laboratory, Upton, NY, 6–9 May 2002.

[66] A.G. Afonin, *et al*., Wide range extracted beam intensity measurement at the IHEP, Proc. 6th European Particle Accelerator Conference (EPAC'98), Stockholm City Conference Centre, 22–26 June 1998.

[67] Z. Pavlovich, *et al*., Studies of beam heating of proton beam profile monitor SEM's, Fermilab Batavia TM-2312, July 2005.

[68] G. Ferioli and R. Jung, Evolution of the secondary emission efficiencies of various materials measured in the CERN SPS secondary beam lines, CERN-SL-97-071-BI, December 1997 and Proc. 3rd European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators (DIPAC 97), Frascati, Italy, 12–14 October 1997, pp. 168–170.

[69] H. Huang, *et al*., Multi-wire beam profile monitor in the AGS, BNL No. 63910, Proc. 17th Particle Accelerator Conference 1997, Vancouver, BC, Canada, 12–16 May 1997.

[70] M. Benedikt, *et al*., Injection matching studies using turn by turn beam profile measurements in the CERN PS, Proc. DIPAC 2001, 5th European Workshop on Diagnostics and Beam Instrumentation, ESRF, Grenoble, France, 13–15 May 2001.

[71] C.M. Bhat, *et al*., Envelope and multi-slit emittance measurements at Fermilab A0 photoinjector and comparison with simulations, FERMILAB-CONF-07-321-AD, June 2007 and Proc. Particle Accelerator Conference (PAC 07), Albuquerque, NM, 25–29 June 2007.

[72] M.A. Clarke-Gayther, A combined function beam emittance and profile measuring system for the ISIS 665 keV H - pre- injector, Proc. 6th European Particle Accelerator Conference (EPAC'98), Stockholm City Conference Centre, 22–26 June 1998.

[73] M. Zhang, Emittance formula for slits and pepper-pot measurement, TM-1988, Fermilab, 1996.

[74] B. Cheymol, *et al*., Design of a new emittance meter for LINAC4, Proc. 9th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators, Mercure Hotel Europe, Basel, Switzerland, 25–27 May 2009.

[75] P. Forck, *et al*., Measurement of the six-dimensional phase space at the new GSI high-current LINAC, Proc. Linac 2000, XX International Linac Conference, Monterey Conference Center and Monterey Double Tree Hotel, Monterey, CA, 21–25 August 2000.

[76] G. Penco, *et al*., Beam emittance measurement for the new full energy injector at ELETTRA, Proc. 11th biennial European Particle Accelerator Conference (EPAC'08), Genoa, Italy, 23–27 June 2008.

[77] G. Arduini, *et al*., New methods to derive the optical and beam parameters in transport channels, *Nucl. Instrum. Meth. Phys. Res. A* **459** (2001) 16–28.

[78] I. Borchardt, *et al*., *Eur. Phys. J. C Particles Fields* **39** (2005) 339–349.

[79] D. Stratakis, *et al*., *Phys. Rev.* **9** (2006) 112801.

[80] A.C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging* (IEEE Press, New York, 1988).

[81] F.E. Hannon, *et al*., Phase space tomography using the Cornell ERL DC Gun, Proc. 11th Biennial European Particle Accelerator Conference (EPAC'08), Genoa, Italy, 23–27 June 2008.

[82] D. Reggiani, *et al*., Transverse phase-space beam tomography at PSI and SNS proton accelerators, Proc. 46th ICFA Advanced Beam Dynamics Workshop on High-Intensity and

High-Brightness Hadron Beams (HB2010), Morschach, Switzerland, 27 September to 1 October 2010.

[83] W. Blokland, *et al*., A new flying wire system for the Tevatron, Proc. 1997 Particle Accelerator Conference, Vancouver, BC, Canada, 12–16 May 1997.

[84] S. Igarashi, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **482** (2002) 32–41.

[85] Ch. Steinbach and M. van Rooij, A scanning wire beam profile monitor, Proc. 1985 Particle Accelerator Conference, Vancouver, BC, Canada, 13–16 May.

[86] C. Fischer, Ionisation losses and wire scanner heating: evaluation, possible solutions, application to the LHC, CERN Geneva, CERN-SL-99-045, Proc. DIPAC 1999, Chester, UK, 1999.

[87] F. Roncarolo and B. Dehning, Transverse emittance blow-up due to the operation of wire scanners, analytical predictions and measurements, CERN-AB-2005-042, June 2005 and Proc. Particle Accelerator Conference (PAC 05), Knoxville, TN, 16–20 May 2005.

[88] M. Koujili, *et al*., Fast and high accuracy wire scanner, Proc. DIPAC09, Basel, Switzerland.

[89] M. Sapinski, *et al*., Carbon fiber damage in accelerator beam, Proc. DIPAC09, Beam Diagnostics and Instrumentation for Particle Accelerators, Basel, Switzerland, 25–27 May 2009.

[90] L. Fröhlich, Thermal load on wirescanners, Proc. 37th ICFA Advanced Beam Dynamics Workshop on Future Light Sources, Hamburg, Germany, 15–19 May 2006.

[91] C. J. Liaw and P.R. Cameron, Carbon wire heating due to scattering in the SNS, Proc. 2001 Particle Accelerator Conference, Chicago, IL, 2001.

[92] P. Elmfors, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **396** (1997) 13–22.

[93] M. Dohlus, *et al*., Report from the HERA Taskforce on Luminosity Optimization: Theory and First Luminosity Scans, DESY HERA 03-01.

[94] M.A. Plum, *et al*., SNS linac wire scanner system, signal levels and accuracy, Proc. the XXI International Linac Conference (LINAC2002), Gyeongju, Korea, 19–23 August 2002.

[95] H. Akikawa, Wire profile monitors in J-PARC, Proc. Linear Accelerator Conference (Linac06), Knoxville, TN, 21–25 August 2006.

[96] W.B. Cottingame, *et al*., *IEEE Trans. Nucl. Sci.* **32** (1985). 1871 - 1873

[97] D.R. Swenson, *et al*., *AIP Conf. Proc.* **319** (1993) 343-352

[98] J. T. Broad and W.P. Reinhardt, *Phys. Rev. A* **14** (1976) 2159–2173.

[99] R. Connolly, *et al*., Laser beam-profile monitor development at BNL for SNS, Proc. LINAC2002, Gyeongju, Korea, 2002.

[100] A. Bosco, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **592** (2008) 162–170.

[101] Y. Liu, Laser wire beam profile monitor at SNS, Proc. 11th biennial European Particle Accelerator Conference (EPAC'08), Genoa, Italy, 23–27 June 2008.

[102] S. Lee, et al., Direct measurements of space-charge-potential in high intensity H^- beam with laser based photo neutralization method, Proc. DIPAC 2005, Lyon, France, 2005.

[103] D. Jeon, *et al*., The laser emittance scanner for 1 GeV H^- beam, Proc. 23rd Particle Accelerator Conference, Vancouver, BC, Canada, 4–8 May 2009.

[104] D.A. Lee, *et al*., A laserwire beam profile measuring device for the RAL Front End Test Stand, Proc. 8th DIPAC 2007, Venice, Mestre, Italy, 20–23 May 2007.

[105] S. Assadi, *et al*., SNS transverse and longitudinal laser profile monitors design, implementation and results, Proc. EPAC 2006, Edinburgh, Scotland, 2006.

[106] D.S.F. Crothers and J.F. McCann, *J. Phys. B* **16** (1983) 165.

[107] W. Hain, *et al*., Beam profile monitors for the HERA proton accelerators, Proc. 2nd European Particle Accelerator Conference, Nice, France, 12–16 June 1990 and DESY HERA 90-11.

[108] J. Krider, Nucl. Instrum. Meth. Phys. Res. A **278** (1989) 660–663.

[109] F. Hornstra, Nondestructive beam profile detection systems for the zero gradient synchrotron, Proc. VI Conference on High Energy Acceleration, Cambridge, MA, 1967, pp. 374–377.

[110] B. Vosicki and K. Zankel, *IEEE Trans. Nucl. Sci.* **22** (1975) 1475-1478.

[111] H. Ishimaru, *IEEE Trans. Nucl. Sci*. **24** (1977) 1681–1682.

[112] C. Böhme, *et al*., Beam test of the fair IPM prototype in COSY, Proc. DIPAC09, Basel, Switzerland, 2009.

[113] B. Dehning, *et al*., Optical design for BIPM imaging system, Proc. DIPAC 2005, Lyon, France, 2005.

[114] K. Wittenburg, a non destructive way to measure betatron mismatch at injection, Proc. 3rd European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators (DIPAC 97), Frascati, Italy, 12–14 October 1997.

[115] D. Liakin, *et al*., Test of a silicon photomultiplier for ionization profile monitor applications, Proc. 8th DIPAC 2007, Venice, Mestre, Italy, 20–23 May 2007.

[116] R.E. Thern, space-charge distortion in the Brookhaven ionization profile monitor, Proc. 1987 Particle Accelerator Conference, Washington, DC, 1987.

[117] A. Jansson, *et al*., The Tevatron ionization profile monitors, FERMILAB-CONF-06-105-AD-CD-E, May 2006 and Proc. 12th Beam Instrumentation Workshop (BIW06), Fermilab, Batavia, IL, 1–4 May 2006.

[118] V. Kamerdzhiev and J. Dietrich, Ionisation beam profile monitor at the cooler synchrotron COSY-JÜLICH, Proc. DIPAC 2003, Mainz, Germany, 2003.

[119] T. Honma, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **490** (2002) 435–443.

[120] P. Forck and A. Peters, *AIP Conf. Proc.* **773** (2004) 179-183

[121] R. Connolly, *et al*., The IPM as a halo measurement and prevention diagnostic, Proc. 29th ICFA Advanced Beam Dynamics Workshop on Beam Halo Dynamics, Diagnostics, and Collimation, in conjunction with the Beam–Beam'03 Workshop, Long Island, NY, 19–23 May 2003.

[122] J. Amundson, *et al*., Calibration of the Fermilab Booster ionization profile monitor, *Phys. Rev*. **6** (2003) 102801.

[123] T. Schotmann, Das Aufloesungsvermoegen der Restgasionisations-Strahlprofilmonitore fuer Protonenstrahlen in PETRA und HERA, Diploma thesis, DESY-HERA 93-09.

[124] W.S. Graves, *Nucl. Instrum. Meth. Phys. Res. A* **364** (1995) 19–26.

[125] R. Williamson, *et al*., Analysis of measurement errors in residual gas ionisation profile monitors in a high intensity proton beam , Proc. 11th Biennial European Particle Accelerator Conference (EPAC'08), Genoa, Italy, 23–27 June 2008.

[126] B. Dehning, *et al*., Simulation of an electron source based calibrating system for an ionisation profile monitor, CERN, Geneva, CERN-AB-2005-073 and Proc. DIPAC 2005, Lyon, France, 2005.

[127] G. Ferioli, *et al*., sensitivity studies with the SPS rest gas profile monitor, Proc. 5th European Workshop on Diagnostics and Beam Instrumentation, ESRF, Grenoble, France, 13–15 May 2001.

[128] T.W. Hardek, *et al*., *IEEE Trans. Nucl. Sci.* **28** (1981) 2219–2221.

[129] T. Kawakubo, *et al*., Non-destructive beam profile monitors in the KEK proton synchrotron, Proc. Workshop on Advanced Beam Instrumentation, Tsukuba, Japan, 22–24 April 1991, KEK Preprint 91-23.

[130] D. Sandoval, *et al*., Fluorescence-based video profile beam diagnostics: theory and experience, Presented at the 5th Annual Workshop on Beam Instrumentation, Santa Fe, NM, 20–23 October 1993.

[131] M.A. Plum, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **492** (2002) 74–90.

[132] P. Forck, *et al*., Profile measurement by beam induced fluorescence for 60 MeV/u to 750 MeV/u heavy ion beams, Proc. 10th Biennial European Particle Accelerator Conference (EPAC'06), Edinburgh, Scotland, 26–30 June 2006.

[133] R.H. Hughes, *et al*., *Phys. Rev.* **123** (1961) 2084–2086.

[134] L.W. Dotchin, *et al*., *J. Chem. Phys.* **59** (1973) 3960-3967

[135] A. Variola, *et al*., *Phys. Rev. STAB.* **10** (2007) 122801

[136] P. Forck, *et al*., Beam induced fluorescence profile monitor developments, Proc. 46th ICFA Advanced Beam Dynamics Workshop on High-Intensity and High-Brightness Hadron Beams, Morschach, Switzerland, 27 September–1 October 2010.

[137] F. Hornstra, A beam induced gas scintillation (BIGS) profile monitor, DESY-HERA 89-04.

[138] P. Ausset, *et al*., Optical transverse beam profile measurements for high power proton beams, Proc. EPAC 2002, Paris, France, 2002.

[139] O.E. Krivosheev and N.V. Mokhov, Tolerable beam loss at high-intensity proton machines, FERMILAB-Conf-00/192, August 2000 and Proc. ICFA Beam Halo and Scraping Workshop, Lake Como, WI, 13–15 September 1999.

[140] T. Wangler, RF Linear Accelerators (John Wiley & Sons, New York, 2008), p. 285.

[141] K. Wittenburg, Beam loss monitoring and control, Proc. 8th European Particle Accelerator Conference (EPAC02), La Villette, Paris, France, 3–7 June 2002.

[142] E.B. Holzer, Commissioning and operational scenarios of the LHC beam loss monitor system. Proc. 39th ICFA Advanced Beam Dynamics Workshop on High Intensity High Brightness Hadron Beams 2006 (HB2006), Tsukuba, Japan, 29 May–2 Jun 2006.

[143] R.E. Shafer, *et al*., The TEVATRON beam position and beam loss monitoring systems, FERMILAB-CONF-83-112-E and Proc. 12th International Conference on High Energy Accelerators, 1988, pp. 609–615.

[144] K. Wittenburg, Beam loss and machine protection, Proc. 33rd ICFA Workshop, Bensheim, Germany, 18–22 October 2004.

[145] K. Wittenburg, The PIN-diode beam loss monitor system at HERA, Proc. 9th Beam Instrumentation Workshop (BIW2000), Boston, MA, 8–11 May 2000.

[146] A. Arauzo, *et al*., LHC beam loss monitors, Proc. 5th European Workshop on Diagnostics and Beam Instrumentation, ESRF, Grenoble, France, 13–15 May 2001.

[147] K. Wittenburg, Beam loss monitors, CAS CERN Accelerator School on Beam Diagnostics, Dourdan, France, 2008, CERN-2009-005.

[148] S. Lee, *et al*., The beam loss monitor system of the J-PARC linac, 3 GeV RCS and 50 GeV MR, Proc. 9th European Particle Accelerator Conference EPAC 2004, Lucerne, Switzerland, 5–9 July 2004.

[149] S. Assadi and A. Zhukov, Beam-loss measurement and simulation of low-energy SNS linac, Proc. LINAC 2006, Knoxville, TN, 2006.

[150] R.E. Shafer, A tutorial on beam loss monitoring, Proc. 10th Beam Instrumentation Workshop 2002, Brookhaven, May 2002.

[151] B. Dehning, *et al*., The LHC beam loss measurement system, Proc. 22nd Particle Accelerator Conference (PAC'07), Albuquerque, NM, 25–29 June 2007.

[152] D. Gassner and R.L. Witkover, Design and testing of the new ion chamber loss monitor for SNS, Proc. 2003 Particle Accelerator Conference, Portland, OR, 12–16 May 2003.

[153] J. Wu, *et al*., Readout process and noise elimination firmware for the Fermilab beam loss system, FERMILAB-CONF-07-095-E, May 2007 and Proc. 15th IEEE Real Time Conference 2007 (RT 07), Batavia, IL, 29 April–4 May 2007.

[154] R.L. Witkover, E. Zitvogel and R. Michnoff, RHIC beam loss monitor system design, Proc. 17th Particle Accelerator Conference (PAC97), Vancouver, BC, Canada, 12–16 May 1997.

[155] N. Nahagawa, *et al*., *Nucl. Instrum. Meth.* **174** (1980) 401–409.

[156] S. Childress, NuMI proton beam diagnostics and control: achieving 2 megawatt capability, Proc. 42nd ICFA Advanced Beam Dynamics Workshop on High-Intensity, High-Brightness Hadron Beams (HB2008), Nashville, TN, 2008.

[157] E. Effinger, *et al*., Single gain radiation tolerant LHC beam loss acquisition card, Proc. DIPAC 2007, Venice, Italy, May 2007, pp. 48–50.

[158] T. Toyama, *et al*., Beam loss monitoring using proportional counters at J-PARC, Proc. 42nd ICFA Advanced Beam Dynamics Workshop on High-Intensity, High-Brightness Hadron Beams (HB2008), Nashville, TN, 2008.

[159] D. McCormick, Fast ion chambers for SLC, Conference Record of the 1991 IEEE Particle Accelerator Conference: Accelerator Science and Technology, vol. 2, 6–9 May 1991, pp. 1240–1242.

[160] P. Michel, *et al*., Beam loss monitoring with long ionization chambers at ELBE, a beam loss monitor with longitudinal resolution, Radiation Source ELBE Annual Report 2002, FZR-375, April 2003.

[161] P. Michel, *et al*., Beam loss detection at radiation source ELBE, Proc. 6th European Workshop on Diagnostics and Beam Instrumentation (DIPAC 2003), 5–7 May 2003, Mainz, Germany.

[162] J. Balsamo, *et al*., *IEEE Trans. Nucl. Sci.* **24** (1977) 1807–1809.

[163] E.R. Beadle, *et al*., The AGS booster beam loss monitor system, Proc. 1991 IEEE Particle Accelerator Conference, San Francisco, CA, 1991, pp. 1231-1233.

[164] N. Nahagawa, *et al*., *Nucl. Instrum. Meth.* **174** (1980) 401–409.

[165] M.A. Clarke-Gayther, *et al*., Global beam loss monitoring using long ionisation chambers at ISIS, Proc. 4th European Particle Accelerator Conference (EPAC '94), London, UK, 27 June–1 July 1994.

[166] K. Wittenburg, *Nucl. Instrum. Meth. A* **270** (1988) 56–61; DESY 87-070 (July 1987); and http://www.microsemi.com/brochures/pindiodes/appendix%20f.pdf

[167] T.I. Meyer, PIN photodiodes for radiation monitoring and protection in the BABAR silicon vertex tracker, *Proc. the Meeting of the Division of Particle and Fields of The American Physical Society*, Columbus, OH, 9–12 August 2000, SLAC-PUB-8651, BABAR-PROC-00/38.

[168] O. Biebel, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **403** (1998) 185-527.

[169] K. Wittenburg, The PIN diode beam loss monitor system at HERA, Proc. 9th Beam Instrumentation Workshop (BIW2000), Boston, MA, 8–11 May 2000.

[170] K. Wittenburg, Reduction of the sensitivity of PIN diode beam loss monitors to synchrotron radiation by use of a copper inlay, DESY HERA 96-06.

[171] F. Ridoutt, Das Ansprechvermögen des PIN Dioden Strahlverlustmonitors, DESY Internal Note PKTR-note No. 91 (1993).

[172] E. Morré, Eine Untersuchung für das Zeus-Experiment, Diploma Thesis, University of Hamburg, August 1992.

[173] H. Burkhardt and I. Reichel, Correction of the LEP beam loss monitors for known saturation effects, CERN SL Note 96-26 OP.

[174] T. Lefevre, *et al*., Beam loss monitoring at the CLIC Test Facility 3, CERN-AB-2004-092, CLIC-NOTE-0611, Proc. 9th European Particle Accelerator Conference (EPAC'04), Lucerne, Switzerland, 5–9 July 2004.

[175] D. Kramer, *et al*., Secondary electron emission beam loss monitor for LHC, Proc. 8th European Workshop on Diagnostics and Beam Instrumentation (DIPAC 2007), Venice, Mestre, Italy, 20–23 May 2007.

[176] E.J. Sternglass, *Phys. Rev.* **108** (1957) 1-12.

[177] V. Agoritsas and C. Johnson, EMI aluminum cathode electron multipliers: CERN TESTS, CERN MPS/CO Note 71-51 (1971).

[178] L. Fröhlich, Experience from the commissioning of the FLASH machine protection system, Proc. 37th ICFA Advanced Beam Dynamics Workshop on Future Light Sources, Hamburg, Germany, 15–19 May 2006.

[179] L. Fröhlich, First operation of the FLASH machine protection system with long bunch trains, Proc. 2006 Linear Accelerator Conference (Linac06), Knoxville, TN, 21–25 August 2006.

[180] C. Leroy, AIP Conf. Proc. **958** (2007) 92–100.

[181] W.R. Leo, *Techniques for Nuclear and Particle Physics Experiments* (Springer-Verlag, Berlin, 1987).

[182] R.C. Fernow, *Introduction to Experimental Particle Physics* (Cambridge University Press, Cambridge, 1986).

[183] S.J. Payne and S. Whitehead, Fine spatial beam loss monitoring for the ISIS proton synchrotron, Proc. 10th biennial European Particle Accelerator Conference (EPAC'06) Edinburgh, Scotland, 26–30 June 2006.

[184] Saint-Gobain, http://www.detectors.saint-gobain.com

[185] W.C. Sellyey, *et al*., Experience with photomultiplier base beam loss monitors (PMBLM) at the Low Energy Demonstration Accelerator (LEDA), Proc. 2001 Particle Accelerator Conference, Chicago, IL, 2001.

[186] J. Perry, *et al.*, The CEBAF beam loss sensors, Proc. 15th IEEE Particle Accelerator Conference, Washington, DC, 17–20 May 1993.

[187] T. Kawakubo, *et al*., High speed beam loss monitor and its deterioration by radiation, Proc. 9th European Particle Accelerator Conference (EPAC'04), Lucerne, Switzerland, 5–9 July 2004.

[188] M. Korfer, *et al*., Beam loss position monitor using Cerenkov radiation in optical fibers, Proc. 7th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators, Lyon, France, 6–8 June 2005.

[189] F. Rüdiger, *et al*., Beam loss position monitoring with optical fibres at DELTA, Proc. 11th European Particle Accelerator Conference (EPAC'08), Genoa, Italy, 23–27 June 2008.

[190] A.P. Zhukov, *et al*., SNS BLM system evolution: detectors, electronics, and software, Proc. 23rd Particle Accelerator Conference, Vancouver, BC, Canada, 4–8 May 2009.

[191] M. Plum, *et al*., Ion-chamber beam-loss-monitor system for the Los Alamos Meson Physics Facility, Proc. 1995 Particle Accelerator Conference, Dallas, TX, 1–5 May 1995.

[192] L. Jensen, Beam instrumentation for the CNGS facility, CERN AB-Note-2006-022 BI (2006).

[193] W.C. Sellyey, Experience with beam loss monitors in the low energy demonstration accelerator (LEDA), Proc. Beam Instrumentation Workshop (BIW2000), 2000.

[194] R. Dolling et al., Beam diagnostics at the high power proton beam lines and targets at PSI, Proc. 7th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators (DIPAC 2005), Lyon, France, 6–8 June 2005.

[195] M. Plum, Beam diagnostic at high intensity storage rings, Proc. Beam Instrumentation Workshop (BIW1993), Santa Fe, NM, 1993.

[196] Xu Mei-Hang, *et al*., *Chin. Phys. C* **33** (2009) 123–126.

[197] P. Thompson, *et al*., RHIC beam loss monitor system commissioning year 00 run, Proc. 9th Beam Instrumentation Workshop (BIW2000), Boston, MA, 8–11 May 2000.

[198] E.B. Holzer, *et al*., Beam loss monitoring system for the LHC, CERN-AB-2006-009 and Proc. IEEE Nuclear Science Symposium and Medical Imaging Conference, vol. 2, El Conquistador Resort, Puerto Rico, 23–29 October 2005, pp. 1052–1056.

[199] M. Wendt, *et al*., Beam instrumentation for future high intense hadron accelerators at Fermilab, Presented at 42nd ICFA Advanced Beam Dynamics Workshop on High-Intensity, High-Brightness Hadron Beams (HB 2008), Nashville, TN, 25–29 August 2008.

[200] V. Shiltsev, Fast pin-diode beam loss monitors at TEVATRON, FERMILAB-TM-2012, July 1997.

[201] R.P. Fliller III, *et al*., Beam diffusion measurements at RHIC, Proc. 2003 Particle Accelerator Conference, 2003.

[202] A.V. Fedotov, Mechanisms of halo formation, Proc. 29th ICFA Advanced Beam Dynamics Workshop on Beam Halo Dynamics, Diagnostics, and Collimation (Halo'03), Long Island, NY, 19–23 May 2003.

[203] P. Camron and K. Wittenburg, Halo diagnostics summary, Proc. 29th ICFA Advanced Beam Dynamics Workshop on Beam Halo Dynamics, Diagnostics, and Collimation (Halo'03), Long Island, NY, 19–23 May 2003.

[204] C.K. Allen and T.P. Wangler, *Phys. Rev. STAB* **5** (2002). 124202

[205] J. Qiang, *et al*., *Phys. Rev. STAB* **5** (2002). 124201

[206] T.P. Wangler, *et al*., Beam halo in proton linac beams, Proc. Linac 2000, 10th International Linac Conference, Monterey, CA, 21–25 August 2000.

[207] C.K. Allen and T.P. Wangler, Parameters for quantifying beam halo, Proc. 2001 Particle Accelerator Conference (PAC2001), Chicago, IL, 18–22 June 2001.

[208] D.A. Bartkoski, *et al*., The development of computational tools for halo analysis and study of halo growth in the spallation neutron source linear accelerator, Proc. 10th European Particle Accelerator Conference (EPAC06), Edinburgh, Scotland, 26–30 June 2006.

[209] J. Amundson, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **570** (2007) 1–9 and FERMILAB-PUB-06-060-CD.

[210] K. Wittenburg, Overview of recent halo diagnosis and non-destructive beam profile monitoring. Proc. 39th ICFA Advanced Beam Dynamics Workshop on High Intensity High Brightness Hadron Beams (HB2006) Tsukuba, Japan, 29 May–2 June 2006; and K. Witteburg, Beam halo and bunch purity monitoring, Lecture given at the CAS course on: "Beam Diagnostics", Le Normont Hotel, Dourdan, France, 28 May–6 June 2008, CERN-2009-005.

[211] A. Browman, *et al*., Halo measurements of the extracted beam at the Los Alamos Proton Storage Ring, Proc. Particle Accelerator Conference 2003, Portland, Oregon, 12–16 May, 2003.

[212] P. Elmfors, *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **396** (1997) 13–22.

[213] A.P. Freyberger, Large dynamic range beam profile measurements, Proc. 7th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators, Lyon, France, 6–8 June 2005.

[214] A.P. Freyberger, Large dynamic range beam profile measurements, Proc. 2003 Particle Accelerator Conference (PAC03), Portland, OR, 12–16 May 2003.

[215] D. Gassner, *et al*., Scintillator telescope in the AGS Extracted Beamline, Proc. 29th ICFA Advanced Beam Dynamics Workshop on Beam Halo Dynamics, Diagnostics, and Collimation (Halo'03), Long Island, NY, 19–23 May 2003.

[216] K. Wittenburg, *et al*., Beam tail measurements using wire scanners at DESY, Proc. 29th ICFA Advanced Beam Dynamics Workshop on Beam Halo Dynamics, Diagnostics, and Collimation (Halo'03), Long Island, NY, 19–23 May 2003.

[217] R. Valdiviez, *et al*., The final mechanical design, fabrication and commissioning of a wire scanner and scraper assembly for halo-formation measurements in a proton beam, Proc. 2001 Particle Accelerator Conference (PAC2001), Chicago, IL, 18–22 June 2001.

[218] J.H. Kamperschroer, *et al*., Analysis of data from the LEDA wire scanner / halo scraper. Proc. Particle Accelerator Conference (PAC2001), Chicago, IL, 18–22 June 2001.

[219] A. Meseck, The electron distribution in HERA and the consequences for the H1 detector after the luminosity upgrade, DESY-THESIS-2000-042 (2000).

[220] N. Simos, *et al*., thermo-mechanical response of the halo intercepts interacting with the SNS proton beam, Proc. 2001 Particle Accelerator Conference, Chicago, IL, 2001.

[221] P.A. Letnes, *et al*., Beam scraping to detect and remove halo in LHC injection, Proc. 11th Biennial European Particle Accelerator Conference (EPAC'08), Genoa, Italy, 23–27 June 2008.

[222] O. Brüning, *et al*., Measuring the effect of an external tune modulation on the particle diffusion in the proton storage ring of HERA, DESY-HERA-94-01, January 1994.

[223] K.H. Mess, *et al*., Measurement of proton beam oscillations at low frequencies, Proc. 4th European Particle Accelerator Conference (EPAC 94), vol. 2, London, 1994, pp. 1731–1733, and Hamburg DESY, Internal Report M-94-03-R.

[224] R.P. Filler, *et al*., Beam diffusion studies art RHIC, Proc. EPAC 2002, Paris, France, 2002.

[225] R. Dölling, Profile, current, and halo monitors of the PROSCAN beam lines, Proc. 11th Beam Instrumentation Workshop (BIW04), Knoxville, TN, 3–6 May 2004.

[226] M.A. McMahan, *et al*., A scintillating fiber beam halo detector for heavy ion beam diagnostics, LBL-34721; CONF-930511-501, PAC 1993.

[227] S.G. Arutunian, *et al*., Petra proton beam profiling by vibrating wire scanner, Proc. 7th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators, DIPAC05, Lyon, France, 6–8 June 2005.

[228] K. Hanke and M. Hori, Design and construction of a beam shape and halo monitor for the CERN SPL, CARE/HIPPI Document-2005-005

[229] SpectraCAM XDR Camera Systems, http://www.photonics.com/Content/ReadArticle.aspx?ArticleID=25896

[230] IMEC, http://www.imec.be

[231] B. Lyot, *Mon. Notices R. Astron. Soc.* **99** (1939) 538.

[232] T. Mitsuhashi, Beam halo observation by coronagraph, Proc. 7th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators (DIPAC05), Lyon, France, 6–8 June 2005.

[233] A. Aleksandrov, *et al*., Beam in gap measurements at the SNS front-end, Proc. Particle Accelerator Conference (PAC 2003), Portland, OR, 12–16 May 2003.

[234] K. Yoshimura, *et al*., Measurements of proton beam extinction at J-PARC, Proc. 1st International Particle Accelerator Conference (IPAC'10), Kyoto International, Conference Centre (KICC), Kyoto, Japan, 23–28 May 2010.

[235] R. Witcover, Considerations in designing a "beam-in-gap" monitor for the spallation neutron source, BNL/SNS Technical Note No. 049, 1998.

[236] R. Thurman-Keup, *et al*., Measurement of the intensity of the beam in the abort gap at the Tevatron utilizing synchrotron light, FERMILAB-CONF-05-139, 2005.

[237] S. De Santis, *et al*., Gated microchannel plate photomultipliers for longitudinal beam diagnostics, Proc. 12th Beam Instrumentation Workshop (BIW06), Fermilab, Batavia, IL, 1–4 May 2006.

[238] J.-F. Beche, *et al.*, Design of an abort gap monitor for the Large Hadron Collider, LARP Note 1, CBP Technical Note 329, 2005.

[239] S. Hutchins, *et al*., Single photon detector tests for the LHC synchrotron light diagnostics, Proc. 7th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators (DIPAC05), Lyon, France, 6–8 June 2005.

[240] M.P. Zorzano and R. Wanzenberg, Intrabeam scattering and the coasting beam in the HERA proton ring, CERN-SL-2000-072 AP.

[241] X.-L. Zhang, *et al*., *Phys. Rev. STAB* **11** (2008) 051002.

[242] K. Ehret, *et al*., *Nucl. Instrum. Meth. A* **456** (2001) 206–216.

[243] S. Spratte, Bestimmung der Wechselwirkungsrate des HERA-B Targets und Untersuchung des Coasting Beam am HERA Protonen-Ring, Thesis, 2000, Desy-Thesis-00-036.

[244] M.K. Unel and R.J. Tesarek, *Nucl. Instrum. Meth. A* **506** (2003) 7–19.

[245] E. Mahner, *et al*., *Phys. Rev. STAB* **11** (2008) 051002.

[246] R. Macek, Recent studies of the electron cloud-induced beam instability at the Los Alamos PSR; Talk at ECLOUD10, Proc. 49th Advanced Beam Dynamics Workshop, Cornell, NY, 8–12 October 2010.

[247] R.A. Rosenberg, *Nucl. Instrum. Meth. A* **453.** 2000, 507-513

[248] R.J. Macek, Electron cloud diagnostics in use at the Los Alamos PSR, Proc. PAC03, Portland, Oregon U.S.A. May 12-16, 2003

[249] C.Y. Tan, The Ecloud measurement setup in the main injector, FERMILAB-CONF-10-508-AD; Proc. 49th ICFA Advance Beam Dynamics Workshop on Electron Cloud Physics (ECLOUD10), Ithaca, New York, 8-12 Oct 2010

[250] K. Sonnad, *et al*., Simulation and analysis of microwave transmission through an electron cloud, a comparison of results, Proc. PAC07, Albuquerque, NM, 2007.

[251] S. De Santis, et al., *Phys. Rev. STAB* **13.** 2010 071002

[252] F. Caspers and F. Zimmermann, Interactions of microwaves and electron clouds, Proc. PAC09, Vancouver, BC, Canada, 2009.

[253] S. De Santis, *et al*., The TE wave transmission method for electron cloud measurements at CESR-TA, Proc. PAC09, Vancouver, BC, Canada, 2009.

[254] D. Moehl, Sources of emittance growth, Proc. CAS 2003, Zeuthen, 2003.

[255] P.J. Bryant, Beam Transfer Lines, Proc. 5th CAS, 7–18 September 1992 and CERN 94-01, Vol. 1.

[256] M.J. Syphers and T. Sen, Notes on amplitude function mismatch, SSCL-604-mc, October 1992.

[257] M. Syphers, T. Sen and D. Edwards, Amplitude Function Mismatch, SSCL-PREPRINT-438, Proc. 1993 Particle Accelerator Conference, Washington, DC, 17–20 May 1993.

[258] C. Bovet, et al., First results from betatron matching monitors installed in the CERN PSB and SPS, CERN-SL-98-037-BI, CERN-SL-98-37-BI, June 1998 and Proc. 6th European Particle Accelerator Conference (EPAC 98), Stockholm, Sweden, 22–26 June 1998.

[259] M. Benedikt, *et al.*, Injection matching studies using turn by turn beam profile measurements in the CERN PS, Proc. 5th European Workshop on Diagnostics and Beam Instrumentation (DIPAC2001), ESRF, Grenoble, France, 13–15 May 2001.

[260] G. Ferioli, *et al*., Sensitivity studies with the SPS rest gas profile monitor, Proc. 5th European Workshop on Diagnostics and Beam Instrumentation (DIPAC2001), ESRF, Grenoble, France, 13–15 May 2001.

[261] A. Jansson, *Phys. Rev. STAB* **5** (2002) 072803.

[262] T. Naito, *et al*., Beta-matching and damping observation by SR monitor at ATF DR, Proc. 6th European Particle Accelerator Conference (EPAC 98), Stockholm, Sweden, 22–26 June 1998.

[263] G. Arduini, *et al*., Energy calibration of the SPS at 450 GeV/*c* with proton and lead ion beams, AB-Note-2003-014-OP, 2003.

[264] A.S. Müller, Measurements of beam energy, Proc. CERN Accelerator School on Beam Diagnostics, Dourdan, France, 28 May–6 June 2008.

[265] K. Hanke and T. Hermanns, Measurement of the energy distribution of the CERN Linac4 160 MeV H^- beam close to the PS booster injection, CERN-sLHC-PROJECTReport-0033, February 2010.

[266] J.C. Dooling, *et al*., A real-time energy monitor system for the IPNS LINAC, Proc. XX International Linac Conference, The Monterey Conference Center, Monterey, CA, 21–25 August 2000.

[267] C. Jamet, *et al*., Phase and amplitude measurements for the SPIRAL2 accelerator, Proc. 9th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators, Mercure Hotel Europe, Basel, Switzerland, 25–27 May 2009.

[268] J. Power and M. Stettler, The design and initial testing of a beam phase and energy measurement for LEDA, Proc. Beam Instrumentation Workshop (BIW98), Stanford, CA, 1998.

[269] L. Weissman, First experience at SARAF with proton beams using the Rutherford scattering monitor, Proc. DIPAC09, Basel, Switzerland, 2009.

[270] P. Forck, Aspects of bunch shape measurements for slow, intense ion beams, Proc. DIPAC09, Basel, Switzerland, 2009.

[271] R. Connolly, Beam-energy and laser beam-profile monitor at the BNL LINAC, Proc. 2010 Beam Instrumentation Workshop (BIW10), La Fonda on the Plaza, Santa Fe, NM, 2–6 May 2010.

[272] Yu.V. Bylinsky, Bunch length and velocity detector and its application in the CERN Heavy Ion Linac, Proc. of the 4th European Particle Accelerator Conference EPAC94, London, 27 June–1 July 1994.

[273] Proc. 5th Workshop in the Framework of CARE-N3-HHH-ABI, Novel Methods for Accelerator Beam Instrumentation; "Schottky, Tune and Chromaticity Diagnostic (with real time feedback), Hotel Prieuré, Chamonix Mont-Blanc, France,11–13 December 2007, CARE-Conf-08-003-HHH. Ed. K. Wittenburg

Vacuum I

Giuliano Franchetti
GSI Darmstadt, D-64291 Darmstadt, Germany

Abstract
This paper is an introduction to the basics of vacuum. It is intended for readers unfamiliar with the topic; more advanced treatments are left to the dedicated CERN Accelerator School on vacuum in accelerators and to the specialized literature. The kinetics of gases, and gas flows through pipes and pumps are reviewed here. The topic of pumps is continued in the paper 'Vacuum II'.

1 Vacuum, mean free path, and beam lifetime

In accelerators, the beam dynamics has the general purpose of controlling the beam particles. For charged particles, this happens through the Lorenz force:

$$\frac{\mathrm{d}m\gamma\vec{v}}{\mathrm{d}t} = q\vec{E} + q\vec{v}\times\vec{B}, \tag{1}$$

where q is the charge of the particle and m is the mass. The time and space structure of the electric and magnetic fields $\vec{E}$ and $\vec{B}$ in an accelerator has the purpose of guiding beam particles along the design trajectory. However, in a particle accelerator there is also present a jungle of unwanted particles, which creates a background that is damaging to beam operation and experiments. These particles are referred to as 'residual gas'. As the particles that make up the 'vacuum' are not controlled by an electromagnetic field, they are in general treated as a gas in the thermodynamic sense, which is therefore characterized by macroscopic quantities such as the pressure P, temperature T, particle density $\tilde{n}$, and composition. The statistical behaviour of vacuum particles is described by kinetic theory [1, 2]. In accelerators, the aim of all of the vacuum systems is to minimize the interaction of the beam with the residual gas. Table 1 lists some typical values of particle densities for different types of vacuum.

Table 1: Examples of particle densities (from Ref. [3])

	Particles m^{-3}
Atmosphere	2.5×10^{25}
Vacuum cleaner	2×10^{25}
Freeze-dryer	10^{22}
Light bulb	10^{20}
Thermos flask	10^{19}
TV tube	10^{14}
Low Earth orbit (300 km)	10^{14}
H_2 in LHC	$\sim 10^{14}$
SRS/Diamond	10^{13}
Surface of Moon	10^{11}
Interstellar space	10^{5}

One fundamental process in a gas is particle–wall collisions. The velocity of a particle after interaction with a wall depends on the particle–wall processes that occur. However, in any case, these collisions are responsible for creating the macroscopic pressure. The SI unit of pressure is the pascal, Pa = $\mathrm{N/m}^2$. Although this is the standard unit, there are also other units in common use, such as the

Table 2: Classification of types of vacuum

Low vacuum	Atmospheric pressure to 1 mbar
Medium vacuum	1 to 10^{-3} mbar
High vacuum (HV)	10^{-3} to 10^{-8} mbar
Ultrahigh vacuum (UHV)	10^{-8} to 10^{-12} mbar
Extreme high vacuum (XHV)	Less than 10^{-12} mbar

bar, torr, and atmosphere, which are related to each other by 1 Pa $= 10^{-2}$ mbar = 7.5×10^{-3} Torr = 9.87×10^{-6} atm. Table 2 shows the classification of types of vacuum in terms of pressure.

Another important process in a gas is particle–particle collisions. This process is the most probable process, as collisions between three or more particles are possible but rather unlikely. In elastic collisions, two fundamental quantities are preserved: energy and momentum. Particle–particle collisions are responsible for creating a distribution of velocities in the particles of a gas. The temperature of the gas is related to the mean kinetic energy of a particle: for monatomic particles, $\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}k_{\mathrm{B}}T$, where $k_{\mathrm{B}} = 1.38 \times 10^{-23}$ J·K^{-1} is the Boltzmann constant. For example, for air, assumed to be composed mainly of nitrogen N_2, at $T = 20°\mathrm{C}$, we find $\sqrt{\langle v^2\rangle} = 511$ m/s. This is not the average velocity v_{a}, which is slightly smaller; in fact, $v_{\mathrm{a}} = \langle v\rangle = 0.92\sqrt{\langle v^2\rangle} = 470$ m/s. When the gas is in equilibrium, the velocity distribution of the molecules follows the Maxwell–Boltzmann distribution

$$\frac{1}{N}\frac{\mathrm{d}N}{\mathrm{d}v} = \frac{4}{\sqrt{\pi}}\left(\frac{m}{2k_{\mathrm{B}}T}\right)^{3/2} v^2 e^{-mv^2/2k_{\mathrm{B}}T}. \tag{2}$$

This distribution is shown in Fig. 1. The most probable velocity is $v_\alpha = \sqrt{2k_{\mathrm{B}}T/m}$, and the average velocity is $v_{\mathrm{a}} = \sqrt{(8/\pi)(k_{\mathrm{B}}T/m)}$. As previously discussed, a gas is characterized by its pressure,

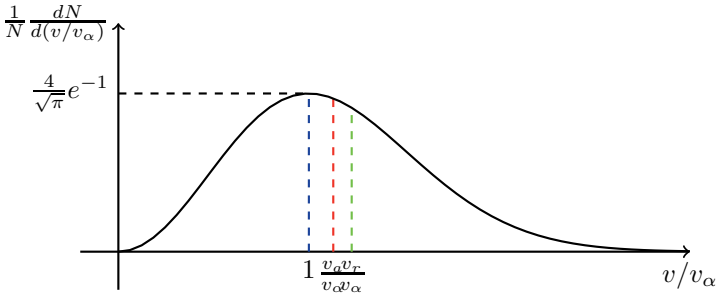


Fig. 1: Maxwell–Boltzmann distribution

temperature, and volume. When the gas is in equilibrium, i.e., in a stationary state, these quantities are related by the equation of state $PV = nR_0T$, where $R_0 = 8.31$ N·m/(mole·K). The pressure is measured in pascals and the volume in m^3, n is the number of moles (1 mole contains a number of particles equal to Avogadro's number N_{A}), $R_0 = N_{\mathrm{A}}k_{\mathrm{B}}$, and T is the absolute temperature, expressed in kelvin.

In vacuum physics, the concept of the mean free path plays an important role in determining the behaviour of a gas. Consider a set of particles at rest, and let the particle density of this distribution be $\tilde{n}$. Now consider N test particles distributed in a plane of area A. If this plane now moves 'orthogonally' through the remainder of the gas, when the plane is displaced by a distance Δl, the test particles will have spanned a volume $A \times \Delta l$ (see Fig. 2). On the other hand, these particles have a cross-section with respect to the remainder of the gas of $\sigma = \pi r^2$, where r is the radius of the cross-section. The particles present in the spanned volume will then cover (with respect to interaction with the test particles) the

available area by an amount $\Delta A = A \times \Delta l\, \tilde{n}\sigma$. Therefore the number of particles that will pass, without collision, through the portion of gas of length Δl is $N - N(A \times \Delta l\, \tilde{n}\sigma)/A$. We obtain the differential equation

$$\frac{\mathrm{d}N}{\mathrm{d}l} = -N \times \tilde{n}\sigma. \tag{3}$$

By integrating this equation, we find the number of surviving test particles $N(l)$ at a distance l. This quantity can be interpreted as the number of particles that did not collide in a distance l, and of these particles, certainly $N(l) \times \Delta l\, \tilde{n}\sigma$ will collide with the remaining gas between l and $l + \Delta l$. Therefore the probability that a particle will travel for a distance l and then collide with the remaining gas in the interval between l and $l + dl$ is $dP(l) = (N(l)/N_0)\tilde{n}\sigma\, \mathrm{d}l$, where N_0 is the initial number of particles in the plane considered. The average distance that a test particle will travel between two collisions is then $\lambda = \int_0^\infty l\, \mathrm{d}P(l) = 1/\sigma\tilde{n}$. Note that in this argument it is assumed that the particles of the gas are at rest, which is not true in a real gas. Maxwell computed the effect of the relative motion of particles of the same species, and this adds a factor to the formula, which becomes

$$\lambda = \frac{1}{\sqrt{2}\sigma\tilde{n}}. \tag{4}$$

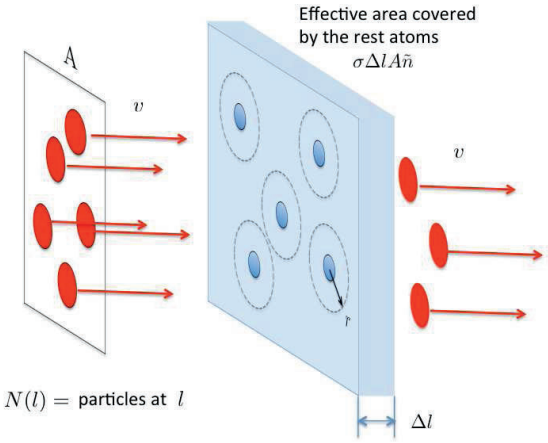


Fig. 2: Conceptual discussion of the mean free path

For example, if we consider air at $T = 20°\mathrm{C}$ and a pressure $P = 1$ atm, we find from the equation of state that $\tilde{n} = P/(k_\mathrm{B}T) = 2.47 \times 10^{25}$ atoms/m^3. The diameter d of an air molecule is 3.74×10^{-10} m [2], from which $\sigma = \pi d^2 = 4.39 \times 10^{-19}$ m^2. Therefore the mean free path is $\lambda = 6.51 \times 10^{-8}$ m. Table 3 shows the relation between the mean free path and the pressure and viscosity for N_2 at $T = 20°\mathrm{C}$, taking $d = 3.15 \times 10^{-10}$ m (obtained from measurements of viscosity [4]).

In synchrotrons and colliders, the circulating beam is disturbed by the presence of residual gas. The importance of the residual gas is of high relevance for circular accelerators, where the beam circulates for a large number of turns. The situation here is of a beam formed by a number of ions of a certain species, which while travelling may collide with residual gas molecules, become ionized, and be lost because they now have the wrong charge state with respect to the machine rigidity. The number of particles surviving a path of length Δl is

$$N(l + \Delta l) = N(l) - N(l)\sigma\tilde{n}\, \Delta l.$$

Assuming that all beam particles have the same longitudinal velocity v_0, we can change the space variable to a time variable, and so we find

$$N(t + \Delta t) = N(t) - N(t)\sigma\tilde{n}v_0\, \Delta t,$$

Table 3: Mean free path, particle density, and viscosity

	P (Pa)	$\tilde{n}$ (m^{-3})	ρ ($kg \cdot m^{-3}$)	η ($m^{-2} \cdot s^{-1}$)	λ (m)
Atmospheric pressure	10^{5}	2.5×10^{25}	1.16	2.9×10^{27}	9×10^{-8}
Primary vacuum	1	2.5×10^{20}	1.16×10^{-5}	2.9×10^{22}	9×10^{-3}
	10^{-1}	2.5×10^{19}	1.16×10^{-6}	2.9×10^{21}	9×10^{-2}
High vacuum	10^{-4}	2.5×10^{16}	1.16×10^{-9}	2.9×10^{18}	9×10^{1}
	10^{-7}	2.5×10^{13}	1.16×10^{-12}	2.9×10^{15}	9×10^{4}
UHV	10^{-10}	2.5×10^{10}	1.16×10^{-12}	2.9×10^{12}	9×10^{7}
XHV	10^{-11}				

and therefore

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -N(t)\sigma\tilde{n}v_0.$$

As the particle density of the residual gas is $\tilde{n} = P/(k_{\mathrm{B}}T)$, this equation becomes

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -N(t)\frac{\sigma P v_0}{k_{\mathrm{B}}T}.$$

The time constant of this equation defines the beam lifetime

$$\tau = \frac{k_{\mathrm{B}}T}{\sigma P v_0}.$$

For example, in the LHC the residual gas contains, among other species, molecules of hydrogen H_2 in thermal equilibrium with cold surfaces at a temperature of 5 K. The cross-section for interaction with the energetic protons circulating at 7 TeV sets the above cross-section to $\sigma = 9.5 \times 10^{-30}$ m^{-2}. Therefore, for a beam lifetime $\tau = 100$ h, we find a requirement for the pressure of H_2 inside the vacuum chamber of $P = 6.7 \times 10^{-8}$ Pa. Clearly, the corresponding requirements for other molecules making up the gas will differ. The pressures foreseen in the LHC for the various components of the gas in the vacuum are shown in Table 4 [5].

Table 4: Pressure of various components of the vacuum in the LHC for a target beam lifetime τ = 100 h

Gas	Nuclear scattering σ (cm^2)	Gas density (m^3)	Pressure (Pa) at 5 K
H_2	9.5×10^{-26}	9.8×10^{14}	6.7×10^{-8}
He	1.26×10^{-25}	7.4×10^{14}	5.1×10^{-8}
CH_4	5.66×10^{-25}	1.6×10^{14}	1.1×10^{-8}
H_2O	5.65×10^{-25}	1.6×10^{14}	1.1×10^{-9}
CO	8.54×10^{-25}	1.1×10^{14}	7.5×10^{-9}
CO_2	1.32×10^{-24}	7.0×10^{13}	4.9×10^{-9}

A special note has to be made here about electrons. Electrons can also be considered as part of the residual gas. Owing to their lightness, electrons are strongly influenced by the electric field of the circulating beam in the accelerator. The electrons in the vacuum chamber are often referred to as the 'electron cloud'. Reviews of the complex processes of build-up of electrons and their interplay with the circulating bunches can be found in Refs. [6, 7, 8].

As previously mentioned, the molecules of the vacuum are subject to collisions with one another, generating a Maxwellian distribution of velocities. One relevant feature of a gas is the 'impingement

rate' J. This quantity measures the number of molecules that hit a unit area of a surface per unit of time. For a Maxwell–Boltzmann velocity distribution, we find [9, 2]

$$J = \frac{1}{4}\tilde{n}v_{\mathrm{a}},$$

where $v_{\mathrm{a}} = \sqrt{(8/\pi)(k_{\mathrm{B}}T/m)}$ is the average velocity of a gas molecule.

2 Gas flows in pipes

The properties of a gas with respect to transport through pipes depend very much on the spatial scale considered, which is set by the size of the pipe or vessel. If the size D of the vessel is much larger than the mean free path λ, then the gas will be dominated by collisions between the gas molecules, and the effect of molecule–wall collisions will be negligible. In this regime, the gas will effectively behave as a continuum, as a local change in properties will be propagated as a wave (like sound in air at atmospheric pressure and room temperature). If instead the size of the vessel is much smaller than the mean free path, the molecules will collide mainly with the walls of the vessel. In this case continuum processes are not possible, and the motion of the particles of the gas is dominated by particle–wall collisions. This regime is referred to as the 'molecular regime'. The Knudsen number K_n characterizes the type of regime in which a gas is found:

$$K_n = \frac{\lambda}{D}\begin{cases} K_n < 0.01 & \text{Continuous regime,} \\ 0.01 < K_n < 0.5 & \text{Transitional regime,} \\ K_n > 0.5 & \text{Molecular regime.} \end{cases}$$

2.1 Throughput and conductances

The creation of a vacuum requires the extraction of air from a vessel, which is done via a system of pumps connected to vessels through pipes.

We now present a short discussion of gas flow in pipes. The flow of a gas through a pipe can be expressed in terms of the number of particles per second $\mathrm{d}N/\mathrm{d}t$ passing through a reference surface across the pipe. On the other hand, measurements of gas flow are better expressed in terms of macroscopic quantities characterizing the thermodynamic state of the gas. If a gas flows in a pipe at a velocity v across an area A, the rate of particles per second $\mathrm{d}N/\mathrm{d}t$ is

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \tilde{n}vA = \frac{P}{k_{\mathrm{B}}T}vA = \frac{P}{k_B T}\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{1}{k_{\mathrm{B}}T}\frac{\mathrm{d}}{\mathrm{d}t}PV. \tag{5}$$

Note that the quantity $Q = PV$ here is a product of a pressure and a volume. Equation (5) can be re-expressed so that the particle flow is

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{1}{k_{\mathrm{B}}T}\dot{Q}.$$

The quantity $\dot{Q}$ is called the throughput and is expressed in Pa·m^3/s. In the absence of adsorption or desorption, the rate at which particles pass through a cross-section of a pipe does not change along the pipe, and hence the throughput does not change either.

It is now useful to introduce the concept of the conductance of a pipe. If there is a flow of particles in a pipe from a section 1 with pressure P_1 to a section 2 with pressure P_2 ($< P_1$), the relation between the throughput and the difference in pressure between the two sections can be expressed as

$$C = \frac{\dot{Q}}{P_1 - P_2}. \tag{6}$$

Here C is the conductance; this is a geometric property of the pipe and of the gas flow. The units of C are m^3/s. For a composite structure formed from several pipes, it is possible to prove, based on the assumption that the throughput is conserved, that for N pipes each having conductance C_i, the conductance C of the composition in series is given by

$$\frac{1}{C} = \sum_{i=1}^{N} \frac{1}{C_i}.$$

For the composition of pipes in parallel, the conductance of the composite structure is

$$C = \sum_{i=1}^{N} C_i.$$

Note that the law of composition of pipes in series can be used to obtain an indication of how the conductance of a long pipe scales with the length. In fact, if C_1 is the conductance of one pipe, by connecting N of these pipes in series we find a conductance $C = C_1/N$. Therefore it is expected that the conductance scales as the inverse of the length of the pipe.

Up to this point, the concept of conductance, alias the relation between the throughput and the pressure gradient, has been discussed without considering the physical behaviour of the gas itself. It is therefore to be expected that a pipe will exhibit different conduction properties according to whether the gas is in a molecular or a continuum regime.

2.2 Molecular flow

Particle–wall collisions characterize the molecular flow of a gas in a pipe. For example, in a vessel of diameter $D = 0.1$ m, the molecular regime $K_n > 0.5$ is obtained for a pressure of $P < 1.3 \times 10^{-3}$ mbar (at room temperature).

Owing to imperfections in the surface of the pipe, the 'gas molecule–wall' collision becomes a very complex process. However, experimental evidence found by Knudsen [10] suggests that a particle hitting a surface at an arbitrary angle with respect to the normal to the surface will emerge after the interaction with an angle not correlated with the incident direction, with a probability that obeys the cosine law. This law states that the probability that a particle will have a direction within a solid angle $\mathrm{d}\omega$ is

$$\mathrm{d}P = \frac{1}{\pi} \cos\theta \, \mathrm{d}\omega.$$

See Fig. 3 for a schematic illustration of the Knudsen law. As a consequence, a particle launched inside

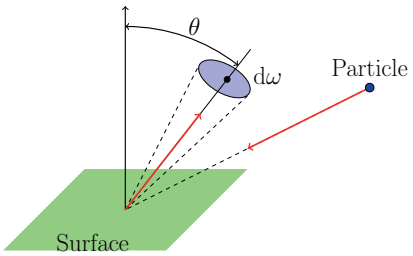


Fig. 3: Description of the cosine law

a pipe will have a certain probability α of passing through the pipe, and hence a probability $1 - \alpha$ of returning back. Here we also assume that no processes of capture by surfaces take place (i.e., no sticking). If N_1 particles per unit time enter the pipe, then $N_1\alpha$ will pass through and $N_1(1 - \alpha)$ will return back.

Therefore, in the molecular regime, there are always two opposite fluxes of particles coexisting inside a pipe.

The simplest example of a conductance is the conductance of an aperture. Consider the situation shown in Fig. 4. Here, two vessels are joined together and connected by an aperture of area A. In the left vessel, the pressure is $P_{\rm u}$, and in the right vessel the pressure is $P_{\rm d}$. The gas has the same temperature T in both vessels. The rate of flow of particles from the vessel at pressure $P_{\rm u}$ to the other vessel is $I_{\rm u} = J_{\rm u}A$. The gas throughput is obtained as $\dot{Q}_{\rm u} = P_{\rm u}\dot{V}_{\rm u}$, where $\dot{V}_{\rm u}$ is the volumetric flow of gas from the vessel at $P_{\rm u}$. The volumetric flow is obtained as $\dot{V}_{\rm u} = I_{\rm u}/\tilde{n}_{\rm u}$, and therefore $\dot{Q}_{\rm u} = P_{\rm u}I_{\rm u}/\tilde{n}_{\rm u}$. The same argument can be repeated for the gas flow from the right vessel to the left. Therefore the total throughput across the aperture is $\dot{Q} = \dot{Q}_{\rm u} - \dot{Q}_{\rm u} = P_{\rm u}I_{\rm u}/\tilde{n}_{\rm u} - P_{\rm d}I_{\rm d}/\tilde{n}_{\rm d}$. Here we also recall that $I = AJ = A\tilde{n}v_{\rm a}/4$, and therefore $I_u/\tilde{n}_{\rm u} = I_{\rm d}/\tilde{n}_{\rm d} = v_{\rm a}A/4$.

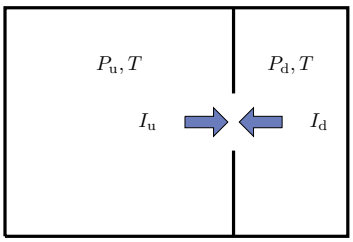


Fig. 4: Schematic illustration of the conductance of an aperture

Finally, we obtain

$$\dot{Q} = A\frac{1}{4}v_{\rm a}(P_{\rm u} - P_{\rm d});$$

comparing this formula with Eq. (6), we find that the conductance of the aperture is

$$C_{\rm a} = A\sqrt{\frac{1}{2\pi}\frac{R_0T}{M}},$$

where M is the molar mass of the gas.

The conductances of all types of pipe are usually referred to the conductance of an aperture C_a. For example, for a long pipe the conductance is [10, 2]

$$C_L = \frac{4}{3}\frac{d}{L}C_{\rm a}.$$

Note that C_L scales as the inverse of the pipe length L, as already inferred from the law of composition of pipes in series.

Observation. The expressions for the conductance in these formulas depend on the geometrical properties of the pipes, and on the type of molecules that the gas is composed of. This property stems from the flow regime, which is molecular here.

2.2.1 Continuum flow regime

For $K_n < 0.01$, there are approximately 100 collisions between particles on average before a particle collides with a wall. In this case perturbations of $\tilde{n}, P, T$ propagate through a continuum medium. In this regime, collisions between particles create the phenomenon of viscosity, which affects the behaviour of gases in pipes.

If a fluid propagates into a pipe at a 'small' speed, the fluid is in the laminar regime and the motion of the particles is parallel to the pipe axis (see Ref. [11] for extensive discussions of flow in pipes). The velocity of the fluid is zero at the walls of the pipe, and is largest at the centre. If the velocity of the fluid

is increased beyond a certain threshold, the fluid changes its properties and its motion becomes turbulent. The quantity that defines the transition from the viscous laminar regime to the turbulent regime is the Reynolds number

$$Re = \frac{\rho v D_{\mathrm{h}}}{\eta},$$

where ρ is the density of the fluid in kg/m^3, v is the average velocity of the gas in m/s, and D_{h} is the hydraulic diameter in metres, given by $D_{\mathrm{h}} = 4A/B$, where A is the cross-sectional area and B is the perimeter of the pipe. Finally, η is the viscosity in Pa·s. If $Re < 2000$, the fluid is in the laminar regime, whereas for $Re > 3000$ the fluid is in the turbulent regime.

The Reynolds number can be expressed in terms of the throughput as

$$Re = 4\frac{\dot{Q}}{B}\frac{M}{R_0 T}\frac{1}{\eta}.$$

For air (N_2) at $T = 20^{\circ}\mathrm{C}$, taking $\eta = 1.75 \times 10^{-5}$ Pa·s, we find $Re = k_{\mathrm{R}}\dot{Q}/B$, where $k_{\mathrm{R}} = 2.615$ s/(m^2·Pa). Therefore the transition to turbulent flow ($Re \sim 2000$) takes place at a transition throughput $\dot{Q}_{\mathrm{T}} = 24d$, where $\dot{Q}_{\mathrm{T}}$ is in mbar·l/s. For a pipe of diameter $d = 25$ mm, we obtain $\dot{Q}_{\mathrm{T}} = 600$ mbar·l/s, which corresponds at atmospheric pressure to a speed of $v = 1.22$ m/s.

2.2.1.1 Laminar regime

One source of complexity in characterizing a fluid flow in a pipe is the compressibility of the fluid. However, it is possible to prove that if the velocity of the fluid has a Mach number $Ma < 0.2$, then a fluid in a pipe behaves as if it were incompressible, i.e., the Bernoulli equation takes a form very similar to that of the equation for incompressible fluids.

It is important to characterize pipes in terms of the conductance when the flow regime is laminar. The conductance can be given under some assumptions about the laminar flow: (1) the fluid is considered incompressible; (2) the flow is fully developed, that is, it reaches a velocity distribution which does not change along a pipe of constant cross-section (see [2] for further discussion); (3) the particle motion is laminar ($Re < 2000$); and (4) the velocity of the fluid at the pipe walls is zero. Under these assumptions, we find that the throughput is given by $\dot{Q} = C(P_{\mathrm{u}} - P_{\mathrm{d}})$, where the conductance is now given by

$$C = \frac{\pi D^4}{128\eta L}\overline{P},$$

where $\overline{P} = (P_{\mathrm{u}} + P_{\mathrm{d}})/2$. Here P_{u} is the pressure upstream, P_{d} is the pressure downstream, D is the diameter of the pipe, L is its length, and η is the fluid viscosity. This finding can be obtained directly from the Hagen–Poiseuille equation [11]. We conclude that in a continuum laminar regime, the conductance depends on the pressure at which the fluid transport occurs.

2.2.1.2 Turbulent regime

In the turbulent regime, for a long pipe, the expression for the throughput becomes

$$\dot{Q} = A\sqrt{\frac{R_0 T}{M}}\sqrt{\frac{D_{\mathrm{h}}}{f_{\mathrm{D}} L}}\sqrt{P_{\mathrm{u}}^2 - P_{\mathrm{d}}^2},$$

where f_{D} is the Darcy friction factor, a quantity which depends nonlinearly on the Reynolds number [12].

3 Sources of vacuum degradation

After the air has been pumped out from a chamber, the vacuum can degrade because new molecules may enter the vessel.

One phenomenon that contributes to spoiling the vacuum is the evaporation and condensation of liquids present on surfaces (because those surfaces have not been cleaned). If there is a spot of liquid on a surface in a high-vacuum vessel, particles evaporate from the surface of the liquid, increasing the pressure in the vessel. At the same time, vapour particles in the vessel impinge on the liquid surface, bringing molecules into the liquid. There are therefore two fluxes of particles: one of evaporation and the other of condensation. The process of condensation depends on the pressure in the vessel, whereas the evaporation process depends only on the temperature. If one waits long enough, the pressure in the vessel will stabilize to the saturated vapour pressure P_{E}, the value of which is given by the Clausius–Clapeyron equation [13]. Therefore we conclude that the fluxes of evaporation J_{E} and condensation J_{C} are equal. Hence the evaporation flux is

$$J_{\mathrm{E}} = P_{\mathrm{E}} N_{\mathrm{a}} \frac{1}{\sqrt{2\pi R_0 M T}}.$$

A spot of liquid in a vacuum chamber will therefore emit a flux of particles J_{E}. This is equivalent to a throughput into the vessel equal to $\dot{Q}_{\mathrm{E}} = A J_{\mathrm{E}} k_{\mathrm{B}} T$, where A is the surface area of liquid exposed to the vacuum in the vessel.

Another source of vacuum degradation is the phenomenon of outgassing. Outgassing is the passage of gas from the walls of a vessel or pipe into the vacuum; it is the release of gas molecules previously adsorbed on surfaces. The locations where gas molecules are captured (and later released) are called adsorption sites. When molecules hit one of these sites, there is a certain probability of capture (see the discussion of 'sticking coefficients' in Ref. [2]). Clearly, when an adsorption site is occupied, it cannot accommodate another molecule, and the amount of gas adsorbed by a surface is proportional to the fraction of adsorption sites occupied, Θ. Typically, a molecule captured on an adsorption site remains there for a time called the 'mean stay time' τ_{d} before being released. Therefore, in a time interval $\mathrm{d}t$, a fraction

Table 5: Mean stay time at T = 20 K

E_{d} (kcal/mole)	Case	τ_{d} (s)
0.1	Helium	1.18×10^{-13}
1.5	H_2 physisorption	1.3×10^{-12}
3–4	Ar, CO, N_2, CO_2 physisorption	1.6×10^{-11}
10–15	Weak chemisorption	2.6×10^{-6}
20	H_2 chemisorption	66
25		3.3×10^{5} ($\sim$ half a week)
30	CO/Ni chemisorption	1.6×10^{9} ($\sim$ 50 years)
40		4.3×10^{16} ($\sim$ half the age of the Earth)
150	O/W chemisorption	1.35×10^{98} (larger than the age of the Universe)

$\Theta\,\mathrm{d}t/\tau_{\mathrm{d}}$ of the total number of sites present on the surface will become free and release gas. The time evolution of Θ, neglecting readsorption, then follows the equation

$$\frac{\mathrm{d}\Theta}{\mathrm{d}t} = -\frac{\Theta}{\tau_{\mathrm{d}}}.$$

The mean stay time is strongly related to the surface temperature and the binding energy of the gas molecule. As shown in [14], we find that $\tau_{\mathrm{d}} = \tau_0 \exp[E_{\mathrm{d}}/(R_0 T)]$, where $\tau_0 = 10^{-13}$ s. Table 5 shows the mean stay times of molecules with different binding energies E_{d}. The number of adsorption sites N_{s} is equal to $A \times 3 \times 10^{15}$, where A is the surface area in cm^2. The outgassing, in terms of throughput,

then becomes

$$\dot{Q}_{\mathrm{G}} = k_{\mathrm{B}} T \frac{N_{\mathrm{s}} \Theta}{\tau_{\mathrm{d}}}.$$

A third source of vacuum degradation is the presence of leaks. If a small hole (i.e., a small channel) is created in a vessel containing a high vacuum, then the throughput of gas from the outside to the inside of the vessel is given by $\dot{Q} \simeq C_{\mathrm{a}} P_0$, where P_0 is the atmospheric pressure (or, more generally, the outside pressure). In the case of air, composed mainly of N_2, at room temperature T = 293 K, for a small hole of diameter $d = 10^{-10}$ m we find a conductance $C_{\mathrm{a}} = 9.17 \times 10^{-19}$ m^3/s. Therefore the throughput is $\dot{Q}_{\mathrm{L}} = 9.17 \times 10^{-14}$ Pa·m^3/s = 9.17×10^{-13} mbar·l/s. For a hole with $d = 10^{-9}$ m, $\dot{Q}_{\mathrm{L}} \simeq 10^{-10}$ mbar·l/s, and therefore 1 cm^3 of air needs 317 years to enter the vessel. Leaks in a vacuum system can be distinguished into the following classes according to the throughput [15]: 'very tight' if $\dot{Q}_{\mathrm{L}} < 10^{-6}$ mbar·l/s, 'tight' if $10^{-6} < \dot{Q}_{\mathrm{L}} < 10^{-5}$ mbar·l/s, and 'with leaks' if $10^{-5} < \dot{Q}_{\mathrm{L}} < 10^{-4}$ mbar·l/s.

Finally, among the sources of vacuum degradation, we should mention the phenomenon of permeation: this process happens when gases are adsorbed on the material of the walls, diffuse through the walls, and are later desorbed into the vessel. The throughput depends on the surface temperature, the thickness and composition of the walls, and the composition of the gas. A discussion of this effect has been given in Ref. [2].

4 Pipes and pumps

The creation of a vacuum is achieved via pumps that are connected to vessels via pipes. An ideal pump behaves in such a way that all of the gas particles that enter the pump inlet never return. The pumping speed of a pump is referred to the volumetric speed S of the gas through the inlet, and is expressed in units of m^3/s. If the inlet of a pump has diameter D, then the gas flow through the ideal pump aperture is given by $I = JD^2\pi/4$, where I is the number of molecules per second passing through the inlet surface. The volumetric pumping speed is then $S = \mathrm{d}V/\mathrm{d}t = I/\tilde{n} = v_{\mathrm{a}} D^2 \pi/16$. For example, for N_2 at $T = 20°\mathrm{C}$, if D = 0.1 m we find a volumetric pumping speed $S_0 = 0.92$ m^3/s. Of course, this pumping speed is an ideal value and the effective pumping speed of a real pump will be smaller. The situation is summarized in Fig. 5. The throughput in the section AA is $\dot{Q} = PS$; it is the same throughout the pipe,

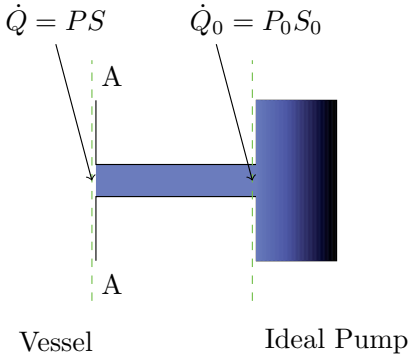


Fig. 5: Effective pumping speed S and its relation to the pumping speed S_0 of an ideal pump

and is equal to the throughput of gas entering the pump, $\dot{Q}_0 = P_0 S_0$. At the same time, $\dot{Q} = C(P - P_0)$, where C is the conductance of the pipe connecting the vessel to the pump. From these relations, we find

$$\frac{1}{S} = \frac{1}{C} + \frac{1}{S_0},$$

which gives the effective pumping speed of an ideal pump. For example, taking a long pipe, $C_L = 4D/(3L)C_\mathrm{a}$, and noting that the volumetric speed of the ideal pump is $S_0 = C_\mathrm{a}$, we find

$$S = \frac{S_0}{1 + 3L/(4D)}.$$

Therefore pumps should be placed as close as possible to vessels in order to exploit the nominal pumping capability of the pump.

5 Making the vacuum: pump-down time and ultimate pressure

We now put together all the different sources of vacuum degradation, which are summarized in Fig. 6. The quantity of gas inside a vessel Q varies according to

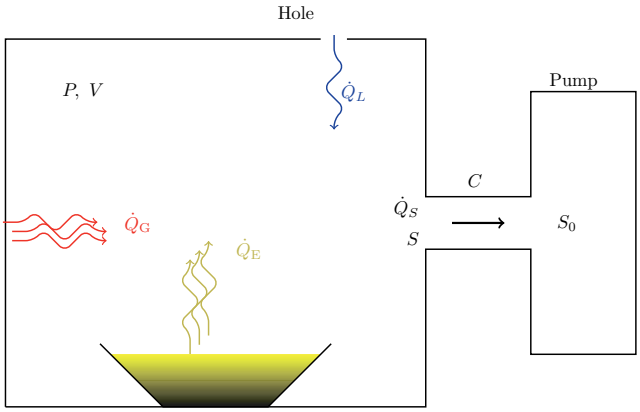


Fig. 6: Summary of sources of vacuum degradation

$$\dot{Q} = \dot{Q}_\mathrm{T} - \dot{Q}_S,$$

where $\dot{Q}_S$ is the throughput removed by the pump, and $\dot{Q}_\mathrm{T}$ is the total throughput entering the vessel because of sources of vacuum degradation. This equation becomes

$$\frac{\mathrm{d}P}{\mathrm{d}t}V = \dot{Q}_\mathrm{T} - \dot{Q}_S, \tag{7}$$

where V is the volume of the vessel. If the sources of vacuum degradation and the pumping speed are independent of the pressure in the vessel, the evolution of the pressure, i.e., the solution of Eq. (7), is

$$P = P_\mathrm{u} + (P_0 - P_\mathrm{u})\mathrm{e}^{-(S/V)t},$$

where P_0 is the initial pressure at $t = 0$, and $P_\mathrm{u} = \dot{Q}_\mathrm{T}/S$. The time constant of the process $\tau_\mathrm{pd} = V/S$ is the pump-down time. From this equation, we see that for $t \to \infty$ the pressure in the vessel converges to P_u, which is called the ultimate pressure. This pressure is determined by the equilibrium between the throughput of the sources of degradation and the throughput removed by the pump.

The vacuum requirements vary from project to project. Table 6 lists examples of the vacuum requirements in the SNS, LHC, and FAIR [16, 5, 17].

6 Pumps

The control and creation of a vacuum is performed via a system of pumps. Pumps are classified according to their principle of operation.

Table 6: Examples of vacuum pressures in some accelerators

SNS	Front end	1×10^{-4} to 4×10^{-7} Torr
	DTL	2×10^{-7} Torr
	CTL	5×10^{-8} Torr
	SCL	$< 10^{-9}$ Torr
	HEBT	$< 5 \times 10^{-8}$ Torr
	Ring	10^{-8} Torr
	RTBR	10^{-7} Torr
LHC		10^{-10}–10^{-11} mbar
FAIR	HEBT	10^{-9} mbar
	SIS100	10^{-12} mbar

6.1 Positive-displacement pumps

The principle of this type of pump is based on the displacement of a volume $\mathcal{V}$ of gas from the vessel to the outside. This process implies that the action of the pump is to seal a volume $\mathcal{V}$ and then to open it outside the vessel. Clearly, when the volume $\mathcal{V}$ of gas is displaced and opened, in order for the gas to flow out, the pressure in the volume has to be larger than the outside pressure $P_{\rm outlet}$. Therefore it is always necessary that the pump should compress the volume $\mathcal{V}$ so that the initial pressure of the gas, equal to the pressure $P_{\rm inlet}$ in the vessel, becomes larger than $P_{\rm outlet}$. If the gas pressure in the vessel is so low that the compression performed by the pump is not enough to reach a pressure larger than $P_{\rm outlet}$, then gas will be transported from the outlet to the inlet! This reasoning shows that a positive-displacement pump will be able to extract gas only if the ratio $P_{\rm outlet}/P_{\rm inlet}$ can be reached by the compression process.

6.1.1 Piston pump

In this type of pump, a piston moves in a cylinder, creating a volume that varies from $V_{\rm min}$ to $V_{\rm max}$. The ideal volumetric pumping speed is $S_0 = V_{\rm max} N_{\rm c}$, where $N_{\rm c}$ is the number of cycles per second of the piston. When the piston creates the minimum volume $V_{\rm min}$, the pressure inside the chamber is equal to $P_{\rm outlet}$. Next, the piston expands the volume and, for an isothermal process, when the volume is $V_{\rm i} = P_{\rm outlet} V_{\rm min}/P_{\rm inlet}$, the gas will start flowing from the vessel into the piston chamber. The volumetric amount of gas which enters the piston chamber is $V_{\rm max} - V_{\rm i}$, and this is the amount of gas later expelled. The effective pumping speed is therefore $S = N_{\rm c}(V_{\rm max} - V_{\rm i})$, that is,

$$S = S_0 \left(1 - \frac{P_{\rm outlet}}{P_{\rm inlet}} \frac{V_{\rm min}}{V_{\rm max}} \right).$$

Figure 7 shows a plot of the volumetric pumping speed versus the ratio of inlet to outlet pressure.

The above discussion shows that the pump stops pumping at a limiting inlet pressure that depends on the compression performed by the pump. The higher the compression, the lower the limiting pressure. If the thermodynamic process characterizing the compression is of a different nature, the dependence of S on $P_{\rm inlet}/P_{\rm outlet}$ will be different: for example, for an isentropic compression we find $S = S_0[1 - (P_{\rm outlet}/P_{\rm inlet})^{1/\gamma} V_{\rm min}/V_{\rm max}]$. The result is, however, that in any case there exists a limiting pressure at which the pump stops functioning. Note that this feature determines an ultimate pressure in a vessel + pump system independently of the presence of sources of vacuum degradation.

In general, from the point of view of the gas flow, a pump is an object that absorbs a throughput $P_{\rm inlet} S_0$ through the inlet but, owing to other uncontrolled processes, is also subject to a back-flow $\dot{Q}_{\rm b}$. The total throughput is then $\dot{Q} = P_{\rm inlet} S_0 - \dot{Q}_{\rm b}$. In order to characterize the back-flow, it is useful to define the *zero-load compression ratio*. This quantity is obtained as follows: the inlet is closed and the pressure $P_{\rm i0}$ is measured at the entrance of the pump. This is the inlet pressure at zero load. The

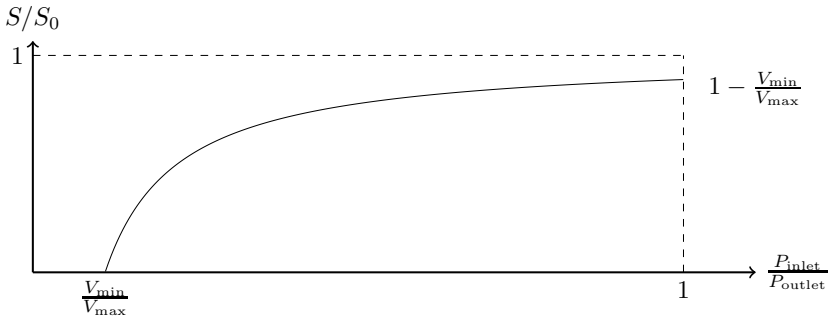


Fig. 7: Pumping speed of a piston pump

zero-load compression ratio is therefore defined as $K_0 = P_0/P_{i0}$, which is a quantity measurable as a function of the outlet pressure P_0 and the pumping speed. As the throughput at the inlet $\dot{Q}$ is zero in this special case, we find that the back-flow throughput $\dot{Q}_b$ is equal to $S_0 P_0/K_0$.

For technical reasons, the compression ratio of a pump cannot be made arbitrarily large. Hence other techniques have been developed to improve the lower limit on pressure.

6.1.2 Rotary pumps

These pumps are characterized by a pumping speed S = 1–1500 m^3/h. The lowest pressure achievable reaches 5×10^{-2} mbar for one-stage pumps, and 10^{-3} mbar for two-stage pumps. An illustration of this type of pump is shown in Fig. 8. These pumps perform the compression via the rotation of

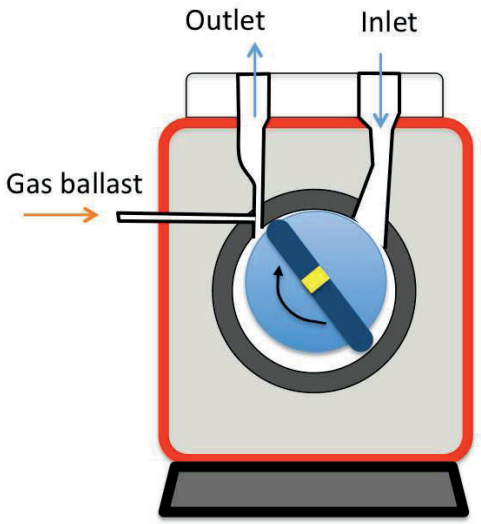


Fig. 8: Illustration of a rotary pump

a rotor, which moves a vane that compresses the gas to a high compression ratio. However, during the compression, there may be a component G of the gas for which the partial pressure P_G becomes too high, resulting in condensation of that component. This problem is avoided by injecting a non-condensable gas during the compression phase. By doing so, the maximum partial pressure P_G can be lowered below the condensation point. This process is called gas ballast [18].

6.1.3 Liquid-ring pumps

Liquid-ring pumps contain a rotating impeller, the axis of which is off centre relative to the casing (see the illustration in Fig. 9). Centrifugal force pushes the liquid against the casing, creating a liquid ring, which seals the impeller, creating vanes that enclose a variable volume. The typical pumping speed is 1–27 000 m^3/h, and gas can be pumped at pressures from 1000 mbar to 33 mbar.

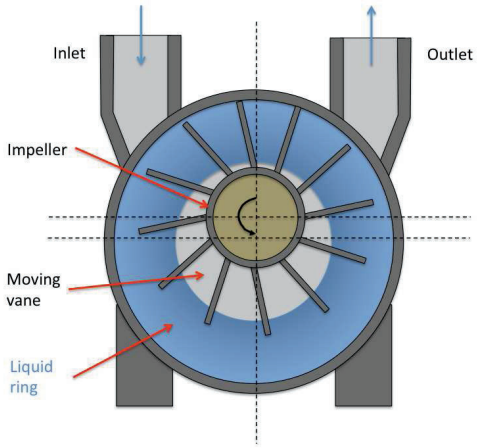


Fig. 9: Illustration of a liquid-ring pump

A problem related to this type of pump arises when the gas expands in the cavities and the pressure becomes lower than the vapour pressure P_s of the liquid (typically water). At that point the liquid boils, but the motion of the impeller then compresses the gas, causing the vapour bubbles to implode, creating shock waves in the liquid medium. This phenomenon is called cavitation, and it causes anomalous functioning of the pump, setting a limit on the pressure at which the pump can work. At $T = 15°C$, the vapour pressure of water is 33 mbar, which is a typical limiting pressure for this type of pump because of the need to avoid cavitation. For a review of liquid-ring pumps, see Ref. [19].

6.1.4 Dry vacuum pumps: the Roots pump

These pumps contain two rotating elements which are separated from the case and from each other by approximately 1 mm. The elements rotate in opposite directions, and as a result of their motion they create a moving vane, which brings a gas volume from the inlet to the outlet. The pumping speed is 75–30 000 m^3/h and the operating pressure is $10–10^{-3}$ mbar. An illustration of this type of pump is shown in Fig. 10. For a general reference, see Ref. [20].

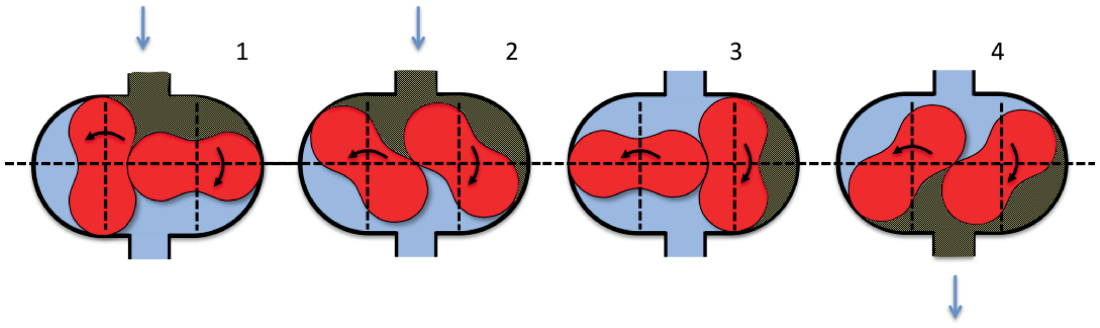


Fig. 10: Illustration of a Roots pump

7 Kinetic vacuum pumps

A different class of pumps, based on a different principle, is that of kinetic vacuum pumps. These pumps give a momentum to gas particles so that they are moved from the inlet to the outlet.

7.1 Molecular-drag pump

This pump is based on the molecular-drag effect. If a gas molecule hits a surface, it will emerge from it in some direction with a probability determined by the cosine law. However, if the surface is in motion with a tangential velocity, the molecule will emerge from reflection from the moving surface with an additional velocity component equal to the speed of the surface. This effect is called the drag effect, and it can be used to create a pump [21]. Consider an long, open channel of transverse cross-section $h \times w$ closed by a surface moving with a velocity U (Fig. 11 (top)). A gas molecule hitting the moving

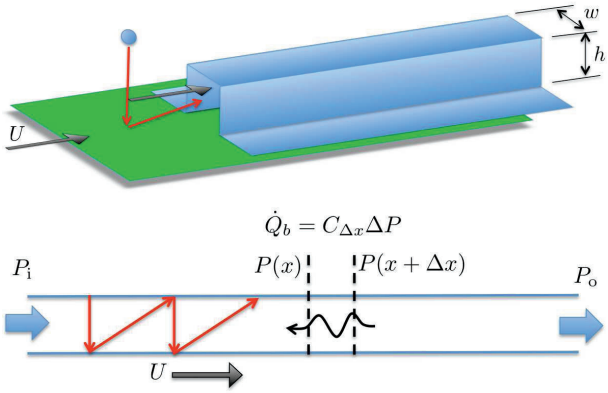


Fig. 11: Principle of the drag pump (top), and schematic illustration of gas flow in the channel of a drag pump (bottom)

surface will acquire a velocity U owing to the drag effect. When the molecule hits any of the other walls of the channel, it will be reflected on average orthogonally to that surface, by the cosine law (see Fig. 11 (bottom)). Therefore a volumetric flow $S_0 = whU/2$ into the channel is established. Taking the back-flow into account, the pumping speed at the inlet is

$$S_i = S_0 \frac{K - K_0}{1 - K_0},$$

where $K_0 = P_{\mathrm{outlet}}/P_{\mathrm{inlet},0}$ is the zero-load compression ratio, and $K = P_{\mathrm{outlet}}/P_{\mathrm{inlet}}$. It can be shown that the compression ratio at zero load takes the form $K_0 = \exp(S_0/C)$, where C is the conductance of the channel. For a long tube, $S_0/C = 3UL/(4hv_\mathrm{a})$, where v_a is the average thermal velocity and L is the length of the channel. For example, for L = 250 mm and h = 3 mm, we have $S_0/C > 10$ and $K_0 \gg 1$, so that we retrieve the form

$$S = S_0 \left(1 - \frac{K}{K_0}\right).$$

An illustration of the functioning of a molecular-drag pump is shown in Fig. 12. The typical pumping speed of a molecular-drag pump is 7–300 l/s, at an operating pressure of 10^{-3}–10^{3} Pa. The ultimate pressure reachable is 10^{-5}–10^{-3} Pa. For references, see Refs. [9, 22].

7.2 Turbomolecular pump

The turbomolecular pump is based on the rotation of a set of blades at a velocity U. The blades are tilted at an angle ϕ. The gas molecules enter the inlet, and the motion of the blades imparts a momentum to

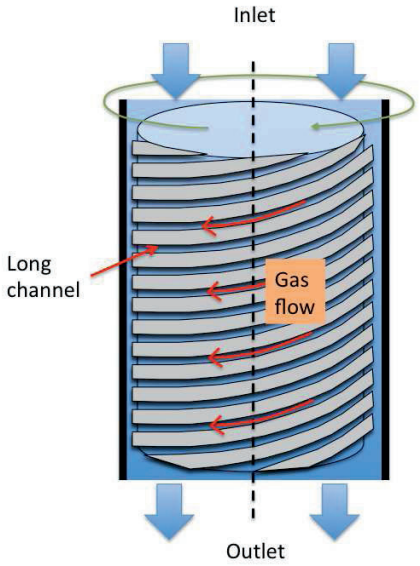


Fig. 12: Illustration of the functioning of a molecular-drag pump

the gas. The situation is illustrated in Fig. 13. When particles leave the set of rotating blades, the gas has acquired a rotational velocity, which makes it difficult to use a subsequent series of rotating blades (as they would move at a similar velocity to the gas). For this reason, a stator consisting of blades at rest is placed after the rotating blades so as to remove the rotational component of the particle velocity. By using this strategy, several rotor + stator blocks can be arranged in a multistage pumping structure.

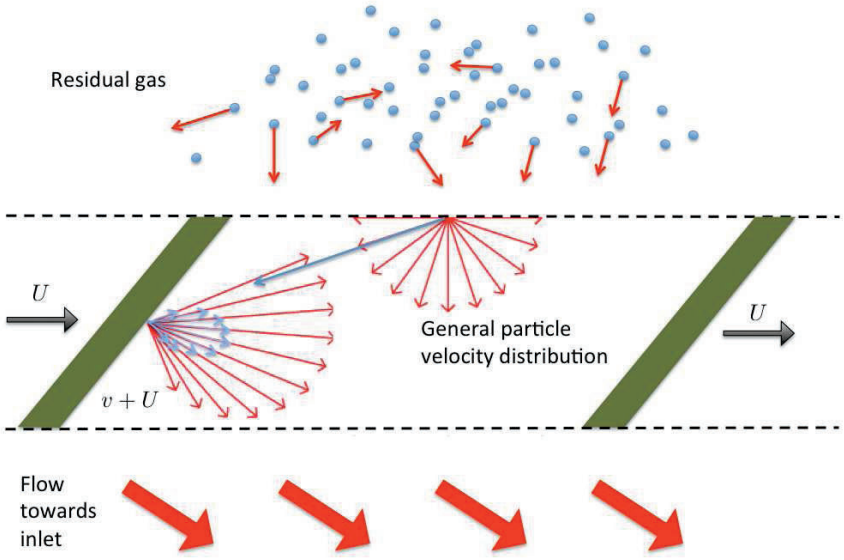


Fig. 13: Principle of the turbomolecular pump

The probability that molecules entering the pump will be pushed out is $W = \dot{N}/(J_{\mathrm{i}}A)$, where J_{i} is the rate of impingement of molecules on the inlet surface. The maximum probability $W_{\max}$ is found when $P_{\mathrm{outlet}} = P_{\mathrm{inlet}}$. Analogously to the case of the molecular-drag pump, the pumping probability W is given by

$$W = W_{\max}\frac{K_0 - K}{K_0 - 1},$$

where K_0 is the compression at zero load. In Ref. [23, 9], it is shown that $K_0 \propto g(\phi)\exp(U/v_{\mathrm{a}})$, and

therefore $K_0 \propto \exp(\sqrt{M})$, where M is the molar mass of the gas. This means that different gas species have different pumping probabilities. In addition, the maximum probability $W_{\max}$ is proportional to $U/v_{\mathrm{a}} \propto \sqrt{M}$. Therefore the maximum pumping speed $S_{\max} = W_{\max} J$ is independent of the molecular mass of the gas, and

$$\frac{S}{S_{\max}} = \frac{K_0 - K}{K_0 - 1}.$$

See Ref. [22, 2] for more details. Figure 14 (left) shows an example of the maximum compression as a function of the foreline pressure for several gas species [2]. The typical pumping speed of these types

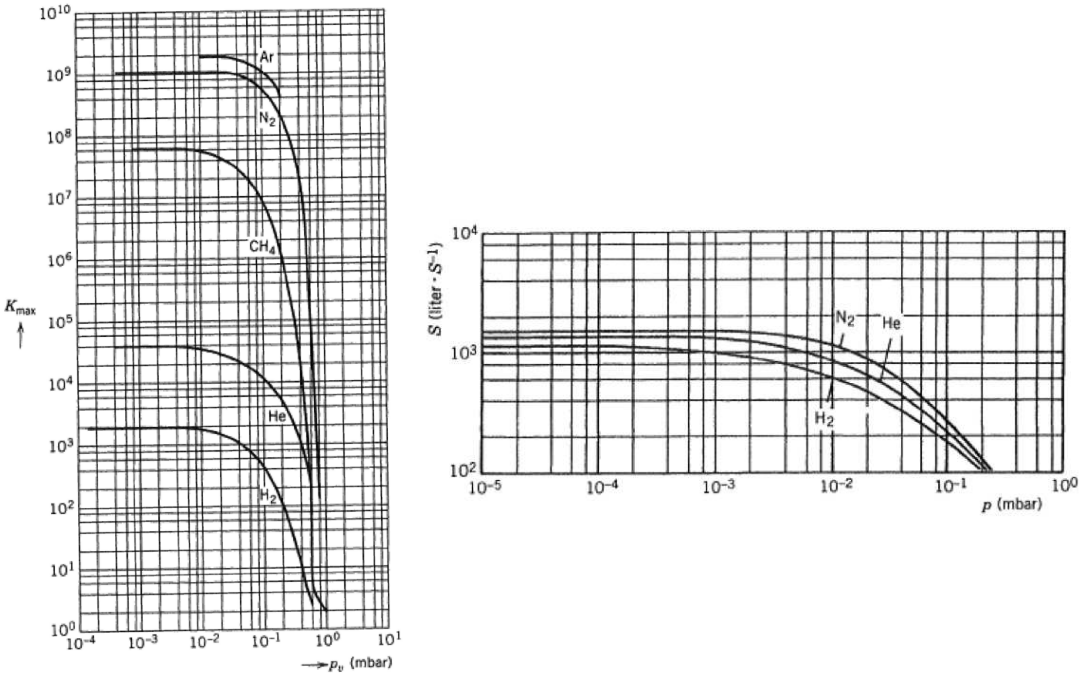


Fig. 14: Maximum compression rate as a function of the foreline pressure (left), and pumping speed (right) (Balzers Pfeiffer)

of pump is 35–25 000 l/s (Fig. 14 (right)), and the ultimate pressure reachable is 10^{-8}–10^{-7} Pa.

Acknowledgements

The author thanks Maria Cristina Bellachioma for comments on this paper and corrections.

References

[1] E.H. Kennard, *Kinetic Theory of Gases, with an Introduction to Statistical Mechanics* (McGraw-Hill, New York, 1938).

[2] J.M. Lafferty, *Foundations of Vacuum Science and Technology* (Wiley, New York, 1988).

[3] R.J. Reid, Vacuum science and technology in accelerators, Cockcroft Institute Lectures, 2010. http://www.cockcroft.ac.uk/education/academic0910.html.

[4] R.C. Weast, Ed., *Handbook of Chemistry and Physics* (CRC Press, Boca Raton, FL, 1975).

[5] LHC, LHC Design Report, Technical report, http://lhc.web.cern.ch/lhc/LHC-DesignReport.html.

[6] G. Rumolo, F. Ruggiero, and F. Zimmermann, *Phys. Rev. ST Accel. Beams* **4**(1) (2001) 012801.

[7] F. Zimmermann, *Phys. Rev. ST Accel. Beams* **7** (2004) 124801.

[8] M.A. Furman and M.T.F. Pivi, *Phys. Rev. ST Accel. Beams* **5** (2002) 124404.

[9] A. Chambers, *Modern Vacuum Physics* (CRC Press, Boca Raton, FL, 2004).

[10] M. Knudsen, *Ann. Phys.* **28** (1909) 75.

[11] R.B. Bird, W.E. Stewart, and E.N. Lightfoot, *Transport Phenomena* (Wiley, New York, 2007).

[12] S.E. Haaland, *J. Fluids Eng.* **105** (1983) 89.

[13] C.H. Collie, *Kinetic Theory and Entropy* (Longman, London, 1982).

[14] P.A. Redhead, *J. Vac. Sci. Technol. A* **13** (1995) 2791.

[15] K. Zapfe, Leak detection, CERN Accelerator School, CERN-2007-003 (2007), p. 227.

[16] J.Y. Tang, SNS vacuum instrumentation and control system, Proc. 8th International Conf. on Accelerator and Large Experimental Physics Control Systems, San Jose, CA, 2001, p. 188.

[17] A. Kraemer, The vacuum system of FAIR accelerator facility, Proc. EPAC 2006, Edinburgh, 2006, p. 1429.

[18] W. Gaede, *Z. Naturforsch.* **2a** (1947) 233–238.

[19] H. Bannwarth, *Liquid Ring Vacuum Pumps, Compressors and Systems* (Wiley-VCH, Weinheim, 2005).

[20] A.P. Troup and N.T.M. Dennis, *J. Vac. Sci. Technol. A* **9** (1991) 2048.

[21] W. Gaede, *Ann. Phys.* **346**(7) (1913) 337–380.

[22] J.F. O'Hanlon, *A User's Guide to Vacuum Technology* (Wiley, Hoboken, NJ, 2003).

[23] C.H. Kruger and A.H. Shapiro, in *Rarefied Gas Dynamics*, Ed. L. Talbot (Academic Press, New York, 1960), p. 117.

Vacuum II

Giuliano Franchetti
GSI Darmstadt, D-64291 Darmstadt, Germany

Abstract
This paper continues the presentation of pumps begun in 'Vacuum I'. The main topic here is gauges and partial-pressure measurements. Starting from the kinetics of gases, the various strategies for measuring vacuum pressures are presented at an introductory level, with some reference to hardware devices. Partial-pressure measurement techniques are introduced, showing that the principles of ion selection have a direct similarity to particle dynamics in accelerators.

1 Pumps

1.1 Kinetic vacuum pumps

1.1.1 Diffusion ejector pump

The principle of the diffusion ejector pump is simple. A heavy molecular liquid is heated to its boiling point and the vapour produced is ejected at high speed from a nozzle, as shown in Fig. 1. The atoms of the liquid, typically an oil, collide with the residual-gas atoms arriving from the inlet, which then receive a downward momentum by collision so that the gas particles flow towards the outlet [1]. The

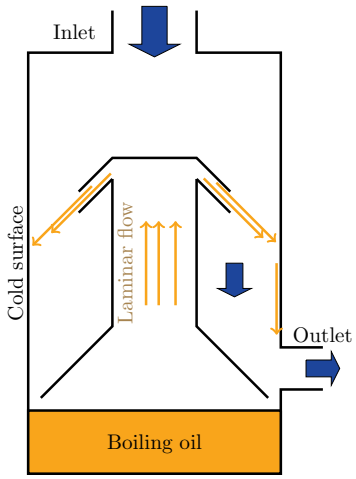


Fig. 1: Schematic illustration of a diffusion ejector pump.

oil jet becomes a skirt, which separates the inlet region from the outlet region of the pump. This layer determines the operating pressure, which for this type of pump is typically 10^{-3}–10^{-8} mbar. This pump is called a diffusion pump because the residual-gas molecules diffuse into the oil jet. Because of the slope of the jet, the diffusion is not symmetric between the outlet and inlet flows. Although this pump was historically one of the first vacuum pumps, the theory describing the way in which it functions is rather complex [1].

This pump is affected by the following problems:

– *back-streaming:* this is due to the fact that the oil jet is not ideal, and therefore some of the oil molecules, owing to thermal spread, can return back towards the inlet;

– *back-migration:* the oil that condenses on the walls of the pump can 'walk back' on the walls rather than running down into the oil reservoir.

These problems are cured by cooling the surfaces of the pump and by inserting a cold cup on top of the nozzle. Back-streaming is controlled by installing a baffle, which allows the oil to be condensed, although it also reduces the conductance of the inlet, affecting the overall pumping speed of the pump [2, 1].

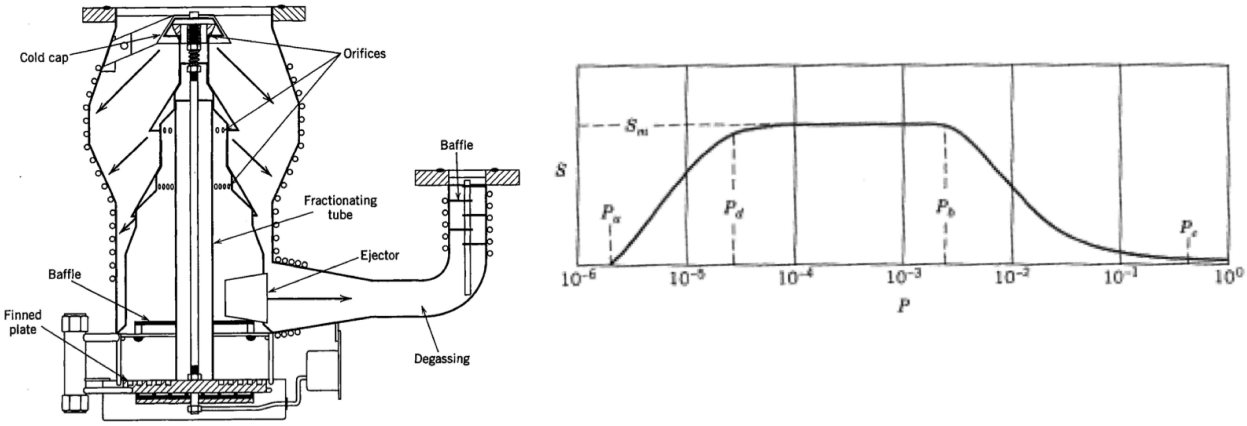


Fig. 2: Left: example of a three-stage diffusion ejector pump (Varian Vacuum Products). Right: pumping speed as a function of the inlet pressure.

Figure 2 (left) shows an example of a three-stage diffusion pump. Figure 2 (right) shows the pumping speed of a diffusion pump with an inlet diameter $D = 0.1$ m, for which a maximum pumping speed $S_{\rm m} \sim 250$ l/s is expected for N_2 [1]. The ultimate inlet pressure $p_{\rm u}$ here is of order 10^{-6} mbar.

2 Capture vacuum pumps

Capture pumps are based on the capture of vacuum molecules by surfaces. This is done by using special materials, called 'getters', which form stable chemical compounds with the gas molecules. There are two main ways to produce a getter surface: (1) depositing an *evaporable getter* film *in situ*, and (2) heating a surface coated with a *non-evaporable getter* (NEG) so that the layer of oxide on the the getter diffuses into the bulk.

2.1 Getter pumps (non-evaporable getters)

Getters are materials which have the capability to adsorb gas molecules by a chemical process [3].

A getter is deposited by coating a surface using a sputtering process. The principle of the use of magnetron sputtering in a pipe is illustrated in Fig. 3 (left). The NEG material is contained in a rod on the axis of the pipe to be coated. A noble gas is injected into the pipe, and a potential difference is established between the pipe and the rod. A longitudinal magnetic field is applied to this system. The combination of a radial electric field and a longitudinal magnetic field forms a trap for electrons, which are created by other processes [1]. The trapped electrons eventually collide with atoms of the noble gas during their motion, and ionize them. In each ionization event, a new electron is formed, which increases the number of trapped electrons (i.e., a discharge occurs), making further ionization of the noble gas easy. The positive ions accelerate in the presence of the radial electric field and hit the rod, causing sputtering. The energy of the collisions depends on the location in which the atoms of the gas are ionized. However,

the energy deposited on the rod is enough to break the bonds of the NEG material, which is then emitted towards the walls of the pipe. Clearly, only neutral atoms of the NEG are able to travel to the walls and stick to the surface. This process, which happens to a large number of NEG atoms, creates a thin film on the inside surface of the pipe. Coating with complex materials such as Ti–Zr–V alloys can be

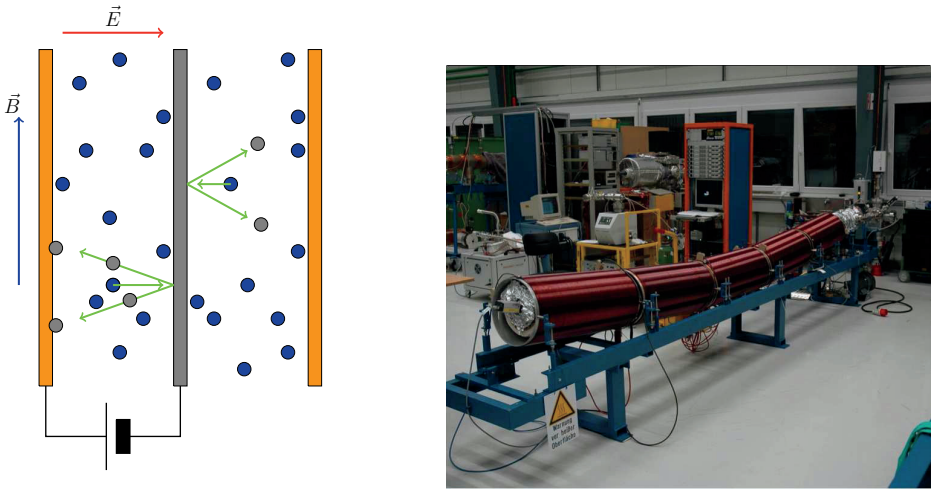


Fig. 3: Left: principle of sputtering. Right: coating of a bent vacuum chamber at GSI [4].

done by using more than one rod; for Ti–Zr–V alloys, three rods, consisting of Ti, Zr, and V, are used. Figure 3 (right) shows the coating technique used at GSI for bent pipes. Solenoids are used to create a bent magnetic field.

After the NEG has been deposited, exposure to air causes the formation of a layer of oxide, which will prevent the NEG from pumping. The oxide layer must be removed to activate the NEG. This process of activation is performed by heating the surface. Increasing the temperature increases the rate of diffusion of the atoms of the oxide layer into the getter, leaving the surface free to capture residual-gas atoms. This diffusion process depends on the composition of the getter.

The pumping process depends on the rate of impingement of the gas on the surface, and the probability that a gas molecule hitting the surface will stick to it. Of course, after a gas atom from the vacuum has been adsorbed by the surface, less of the surface will remain free for the pumping process. This results in a reduction of the pumping speed of the surface. Therefore the pumping speed of a surface is related to the sorption capacity. Figure 4 shows an example of a plot of the sorption capacity versus the pumping speed [5].

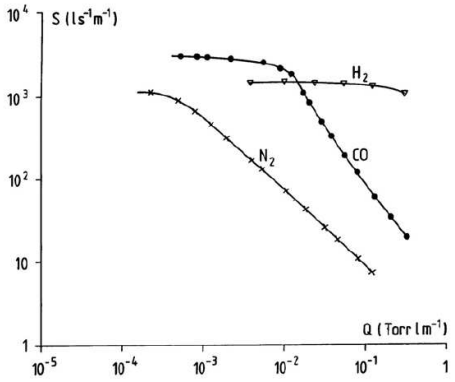


Fig. 4: Relation between pumping speed and sorption capacity.

2.2 Sputter ion pumps (evaporable getters)

Sputter ion pumps (SIPs) are devices that create a sputtering process of a getter material [6]. An SIP contains a cylinder, which forms an anode, and two plates at opposite ends, which form a cathode (Fig. 5 (left), from [1]). Owing to this configuration, the electric field in the pump is complex. In addition, a longitudinal magnetic field is applied. The resulting electric and magnetic fields, for suitable values of E and B, create a trap for electrons, which are spontaneously present in the vacuum. Between the anodes and the cathode, residual-gas atoms can circulate, and the electrons will eventually ionize some gas atoms during their motion. The electrons newly produced by this ionization remain trapped in the $\vec{E}, \vec{B}$ field, whereas the ionized gas atoms are too slow to experience a significant force from the magnetic field. As a result, the ionized residual gas is accelerated by the $\vec{E}$ field and eventually hits the cathodes, which are typically covered with titanium. Therefore a sputtering process occurs, and the titanium is sputtered around the pump (see Fig. 5 (left), taken from [1]). The more the pump captures vacuum gas atoms, the more sputtering takes place, generating more surfaces coated with fresh titanium. An example of the pumping speed of an SIP is shown in Fig. 5 (right).

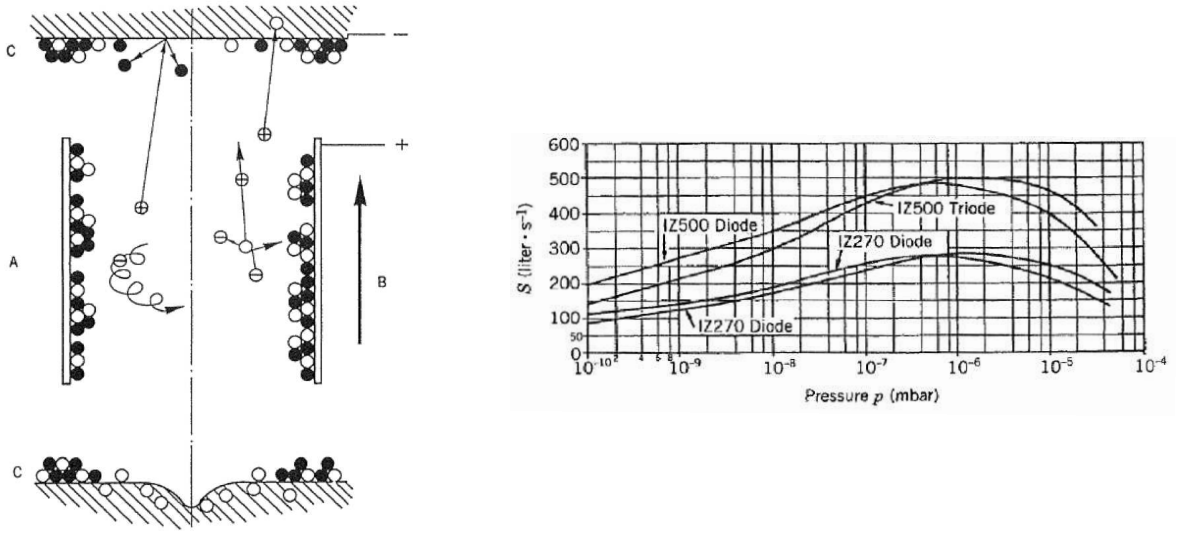


Fig. 5: Left: schematic illustration of a sputter ion pump. Right: pumping speed versus pressure at the inlet, from Ref. [1], p. 342.

2.3 Cryopumps

Cryopumps make use of the dispersion forces between molecules and the internal surfaces of the pump. These forces are stronger than the forces between the molecules. When a gas molecule hits a surface, it may stick because of this mechanism.

Figure 6 shows a schematic illustration of a cryopump. The cold surfaces form a container attached to the vessel. The residual gas in the vessel is at a temperature $T_{\rm w}$ and pressure $P_{\rm w}$, and its particle density is $\tilde{n}_{\rm w}$, where the subscript 'w' stands for 'warm'. Similarly, the particles in the cold region are characterized by a temperature $T_{\rm c}$ and pressure $P_{\rm c}$, and the particle density is $\tilde{n}_{\rm c}$. The pump works at low pressure. In the aperture between the pump and the vessel there are two fluxes of particles, the flux $I_{\rm w}$ of particles moving from the warm to the cold section and the flux $I_{\rm c}$ from the cold to the warm section. Let A be the area of the inlet, so that the flux of particles $I_{\rm w}$ is given by

$$I_{\rm w} = \frac{P_{\rm w} A}{4 k_{\rm B} T_{\rm w}} v_{\rm w},$$

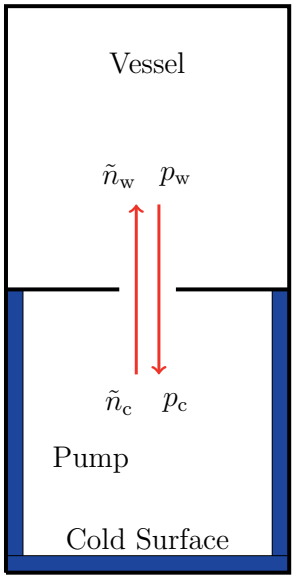


Fig. 6: Schematic illustration of a cryopump.

and the flux I_c is given by

$$I_c = \frac{P_c A}{4k_B T_c} v_c,$$

where v_w and v_c are the average thermal velocities of the molecules in the warm and cold regions, respectively. Note that if $I_w = I_c$, the net flow of particles through A is zero, even though $P_w \neq P_c$ (in the continuum regime, the same condition would require that the two pressures were equal). The condition $I_w = I_c$ is reached if

$$\frac{P_c}{\sqrt{T_c}} = \frac{P_w}{\sqrt{T_w}}.$$

This effect is called thermal transpiration. When the condition for thermal transpiration is not fulfilled, we find a net flow through A,

$$I_{net} = \frac{A P_w v_w}{4k_B T_w} \left(1 - \frac{P_c}{P_w} \frac{v_c T_w}{v_w T_c}\right).$$

Now the maximum flux is

$$I_{max} = \frac{A P_w v_w}{4k_B T_w},$$

and therefore

$$I_{net} = I_{max} \left(1 - \frac{P_c}{P_w} \frac{v_c T_w}{v_w T_c}\right).$$

Defining

$$P_w(\text{ultim}) = P_c \sqrt{\frac{T_w}{T_c}},$$

we obtain

$$I_{net} = I_{max} \left(1 - \frac{P_w(\text{ultim})}{P_w}\right).$$

The issue is to establish what $P_w(\text{ultim})$ is; that is, what P_c is when the temperature inside the pump is T_c. This is relatively easy in the case of cryocondensation, where the pressure inside the cold region is given by the vapour pressure of the substance. Figure 7 [2] shows the dependence of the vapour pressure on temperature for typical residual gases. For example, at T = 4 K, the vapour pressure of H_2 is

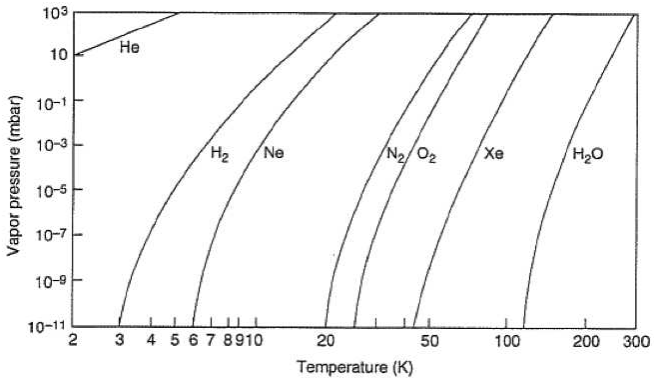


Fig. 7: Vapour pressures of gases at cryogenic temperatures.

$P_{\mathrm{H_2}} = 10^{-7}$ mbar, which gives $P_{\mathrm{w}}(\mathrm{ultim}) = 8.5 \times 10^{-7}$ mbar.

We conclude by showing in Fig. 8 a summary of the various types of pumps and their ranges of application.

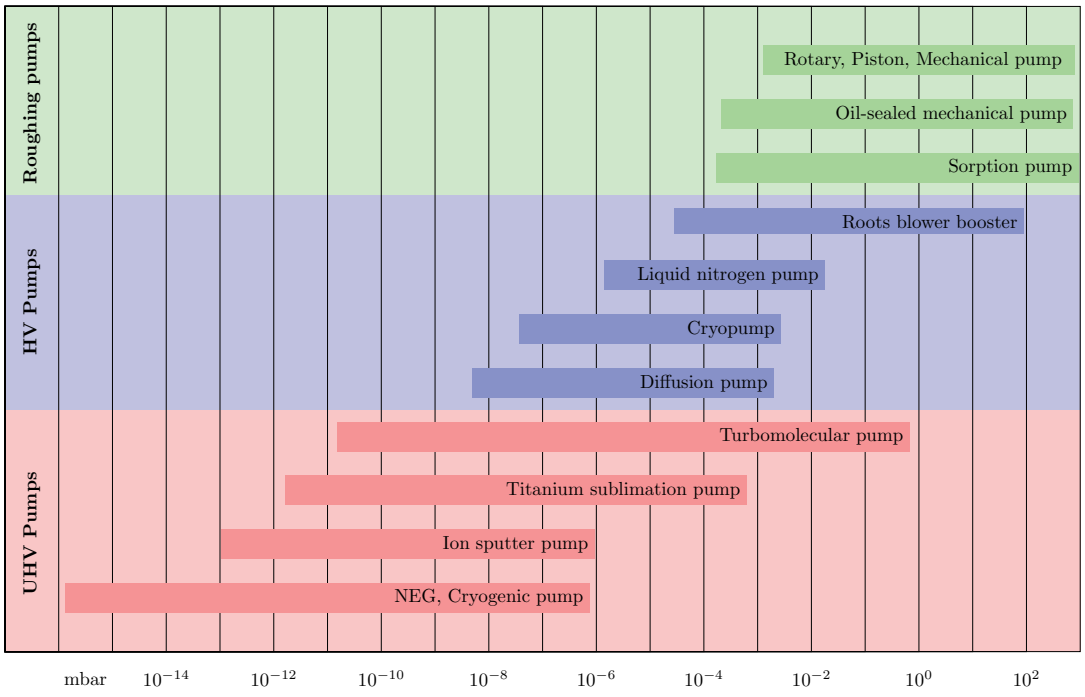


Fig. 8: Summary of the working ranges of the main types of pumps (adapted from [7]).

3 Gauges

Gauges are devices for measuringe the total pressure of the residual gas in a vacuum. Each type of gauge is best adapted to a certain pressure range, which depends on its principle of operation. We now review the main types of gauges.

3.1 Liquid manometers

Liquid manometers are based on the displacement of a liquid in a U-tube when the pressures in the two branches differ. The difference in pressure is given by $P_2 - P_1 = h\rho g$, where h is the relative

displacement of the liquid in the manometer, ρ is the density of the liquid, and g is the gravitational acceleration. Normally the liquid is mercury, because of its high density. The accuracy of this method relies on a knowledge of ρ and g. In addition, the surface tension of the mercury depresses the surface of the liquid, making it difficult to assess the mercury level with a precision better than ± 0.1 mm. Another issue concerning the use of mercury is the serious health hazard that it presents if not handled with care.

3.2 McLeod gauges

An improvement is obtained with the McLeod gauge [8]. This gauge consists of a U-tube, which has one closed arm (the first arm), while the other arm is connected to the vessel (Fig. 9). The bottom of the U-tube is connected to a mercury reservoir. At the beginning of the measurement, the reservoir is lowered so that the mercury level is below the lowest point of the U-tube. Then the reservoir is slowly raised. At a certain point, the mercury seals the first arm from the second. Therefore the number of gas molecules in the first arm is fixed, and a further raising of the reservoir causes the level of the mercury (measured from the closed end) to rise from h_0 to Δh. This creates a compression of the gas in the first arm according to $P_2 h_0 A = P_2' \, \Delta h \, A$ (the compression is isothermal). The measurement is performed by

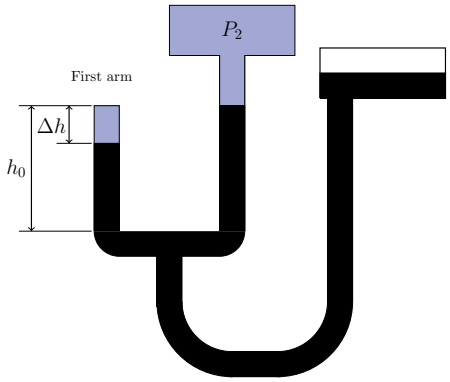


Fig. 9: Schematic illustration of the McLeod gauge.

raising the reservoir so that the level of mercury in the second arm is brought to the same height as the end of the first arm (see Fig. 9). At this point the pressure in the first arm on the surface of the mercury is P_2', while the pressure in the other arm at the same height is $\Delta h \, \rho g + P_2$. By equating the two pressures, we find the pressure P_2 as

$$P_2 = \rho g A \frac{\Delta h^2}{V}.$$

Here $V = h_0 A$ is the initial volume of gas in the first arm when the mercury seals it, and A is the cross-sectional area of the the first arm. It is also assumed that $V \gg \Delta h \, A$. Therefore, in this gauge, the change in pressure is quadratic in Δh.

In a second mode of use of this gauge, the height of the reservoir is changed so as to keep the distance between the surface of the mercury in the first arm and the closed end of that arm constant at $\Delta h = d$. Then it is measured the difference of height h between the mercury level in the first arm and the mercury level in the second arm (which is allow to vary). In this case the response is

$$P_2 = \frac{A h \rho g d}{V}.$$

Note that large volumes facilitate the measurement of small pressures in the first arm (the product $P_2 V$ becomes large), and therefore a bulb is constructed in the first arm to increase V (see Fig. 6.3 on p. 380 of Ref. [1]).

3.3 Viscosity gauges

Viscosity gauges are based on the effect of viscosity. This type of gauge uses a rotating sphere suspended magnetically [9, 10]. Gas molecules hitting the surface of the sphere take away rotational momentum (Fig. 10 (left)). After the molecules have hit the internal walls of the gauge, thermalization removes the angular momentum of the gas. The overall effect of the gas is to reduce the angular velocity according

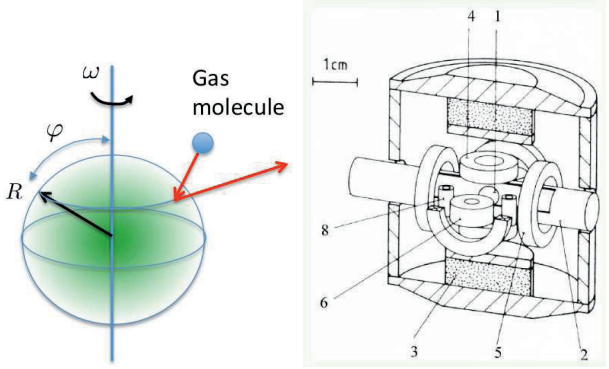


Fig. 10: Left: schematic illustration of the spinning-rotor gauge. Right: technical details of the construction: 1, rotor; 2, vacuum tube; 3, permanent magnets; 4, two coils for vertical stabilization; 5, four drive coils; 6, two detection coils; 8, four coils for horizontal stabilization. From Ref. [11].

to the rate of collision of particles with the surface, which is proportional to the gas pressure P. It is possible to prove that the pressure P is given by

$$P = -\frac{\pi}{10}\frac{\rho R v_{\mathrm{a}}}{\delta}\left(\frac{1}{\omega}\frac{\mathrm{d}\omega}{\mathrm{d}t} + 2\alpha\frac{\mathrm{d}T}{\mathrm{d}t}\right), \tag{1}$$

where ρ is the density of the sphere, R is the radius of the sphere, v_{a} is the thermal velocity of the residual-gas atoms, ω is the angular velocity of the rotating sphere, α is the coefficient of thermal expansion of the sphere, T is the temperature, and δ is an accommodation factor (normally close to unity). The second term in the brackets is not related directly to the pressure of the residual gas, but is related to the change in temperature of the sphere. The dependence on the temperature in Eq. (1) stems from thermal expansion: if the sphere expands, then the conservation of angular momentum will reduce the angular velocity. A drawing of the technical construction of this gauge is shown in Fig. 10 (right).

In practice, this gauge is operated by measuring the angular velocity over an interval of time. The smaller this interval is, the larger is the error in the measurement, as shown in Fig. 11 (from Ref. [12]).

3.4 Thermal-conductivity gauges

These gauges are based on the exchange of energy by thermal conduction. For a gas in the molecular regime, the particles of the gas collide mainly with the walls rather than with each other. When a particle belonging to a gas at temperature T_{g} hits a wall that has a temperature T_{w}, the particle may or may not acquire the same temperature as the wall when it leaves the wall; this depends on how well the particle is 'accommodated' to the wall. The situation can be summarized by introducing an accommodation factor

$$\alpha = \frac{T_{\mathrm{r}} - T_{\mathrm{g}}}{T_{\mathrm{w}} - T_{\mathrm{g}}},$$

where T_{r} is the temperature of the particle after hitting the wall (i.e., after reflection). If $\alpha = 1$, then $T_{\mathrm{r}} = T_{\mathrm{w}}$, whereas if $\alpha = 0$, then $T_{\mathrm{r}} = T_{\mathrm{g}}$.

Thermal-conductivity gauges typically contain a wire surrounded by a cylinder filled with the residual gas of the vessel. The molecules of the gas hit the wire and the cylinder repeatedly, transporting

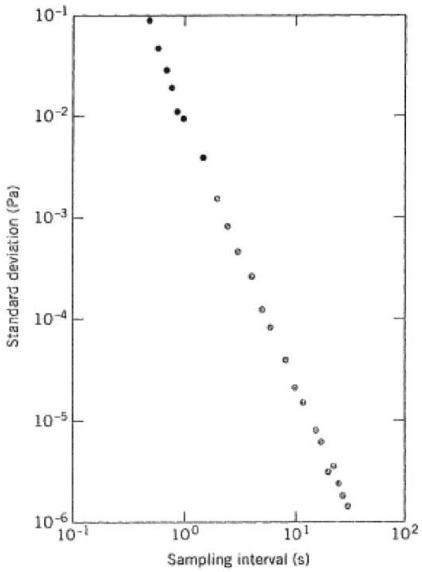


Fig. 11: Measurement error of a spinning-rotor gauge: variation of the random noise with integration time [12].

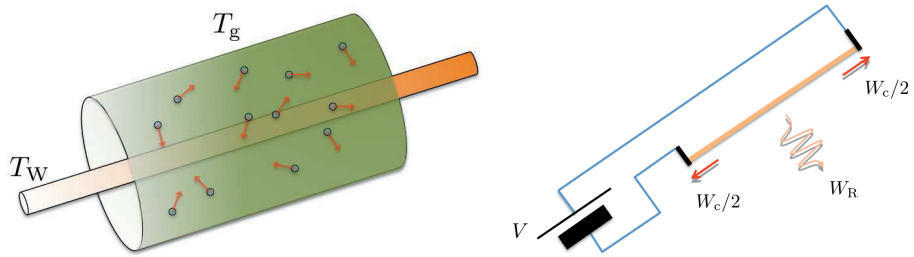


Fig. 12: Left: schematic illustration of a thermal-conductivlty gauge. Right: sources of energy loss.

thermal energy from the hot wire to the colder cylinder. The situation is summarized in Fig. 12 (left). An analysis of the kinetics of the gas and of the energy transport allows the following flux of energy to be derived [13]:

$$E_{\mathrm{G}} = \frac{1}{4}\frac{\gamma+1}{\gamma-1}\sqrt{\frac{2k_{\mathrm{B}}}{\pi m T_{\mathrm{g}}}}\alpha(T_{\mathrm{w}} - T_{\mathrm{g}})P. \tag{2}$$

Here, $\gamma = c_p/c_v$ is the ratio of specific heat capacities at constant pressure and constant volume. By measuring the power loss $W_{\mathrm{G}} = E_{\mathrm{G}}S_{\mathrm{w}}$, where S_{w} is the surface area of the wire, the pressure of the residual gas can be found. The correct application of Eq. (2), however, requires careful account to be taken of all sources of energy loss, which are the following (see Fig. 12 (right)).

- *Energy loss because of the gas molecules*. This contribution is calculated from Eq. (2).
- *Energy loss by radiation*. The hot wire radiates a power equal to $W_{\mathrm{R}} = \epsilon\sigma(T_{\mathrm{w}}^4 - T_{\mathrm{g}}^4)S_{\mathrm{w}}$, where $\sigma = 5.673 \times 10^{-8}$ W·m^{-2}·K^{-4} is the Stefan–Boltzmann constant and ϵ is the emissivity.
- *Energy loss by heat conduction along the wire*. The wire is hot because of the passage of current through it; therefore locally, in each piece of the wire, there is heat generation simultaneously with heat transport via conduction in accordance with Fourier's law, and hence $\mathrm{d}W_{\mathrm{c}}/\mathrm{d}l = -GA\,\mathrm{d}T/\mathrm{d}l$, where l is the length of the wire, G is the coefficient of thermal conductivity of the wire, and A is its cross-section.

3.4.1 Pirani gauges

The use of a thermal gauge requires a knowledge of the temperature of the wire; the temperature of the cylinder is easily controlled. Typically, the gauge is used in a Wheatstone bridge as illustrated in Fig. 13 [14]. The compensating tube is kept at a constant temperature and the voltage V is varied to balance the bridge. For $R_1 = R_3$, if the temperatures of the two wires are not equal, their resistivities also differ, leading to $R_2 \neq R_4$, and hence M measures a potential difference. When the voltage V is changed, the temperature of the wire in the gauge is changed, and balancing the bridge ensures that $R_2 = R_4 = R$. At that point the temperature of the wire in the gauge is known (it is equal to that of the compensating wire kept at a constant temperature), and the power dissipated is measured as $W = V^2/(4R)$. This gives $W = W_{\rm G} + W_{\rm R} + W_{\rm c}$, from which $W_{\rm G}$ is found if the other contributions $W_{\rm R}, W_{\rm c}$ are negligible.

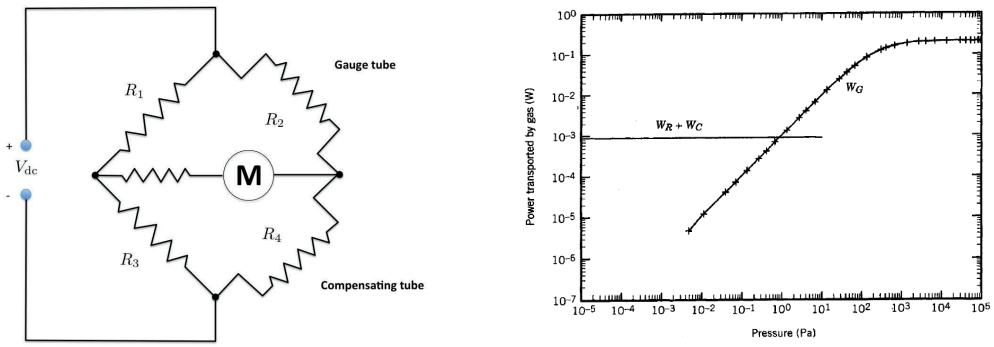


Fig. 13: Left: schematic illustration of a Pirani gauge. Right: example of power transported by the gas vs. vacuum pressure.

Figure 13 (left) (p. 408 of Ref. [1]) shows an example of the power dissipated in a Pirani gauge. A linear dependence is preserved until a continuum regime is reached, where the thermal conductivity $\lambda = \eta c_v$ becomes independent of the pressure in the cylinder, and hence the power loss becomes independent of the pressure. Here, η is the viscosity and c_v is the heat capacity per unit mass at constant volume.

3.5 Principle of ionization gauges

Whereas in the previous type of gauge the pressure is measured by exploiting particle–wall interactions, ionization gauges measure collisions between particles; the collision rate is proportional to the particle density of the residual gas, and hence to the pressure. As the mean free path is larger than the size of the vessel, these collisions are between the particles of the residual gas and other particles (not belonging to the residual gas). The measurement of the collisions is performed via an ionization process, in which a beam of electrons is sent through a region filled with residual gas. These electrons are accelerated so as to have enough energy to ionize the residual-gas atoms. The ionized atoms and electrons are then subjected to a transverse electric field, which accelerates the ions and electrons in opposite directions, creating a current i_+ proportional to the ionization rate in the region containing the residual gas. The principle is shown in Fig. 14 (top).

The method becomes useful when the current of electrons i_- can be correlated with the ionization current i_+ via the relation

$$\frac{i_+}{i_-} = KP, \tag{3}$$

where $K = (\sigma_{\rm i} L)/(k_{\rm B} T)$ is called the sensitivity. Here, $\sigma_{\rm i}$ is the cross-section of the electrons for ionization of the residual gas, and L is the path length travelled by the electrons. Note that this length cannot be too large, because Eq. (3) is valid only in the molecular regime. Note also that the cross-section $\sigma_{\rm i}$ varies according to the type of atoms and the energy of the projectile electrons. The maximum production rate is found in the range of 100–200 eV (see Fig. 14 (bottom), taken from Ref. [15]).

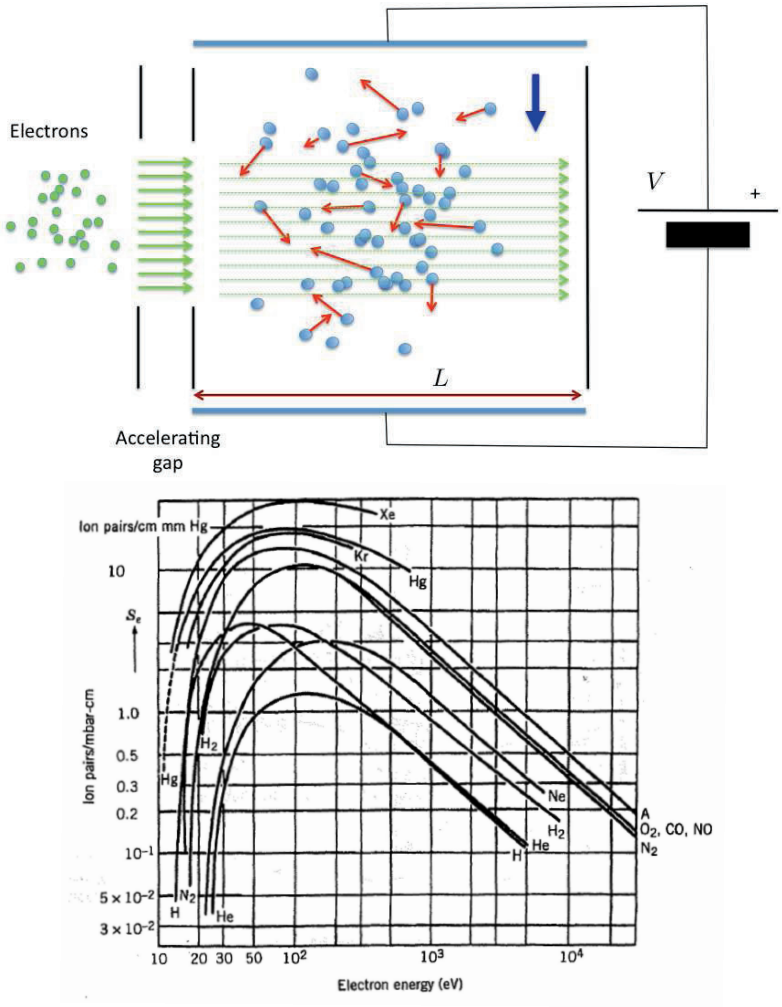


Fig. 14: Top: principle of ionization gauge. Bottom: rate of production of ion pairs by ionization with electrons.

3.6 Hot-cathode gauges

Hot-cathode gauges are based on the production of projectile electrons via thermionic emission from a hot filament. The electrons emitted are accelerated by a potential difference created by the filament and a grid. The electrons are accelerated radially, reaching their maximum speed at the grid; hence the filament–grid potential difference must be large enough to create an ionization process in the residual gas. During their travel, many electrons pass through the grid and then experience an outwards confining electric field created by an anode outside the grid, which bounces them back towards the grid. In this way, the electrons oscillate around the grid, having their maximum energy at the position of the grid, creating an area where residual-gas atoms can be ionized (see Fig. 15). The residual-gas atoms can be ionized either in the external region outside the grid or in the internal region inside it. Both regions produce ions and electrons, but only the ionization that takes place between the anode and the grid produces easily detectable information about the collision rate. In fact, the ionization process in the internal region adds electrons to those already present produced by the hot cathode. But, as can be seen from Eq. (3), the sensitivity is small, in order to keep the gauge in a molecular regime, and hence the current produced by ionization processes, i_+, is much smaller than that created by the hot cathode, i_-.

The pressure range in which this gauge can be used has (1) an upper limit, determined by the deviation from the linear response, and (2) a lower limit, determined by the effect of soft X-rays [16].

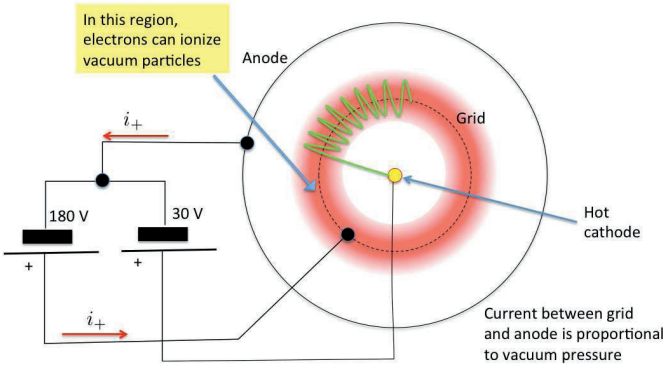


Fig. 15: Schematic illustration of hot-cathode gauge

The electrons created by the cathode eventually hit the grid during their oscillatory motion around it. In this process, the energy of the electrons may be converted to X-rays, which are emitted from the grid in all directions. The X-rays emitted outwards, when they hit the anode, can create new electrons, which are added to those already present as a result of the ionization of the residual gas (see Fig. 16). As a consequence, the effective current measured by the device is $i_+ + i_\mathrm{r} = KPi_- + i_+$, where i_r is the residual current generated by the soft X-rays. This effect limits the use of the gauge according to the strength of i_r, and sets a lower limit typically equal to 10^{-7} mbar.

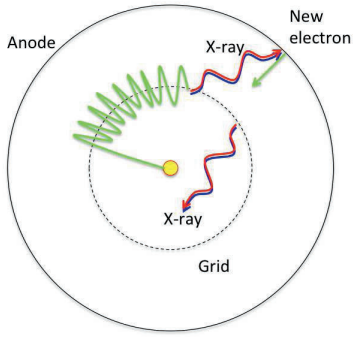


Fig. 16: Visualization of the mechanism of the X-ray limit

3.7 Bayard–Alpert gauges

Bayard and Alpert introduced a substantial improvement [17] to the ionization gauge. Their strategy was to reduce the anode to an object with a very small surface area, i.e., a wire, and put it in the place previously occupied by the cathode. The cathode is moved outside the grid, as shown in Fig. 17 (left). The cathode emits electrons, which again oscillate around the grid. Now, however, the X-rays emitted in all directions have a very low probability of hitting the anode, which is a thin wire, and therefore the X-ray limit is easily reduced by a factor of 100–1000. Figure 17 (right) shows the original drawing of the Bayard–Alpert gauge [17].

3.8 Penning gauges (cold-cathode gauges)

The Penning gauge contains two parallel conducting plates and, between them, a conducting cylinder with its axis orthogonal to the plates (Fig. 18 (top)). A potential difference is applied between the cylinder and the two plates and, at the same time, a longitudinal magnetic field is applied. The configuration of electric and magnetic fields, for suitable values of $\vec{E}$ and $\vec{B}$, becomes a trap for electrons. Cosmic radiation or other some process generates an electron inside this Penning trap, and this electron remains

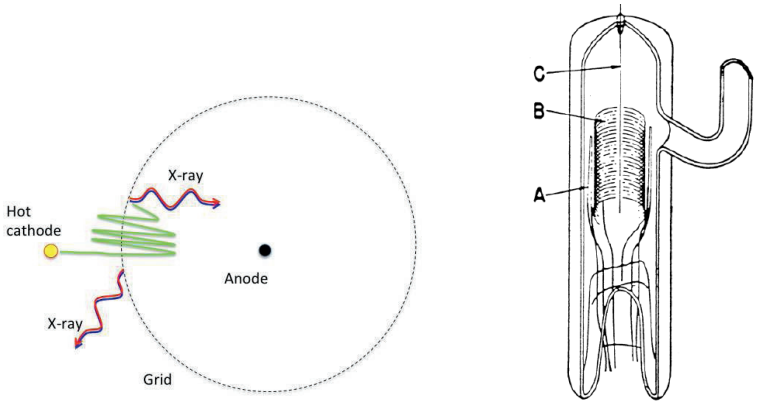


Fig. 17: Right: principle of the Bayard–Alpert gauge. Left: a drawing of the original device [17].

there, performing a complex motion. At the same time, residual-gas molecules move throughout the gauge, and after some time the trapped electron will hit and ionize a neutral residual-gas molecule. The newly produced electron is trapped as well. The positive ion does not have a large velocity, because of its large mass. Consequently, the magnetic component of the Lorentz force $e\vec{v} \times \vec{B}$ has a negligible effect compared with the force exerted by the electric field. Therefore, whereas the new electron remains trapped, the ion follows the electric field and hits the plates (the cathodes). The situation is then that as the ionization process progresses, more electrons are trapped and favour the ionization process further. The trapped electrons form a negative space charge, while the positive ions remain on the surface of the plates. This process of formation of a discharge goes on until the electrostatic potential generated by the space charge is so large that the trapping of new electrons is not possible. From this moment onwards, every new ionization event in the residual gas produces an electron–ion pair, which can be detected when the new electron reaches the anode (or another electron in the discharge reaches the anode).

The gauge becomes operational only when the discharge is fully formed. This process can take a long time when the vacuum is very high. In this case electrons can be created by an auxiliary source for the purpose of feeding the discharge, which eventually becomes self-sustaining. Figure 18 (bottom) shows an example of the dependence of i_+ on P. The response is linear down to 10^{-9} mbar [18].

A summary of the pressure ranges in which the various types of gauges can function is shown in Fig. 19 (adapted from Ref. [7]).

4 Partial-pressure measurements

Partial-pressure gauges allow the determination of components of the gas and their partial pressures. These gauges are typically composed of three parts: (1) an ion source, (2) a mass analyser, and (3) an ion current detection system. The ion source causes ionization of the residual gas, and ions of different species, produced in proportion to their partial pressure, are transported into the mass analyser, which selects the appropriate ion species by filtering the ions to select the correct charge-to-mass ratio. The flux of selected ions is finally measured, providing a quantitative measure of the particle density of that component, i.e., of the partial pressure.

4.1 Ion sources

Ion sources ionize the residual gas via electron impact. The rate of ion production is proportional to the density of each ion species. The electron impact ionization process transfers kinetic energy from electrons to the residual-gas molecules. This process is complex and can generate more than one electron per atom, but the most important process is $\mathrm{M} + \mathrm{e}^- \rightarrow \mathrm{M}^+ + 2\mathrm{e}^-$. The minimum potential necessary for the ionization of the gas atoms is called the ‘appearance potential’ and is approximately 15 eV. Projectile

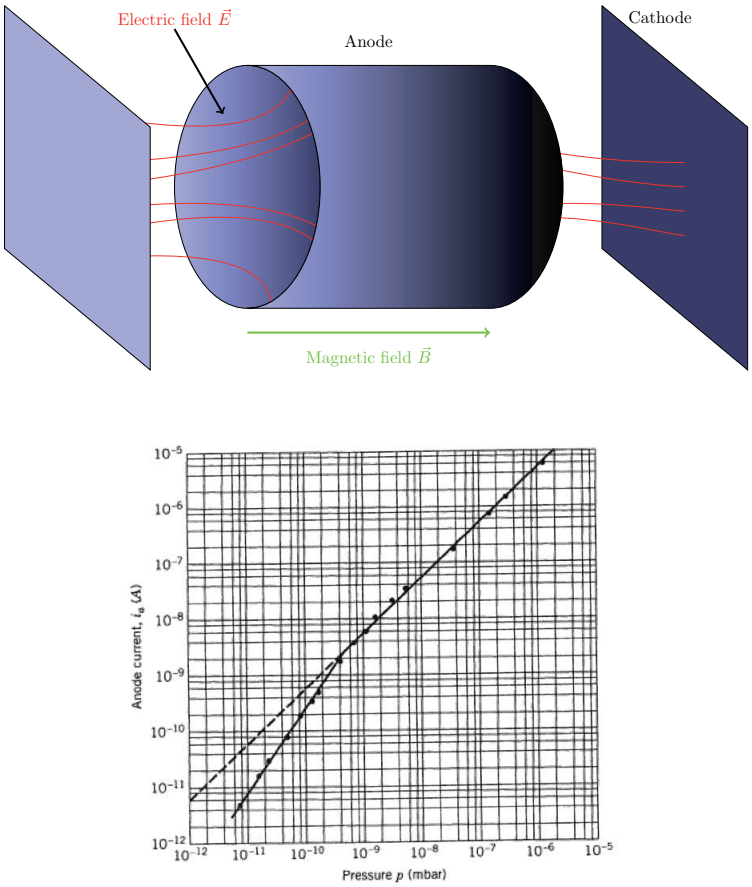


Fig. 18: Top: schematic illustration of a Penning gauge. Bottom: sensitivity of a Penning gauge.

electrons at the appearance potential produce a low ionization rate, as shown in Fig. 14 (right). The maximum production rate varies from gas to gas, but is centred around approximately 100 eV.

A simplified schematic illustration of an ion source is shown in Fig. 20 (left). The electrons produced via thermionic emission are accelerated by the potential difference between the filament and the grid (at a potential V_g). The ion collector provides an electric field to keep the electrons oscillating around the grid, creating an ionization region (light grey area). The ionized atoms inside the grid feel the electric field created by the collector and are accelerated in that region. This scheme does not allow mass analysis, but Fig. 20 (right) shows a modification in which a similar scheme allows the residual gas to be ionized by making the projectile electrons oscillate in a region of the vessel (an external grid prevents the electrons from travelling into the vessel). The ions are extracted from the ionization region and enter the mass analyser. As previously discussed, the current of ions is proportional to the current i_- of the projectile electrons, via the relation $i_+ = i_- \sigma_\mathrm{i} F P/T$, where F is an ion transmission factor depending on the geometry, and P and T are the pressure and temperature of the gas.

4.2 Ion detection

The ion detection may be based on either a Faraday cup (see Refs. [19, 1]) or a secondary-electron multiplier. We discuss the second of these approaches here.

The most important parameter of a secondary-electron multiplier is the 'gain', defined as the ratio of the number of electrons produced at the output to the number of ions entering the multiplier. The device consists a tube, the ends of which have a potential difference V_0 between them (see Fig. 21). As

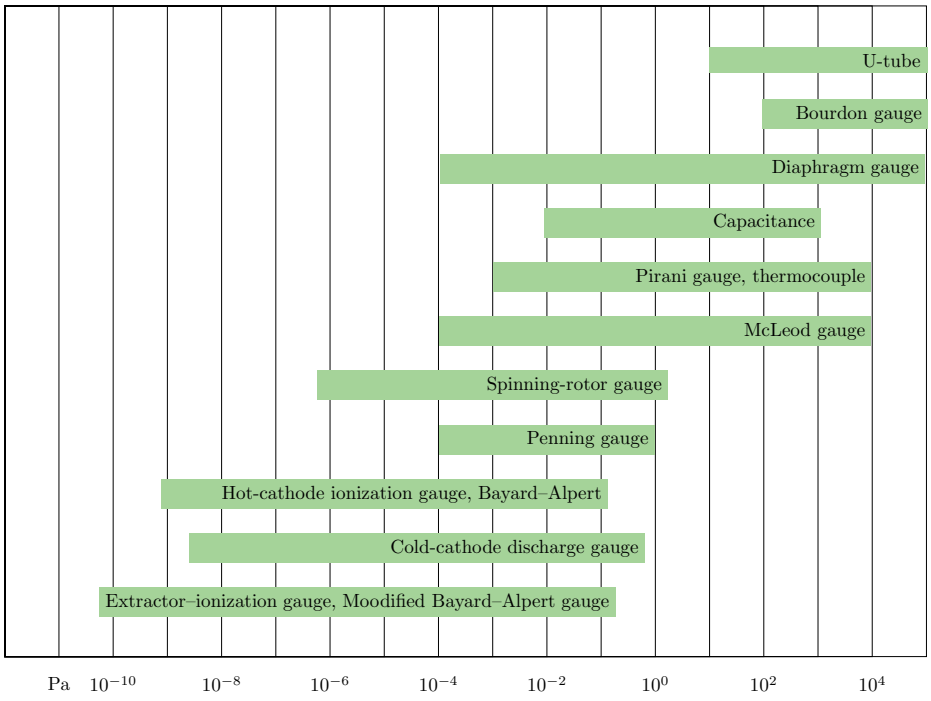


Fig. 19: Summary of pressure ranges of function of gauges

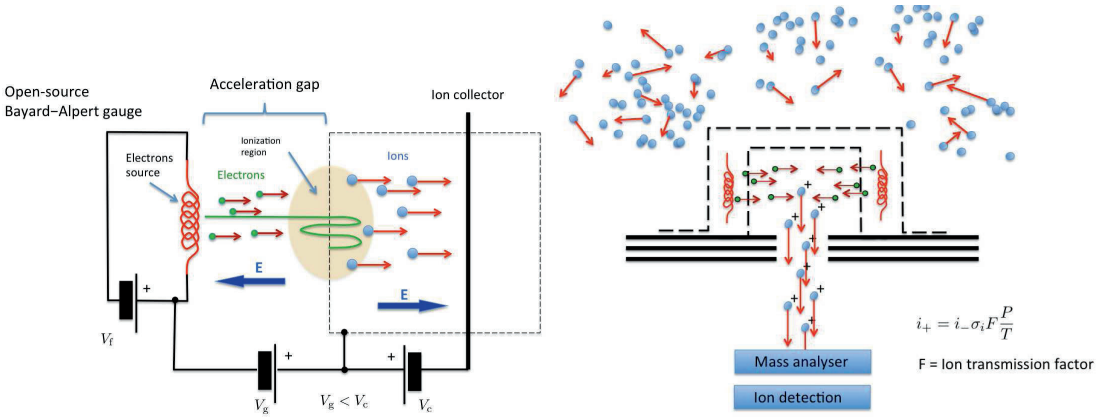


Fig. 20: Left: example of ion source. Right: adaptation for partial-pressure measurement.

the tube has a certain resistivity (that of the dynode material from which it is constructed), a longitudinal electric field is established in it. The principle of operation is the following. When an ion enters the tube, it causes the emission of secondary electrons when it hits the wall of the tube; the ratio of secondary electrons to ions, δ, is called the secondary-electron yield. These electrons are emitted with a certain velocity, and during the time needed for them to hit the tube wall again, they are accelerated by the longitudinal electric field and gain enough energy to create new electrons at the next impact with the wall. This process becomes an avalanche process and provides at the end of the tube a large number of electrons, which can easily be measured. The gain is given by [20, 1]

$$G = \left(\frac{KV_0^2}{4V\alpha^2} \right)^{4V\alpha^2/V_0},$$

where $\alpha = L/d$, L is the length of the tube, d is the diameter of the tube, V_0 is the applied voltage, V is

the initial energy of the electron, $K = \delta V_c$ (where δ is the secondary-emission coefficient), and V_c is the collision energy.

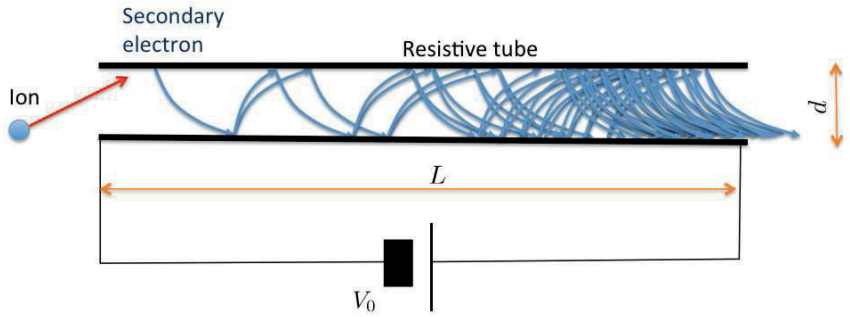


Fig. 21: Schematic illustration of a secondary-electron multiplier.

4.3 Mass analysers

The mass analyser filters the ionized particles according to the ratio of charge to mass. There are several types of mass analyser, classified according to the technique used to perform the filtering process. See Ref. [1] for a complete review.

4.3.1 Quadrupole mass spectrometer

The quadrupole mass spectrometer contains four parallel rods, connected electrically to an alternating voltage source [21]. A schematic illustration of the device is shown in Fig. 22. If r_0 is the distance of each of these rods from the axis, the equation of motion of a particle with mass M and charge e travelling through the channel between the rods is

$$\begin{aligned} \frac{\mathrm{d}^2x}{\mathrm{d}t^2} &= \frac{e}{M}\frac{1}{r_0^2}(U + V\cos\omega t)x, \\ \frac{\mathrm{d}^2y}{\mathrm{d}t^2} &= -\frac{e}{M}\frac{1}{r_0^2}(U + V\cos\omega t)y, \\ \frac{\mathrm{d}^2z}{\mathrm{d}t^2} &= 0, \end{aligned}$$

where U and V are constants. By defining $a = 4eU/Mr_0^2\omega^2$ and $q = 2eV/Mr_0^2\omega^2$, this equation can be

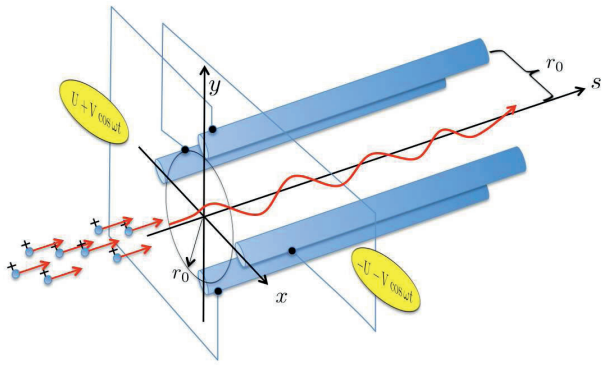


Fig. 22: Schematic illustration of a quadrupole mass spectrometer.

rescaled into the form

$$\begin{aligned} \frac{\mathrm{d}^2x}{\mathrm{d}\theta^2} &= (a + 2q\cos 2\theta)x, \\ \frac{\mathrm{d}^2y}{\mathrm{d}\theta^2} &= -(a + 2q\cos 2\theta)y, \\ \frac{\mathrm{d}^2z}{\mathrm{d}\theta^2} &= 0, \end{aligned}$$

where the equations for x and y are Mathieu equations [22]. The variable θ is defined by $2\theta = \omega t$. The properties of the transport of the ions along the channel between the rods depend on the parameters a, q.

The motion of the ions in the transverse plane may be stable or unstable. If $q = 0$, the four rods act like a constant focusing channel, which is focusing in the horizontal plane if $a < 0$. Clearly, however, transport is impossible when $q = 0$, as one of the two transverse planes is always defocusing. The presence of the term q changes the stability properties of the channel. Usually, the stability of a Mathieu-driven particle transport is referred to an infinitely long channel. For practical purposes, a length corresponding to 100 transverse linear oscillations can typically be considered as long enough for this.

The stability and instability of transport along a channel driven in accordance with a Mathieu equation are summarized in the stability charts shown in Fig. 23. In the left chart, the shaded region

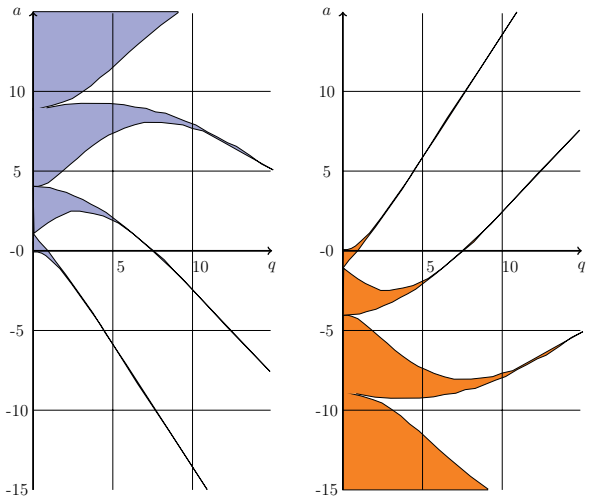


Fig. 23: Stability charts of the Mathieu equation. The left chart shows the stability in the vertical plane, and the right chart the stability in the horizontal plane.

shows the parameters q, a for which the motion in the y-plane is stable. Similarly, the right chart shows the stability in the x-plane. The horizontal axis is q, and the vertical axis is a. Depending on the pair of parameters (q, a), the motion can be either stable or unstable in the horizontal plane, and similarly for the vertical plane. Only when the motion is stable in both planes will the ions pass through the channel between the rods and be detected by the ion detector. A particular use of the stability chart is shown in Fig. 24.

Here, the triangle-like region where the the two shaded areas overlap is the region of the parameters (q, a) which guarantees stability of the motion in both planes. The tip of the triangular region has values $q_0 = 0.706, a_0 = 0.237$. The mass analyser is used by dynamically varying the parameters q, a so as to sweep along a line in the chart, crossing the tip of the stability region. This process is shown by the straight line in Fig. 24. Given a certain species of ions, characterized by a mass M_1 and charge e, there

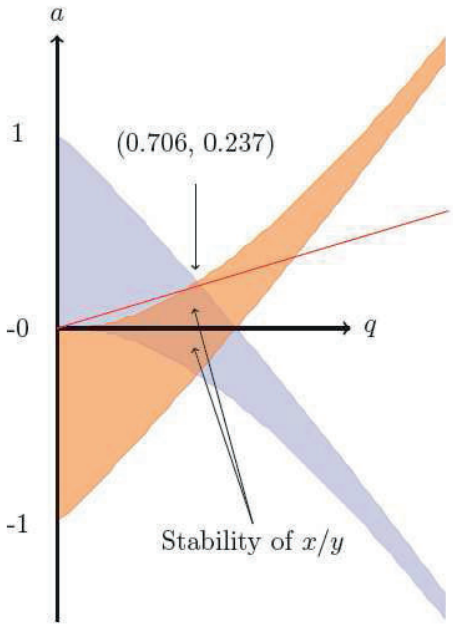


Fig. 24: Stability region for vertical and horizontal motion in the Mathieu stability chart.

are only two values of U_1 and V_1 such that $q = q_0$ and $a = a_0$. Therefore, by varying V and keeping the ratio V/U constant, the parameters q, a are spanned as shown in Fig. 24 as the tip of the stability region is passed through. The tip of the stability region is reached at $V = V_1$, and therefore we obtain

$$\frac{M_1}{e} = \frac{2V_1}{q_0 r_0^2 \omega^2}.$$

The voltage V_1, which corresponds to passage through the tip, is determined by the ion detector. Current peaks for the ions that are transported through the mass analyser can be identified in a diagram of V versus i_+. In this way, the current i_+ of those ions which have a mass-to-charge ratio M_1/e is measured. The current i_+ is proportional to the partial pressure of the gas with mass M_1.

The precision of the mass identification is given by the resolving power $M/\Delta M$, where ΔM is the uncertainty of the device with respect to filtering neighbouring ions. This error is related to the time needed in the quadrupole for unstable motion to develop. The escape of the ions is typically related to the number of oscillations made by the ions as they pass along the channel between the four rods. It is found that [21]

$$\frac{M}{\Delta M} \propto \frac{M\omega^2 L^2}{2eV_z}. \tag{4}$$

The term eV_z represents the kinetic energy that the ions have at the entrance to the channel. The time of transit is $L/v_z = \sqrt{M/2eV_z}L$, and the number of oscillations inside the channel before the particle leaves is $\sqrt{M/2eV_z}L\omega$, and hence the left-hand side of Eq. (4) shows that the resolving power is proportional to the square of the number of transverse oscillations.

4.3.2 Magnetic-sector analyser

A different strategy for analysing the mass of ions is implemented by the magnetic mass analyser. The principle is shown in Fig. 25 (from Ref. [1], p. 462). The ions are transported to the entrance of a magnetic sector, with a bending angle of 90° for example. The magnetic field B is varied with time by ramping it from zero to some maximum value. Only ions which have the correct ratio M/q can pass

through the magnet and reach the detector. Therefore, in a graph of B versus i_+, a spike in the current i_+ is observed when M/q satisfies the condition

$$\frac{M}{q} = \frac{R^2 B^2}{2E_\mathrm{k}/q}.$$

Here E_k is the kinetic energy of the ions, and q is their charge. The resolving power of the device is

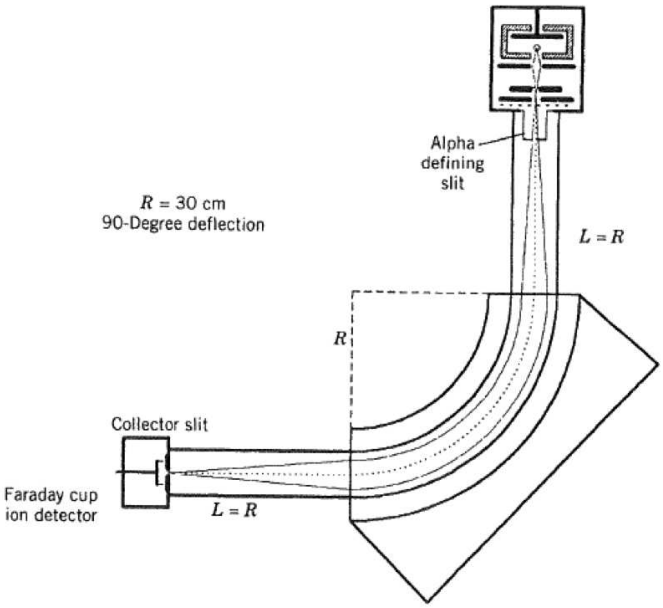


Fig. 25: 90° magnetic-sector mass spectrometer.

related to the trajectory of the ions, which needs to pass through the collector slit. The resolving power $M/\Delta M$ is given approximately by

$$\frac{M}{\Delta M} \simeq \frac{R}{W_\mathrm{source} + W_\mathrm{collector}},$$

where W_source and $W_\mathrm{collector}$ are the widths of the slits at the source and collector [1].

4.3.3 Omegatron

The omegatron is a device containing two parallel plates, which provide an RF electric field. A magnetic field B is applied orthogonally to the electric field [23]. A schematic illustration of the omegatron is shown in Fig. 26 [24, 25]. The principle of operation is the following. The hot filament is a source of electrons, which cause ionization in the vacuum gas. Once an atom of gas is ionized, the ion is subject to the RF electric field. The magnetic field alone would cause a circular motion of the ions at a constant velocity v, with a radius of rotation $r = Mv/(qB)$. Therefore the revolution time is $\tau = 2\pi M/(qB)$. If the frequency of the electric field is $1/\tau$, then a resonant process takes place, and the ion gains energy and spirals outwards, eventually intercepting the collector. Again, a measurement of the current between the filament and the collector provides the ionization rate of atoms with a specific mass-to-charge ratio. A scan of the frequency allows the determination of those ions which can resonate, and hence their mass.

Owing to the complex resonant dynamics, ions with a mass-to-charge ratio that is not exactly the selected value can sometimes undergo large oscillations and reach the ion collector. This possibility determines the resolving power of the device, which can be calculated as

$$\frac{M}{\Delta M} = 4.8 \times 10^{-5} \frac{R_0 B_0^2}{E_0} \frac{1}{M},$$

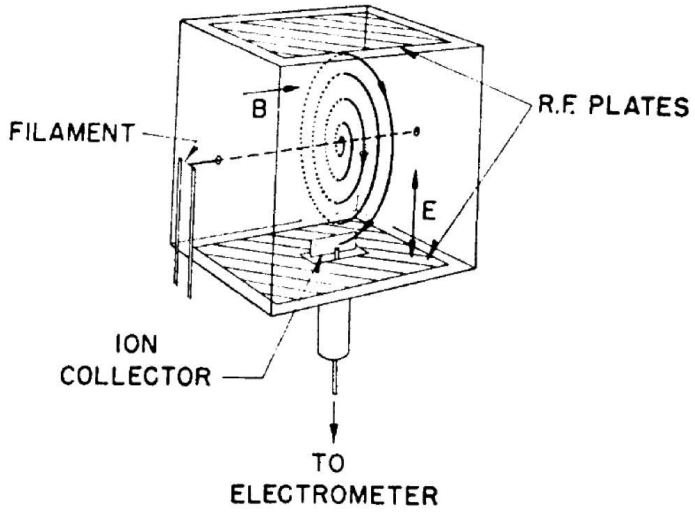


Fig. 26: Schematic illustration of the omegatron [24].

where R_0 is the distance from the collector to the filament (in cm), E_0 is the amplitude of the electric field (in V/cm), B_0 is the strength of the magnetic field (in gauss), and M on the right-hand side is the mass of the ion in atomic mass units [23].

Acknowledgements

The author thanks Maria Cristina Bellachioma for comments on this paper and corrections.

References

[1] J.M. Lafferty, *Foundations of Vacuum Science and Technology* (Wiley, New York, 1988).

[2] A. Chambers, *Modern Vacuum Physics* (CRC Press, Boca Raton, FL, 2004).

[3] C. Benvenuti. Molecular surface pumping: the getter pumps, CERN Accelerator School on Vacuum Technology, 1999.

[4] M.C. Bellachioma, J. Kurdal, M. Bender, H. Kollmus, A. Kraemer, and H. Reich-Sprenger, *Vacuum* **82**(4) (2007) 435–439.

[5] C. Benvenuti, Getter pumping, CERN Accelerator School, CERN-2007-003 (2007), p. 313.

[6] L.D. Hall, *Rev. Sci. Instrum.* **29** (1958) 367.

[7] N. Marquardt, Introduction to the principles of vacuum physics, CERN Accelerator School on Vacuum Technology, 1999, p. 1.

[8] H. McLeod, *Philos. Mag.* **48** (1874) 110.

[9] J.K. Fremerey, *Rev. Sci. Instrum.* **44** (1973) 1396–1397.

[10] J.K. Fremerey, *Vacuum* **32**(10–11) (1982) 685–690.

[11] J.K. Fremerey, *J. Vac. Sci. Technol. A* **3** (1985) 1715.

[12] F.J. Redgrave and S.P. Downes, *Vacuum* **38**(8–10) (1988) 839–842.

[13] E.H. Kennard, *Kinetic Theory of Gases, with an Introduction to Statistical Mechanics* (McGraw-Hill, New York, 1938).

[14] M. Pirani, *Dtsch. Phys. Ges. Verh.* **8** (1906) 686.

[15] A. Von Engel, *Ionized Gases*, AVS Classics in Vacuum Science and Technology (AIP, New York, 1997).

[16] D. Alpert, *Vacuum Science and Technology: Pioneers of the 20th Century: History of Vacuum Science and Technology* (American Vacuum Society, New York, 1993).

[17] R.T. Bayard and D. Alpert, *Rev. Sci. Instrum.* **21**(6) (1950) 571.

[18] S.L. Rutherford, "Sputter-Ion Pumps for Low Pressure Operation," 10th National Vacuum Symposium, 1963, (MacMillan, New York, 1963) p. 185.

[19] G.F. Metcalf and B.J. Thompson, *Phys. Rev.* **36** (1930) 1489–1494.

[20] J. Adams and B.W. Manley, *IEEE Trans. Nucl. Sci.* **13** (1966) 88–99.

[21] P.H. Dawson, *Quadrupole Mass Spectrometry and Its Applications*, AVS Classics in Vacuum Science and Technology (AIP Press, New York, 1995).

[22] E. Mathieu, *J. Math. Pures Appl.* (1868) 137.

[23] H. Sommer, H.A. Thomas, and J.A. Hipple, *Phys. Rev.* **82** (1951) 697–702.

[24] J.H. Leck, Partial pressure measurement, CERN Accelerator School, 1999, pp. 89–98.

[25] D. Alpert and R.S. Buritz, *J. Appl. Phys.* **25**(2) (1954) 202.

Fundamental of cryogenics (for superconducting RF technology)

Paolo Pierini
INFN Sezione di Milano, Laboratorio Acceleratori e Superconduttività Applicata, Milano, Italy

Abstract
This review briefly illustrates a few fundamental concepts of cryogenic engineering, the technological practice that allows reaching and maintaining the low-temperature operating conditions of the superconducting devices needed in particle accelerators. To limit the scope of the task, and not to duplicate coverage of cryogenic engineering concepts particularly relevant to superconducting magnets that can be found in previous CAS editions, the overview presented in this course focuses on superconducting radio-frequency cavities.

1 Cryogenics and CAS

Several previous CERN Accelerator Schools (CASs) [1–3] have extensively covered many of the theoretical and technological aspects of cryogenics, which is the practice of reaching and maintaining low temperatures. Cryogenic engineering is the technical discipline needed to guarantee the operational environment of superconducting devices, such as the cavities and magnets for particle accelerators.

In this short contribution I will illustrate a few key concepts and provide a few "practical engineering" considerations concentrated on the case of superconducting radio-frequency (RF) linear accelerators. Superconductivity in magnets (and the necessary cryogenic implications) is another broad topic, somewhat less relevant to the topic of this course (high-power hadron machines) and extensively covered in previous CAS courses [1–3].

A separate contribution to this course (by H. Podlech) is dedicated to the systematic comparisons between the room-temperature normal conducting RF technology and the superconducting RF technology, while this contribution concentrates on cryogenic concepts in general and design considerations on cryostats and cryogenics for superconducting RF linear accelerators. This review is of course incomplete in many areas (such as the cryogenic instrumentation) and intentionally "light" on the most theoretical topics that could easily require an entire course.

2 Superconducting RF cavities and their cryogenic requirements

When a DC field is applied to a superconducting device, electrons condensed in Cooper pairs carry all of the current, and the electric resistance vanishes. All electrons are condensed into pairs only at $T = 0$ K, and the fraction of paired electrons decreases exponentially until the material reaches its critical temperature, above which a purely resistive behaviour is shown.

In the case of time-dependent RF fields and currents, dissipation takes place at all temperatures above 0 K, due to the fact that unpaired electrons feel the effect of the RF field and, differently from the frictionless motion of the Cooper pairs, generate currents leading to resistive losses.

In any superconducting radiofrequency (SCRF) resonator, which is a device aiming at creating a pattern of electric and magnetic fields in a region of space enclosed by a metallic boundary, power is dissipated on the metallic cavity walls, according to their surface resistance R_s and to the surface magnetic field intensity:

$$P_{diss} = \frac{1}{2} R_s \int_S |H|^2 dS \tag{1}$$

It is therefore from exploring the behaviour of the surface resistance with respect to temperature and frequency for the superconducting and normal conducting cases that the advantages of superconductivity can be assessed.

2.1 The advantages of RF superconductivity

RF fields at a frequency f penetrate into a normal conducting metal (of electrical conductivity σ) by the skin depth δ, thus leading to a surface resistance given by

$$R_s = \frac{1}{\delta\sigma} = \sqrt{\frac{\pi f \mu_0}{\sigma}} \tag{2}$$

Dissipation in a normal conducting device therefore depends on material properties (through σ) and has a frequency dependence as $f^{\frac{1}{2}}$.

The Bardeen, Cooper and Schrieffer (BCS) theory of superconductivity can be used to develop expressions for the surface impedance of superconductors on the base of several fundamental material properties (as the mean free path, coherence length, etc.) [4]. For the case of Nb at temperatures below its critical temperature and frequencies below the terahertz range these complex expressions can be rewritten as the following "engineering" approximate formula:

$$R_{BCS}(T, f)\ [\Omega] = 2 \times 10^{-4} \frac{1}{T} \left(\frac{f}{1.5}\right)^2 \exp\left(-\frac{17.67}{T}\right) \tag{3}$$

where T is expressed in Kelvin and f is expressed in gigahertz [4].

The ratio between Eqs. (3) and (2) is shown in Fig. 1, for frequencies up to 3 GHz and the two particular temperatures of 4.2 K (boiling temperature of liquid He at ambient pressure) and 2.0 K (subatmospheric liquid He in the superfluid phase). This plot shows that in the range between 300 MHz and 1.5 GHz the use of superconductivity can significantly reduce the power dissipated on the resonator walls by four to seven orders of magnitudes (depending on operating temperature and frequency).

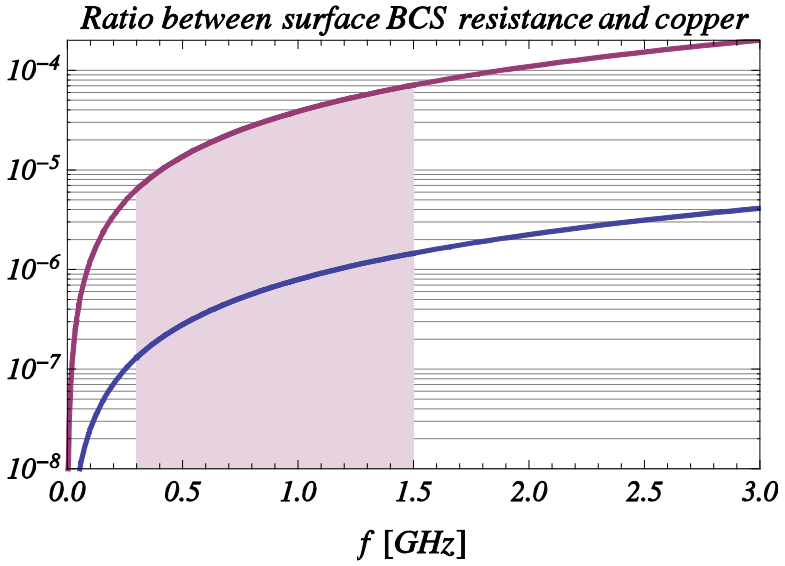


Fig. 1: Ratio between the BCS surface resistance term given by Eq. (3) to the copper surface resistance given by Eq. (2), for T = 2 K (bottom blue curve) and 4.2 K (upper purple curve). The "practical" SRF frequency region 300 MHz to 1.5 GHz has been highlighted.

2.2 SCRF Cavity fabrication technology

The situation illustrated by Fig. 1 represents an idealized condition, in real cases the surface resistance of a practically achievable superconductor is only partially due to the BCS contribution and can be written more generally as

$$R_s(T, f, H) = R_{BCS}(T, f) + R_{mag}(H, f) + R_{residual} \tag{4}$$

where the first term after the BCS contribution represents the increased losses due to trapped DC magnetic fields (e.g. Earth's magnetic field, which needs to be properly shielded from the cavity environment) and the second is the so-called "*residual resistance*" contribution, accounting for several sources, such as chemical residues on the surfaces, foreign material inclusions, condensed gases or hydrides and oxides layers. To reach surface resistance values close to the BCS contribution, stringent material and procedures quality measures need to be implemented and strictly followed for the fabrication of a successful resonator made of a bulk superconductor material, such as Nb.

This is a broad topic, and here it is sufficient to mention that high-purity Nb material, free from inclusion and defects, should be used for the cavity RF surfaces. Clean joining technologies leading to no foreign material inclusions need to be followed during the resonator fabrication (most important is the electron beam welding technique). An aggressive chemistry, or electrochemistry, needs to be performed on all of the RF surfaces, to completely remove the surface layer damaged during the fabrication process, and all final treatments and cavity preparation for operation should take place in a clean room environment. A review of the SCRF Cavity technology can be found in Ref. [5].

2.3 The need for a cryoplant and the Carnot theorem

In addition to the cavity fabrication issues described above, one has also to realize that although superconductivity allows a dramatic decrease in the resonator power consumption, the power is deposited at the extremely low temperatures of operation.

Therefore, special measures are needed to achieve, guarantee and preserve the low-temperature environment required for the onset of superconductivity. In short, a *cryogenic infrastructure* (the cryoplant) is needed for the production and handling of the coolant, and for the removal of heat deposited at low temperatures.

The cryoplant is a thermal machine that performs work at room temperature to extract heat at low temperatures. Its two main functions are to bring the devices to the nominal temperatures and to keep them cool by removing any heat deposited at low temperatures. For a thermal machine that operates between the ambient temperature T_{amb} and the cold temperature T_{cold} to remove the heat (Q_{in}) from the cold temperature environment by means of performing work W at the ambient temperature, the following relation holds:

$$Q_{in} \leq W \frac{T_{cold}}{T_{ambient} - T_{cold}} \tag{5}$$

where the equality is valid for an ideal reversible process and the factor $T_{cold}/(T_{ambient} - T_{cold})$ represents the efficiency of the Carnot cycle between these temperatures. The inequality of Eq. (5) takes into account the efficiency of the real thermal machine, in which irreversible processes take place, which is in the range 25 % to 30 % for 4.2 K operation and 15 % to 20 % for 2 K. We can summarize these considerations in the following statements:

- to remove 1 W at 4.2 K, approximately 250 W are needed at room temperature;
- to remove 1 W at 2 K, approximately 750 W are needed at room temperature.

Even taking into account this overall thermal efficiency, however, Fig. 1 clearly shows that there is still an overall advantage for RF superconductivity in the moderate frequency region (up to ~ few gigahertz).

Niobium, with a critical temperature of 9.2 K, is the currently available material for fabrication of bulk superconducting resonators [4]. Table 1 shows the normal boiling point (temperature at which the liquid vapour pressure equals the atmospheric pressure) for various cryogenic fluids. Clearly, for the operation of SCRF resonators, the only available option is the use of helium.

Table 1: Normal boiling point of various fluids

	$_4$He	H_2	Ne	N_2	Ar	O_2
Normal boiling point temperature, K	4.22	20.28	27.09	77.36	87.28	90.19

2.4 Heat pumps as entropy pumps (get ready for non-idealities)

Entropy is the correct function of state that allows proper understanding and description of cryoplants, and the assessment of the non-ideality level of the process [6]. The Second Law of thermodynamics states that in any process the entropy output is always greater than the entropy input (and equal to that in the ideal, practically unreachable, reversible case).

Cryogenic systems can be described as entropy pumps [6], which transfer entropy from the cold region into the warmer environment (if we perform the necessary work). The top half of Fig. 2 represents a cryogenic plant as an entropy pump between the cold temperature, T_{in}, and the hot temperature, T_{out}, environments. To extract the heat flowing into the cold region, Q_{in}, entropy needs to be "transported" to the hot region and released there as a heat output at the hot temperature (as $Q_{out} = S_{out}\, T_{out}$).

On the top left part of Fig. 2 the ideal reversible process is illustrated. In this case entropy is conserved from the cold to the hot environments, thus the heat output is amplified by the ratio T_{out}/T_{in}. The energy conservation principle requires that the difference between the heat released into the hot environment and the heat deposited into the cold region is provided as work performed by the machine. On the top right part of the figure the irreversible case is shown, and the cryoplant is shown to introduce non-idealities, thus incrementing the entropy flowing from the cold to the hot environments. This increased entropy is then released as heat at the hot temperature level and additional work needs to be performed to make up for this non-ideality of real processes.

The bottom part of Fig. 2 shows a numerical example of the concepts expressed before for a cryoplant operating between 2 K and 300 K. A heat load of 1 W is deposited at 2 K, corresponding to an input entropy of 0.5 W/K. In the ideal case this entropy is preserved to the output and a heat load of 0.5 W/K × 300 K = 150 W is released to the ambient. A work of 149 W is thus needed by the ideal reversible process. If we now assume that non-idealities in the real process are such that entropy is incremented by five times (i.e. we have an internal entropy source of 2 W/K), then the output entropy of 2.5 W/K at the highest temperature corresponds to a heat rejection in the environment of 750 W, nearly all to be provided as work by the machine.

The overall efficiencies referred in the previous paragraphs are therefore related to the fact that real large helium cryoplants have internal entropy sources accounting for a multiplication of the incoming entropy from the load by a factor of ~3 at 4.2 K, and by a factor of ~5 at 2 K.

In a real cryoplant there are a large number of entropy sources contributing to the reduction of efficiency with respect to the ideal reversible case and accounting for the factors discussed above. A detailed review is presented in Ref. [6], together with an analysis of the various thermodynamic cycles used in cryoplants and the description of their main technological components. The largest single source of entropy production is typically associated with the gas compression stages (e.g. to account

for gas recooling in multistage compression schemes by discharging heat to the environment). Entropy can also be introduced in the expansion stages (by spurious heat loads by conduction or radiation or by warm gas leaks), in heat exchangers and by heat leaks into the cold parts of the plant (e.g. by instrumentation cables or gas leaks).

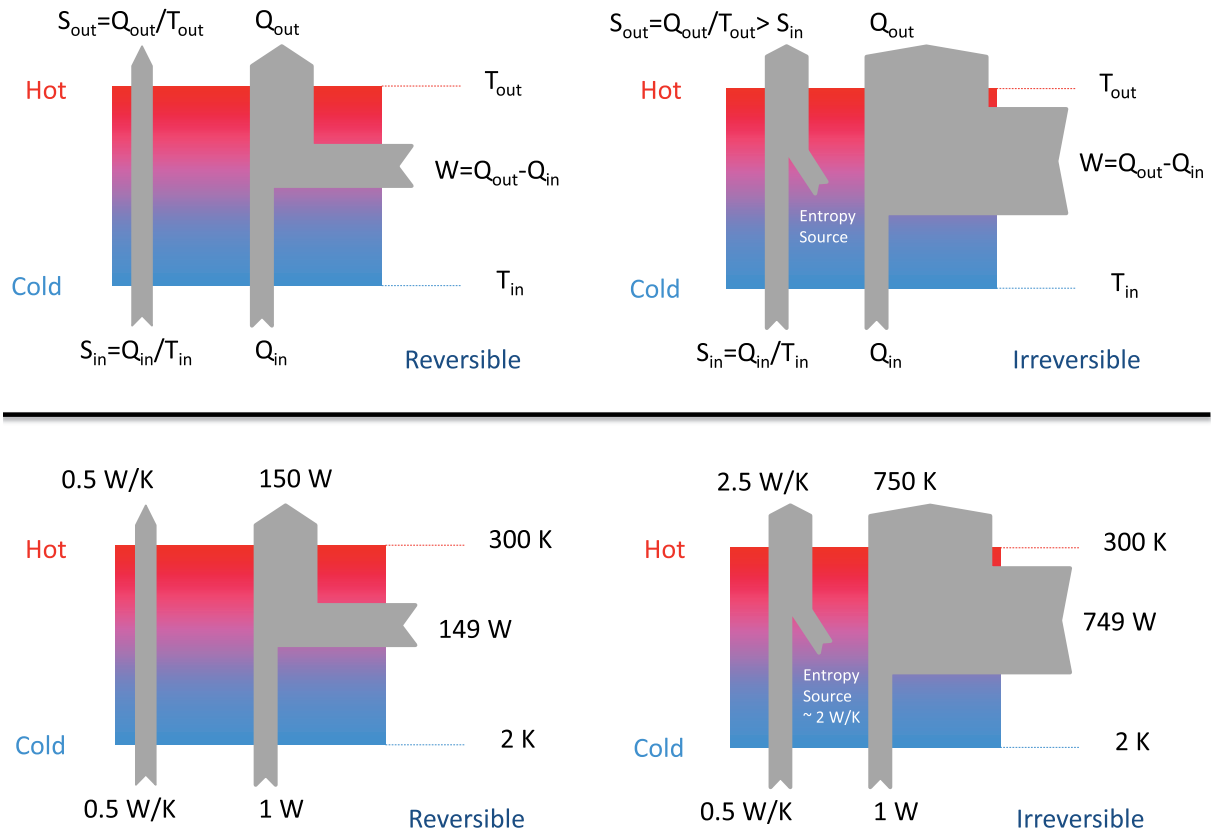


Fig. 2: Cryoplants as entropy pumps. The top part of the figure shows the entropy and heat flow for the ideal and irreversible cases. The bottom part illustrates a numerical example. For a full explanation of the figure refer to Section 2.4.

2.5 From a conceptual helium refrigerator to real machines

A conceptually simple cryogenic cycle aimed at removing a thermal load from a low-temperature region consists of two main stages: a *compression* stage, where the entropy content of the fluid is reduced at the expense of the work performed to compress the gas; and an *expansion* stage, which cools the fluid, either at the expense of internal forces or removing energy as work. In the compression stage the fluid is prepared to receive the entropy content of the load and entropy is then released to the environment by a heat exchanger right after compression. The expansion stage cools the fluid to a slightly lower temperature than the load, to extract entropy from it. These two stages are typically separated by a temperature staging device (as a counterflow heat exchanger) which guarantees that the two processes take place at different temperatures and prepares the fluid for optimal compression and expansion conditions. Heat exchangers allow transferring heat (i.e. entropy) from the cold region to the cycle and discarding it from the cycle to the environment.

The most common compound cycle used in modern cryoplants is the so-called Claude Cycle, displayed in Fig. 3 in its simplest variant, which can be optimized using several heat exchanges and expansion stages (typically using turbines) for efficiency.

A review of helium liquefaction plants and of the several variants of their cycle is outside the scope of this introductory course and the reader can find excellent coverage in specialized contributions of other CAS courses on superconductivity [6, 7], which also introduce all of the necessary thermodynamics concepts for their discussion.

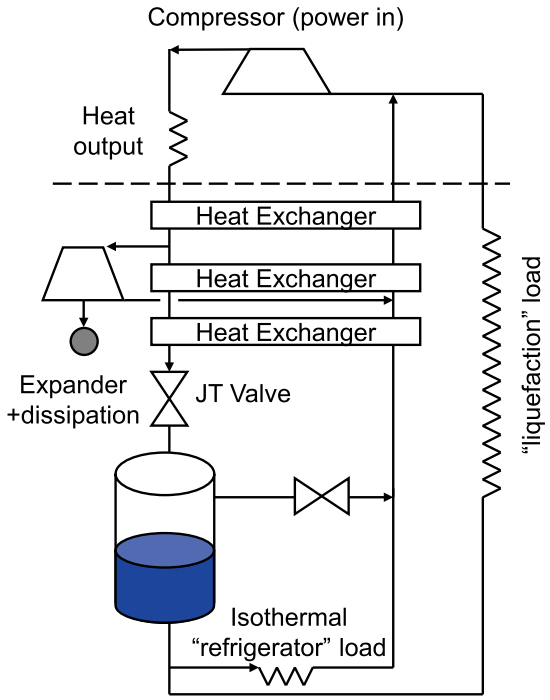


Fig. 3: Schematic view of a cryoplant based on the Claude cycle

2.6 Heat removal from the cold region: operating modes for a superconducting device

A major task of the cryoplant after cooling the superconducting devices is to guarantee the operating temperature by removing the heat deposited in the cold region. In general, heat is removed increasing the energy content of the cooling fluid (whether a liquid or a vapour). The cooling capacity is directly proportional to the fluid mass flow $\dot{m}$ and the enthalpy difference ΔH between the input and output fluid states:

$$P\,[\mathrm{W}] = \dot{m}\,[\mathrm{g/s}]\,\Delta H[\mathrm{J/g}] \tag{6}$$

Superconducting RF cavities are usually cooled in isothermal conditions either in pool boiling Helium I at atmospheric pressure (or slightly above, 4.2–4.5 K operation, as in HERA, LEP or KEKB) or in a saturated Helium II (superfluid) subatmospheric bath (2 K operation at 31.29 mbar, below the 2.17 lambda point, as in CEBAF, TTF, SNS, and the foreseen mode for ILC and ESS). In this isothermal cooling mode the heat absorbed by the load is spent in the phase transition from the liquid phase into the vapour phase (latent heat is 20.3 J/g for pool boiling He I at 1 atm, 4.2 K and 23.4 J/g for saturated Helium II bath at 31 mbar, 2 K).

The pressure–temperature phase diagram for He is shown in Fig. 4, where the typical regions of cooling mode of superconducting devices are highlighted. Figure 4 shows also cooling modes for superconducting magnets [9], usually cooled by forced flow of sub-cooled or supercritical He I (such as Tevatron, HERA and the SSC) or by pressurized He II (LHC). For magnets the operation with pressurized helium (single phase, either He I or superfluid He II) gives the maximum penetration of the coolant into the magnet coils, for increased heat transfer and stability, but leads to the need for a proper handling of the temperature rises induced along the cooling channels. For the thin-walled superconducting resonators, however, the pool boiling mode (especially in the case of a subatmospheric bath), in addition to the obvious advantage of a lower operating temperature [see

Eq. (3)], has the benefits of increased pressure stability, large local heat transfer capabilities to accommodate hot spots and isothermal cooling conditions. Saturated Helium II low-pressure (~30 mbar) operation is particularly important for the case of high-loaded quality factor (i.e. small bandwidth) resonators, to minimize field phase and amplitude perturbations induced by pressure variation in the bath.

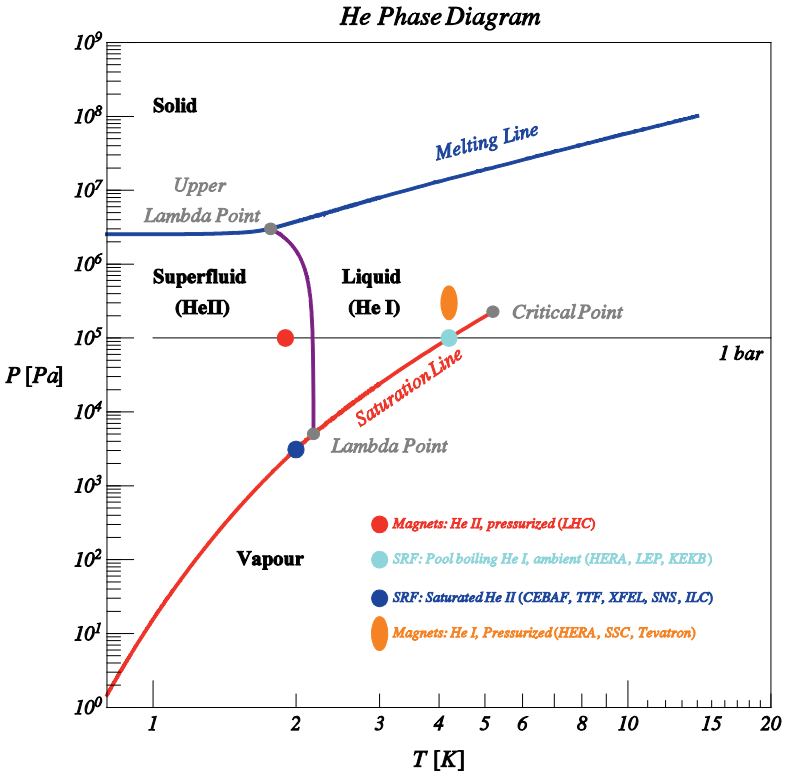


Fig. 4: Pressure–temperature phase diagram of helium

On the basis of the considerations expressed above, saturated He II operation with a two-phase mixture along the saturation curve is usually the preferred operation mode of high-field high-quality SCRF resonators. This operation mode, however, introduces a few additional complexities: a pumping system to establish the low-pressure operation; the need to handle pressure increase in the cryogenic piping at large mass flows not to perturb the operation temperature; the potential risk of air leaks into the low-pressure environment; and the operation of RF and electrical feedthroughs close to the minimum of the Paschen curve (e.g. in vertical test stations) where breakdown conditions are greatly enhanced.

Ultimately, heat is extracted from the cold region by evaporation of vapours from the saturated bath, which are then carried away. Since the latent heat of helium is quite small (~23.4 J/g at 2 K, for reference N_2 is ~200 J/g at 77 K), large heat deposition in the saturated He II bath imply large mass flows of low-pressure gas, i.e. a large volume flow. To reduce the volume flow large installations include cold compressors to increase pressure conditions of the cold gas before it reaches room temperature [8]. These devices also reduce the need of subatmospheric piping in the system, thus decreasing the possibility of air inleaks.

3 Cooling (and maintaining at cold) accelerator components

Physicists and engineers designing prototypical superconducting accelerator components usually concentrate mostly on the component design (either cavities or magnets) and later “jacket” them into helium vessels and cryostats for their testing.

For a large superconducting accelerator facility, however, the cooling mode of operation, the heat transfer mechanism in the operational environment, provisions for cooldown and warmup procedures and transient operation conditions need to be considered early in the component design and integrated into its supporting infrastructure, i.e. the cryogenic system. All of these considerations can affect the complexity (or, conversely, the simplicity) of the cooling system and are needed for a proper trade-off optimization between component and support system design.

3.1 Heat transfer mechanisms

One important consideration in the design of superconducting accelerator components operating at cold temperatures is the proper understanding and handling of heat transfer processes, especially those not strictly related to the loads associated with the proper device function (e.g. RF losses in a cavity will always occur at the cold operating temperature). In particular, spurious heat leaks due to mechanical supports, ancillary equipment as RF couplers and cavity tuning devices, or cabling for instrumentation and diagnostics, should be adequately minimized.

As explained in the previous sections, removal of power deposition at cold temperatures implies much larger power consumption at room temperature. Extending the considerations on thermal machine efficiencies outlined in Section 2.3 at intermediate temperature levels we have the following (approximate) efficiencies:

- 1 W deposited at 2 K requires approximately 750 W at room temperature;
- 1 W deposited at 4.2 K requires approximately 250 W at room temperature;
- 1 W deposited at 70 K requires approximately 12 W at room temperature.

These rough considerations suggest that it is crucial to intercept any spurious thermal flux from the room-temperature environment before it reaches the coldest region. This consideration applies to all heat transfer mechanisms that can take place in the operational environment by conduction, convection or radiation. For a detailed overview of heat transfer mechanisms, refer to Refs. [10, 11].

3.1.1 Conduction

Heat is transported by conduction mechanisms inside solid or stagnant fluids, by processes occurring at the atomic scale. Conduction obeys Fourier's law, stating that the heat flow $\dot{Q}$ through a surface S of a material with thermal conductivity k in the presence of a temperature gradient ∇T is given by

$$\dot{Q} = -k(T)S\nabla T \tag{7}$$

In the simple single-dimensional case and for a material of length L and cross section S where the two end points are placed at the temperatures T_{hot} and T_{cold} can write the conduction equation in its integral form:

$$\dot{Q} = \frac{S}{L}\left(\int_{T_{ref}}^{T_{hot}} k(T)dT - \int_{T_{ref}}^{T_{cold}} k(T)dT\right) = \frac{S}{L}(K(T_{hot}) - K(T_{cold})) \tag{8}$$

where $K(T)$ is the thermal conductivity integral (evaluated from a reference temperature T_{ref}), usually found in literature or available from specialized software packages containing material properties at cryogenic temperatures [12].

Thermal conductivity (and its integral) varies greatly with temperature and with material. Proper choice of materials and thermal intercept strategies for the conduction paths to the cold environment is therefore necessary in the design of the SCRF components.

SCRF cavities need several penetrations from the room-temperature operation to provide structural support and frequency regulation, to provide the RF power for beam acceleration, to damp

and extract spurious RF components at higher frequencies that may lead to detrimental effect on the beam characteristics, and cables for diagnostics and control. Direct connections from the room-temperature environment to the cold region should be avoided and, when needed, proper use of low thermal conduction materials should be made, providing thermalization at intermediate temperatures to intercept the heat flux under more favourable conditions.

Figure 5 shows thermal conductivity as a function of temperature for a few materials commonly used in cryogenic applications, and Figure 6 shows the thermal conductivity integrals for the same materials (with T_{ref} =2 K) as a function of the higher temperature T.

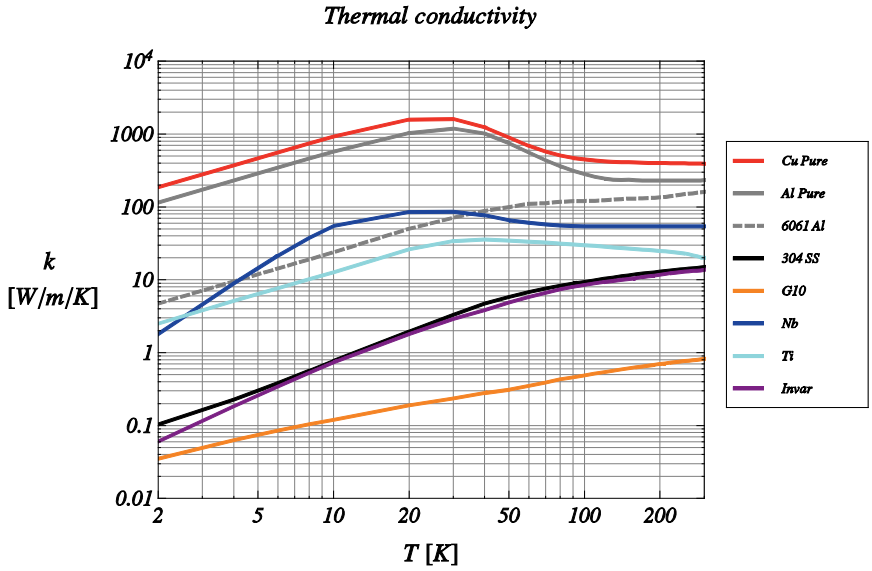


Fig. 5: Thermal conductivity of common materials used for cryogenic applications. Copper, aluminum and niobium curves refer to pure materials, showing the increased conductivity at the phonon peak. This characteristic is not shown by alloys (see e.g. 6061-T6 Al alloy). Data for this figure and Figs. 6 and 8 are from the CRYOCOMP package [12].

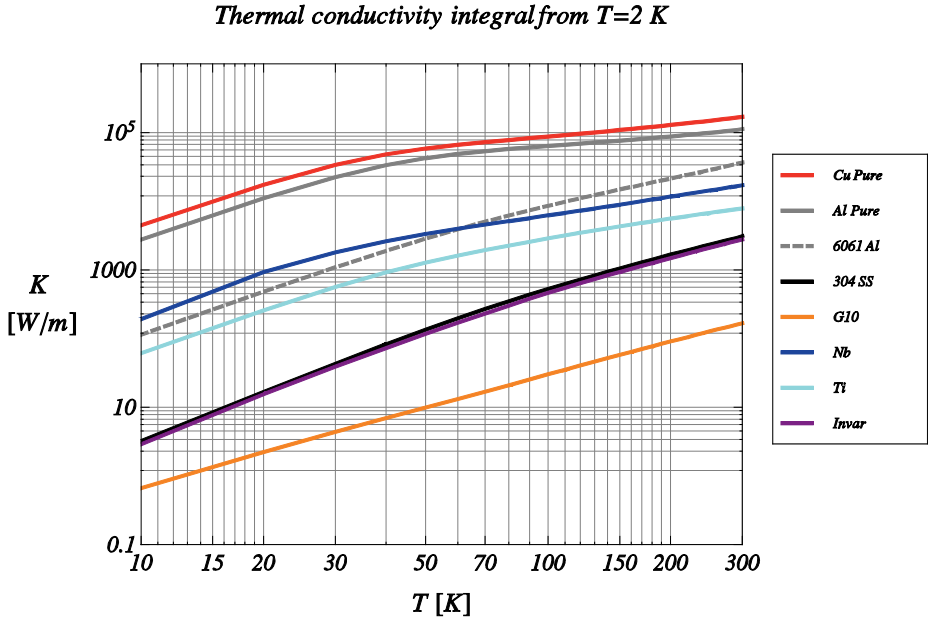


Fig. 6: Thermal conductivity integrals (T_{ref} = 2 K) for the same materials as Fig. 5.

3.1.2 Convection

Macroscopic fluid movement is the mechanism responsible of the heat transfer between the "wet" surfaces exposed to the fluid. The general law that can be used to describe convection heat transfer between a surface S at temperature $T_{surface}$ with a bulk fluid at temperature T_{fluid} is the Newton law of cooling [10, 11]:

$$\dot{Q} = hS(T_{surface} - T_{fluid}) \tag{9}$$

where h is the convection coefficient. In most situations it is impossible to derive an analytical formulation for the convection exchange coefficient, and the description of the heat transfer would require the numerical solution of partial differential equation associated with the fluid motion. Indeed Eq. (9) is a definition of the convection coefficient, which depends on the geometrical configuration of the flow and many fluid properties (its velocity, viscosity, thermal conductivity, specific heat and density). Determination of h is usually performed from experimental empirical correlations between the dimensionless groups typically used in fluid mechanics to perform dimensional analysis. The most important of these dimensionless quantities are as follows (for extensive treatment, refer to Ref. [11].

The Reynolds number (*Re*), which represents the ratio between the inertia and viscous forces, and therefore allows the laminar ($Re < 2000$) and turbulent ($Re > 10^4$) regimes for flow conditions to be defined

$$Re = \frac{\rho v D}{\mu} \tag{10}$$

The Nusselt number (*Nu*), which represents the ratio between the convection and conduction heat exchange mechanisms,

$$Nu = \frac{hD}{k} \tag{11}$$

The Prandtl number (*Pr*), which is a property of the fluid, and represents the ratio between its ability to transport momentum and to transfer heat,

$$Pr = \frac{\mu C_p}{k} \tag{12}$$

In these definitions μ, ρ, C_p, k, v indicate, respectively, the fluid dynamic viscosity, density, specific heat, thermal conductivity and fluid velocity evaluated at the fluid state, and D is the characteristic dimension of the flow.

Empirical correlations, generally developed for non-cryogenic fluids, relate these parameters (in general expressed as $Nu = f(Pr, Re)$) for different geometry and flow configurations, and can be then used to determine the convection coefficient to be used in Eq. (9) to describe convective heat transfer. In general, except for the case of He II, the same correlations developed for non-cryogenic fluids can be used, with the caution to use them in their proper regions of validity (typically the flow regime and the geometrical configuration of the fluid/material interface), and to evaluate the fluid properties at the correct temperature and pressure conditions for the cryogen.

As an example, for turbulent liquid and gas single-phase internal flows the Dittus–Boelter correlation states that

$$Nu = 0.023\, Re^{0.8} Pr^{1/3} \tag{13}$$

So, to properly describe the convective thermal exchange in this case, first the turbulent flow regime needs to be assessed (by the calculation of Re as given by Eq. (10)), then the fluid properties are used to evaluate Pr and the value of Nu is derived from the correlation expressed by Eq. (13). When Nu is determined, the heat convection coefficient can be derived from the definition in Eq. (11) and the fluid properties. A number of correlations similar to Eq. (13) describe other typical geometrical and flow conditions usually found in cryogenic systems [10, 11].

Convection is one of the physical mechanisms by which we are able to extract heat from our devices and route it to the cooling fluids in the cryogenic piping. Analysis of convection exchanges are therefore important to make proper provision for the cooldown and warmup procedure and for the good behaviour of the thermal intercept circuits needed to prevent conduction load to the cold mass. This is also important for the piping that needs to extract the incoming radiation from the cold mass thermal shields (see below). An insufficient heat transfer coefficient in the presence of high heat loads would drive the pipe surface temperatures to higher values to sustain the transfer (by Eq. (9)), leading to increased temperatures for the heat interceptions of the shields.

3.1.3 Radiation

Heat is transported in the form of electromagnetic radiation emitted by surfaces, in the absence of any supporting fluid or medium. The total radiation flow emitted in all directions impinging on a surface S by a body of emissivity ε at the temperature T_{hot} is given by

$$\dot{Q} = \varepsilon S \sigma_{SB} T_{hot}^4 \tag{14}$$

where $\sigma_{SB} = 5.67 \times 10^{-8}$ W m^{-2} K^{-4} is the Stefan–Boltzmann constant.

Therefore, unshielded thermal radiation from the room-temperature environment reaching negligible temperatures would cause a load of approximately 500 W m^{-2}. A surface at the liquid nitrogen temperature would still cause 2 W m^{-2} of radiation load. A suitable strategy to prevent these loads to reach the cold environment is needed.

The flow collected by a surface S of emissivity ε and temperature T_{cold} coming from a black parallel surface at temperature T_{hot} is

$$\dot{Q} = \varepsilon S \sigma_{SB} (T_{hot}^4 - T_{cold}^4) \tag{15}$$

this general formula can then be extended for different geometries and to account for surfaces of different emissivities, and formulas are given in Refs. [10, 11]. For the simple case of two parallel plates of different emissivities, the quantity ε in Eq. (15) can be substituted by the combined emissivity factor $\varepsilon_{hot}\varepsilon_{cold}/[\varepsilon_{hot} + (1 - \varepsilon_{hot})\varepsilon_{cold}]$. Material, temperature and surface finishing have a strong impact on surface emissivity, as can be seen from Table 2 [10, 11].

Table 2: Total emissivity for different temperature for selected material and surface conditions [10, 11]

Material	**Emissivity**		
	***T* = 4.2 K**	***T* = 77 K**	***T* = 300 K**
304 Stainless steel, as fabricated	0.12	0.34	
304 Stainless steel, mechanical polish	0.074	0.12	0.16
Aluminum, electropolished	0.04	0.08	0.15
Aluminum, mechanical polish	0.06	0.10	0.20
Aluminum, 7 μm oxide layer			0.75
Copper, as fabricated	0.062	0.12	
Copper, mechanical polish	0.054	0.07	0.10
Copper, with black paint (80 μm)	0.892	0.91	0.935
Aluminum coating on Mylar (both sides)		0.009	0.025

Radiation effects can therefore be somewhat mitigated by a proper use of material and surface finish condition. A more important measure for the management of radiative load in cryostats is to intercept the thermal flux impinging on the cold surfaces from the room temperature environment with one (or more) thermal screens actively cooled at intermediate temperatures from the two surfaces. Typically in all cryostats for superconducting magnets and cavities a shield make of a good thermal

conducting material (typically Cu or Al) intercepts the ambient radiation flux at temperatures in the range from 40 K to 80 K.

A second very effective measure (often combined with the thermal shielding) to protect the surfaces from radiation load is to wrap them with many "floating" radiation-cooled reflective screens, interposed between the hot and cold surfaces. This is the concept of multilayer insulation (MLI) [8], in which several (typically 10 to 30) foils of reflective aluminium (or aluminized/double aluminized polyester films) are separated by a thermal insulating spacer material (as glass-fibre or polyester or paper foils) and "wrapped" around the cold surfaces to decrease the impinging thermal flux. The packing density of the layers affects performances, but even more important are the installation procedures. In particular, it is important to avoid "holes" leaving a direct line of sight to the cold environment and to prevent thermal short circuits between the layers, which have an impact on the concept of "floating" screens.

With MLI insulation [8, 10] the radiative load can be brought to the following levels:

- 0.5 W/m^2 to 1.5 W/m^2 from the room temperature environment to the intermediate shield temperatures (~40–80 K),
- 0.05 W/m^2 to 0.1 W/m^2 from the thermal shields to negligible temperatures.

To achieve these values proper care need to be taken around the needed penetrations in the temperature shields that need to accommodate pumping lines, support, current leads, RF feed and cabling, avoiding the exposure through holes of a direct line of sight from the "blackbody" ambient temperature environment.

3.2 Putting it all together: the cryostat/cryomodule environment

The cryogenic considerations expressed in Section 1, the heat transfer mechanism described in this section and the further engineering consideration exposed in the following pages play a large role in the design of one important item in a superconducting RF linac: the cavity cryomodule.

3.2.1 Functions of a cryostat/cryomodule

Cryomodules are the *modular building blocks* of all superconducting linear accelerators and need to fulfil the following main functions:

(i) provide mechanical support for the cavities (and possibly focusing elements);

(ii) meet alignment tolerances according to beam dynamics specifications;

(iii) create and maintain (efficiently) the cold environment for the cavity operation.

The linac cryomodules are also an important part of the cryogenic plant for any linac operation, since they are the regions where the major heat loads are located.

A conceptual "cartoon" view for a cryomodule of a superconducting RF linac is illustrated in Fig. 7. The cavities, placed in the central region, are supported to the external enclosure (the vacuum vessel), which is put under vacuum to inhibit gas-driven convective and conductive phenomena. Several penetrations connect the cold mass containing the cavities to the external environment (RF couplers, diagnostic and instrumentation cabling). The cold mass is wrapped by one or more layers of thermal shielding, with the primary role of intercepting the thermal radiation reaching the cold temperatures and the secondary role of convenient "manifolds" to provide thermal intercept to the conduction paths represented by the penetrations. Finally, a number of circuits provide the flow of the coolants needed for maintaining the cold mass and shield temperature levels, with the correct fluid conditions and flow rates to remove the estimated heat deposition (according to Eq. (6)).

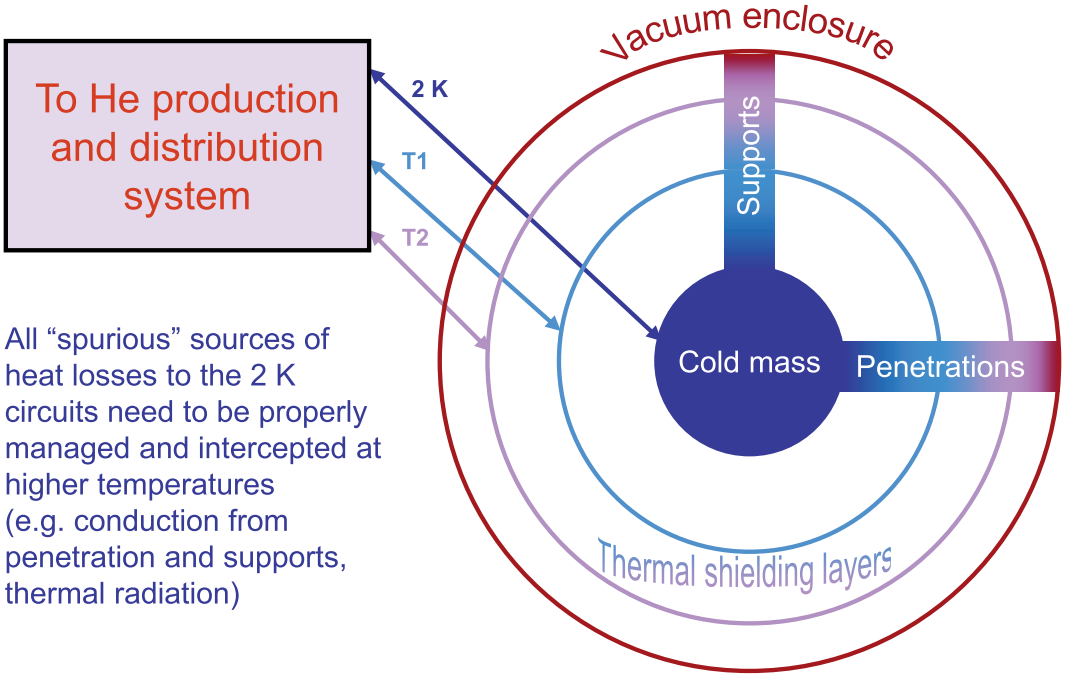


Fig. 7: A "cartoon" view of a SCRF cryomodule

3.2.2 Engineering practices for cryomodule design

Several engineering "practices" are therefore needed to develop a linac module concept, listed here along with their main goals:

- **Thermal design**
  - Minimization of spurious heat loads at the cold temperatures, especially important for large accelerator complexes.
  - Heat removal at various temperature levels, including provisions for thermal shielding and interception of the conduction paths.
  - Capabilities for cooldown and warmup, where the large enthalpy content of the cold mass needs to be carried away or restored in a short time without inducing large thermal gradient (and, thus, structural deformations).
- **Mechanical design**
  - Stable supporting of the cold mass, with minimal thermal losses.
  - Handling of gravity, vacuum and pressure loads.
  - Robustness with respect to thermal stresses induced by thermal gradients occurring during transient conditions and operation.
  - Provisions for implementation of reliable alignment of the sensitive components, and their preservation under differential thermal contractions.
- **"Hydraulics" and piping**
  - Integration of the cryostat cooling circuits in the cryogenic system design.

It is also important to note that the above tasks are not independent design actions that can be optimized independently from each other, since often the optimization strategies in the different

domains would be conflicting. The most promising and stable mechanical support structure can very likely result in huge heat loads at the cold temperature, and therefore a coupled analysis and overall optimization is often required.

The following sections illustrate a few additional issues that need to be addressed in the design of cryogenic components.

3.2.3 Differential contractions

Designers of cryogenic components for accelerators face an additional complication given by the huge variation in the thermal contraction coefficients of different materials. Special care needs to be taken to account for this fact by selecting compatible materials when possible, allowing relative movements and avoiding to mechanical over-constrain the system, to prevent the occurrence of severe thermal stresses during cool-down, potentially capable of causing mechanical failures. Figure 8 shows the total linear contraction from room temperature as a function of the final temperature for selected materials. From this plot one can immediately see that at low temperatures:

(i) The only material with a behaviour similar to Nb is Ti (and this explains why the superconducting cavity helium reservoirs are often manufactured using Ti).

(ii) Stainless steel contracts twice as much as Nb, so relative displacement between the cold cavities and inner cryostat components is always a concern and needs to be handled.

(iii) Aluminum (pure and its alloys) has a thermal contraction 30 % higher than SS. Since the thermal shields are often made by Al or Cu for their good conduction properties, relative movement with respect to the cryostat structural components has to be allowed.

(iv) Special alloys (e.g. Invar) can be manufactured with very low thermal expansion coefficients. These material are important to provide "fixed" temperature-independent points for the positioning of critical components which need to have an interface at warm positions (e.g. in SCRF cavities the RF couplers connected to the vacuum vessel and bringing the RF power to the cold cavities).

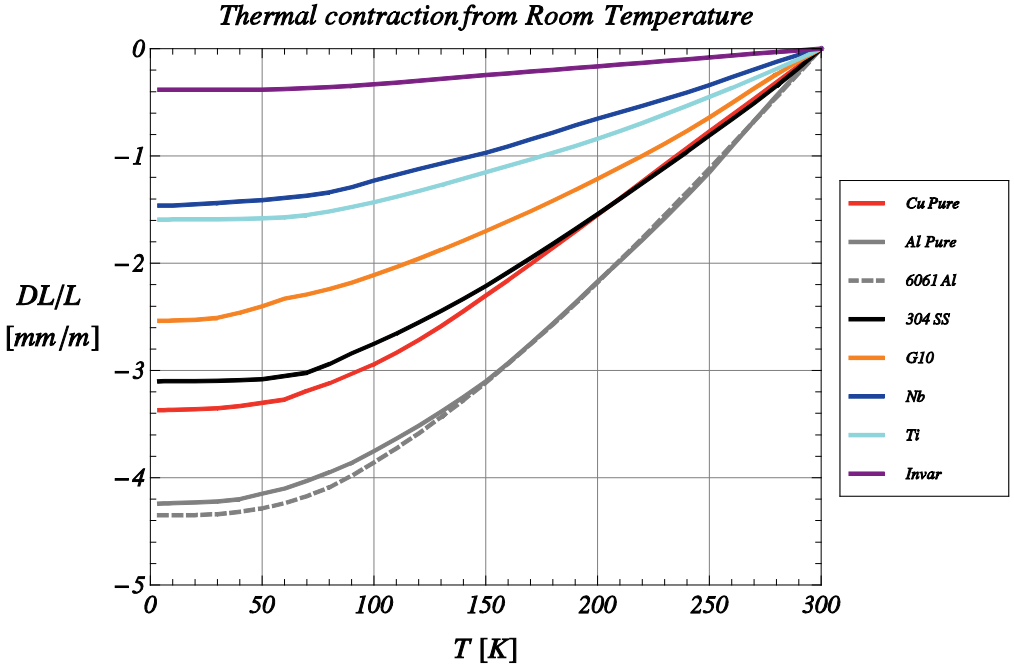


Fig. 8: Total thermal contraction from room temperature as a function of the final temperature level for different structural material for SCRF components

The effect of differential thermal contractions and relative component movement has to be taken fully into account in the design of the cryomodule, the container object for the superconducting cavities, as they have an impact on the final position of the components at the operating temperature, which needs to be compatible with the alignment accuracies required by beam dynamics analysis of the accelerator application.

3.2.4 Provisions for cooldown and warmup transients

When designing cryostats for superconducting devices, provisions have to be made for the necessary transient conditions to achieve operation: cooldown (and warmup). Very often the cryostat needs to provide separate helium lines for cooldown and filling operations. Filling lines from the top of the device can lead to interruptions of the liquid flow due to the outgoing vapours and either need to be placed at the bottom of the cold mass or require a separate line to vent the vapours.

Stratification of helium flow with a thermal gradient can lead to impractical long times for the cryostat warmup to ambient temperature, and electrical heaters or warmup lines to bring warm vapours need to be planned in the design.

3.2.5 Pressure drops

In a pipe of diameter D and length L where a fluid with uniform density ρ and velocity v flows at a mass flow rate $\dot{m}$ in the turbulent regime, neglecting elevation changes, pressure experiences a drop ΔP given by

$$\Delta P = \frac{8}{\pi^2}\frac{\dot{m}^2}{\rho D^5} L f \tag{16}$$

where f is the fluid friction coefficient (depending on geometry, flow conditions and pipe roughness), which can be found tabulated in the literature [8].

Equation (16) is particularly important for the case of subatmospheric operation in He II, for the description of the pressure increase in the cryogenic circuit collecting the 2 K vapours from the cavities. The mass flow of the gas extracting the RF power deposited on the cavities can lead to a pressure increase at the cavity level, thus to a temperature increase (for operation along the saturation curve $T = T(P)$). An insufficient piping sizing would therefore lead to the inability to maintain the correct operating temperatures of the cavities, increasing dissipation according to Eqs. (1) and (3).

4 Case study: TTF/XFEL/ILC modules

One particularly significant example of a state-of-the-art cryostat for 2 K operation is the TTF cryomodule [13] (and its variants). Its design has been conceived in the 1990s for the concept of the TESLA superconducting linear collider and has been used for the construction of the Tesla Test Facility (TTF), now operating as the FLASH free electron laser user facility in DESY, Hamburg. The successful TTF design, with minimal modification, has been later adopted for the accelerator of the European XFEL Project, which will operate a ~2 km linac composed of 100 modules of this type. The concept has also evolved, with a few variations, into the baseline for the International Linear Collider (ILC) project, and adapted to other current or proposed projects (e.g. Cornell ERL, FNAL Project-X).

The module design is illustrated here to "summarize" the interplay of design issues at various system levels for accelerator/cryosystem and cryomodule.

4.1 The TESLA requirements

The TESLA 33 km collider proposal set the main initial requirements for the design of its modular block, as described in the following.

4.1.1 High filling factor

The energy reach of the TESLA 500 GeV to 800 GeV collider forced the maximization, to the maximum extent, of the ratio between the real estate gradient (i.e. total energy gain divided by overall accelerator length) and the cavity gradient performances (i.e. cavity energy gain divided by the nominal RF cavity length).

Thus, the design called for long cryomodules (containing many cavities) and to the proposal to connect them in long cryo-units, separated by short interconnections.

4.1.2 Moderate cost per unit length

Again, the scale of the project imposed a simple functional design, based on proven and reliable technology, readily available in the industrial context.

In particular, the cheapest allowable materials respecting operational load requirements have been selected, and the design effort has been directed to achieve the smallest number of machining steps per component.

To minimize operation costs, very small static losses are requested in the design (where static here means spurious losses to the cold environment in the absence of the dynamic loads due to RF cavity excitation). In particular, to avoid radiative load at 2 K levels, a double thermal shield has been foreseen in the design, with the outer shield operating around 70 K and an inner shield operating around 5 K. The double shields concept also allows a practical way to heat sink all conduction paths at these same intercept temperatures of the shield cooling circuits.

4.1.3 Effective cold mass alignment strategy

The module is designed to preserve, after cooldown, the alignment performed at room temperature. An extensive part of the engineering R&D phase at TTF has been dedicated to the assessment of the alignment reproducibility and stability, developing the necessary diagnostic methods and instrumentation, and performing the necessary experimental validation [14].

4.1.4 Effective and reproducible assembly procedure

The TESLA collider collaboration pushed the technology of bulk Nb superconducting cavities to unprecedented records. One important ingredient towards this achievement was the controlled cavity handling in a class 10 clean room up to the closure of the cavity string end gate valves. Thus, in the TTF assembly scheme the clean room preparation of the cavity string is completely separated from the module assembly. No cryomodule parts enter in the clean environment of the cavity assembly facility, thus allowing contamination of the cavity surfaces to be avoided or unnecessary expensive cleanliness requirements on cryomodule components to be imposed.

A reliable and cost-effective assembly scheme has therefore been implemented in the early stages of the development of the concept, leading to the definition of the necessary assembly toolings in parallel with the module design.

4.2 Consequences and cryomodule concept

The combined request of a high filling factor (to limit machine size) and low static heat losses (to limit operational costs) led to the integration of the cryomodule concept into the design and optimization of the overall cryogenic infrastructure of the collider. In particular, each cold–warm transition along the beam line and each cryogenic distribution box into the module require tunnel space and introduce additional heat losses. Thus, long cryomodules with many cavities (and focussing magnets) were preferred, cryogenically connected to form “cryo-strings” to minimize the necessity of cryogenic

feeds. As a consequence, the helium distribution lines were integrated within the cryomodule environment. The TTF cryomodule contains eight cavities and a magnet package in ~12 m.

The limit of the length of each cryomodule unit is then set by its fabrication aspects (in terms of the need for large assembly tooling and precision machineries required for milling operations), module handling and transport considerations, and the foreseen capability to provide and guarantee the alignment within the required tolerances. Practically, this limit is in the 10 m to 15 m range.

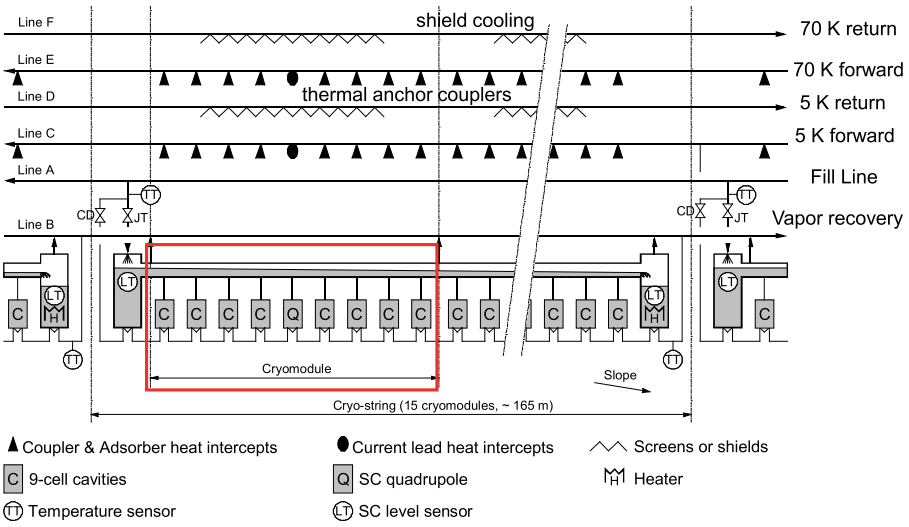


Fig. 9: Cryogenic circuits of the ILC cryo-string (courtesy of T. Peterson). The red box delimits the cavity and quadrupole components within one cryomodule.

The same concept of the TESLA [15] and XFEL cryo-strings briefly described above has been exploited for the cryogenic scheme of the ILC cryo-string [16], shown in Fig. 9. The various cryogenic lines that provide heat interception for the conducting paths (shown with a filled black triangle, on lines C and E) and thermal shields cooling (wavy lines, on lines D and F) are displayed in the upper part of the figure. The lower part of the figure shows that from the cryogenic point of view all cryomodules (one cryomodule unit is highlighted by the red box) are connected into a single continuous line of two-phase He II filling all of the cavities. This two-phase line is fed once per cryo-string by a subcooled pressurized He II line (line A). Vapour coming from the cavities is collected by a single gas recovery line (line B) running along the 12–15 module strings.

In this concept the cryogenic distribution for the whole cryo-string is integrated into the cryomodule, for static losses minimization. As the RF heat loads increase with the number of cavities in the modules and the cryogenic lines within the module serve the dozen or more modules connected in the cryo-string, the size of the cryogenic piping needs to be increased with respect to the case of a single, individually fed module (to guarantee the correct convective heat exchange and to contain the pressure drops). All of the cryogenic lines shown in Fig. 9 are integrated in the mechanical and thermal design of the cryomodule. Piping connections between adjacent cryomodules are welded, to guarantee continuity and avoid potential leaks from flanged connections.

To remove the RF power dissipated along one cryo-string formed by several cryomodules, a large mass flow of He gas is needed and, therefore, according to considerations expressed by Eq. (16), a large diameter is needed to reduce the pressure drop in the He gas return pipe (HeGRP). To combine a mechanical function, the HeGRP was dimensioned with an even larger diameter than strictly required by the handling of the pressure drop, so that it can act as the main structural backbone for the module string. Cavities and quadrupole are thus supported by this large structural backbone, which is stably thermalized at 2 K by the collected vapours.

The HeGRP is then supported by means of three low conductivity suspension composite posts to the vacuum vessel. The central support post is rigidly connected at the module vessel centre, whereas the two end ones are allowed to slide to recover differential contraction.

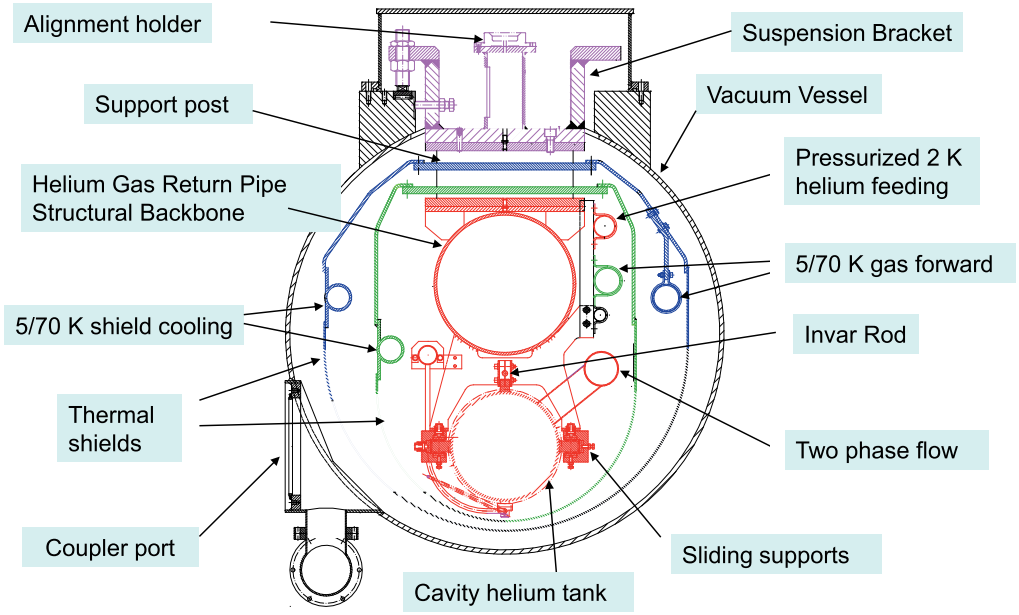


Fig. 10: The cross section of the TTF cryomodules

Figure 10 shows a cross section of the TTF/XFEL/ILC cryomodule concept, illustrating all of the characteristics described before:

(i) The large HeGRP supporting the string of eight cavities and a magnet package is clearly visible above the cavities in the 2 K region (shown in red). Below the HeGRP and above the cavity position the long invar rod provides longitudinal cavity fixing, while the cavity is supported by means of sliding supports [18]. To the right-hand side the two-phase pipe is used to fill the cavities with saturated He II at 2 K. At each module interconnection the He II vapours are brought to the HeGRP. The whole cold mass is supported by the composite G10 support post, and the conduction path is intercepted at the two thermal shield temperatures. The suspension brackets at the top of the support posts slide on linear rollers to accommodate differential contraction between ~12 m HeGRP and the vessel.

(ii) The two thermal shields at 70 K (blue) and 5 K (green), with integrated cooling pipes, protect the cold mass from the thermal radiation of the room-temperature vacuum vessel. Radiation load from the 5 K shield to the 2 K region is reduced to negligible levels. The thermal shields are cooled directly by an extruded aluminium pipe (shown in the left part of the picture), directly welded to the shield parts by means of a stress-relieving welding scheme [17]. The shields also act as thermal intercepts for the support posts and provide cable thermalization and effective coupler thermal intercepts via short thick copper braids.

(iii) The cryogenic lines bringing the cooling fluids forward through the modules downstream along the cryo-string (5 K and 70 K circuit and pressurized He II line) are shown in the top right part of the figure, and are thermally insulated from the 2 K cold mass by low thermal conductivity G10 supports. Heat load is absorbed by the pipe welded to the shields on the return path.

4.3 TTF/TESLA module achievements

In addition to the main design considerations summarized in this section, the design of the TTF/TESLA modules has provided engineered solutions for many of the items and cryostat design tasks outlined in the previous sections.

The cavity support from the HeGRP is realized by means of a sliding scheme [18] that allows the large differential contraction of the 12 m stainless steel HeGRP, completely decoupled from the cavities (which are built with Nb and Ti). The cavities, in fact, need to be fixed longitudinally at the warm position of the coupler ports on the vacuum vessel, to minimize stresses of the fragile ceramic RF window components in the couplers. Thus, a long invar rod (with a small coefficient of expansion) is used to clamp the cavity in this longitudinal position, which varies only slightly during the thermal cycle. Similarly, the support posts can slide on the vacuum vessel to allow the substantial length reduction of the HeGRP after cooldown.

The suspension and sliding cavity mechanisms, combined with a cold mass alignment strategy that relies on the cavity referencing to the HeGRP, has demonstrated the ability to achieve the necessary alignment reproducibility [14].

The thermal design has been verified by measurements of the integral heat loads of the module and fulfils the low static loads goals set by the TESLA Project.

5 Concluding remarks

This short lecture cannot cover all aspects of the cryogenic engineering concepts needed for the development of components for SCRF linacs, and a personal perspective has been offered here.

The main message for this lecture is that for a superconducting RF linac the design of its most critical component, the RF cavity, is only the starting point. A lot of physics considerations and detailed engineering need to be properly addressed to achieve the design of the supporting systems that need to provide its operating conditions. A key element of these systems is the cryomodule, the modular building block of the accelerator.

The overall design choices for the accelerator complex (especially for large machines) have strong constraints and implications on the cryomodule design, driving its conceptual definition, as reviewed in the case study.

Finally, plans for providing adequate mechanisms for cooling to nominal levels, heat removal during operation, control/preservation of alignment, countermeasures to prevent thermal stresses and avoid spurious heat leaks should be developed early in the cryostat and facility designs.

Acknowledgements

The author wants to thank all contributors to previous CAS courses dedicated to superconductivity in particle accelerators, in which the interested reader may find good coverage of topics only briefly mentioned here.

Particular thanks go to Carlo Pagani and many colleagues at INFN-Milano and at DESY-MKS, and to Tom Peterson, FNAL, for providing an endless repository of material to feed my curiosity and my slides.

References

[1] S. Turner, Ed., *CERN Accelerator School: Superconductivity in Particle Accelerators*, CERN 89-04, 1989.

[2] S. Turner, Ed., *CERN Accelerator School: Superconductivity in Particle Accelerators*, CERN 96-03, 1996.

[3] S. Russenchuck and G. Vandoni, Eds., *CERN Accelerator School: Superconductivity and Cryogenics in Particle Accelerators*, CERN 2004-08, 2004.

[4] H. Padamsee, J. Knobloch and T. Hays, *RF Superconductivity for Accelerators*, 2nd edition (New York, Wiley-VCH, 2008).

[5] H. Padamsee, *RF Superconductivity: Volume II: Science, Technology and Applications* (New York, Wiley-VCH, 2009).

[6] J. Schmidt, Cryogenics, S. Turner, Ed., *CERN Accelerator School: Superconductivity in Particle Accelerators*, CERN 89-04, 1989, p. 265.

[7] J.L. Olmes and C. Palmy, Cryogenics, S. Turner, Ed., *CERN Accelerator School: Superconductivity in Particle Accelerators*, CERN 89-04, 1989.

[8] J. Weisend II, Ed., *Handbook of Cryogenic Engineering* (Taylor and Francis, London, 1998).

[9] P. Lebrun and L. Tavian, The technology of superfluid helium, S. Russenchuck and G. Vandoni, Eds., *CERN Accelerator School: Superconductivity and Cryogenics in Particle Accelerators*, CERN 2004-08, 2004, p. 375.

[10] G. Vandoni, Heat transfer, S. Russenchuck and G. Vandoni, Eds., *CERN Accelerator School: Superconductivity and Cryogenics in Particle Accelerators*, CERN 2004-08, 2004, p. 325.

[11] R. Barron, *Cryogenic Heat Transfer* (Taylor and Francis, London, 1999).

[12] CRYOCOMP software package http://www.eckelsengineering.com/ and METALPAK http://www.htess.com.

[13] C. Pagani, *et al*., *Adv. Cryog. Eng.* **43A** (1998) 87–96.

[14] A. Bosotti, *et al*., Analysis of the cold mass displacements at the TTF, Proc. of EPAC2004, Lucerne, Switzerland, p. 1681.

[15] S. Wolff, *et al*., The TESLA cryogenic distribution system, Proc. of the HEACC2001, Tsukuba, Japan (TESLA Report 2001-37).

[16] ILC Reference Design Report: Accelerator, ILC-REPORT-2007-1, August 2007.

[17] C. Pagani, *et al*., *Adv. Cryog. Eng.* **43A** (1998) 307–314.

[18] D. Barni, *et al*., *Adv. Cryog. Eng.* **45A** (2000) 905–911.

Ion sources for high-power hadron accelerators

Daniel C. Faircloth
Rutherford Appleton Laboratory, Chilton, Oxfordshire, UK

Abstract
Ion sources are a critical component of all particle accelerators. They create the initial beam that is accelerated by the rest of the machine. This paper will introduce the many methods of creating a beam for high-power hadron accelerators. A brief introduction to some of the relevant concepts of plasma physics and beam formation is given. The different types of ion source used in accelerators today are examined. Positive ion sources for producing H^+ ions and multiply charged heavy ions are covered. The physical principles involved with negative ion production are outlined and different types of negative ion sources are described. Cutting edge ion source technology and the techniques used to develop sources for the next generation of accelerators are discussed.

1 Introduction

1.1 Ion source basics

An ion is an atom or molecule in which the total number of electrons is not equal to the total number of protons, thus giving it a net positive or negative electrical charge. The name ion (from the Greek ιον, meaning "going") was first suggested by William Whewell in 1834. Michael Faraday used the term to refer to the charged particles that carry current in his electrolysis experiments.

Ion sources consist of two parts: a plasma generator and an extraction system.

The plasma generator must be able to provide enough of the correct ions to the extraction system. There are numerous ways of making plasma: electrical discharges in all of their forms; heating by many different means; using lasers; or even being hit by beams of other particles. The key factor is that the plasma must be stable for long enough to extract a beam for whatever the application requires.

The extraction system must be able produce a beam of the correct shape and divergence angle to the next phase of the accelerator by extracting the correct ions from the plasma and removing any unwanted ions, electrons or neutral particles.

There is a large range of different ion sources out there with many varied applications. Some need to produce ions from tiny samples so they can be analysed and measured. Some need to produce ions for industrial processes such as coating, etching or implanting. Some will go into space to provide thrust for satellites and spaceships. Fusion research demands ion sources that generate huge currents of hundreds of amps with beam cross sections measured in square metres. Radioactive rare isotope ion sources for fundamental research need an entire accelerator facility as one of their key components.

This paper will concentrate on sources for high-power hadron particle accelerators and so will focus on high-current, low-emittance sources. All of the sources in this paper can be configured to produce singly charged positive ions (e.g. H^+, D^+, Li^+), some are better suited to produce multiply charged heavy positive ions (e.g. Pb^{27+}) and some can create significant quantities of negative ions (e.g. H^-). There are lots of hybrid sources that combine features from different types of source. For the sake of simplicity this paper attempts to concentrate only on the archetypal source types.

1.2 History

The first low-pressure discharges were produced by Heinrich Geißler and Julius Plücker in Germany in the mid 1850s. Geißler was a glass blower and inventor commissioned by Plücker to make evacuated glass tubes for his experiments on electric discharges at the University of Bonn. Geißler and Plücker invented a mercury displacement pump that could produce previously unattainable low pressures of less than 100 Pa (1 mbar). The tubes, with electrodes at either end, could be filled with different gases and then evacuated. When a current was passed through them a glow discharge was formed. This allowed Plücker to perform the first experiments in plasma physics. He demonstrated that the plasma could be affected by magnetic fields. The ethereally glowing Geißler tubes (as they became known) were popular mid 19th century entertainment devices, but they also opened the door to the experimentalists that would usher in the atomic age.

In 1869 Johann Hittorf spotted cathode rays in a Geißler tube, but it was William Crookes in early 1870s London that first produced them without the glow discharge. Crookes used a modified Geißler tube and an improved mercury pump made by Hermann Sprengel. He was able to obtain pressures as low as 1 Pa (1×10^{-5} mbar). At these very low pressures the glow discharge stops, leaving a pure cathode ray (electron) source.

Thermionic emission of electrons had first been observed by Fredrick Guthrie in 1868. He noticed that red hot metal balls lost charge. Hittorf and various other German researchers also investigated the phenomenon, but it was Thomas Edison that really developed the idea when trying to work out how to improve his light bulbs in 1880.

British and German researchers continued experimenting with vacuum tubes and in 1886 Eugen Goldstein discovered that a perforated cathode (with holes in it) could also emit a beam: he called them anode rays or canal rays (because they emerge from channels in the cathode) and these rays turned out to be positive ions. A schematic of a canal ray ion source is shown in Fig. 1.

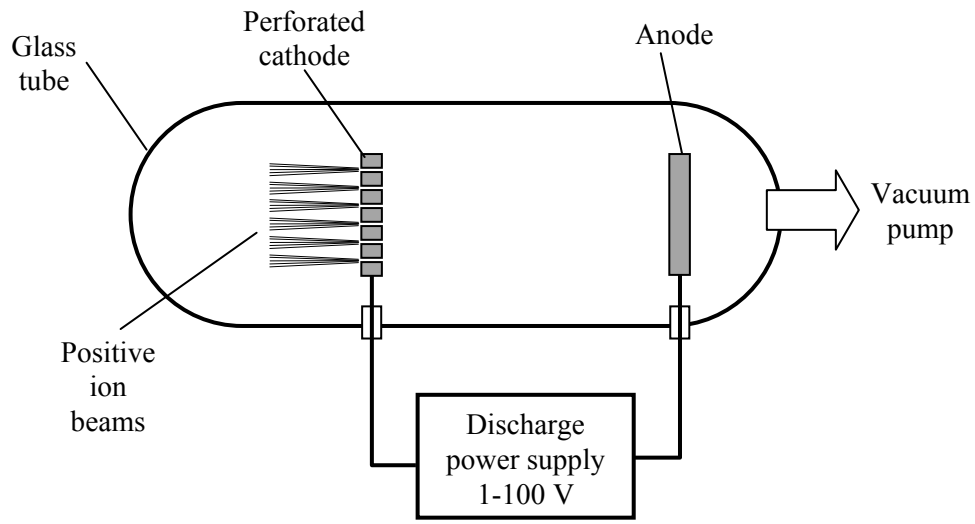


Fig. 1: Schematic of a canal ray tube: the first positive ion source

Eventually, in 1897 J.J. Thomson proved that the cathode rays were actually negatively charged particles which were later named electrons. Experiments with magnetic and electrostatic deflection of the newly produced beams of particles led to new theories on the nature of matter. In the early 20th century, a drive to understand the structure within the atom caused researchers to try and further accelerate beams of particles. Different devices and machines were developed and the ion source as we think of it today was born as a means to produce beams of particles.

2 Plasma

2.1 Introduction

Plasma is the fourth state of matter: if you keep heating a solid, liquid or a gas it will eventually enter the plasma state. Plasma consists of both negatively and positively charged particles in approximately equal proportions, along with un-ionized neutral atoms and molecules. The charged particles consist of positive ions, negative ions and electrons. The particles are always interacting with each other. Collisions can cause ionization or neutralization. Atoms and molecules can be put into excited states and can absorb and emit photons.

The physics of plasmas can be extremely complex. What follows are some of the key concepts relating to ion sources. Plasmas exist in nature wherever the temperature is high enough. Some examples are shown in Fig. 2. The two basic parameters that define a plasma are density and temperature.

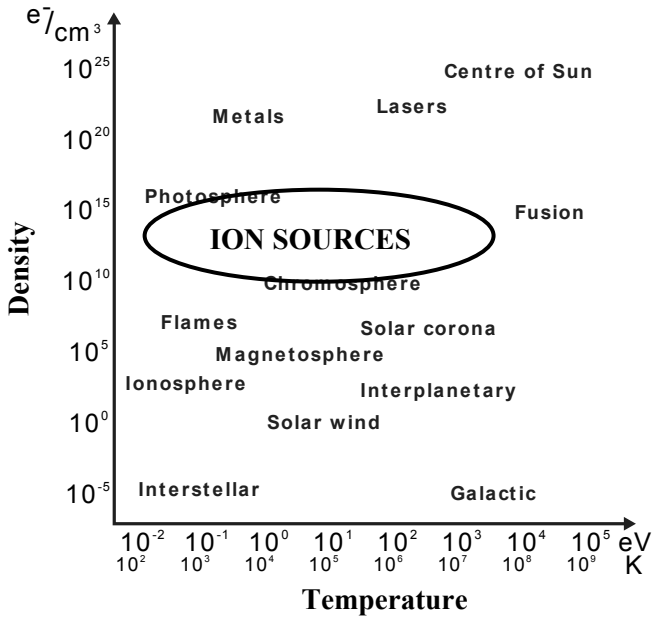


Fig. 2: Different types of plasma

2.2 Basic plasma parameters

2.2.1 Density, n

The most basic parameter is the density of each of the constituents in the plasma. It is usually written as n with subscript to represent the type of particle and is expressed in number of particles per cubic meter. Some older papers give density in particles per cubic centimeter.

n_e = density of electrons
n_i = density of ions
n_n = density of neutrals

2.2.2 Temperature, T

The temperature of the plasma is a measurement of how fast each of the particles is going, otherwise known as the particle kinetic energy. The Boltzmann constant gives 11 600 K = 1 eV. The temperature is usually expressed in electronvolts:

T_e = temperature of electrons
T_i = temperature of ions
T_n = temperature of neutrals

The temperatures of the electrons, ions and neutrals can be different. Different types of plasma produced in different ion sources can have very different ion and electron temperatures. For example, electron cyclotron resonance (ECR) ion sources (see Section 4.4.2), where the plasma is heated by accelerating the electrons, can have $T_e > 1$ keV and $T_i < 1$ eV.

2.2.3 Charge state, q

The charge state of the ions is also important when defining the properties of a plasma. The charge state of an ion indicates how many electrons have been removed from it. Singly charged ions have a charge state $q = +1$. Not all ions will be singly charged, some ions will be multiply ionized (e.g. Pb^{3+} which has a charge state $q = +3$). Some ions will be negatively charged (e.g. H^- which has a charge state $q = -1$). The densities, n, of each charge state can be very different.

Some sources are designed to produce beams of ions with very high charge state, such as Ag^{32+} ions from an electron beam ion source (EBIS; see Section 4.5).

2.3 Ionization energy

The ionization energy is the energy in electronvolts required to remove an electron from an atom. The larger the atom, the easier it is to remove the outermost electron. The second electron is always harder to remove, the third even harder and so on. In most ion sources the energy for ionization comes from electrons impacting on neutral atoms or molecules in electrical discharges. The electrons receive their energy from being accelerated by the field applied to the discharge.

2.4 Temperature distributions

Obviously not all electrons in plasma will have the same temperature and the same is true for the ions of the same species. The numbers T_e, T_i and T_n are merely averages. If the plasma is in thermal equilibrium then the distribution will be Maxwellian and obey Maxwell–Boltzmann statistics.

Using the standard equations the mean speeds of the ions can be calculated to be:

$$\text{velocity of electrons, } \bar{v}_e = 67\sqrt{T_e} \tag{1}$$

$$\text{velocity of ions, } \bar{v}_i = 1.57\sqrt{\frac{T_i}{A}} \tag{2}$$

where A is the ion mass in atomic mass units.

Often the plasma is in a magnetic field. The ions and electrons will spiral around the magnetic field lines and slowly move along them. Hence, the particle velocities (temperatures) will not be the same in all directions. The particle temperatures are defined as $T_{i\parallel}$ is the ion temperature parallel to the magnetic field and $T_{i\perp}$ is the ion temperature perpendicular to the magnetic field.

2.5 Quasi-neutrality

Plasma is generally charge neutral, so all of the charge states of all the ions adds up to the same number as the number of electrons:

$$\sum q_i n_i = n_e \tag{3}$$

2.6 Percentage ionization

The percentage ionization is a measure of how ionized the gas is, i.e. what proportion of the atoms have actually been ionized:

$$\text{percentage ionization} = 100 \times \frac{n_i}{n_i + n_n} \tag{4}$$

When the percentage is above 10% the plasma is said to be highly ionised and the interactions that take place within are dominated by plasma physics. Less than 1% ionisation and interactions with neutrals must be considered.

2.7 Electrical discharges

2.7.1 Overview

The driving field applied to a discharge is often the electrical field; however the magnetic field can also be used in the case of inductively coupled discharges. Inductively coupled discharges require a time varying field and so are more difficult to analyse. It is best to start with an explanation of a DC electric-field-driven discharge between two electrodes. The general current and voltage characteristics of such a discharge are summarized in Fig. 3. The exact shape of the curve depends on the type of gas, pressure, electrode geometry, electrode temperatures, electrode materials, and any magnetic fields present.

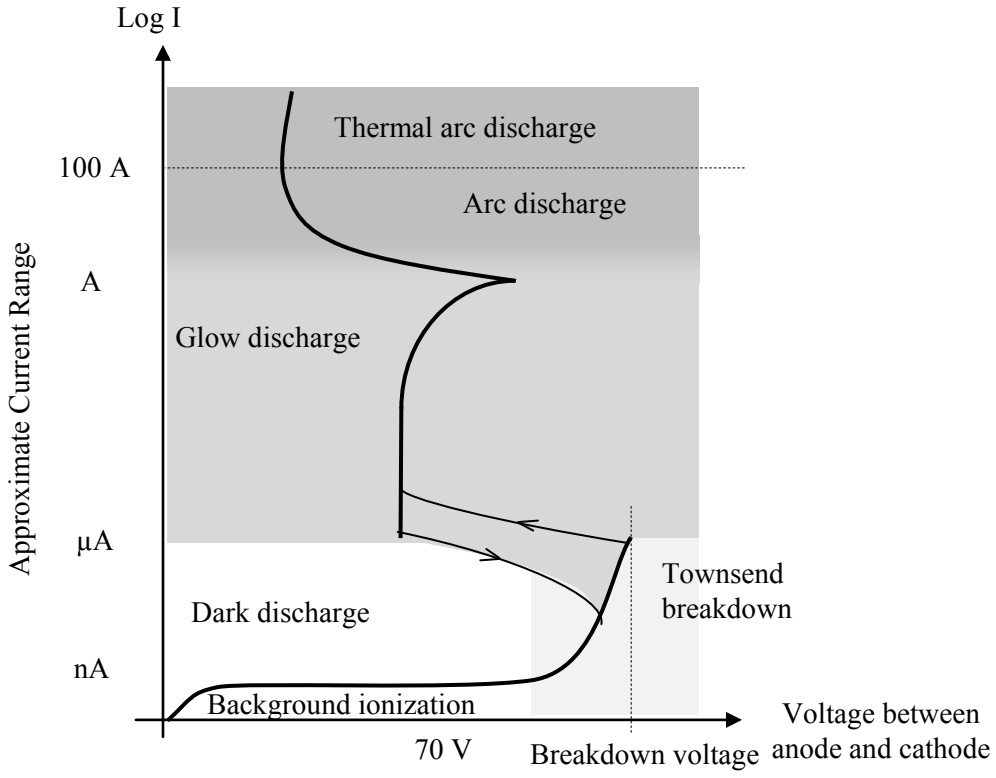


Fig. 3: The current voltage characteristics of a typical electrical discharge

2.7.2 Dark discharge

At low voltages the current between two electrodes is very small, but it slowly increases as the voltage between the electrodes increases as shown in the bottom left corner of Fig. 3. This tiny current comes from ions and electrons produced by background ionization. These are swept out of the gap by the electric field between the electrodes that is created by the applied voltage. There are only enough charge carriers produced by background radiation for a few nanoamps of current, so the current quickly saturates. The voltage can then be increased with no increase in current. The ions and electrons are pulled towards the electrodes through the gas molecules interacting with them as they go.

2.7.3 Townsend breakdown

Eventually the applied electric field is high enough to accelerate the electrons to the ionization energy of the gas. At this point the current rapidly increases as shown in bottom right corner of Fig. 3. The electrons ionize the neutral atoms and molecules, producing more electrons. These additional electrons are accelerated to ionize even more atoms producing even more free electrons in an avalanche breakdown process known as Townsend breakdown. This runaway process means the voltage needed to sustain the discharge drops significantly. The discharge has entered the glow discharge regime.

2.7.4 Glow discharge

The glow discharge is so called because it emits a significant amount of light. Most of the photons that make up this light are produced when atoms that have had their orbital electrons excited by electron bombardment, relax back to their ground states. Photons are produced in any event that needs to release energy, for example when ions recombine with the free electrons and when vibrationally excited molecules relax.

A glow discharge is self-sustaining because positive ions that are accelerated to the cathode impact, producing more electrons in a process called secondary emission. This is why there is a hysteresis in the current versus voltage curve at the glow-to-dark discharge transition.

The current in a glow discharge can be increased with very little increase in discharge voltage. The plasma distributes itself around the cathode surface as the current increases. Eventually the current reaches a point where the cathode surface is completely covered with plasma and the only way to increase the current further is to increase the current density at the cathode. This causes the plasma voltage near the cathode to rise.

2.7.5 Arc discharge

The increased current density leads to cathode heating and eventually the cathode surface reaches a temperature where it starts to thermionically emit electrons and the discharge moves into the arc regime with a negative current versus voltage characteristic.

The current increases until plasma is almost completely ionized (there are few neutral particles left). Eventually the current density in the plasma reaches a point where the ions have the same average velocity as the electrons: they have reached thermal equilibrium. The discharge enters the thermal arc regime where the discharge voltage rises as the current increases.

2.7.6 Importance of the power supply

The power supply used to produce the discharge will have a large effect on the type of discharge produced. The discharge current and voltage obtained will be where the power supply load curve intersects the characteristic shown in Fig. 3. The gradient of the discharge characteristic at the intersection point determines whether the discharge is stable or not. Most ion sources operate in the glow regime.

2.8 Paschen curve

The breakdown voltage of any gas between two flat electrodes depends only on the electron mean free path and the distance between the electrodes. The mean free path is the average distance particles travel before hitting other particles. It is directly related to pressure. Figure 4 shows how the breakdown voltage of hydrogen varies with the product of pressure, *p*, and distance, *d*, between electrodes. This was first stated in 1889 by Friedrich Paschen [1].

At very low pressures, the mean free path between collisions is longer than the distance between the electrodes. So although the electrons can be accelerated to ionising energies, they are unlikely to hit anything other than the anode. This means that the breakdown voltage is very high at very low pressures.

At very high pressures the mean free path is very short. This means that the electrons never have enough time to be accelerated before hitting another particle. This means at the breakdown voltage is high at very high pressures.

Between these two extremes is a minimum whose position depends on the type of gas and the electrode material. This "Paschen minimum" leads to a counterintuitive phenomenon: operating just below this minimum, electrodes further apart will have a lower breakdown voltage than those closer together. This is because in a longer gap there is more space for the electron avalanches to develop.

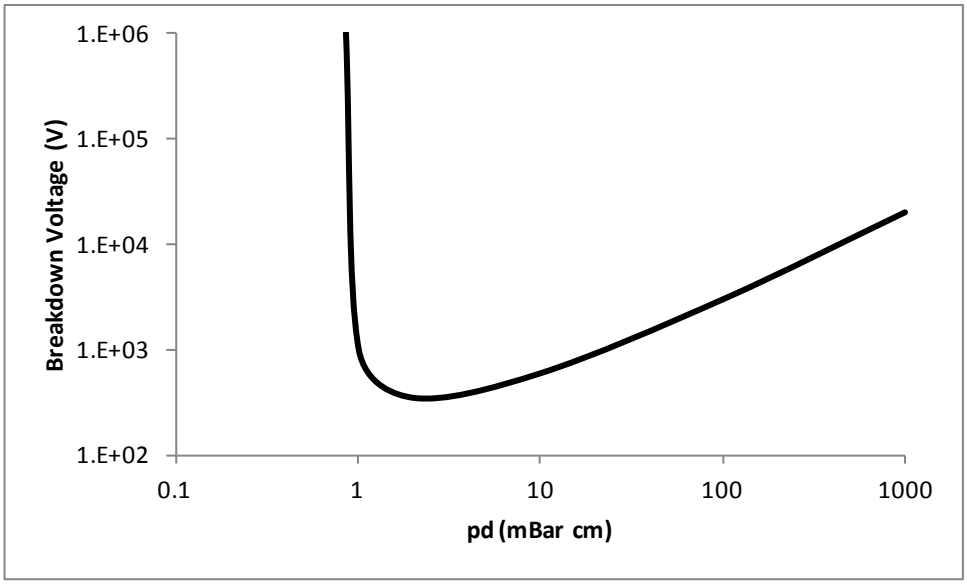


Fig. 4: The Paschen curve for hydrogen

2.9 Collisions

Collisions between particles in a plasma are fundamentally different from collisions in a neutral gas. The ions in a plasma interact by the Coulomb force: they can be attracted or repelled from a great distance. As an ion moves in a plasma its direction is gradually changed as it passes the electric fields of neighbouring particles, whereas in a neutral gas the particles only interact when they get so close to each other that they literally bounce off each other's outer electron orbitals. In a neutral gas the average distance the particles travel in a straight line before bouncing off another particle is referred to as the mean free path. In a plasma, the mean free path concept does not work because the ions and electrons are always interacting with each other by their electric fields. Instead a concept called "relaxation time" is invoked: this is the time it takes for an ion to change direction by 90°. The relaxation time τ_θ can also be described as "the average 90° deflection time". In a plasma there are different relaxation times between each of the different particle species.

2.10 Work function

In any solid metal, there are one or two electrons per atom that are free to move from atom to atom. This is sometimes collectively referred to as a "sea of electrons". Their velocities follow a statistical distribution, rather than being uniform. Occasionally an electron will have enough velocity to exit the metal without being pulled back in. The minimum amount of energy needed for an electron to leave a surface is called the work function. Specifically the work function is the energy needed to move an electron from the Fermi level into vacuum. The work function is characteristic of the material and for most metals is of the order of several electronvolts.

2.11 Thermionic emission

Thermionic emission is the heat-induced flow of charge carriers from a surface or over a potential-energy barrier. This occurs because the thermal energy given to the carrier overcomes the work function of the metal. Thermionic currents can be increased by decreasing the work function. This often-desired goal can be achieved by applying various oxide coatings to the wire.

In 1901 Owen Richardson found that the current from a heated wire varied exponentially with temperature. He later proposed this equation:

$$J = A_G T^2 e^{\frac{-W}{kT}} \tag{5}$$

where J is the electron current density on the surface of the cathode, W is the cathode work function and T is the temperature of the cathode.

Here A_G is given by

$$A_G = \lambda_R A_0 \tag{6}$$

where λ_R is a material-specific correction factor that is typically of order 0.5 and A_0 is a universal constant given by

$$A_0 = \frac{4\pi m k^2 e}{h^3} = 1.20173 \times 10^6 \text{ Am}^{-2}\text{K}^{-2} \tag{7}$$

where m and e are the mass and charge of an electron and h is Planck's constant.

2.12 Magnetic confinement

Charged particles will rotate around magnetic field lines: this means that they tend to travel along magnetic field lines, by spiralling along them. This effect can be exploited to confine plasma in an ion source. A dipole field will confine particles in the direction of the magnetic field. This can be used to confine electrons between two parallel cathodes, as employed in the Penning ion source (Section 5.3.3). A solenoid field will keep charged particles confined axially. Solenoidal fields are used in duoplasmatrons (Section 4.3.2), microwave ion sources (Section 4.4), EBISs (Section 4.5) and vacuum arc ion sources (Section 4.7).

A multicusp field is composed of alternating north and south poles (see Section 5.4). This arrangement is used around the edge of plasma chamber to confine both electrons and ions and prevents them from hitting the walls of the chamber. A specific type of multicusp field (the hexapole field) is used in to increase the confinement time in ECR ion sources (Section 4.4.2).

2.13 Debye length

Named after the Dutch scientist Peter Debye, the Debye length, λ_D, is the distance over which the free electrons redistribute themselves to screen out electric fields in plasma. This screening process occurs because the light mobile electrons are repelled from each other whilst being pulled by neighbouring

heavy low-mobility positive ions, thus the electrons will always distribute themselves between the ions. Their electric fields counteract the fields of the ions creating a screening effect. The Debye length not only limits the influential range that particles' electric fields have on each other but it also limits how far electric fields produced by voltages applied to electrodes can penetrate into the plasma. The Debye length effect is what makes the plasma quasi-neutral over long distances.

The higher the electron density the more effective the screening, thus the shorter this screening (Debye) length will be.

The Debye length is given by

$$\lambda_D = \sqrt{\frac{\epsilon_0 k T_e}{n_e {q_e}^2}} \tag{8}$$

Where:

λ_D is the Debye length,

ϵ_0 is the permittivity of free space,

k is the Boltzmann constant,

q_e is the charge of an electron,

T_e is the temperatures of the electrons

n_e is the density of electrons

λ_D is of the order 0.1 – 1 mm for ion source plasmas.

2.14 Plasma sheath

The screening effect of the plasma creates a phenomenon called the plasma sheath around the cathode electrode. The plasma sheath is also called the Debye sheath. The sheath has a greater density of positive ions, and hence an overall excess positive charge. It balances an opposite negative charge on the cathode with which it is in contact. The plasma sheath is several Debye lengths thick.

The quasi-uniform plasma potential is closest to the anode voltage and the largest potential drop in a plasma is across the plasma sheath near the cathode.

A related phenomenon is the double sheath. This occurs when a current flows in the plasma.

2.15 Particle feed methods

A supply of material to be ionized must be provided to the plasma. If the material is a gas it can be introduced via a needle valve or mass flow controller. If the source is pulsed the gas is usually also pulsed to help maintain low pressures in the source, which is usually achieved with a fast piezo-electric valve. Some gasses are very corrosive so compounds of the element are used instead.

Some materials can be heated in ovens (e.g. caesium, see Section 5.2.6). Other solid materials with low vapour pressure are more suited to ionization by a laser (Section 4.6) or in an arc discharge (Section 4.7). Some sources (such as the EBIS, see Section 4.5) are often fed by another ion source.

It is important to prevent unionized material and excess ions entering the next stage of the accelerator, this is achieved by having a high pumping speed and vessel constrictions with baffles and cold traps.

3 Extraction

3.1 Introduction

The purpose of the extraction system is to produce a beam from the plasma generator and deliver it to the next acceleration stage. The basics of extraction are very simple: apply a high voltage between an ion emitting surface and an extraction electrode with a hole in it. The extraction electrode can also be called the acceleration electrode or the ground electrode (if the plasma is produced in a discharge on a high-voltage platform).

3.2 Meniscus emitting surface

In plasma sources the ion emitting surface is the edge of the plasma itself. At the extraction region the plasma is bounded by an electrode with a hole in it. This electrode is variously called the outlet electrode, aperture electrode or plasma electrode and it is often at the same potential as the plasma anode. The edge of the plasma sits across this hole and is called the plasma meniscus. It is the boundary layer between the discharge and the beam. The shape of the meniscus depends on the local electric field and the local plasma densities. Figure 5 shows how the plasma meniscus can change from being convex to concave as the plasma density decreases. The trajectories of the particles depend on the meniscus shape, so it is important to run the ion source with operating conditions that provide an optimum meniscus shape. This is called the "matched case" and is found by varying the plasma density and extraction potential until the beam is well transported. It is important to point out that the diagrams in Fig. 5 do not include space charge effects that cause the beam to diverge (see Section 3.7).

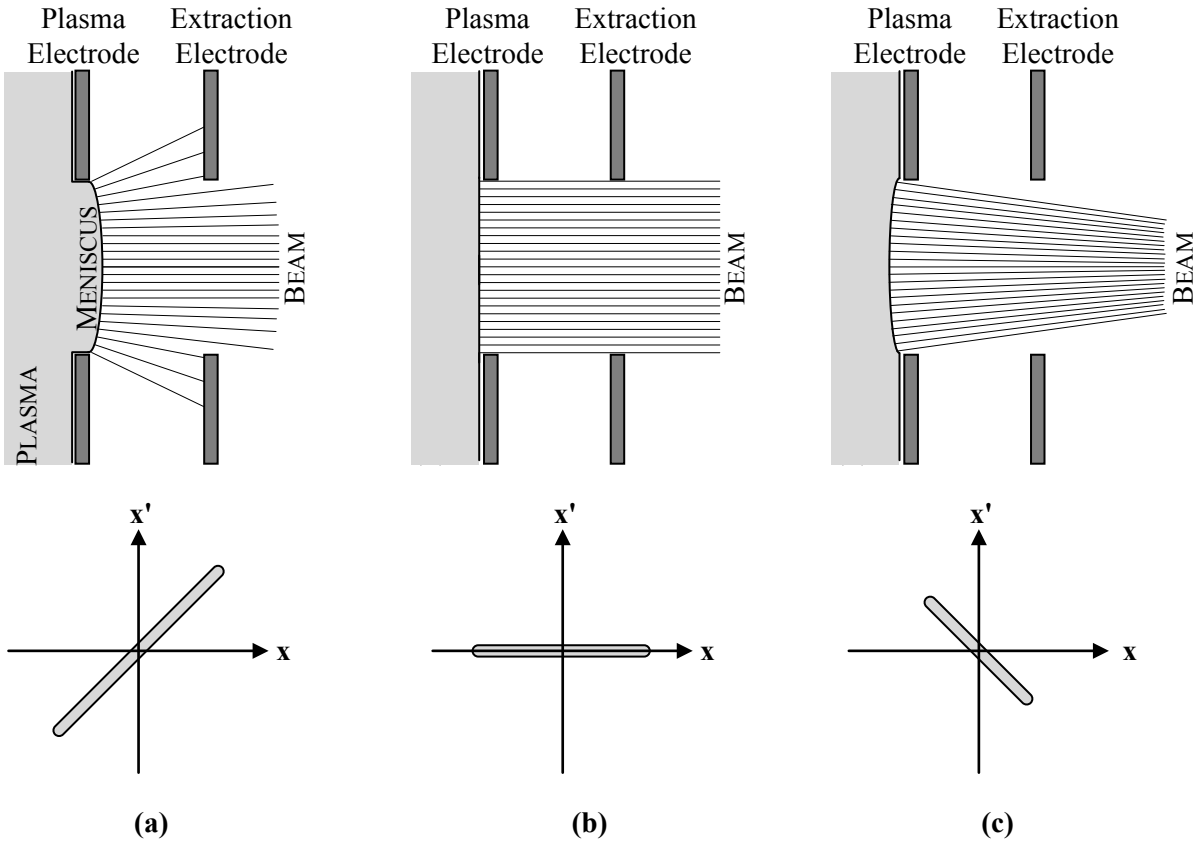


Fig. 5: Three different plasma meniscus shapes and their corresponding emittance phases space plots: (a) convex, high plasma density; (b) flat, medium plasma density; (c) concave, low plasma density

3.3 Solid emitting surface

In surface converter sources the ions of interest are actually produced on a solid surface inside a plasma chamber (see Section 5.5). This has a great advantage over meniscus emission in that the exact shape of the surface can be precisely defined. Solid emission surfaces are concave with a radius of curvature approximately centred on the exit aperture in the plasma chamber. This causes the ions produced on the emission surface to be focused at the exit aperture, resulting in a high-quality, low-divergence beam.

3.4 Emittance

For high-power particle accelerators it is essential that the beam produced by the ion source has a low divergence angle. This allows the beam to be transported and accelerated easily by the rest of the machine without losing any beam.

In particle accelerators a way of specifying the divergence of a beam is emittance. Emittance is a measurement of how large a beam is and how much it is diverging. It is measured in millimetre-milliradians and is the product of beam size and divergence angle. Often emittance is normalized to beam energy because a beam that has had its longitudinal velocity increased by acceleration will still have the same transverse velocity, thus its divergence angle will be reduced. Emittance is normalized to allow emittances to be compared at different energies in different accelerators.

For any beam, two emittances are given: horizontal and vertical. These can be just single values or complete phase space diagrams. Phase space diagrams are plots of divergence angle versus transverse position. Figure 5 gives examples of phase space plots for divergent, parallel and convergent beams. If the horizontal axis of the emittance phase space diagram is x then the vertical axis is usually given as x' (where x' is $\frac{dx}{dz}$ or $\frac{v_x}{v_z}$), this is because the ratio of transverse and longitudinal velocities of the particles defines the divergence angle. The units of x' are radians and are usually expressed in milliradians.

The emittance of a beam is the area enclosed by its phase space plot divided by π. For real beams this statement needs clarification. Real beams have halos: outlying particles that have much larger divergence angles and positions than the core of the beam. They are created when some particles at the edge of the beam experience fringe fields or other non-uniformarities that cause them to separate further from the core of the beam. The halo particles are in the minority; most of the particles are in the dense core of the beam. If these outlying particles are included in the total phase space area calculation the beam will have a huge emittance. To get round this emittances are either quoted as the 95% emittance or root mean square (r.m.s.) emittance. The 95% emittance is the area that encloses 95% of all of the particles. The r.m.s. emittance is calculated from the r.m.s. values of all of the positions and angle measurements. Both methods give a realistic measurement of the overall beam divergence without being unfairly enhanced by the halo particles.

The unit of emittance is actually distance. This makes sense because the dimensions of phase space are millimetres and a dimensionless angle, hence areas in phase space have units of mm. Ion source emittances, however, are very often expressed in units of π·mm·mrad. This is because the particle distributions in phase space plots often have an ellipse drawn round them to define the beam boundary. The area of an ellipse is π multiplied by the product of the length of its two semi-axes. Later, in the rest of the accelerator, higher-energy beams usually do tend to have elliptical phase space distributions. Close to the ion source large aberrations still exist and the beam shape in phase space is often far from elliptical, so a r.m.s integration is the best method to calculate the emittance. It is counterintuitive that r.m.s. ion source emittances are expressed in π·mm·mrad (which is indicative of an ellipse calculation). Units of mm or π·mm·mrad do not change the emittance value quoted and can be used interchangeably. Care must be taken, however, when comparing emittances from older papers that use mm·mrad, in this case the emittance is actually π times larger than if quoted in mm.

3.5 Energy spread

In a beam not all of the particles have the same energy. The energy distribution of the particles is a measurement of the range of different particle velocities in the beam. It is effectively the longitudinal emittance of the beam measured in electronvolts. It is more commonly defined as the full-width–half-maximum of the energy distribution in electronvolts. It is sometimes also referred to as the momentum spread. In emittance phase space plots a proportion of the "thickness" of the phase space distribution is caused by energy spread.

The energy spread causes the beam to spread out in time, so this limits the minimum pulse length achievable. The origin and size of the energy spread is different for different types of source. It can be caused by variations in potentials or temperatures on the plasma production surface, oscillations in the plasma or unstable extraction voltages. The energy spread can range from less than 1 eV to as much as 100 eV.

Energy spread is important because it will produce transverse emittance growth as the beam passes though magnets and accelerating gaps. The beam emittance can be transferred between longitudinal and vertical directions and vice versa.

3.6 Brightness

The brightness of a beam is another key beam parameter. It is the beam current, I, divided by the emittances:

$$B = \frac{I}{\varepsilon_x \varepsilon_y} \tag{9}$$

Unfortunately there are several ways to define brightness: they all have the same basic form as Eq. (9) but they have each have different scale factors based on multiples of π and 2. The reader should take caution when comparing brightness from different authors. Also sometimes the emittances are normalized to energy, which gives an emittance-normalized brightness.

3.7 Space charge

Space charge effects are critical in ion source design. For high-brightness, low-energy beams electrostatic forces are a key factor. The beam will blow-up under its own space charge so it is critical to get the beam energy up to at least 10 keV as fast as possible to minimize the effect which is worse at low energies.

A phenomenon known as "space charge compensation" or "space charge neutralization" is essential for high current ion sources. The pressure in the vacuum vessel directly after extraction will be higher than in the rest of the accelerator because of gas loading from the ion source itself. The beam ionizes the background gas as it passes through it. If the beam is positive it repels the positive background ions and draws in the negative ions and electrons. If the beam is negative it repels the negative background ions and electrons and draws in the positive ions. The effect is to neutralize the beam, reducing its space charge and reducing the beam blow-up. Beams can be almost 100% space charge compensated, meaning they see almost no beam blow-up.

Space charge compensation is a complex dynamic process, for pulsed beams it can take about 100 μs to build up the compensation particles so the start of a pulsed beam will have a transient change in emittance. For very long beam pulses (> 1 ms) the beam can actually lose its compensating particles by diffusion a process known as decompensation.

Space charge compensation is not possible in accelerating gaps because any compensating particles produced are swept out of the gap by the field. For the short time compensating particles remain in the gap they are not effective at compensating the beam because they are moving too fast.

3.8 Child–Langmuir law

There is an absolute limit to the current density that can be extracted from a plasma. There comes a point where the space charge of the beam being extracted actually cancels out the extraction field, making it impossible to extract a higher current density. The current density where this happens can be calculated from the Child–Langmuir equation:

$$j = \frac{\frac{4}{9}\epsilon_0\sqrt{\frac{2q_i}{m_i}}V^{\frac{3}{2}}}{d^2} \tag{10}$$

where j is the current density in A m^{-2}, q_i is the ion charge in coulombs, d is the extraction gap in metres, m_i is the ion mass in kg and V is extraction voltage in V.

If more useful units are used, Eq. (10) becomes

$$j = \frac{1.72\sqrt{\frac{q}{A}}V^{\frac{3}{2}}}{d^2} \tag{11}$$

where q is the ion charge state, A is the ion mass in atomic mass units, d is the extraction gap width in cm, V is the extraction voltage in kV and j the current density is now in mA cm^{-2}.

These equations are true for space charge limited conditions, i.e. where the plasma generator has plenty of ions to give, but space charge limits the current. If the plasma cannot give any more ions then the source is no longer space charge limited and the current versus voltage relationship shown in Eq. (11) no longer follows.

3.9 Perveance

The perveance, P, of an ion source is a measurement of how space charge limited the source is. It is defined as

$$P = \frac{I}{V^{\frac{3}{2}}} \tag{12}$$

where I is the beam current.

It is the constant of proportionality in Eq. (12), it should be constant as the extraction voltage is increased. The voltage where P starts to decrease is an indication that the plasma can no longer supply enough particles to the extractor. The word perveance comes from the Latin “pervenio” meaning to attain. Perveance is also called “puissance” in some texts; this is actually a better word as it is French for strength or ability, and perveance refers to the strength or ability of the plasma to deliver ions.

3.10 Pierce extraction

The shape of the electric field in the extraction gap will shape the beam as it is extracted. The Pierce electrode geometry is an attempt to produce an absolutely parallel beam. The idea is to produce an extraction field that has a zero transverse value at the edge of the beam, thus not having any focusing effect on the beam. The standard Pierce geometry consists of a plasma electrode at an angle of 67.5° to the beam axis. The extraction electrode is curved along an equipotential line to the solution of the equation that gives zero traverse field at the beam edge.

In reality, a completely parallel beam is impossible.

3.11 Suppressor electrode

In accelerating gaps for positive ions, the electrons will be accelerated in the opposite direction and into the ion source. This is not desirable so an electron suppressor electrode is often added just before

the ground electrode. The suppressor electrode is biased slightly more negative than the ground electrode. Any electrons heading into the acceleration gap from the ground electrode side will be reflected back as shown in Fig. 6.

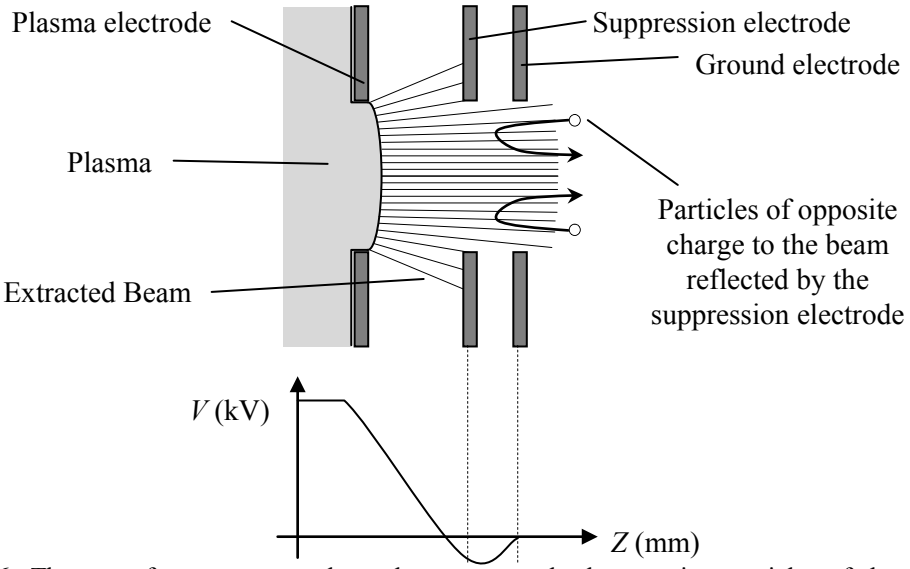


Fig. 6: The use of a suppressor electrode to prevent back-streaming particles of the opposite charge from entering the ion source

In negative ion sources, protons will be accelerated back into the source instead of electrons so the suppressor is biased with a positive voltage. Back-streaming particles can damage the source by sputtering so it is important to suppress them.

3.12 Negative ion extraction

One of the main challenges with negative ion source design is how to deal with the co-extracted electrons. Extracting electrons with the negative ions is obviously unavoidable because they both have the same charge. The ion source engineer must first try to minimize the amount of co-extracted electrons, then find a way to separate and dump the unwanted electrons from the negative ion beam. In some cases the electron current can be 1000 times greater than the negative ion current itself.

3.13 Low-energy beam transport

It could be argued that the ion source extraction system should include the low-energy beam transport (LEBT) system as well. Beam halo and emittance effects mean that the current measured directly after extraction is not a true measure of the beam current that can be transported to the next stage of the accelerator. Often there is significant collimation of the beam on the way through the LEBT and a large proportion of beam current can be lost.

The whole ion source usually sits on a high-voltage platform. The beam is accelerated to ground (this is why the last electrode of the extraction system in Fig. 6 is labelled as the ground electrode), then the beam enters the LEBT. LEBTs can be magnetic or electrostatic or a combination of both. It is common to use between one and four focusing elements. These can be electrostatic Einzel lenses or electromagnetic solenoids and quadrupoles. In ion sources that use caesium vapour it is better to use an electromagnetic LEBT to prevent sparking.

4 Positive ion sources

4.1 Introduction

Researchers had been experimenting with beams of positive ions (or canal rays as they called them) since 1886 when Eugen Goldstein discovered that they were emitted from holes in the cathode. The problem with canal ray sources was that the beam energy could not easily be varied and they had a huge energy spread- as large as the discharge voltage.

4.2 Electron bombardment sources

The first real positive ion source was developed by Arthur Dempster at the University of Chicago in 1916 [2]. The basic design is shown in Fig. 7. It is the first source to introduce an extraction electrode.

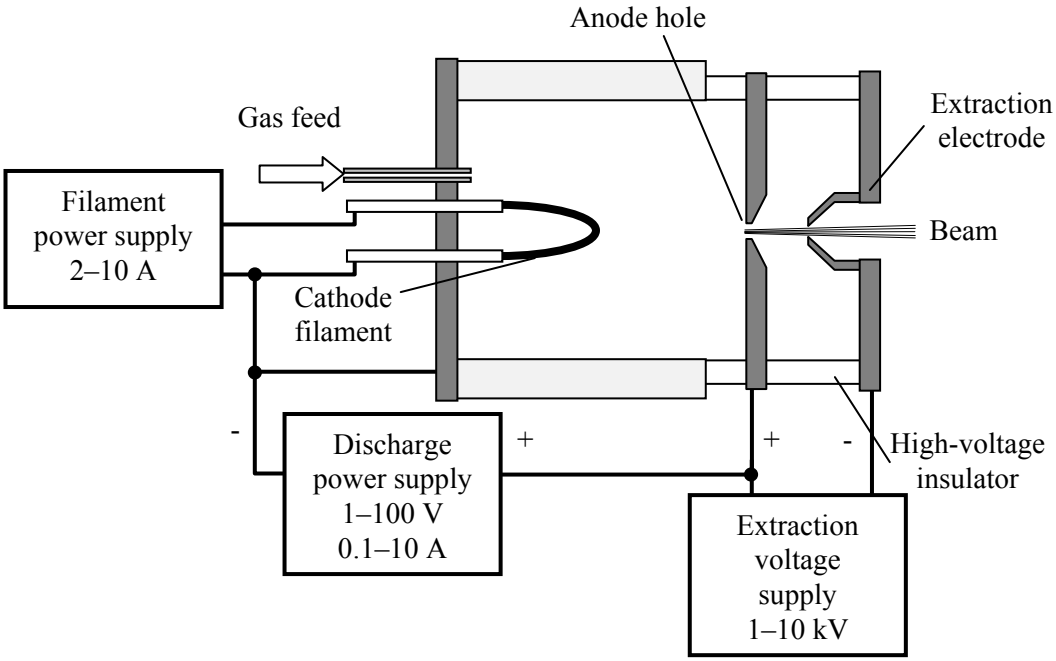


Fig. 7: Schematic of an electron bombardment source

The cathode is heated by passing a current so that it thermionically emits electrons. The electrons are accelerated by the discharge power supply voltage. As long the discharge voltage is higher than the ionization energy of the gas fed into the source, the electrons will be able to ionize the gas by impact ionization. The anode has a small hole in it, opposite which there is an extraction electrode. A negative high voltage is applied to the extraction electrode. The positive ions produced near the anode hole are extracted from the source. Beam currents of about 1 mA can be produced.

Electron bombardment sources are very cheap and easy to produce, they can be used to produce positive beams from almost every element, but they cannot generate beam currents of the magnitude required for high-power accelerators.

4.3 Plasmatrons

4.3.1 Introduction

The plasmatron was first developed by the prolific aristocratic German inventor Manfred von Ardenne in the late 1940s. It is a development of the electron bombardment source. To increase the beam current, a conical shaped intermediate electrode is positioned between a heated filament cathode and

an anode as shown in Fig. 8. The purpose of the conical intermediate electrode is to "funnel" the plasma down to a higher-density region near the anode extraction hole. A plasma double sheath forms on the conical intermediate electrode. The higher plasma density near the extraction region allows more ions to be extracted.

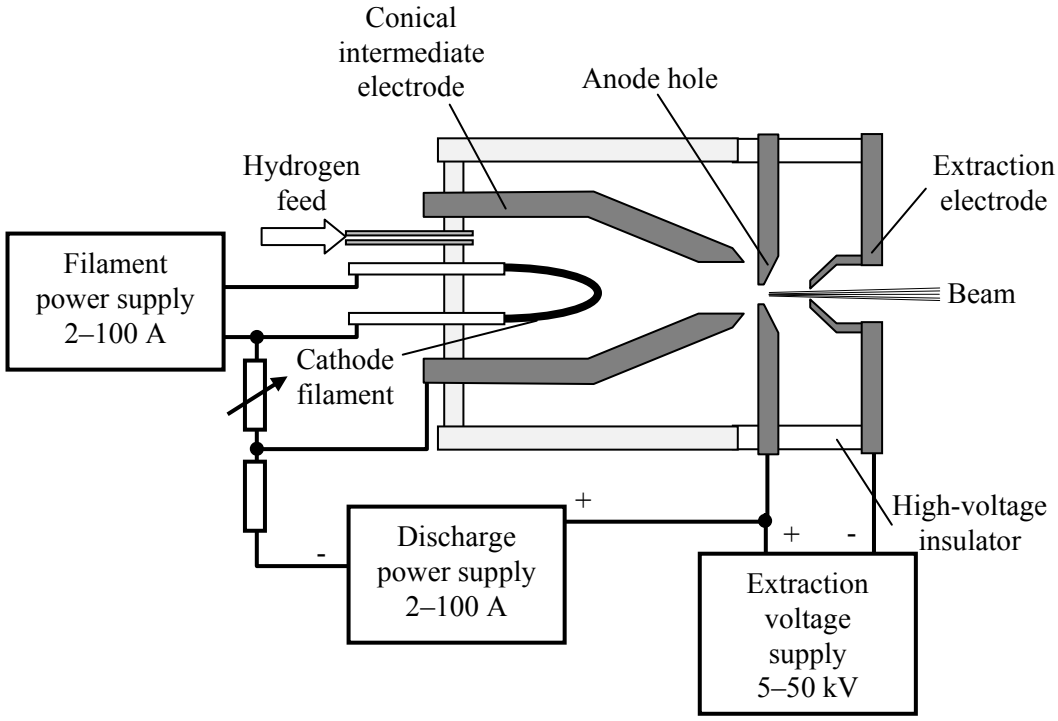


Fig. 8: A schematic of a plasmatron source

4.3.2 Duoplasmatron

Von Ardenne continued to develop the plasmatron and in 1956 he invented the Duoplasmatron, shown in Fig. 9. It is effectively the same as a plasmatron but the conical intermediate electrode is made of soft iron. The plasma chamber is positioned inside a solenoid. The conical soft iron intermediate electrode squeezes the axial magnetic field lines and concentrates them just in front of the anode. The squeezing of the magnetic field lines and the funnelling effect of the cone create a very high plasma density just in front of the extraction hole. This greatly increases the ion density and allows very high positive ion currents of up to 1.5 A to be extracted. The ions streaming through the anode hole are too dense to allow the extraction of ion beams with uniform distribution and low emittance so the plasma is allowed to expand into an expansion cup before being extracted. To prevent back-streaming electrons a suppressor electrode is used. The name duoplasmatron comes from the two (duo) different plasma densities that exist in the source.

The duoplasmatron is probably one of the most common types of positive ion source because it makes very high beam currents, is cheap and easy to maintain and works with a wide range of gases. Lifetimes are limited to a few weeks at high currents and duty factors or with heavy ions because of filament sputtering. Filaments can be easily replaced. At lower currents lifetimes can be much longer.

CERN have used a duoplasmatron on LINAC2 for over 30 years which now ultimately fills the Large Hadron Collider (LHC). It reliably produces 300 mA beams of protons in pulses up to 150 μs long at 1 Hz with lifetimes stretching to several months.

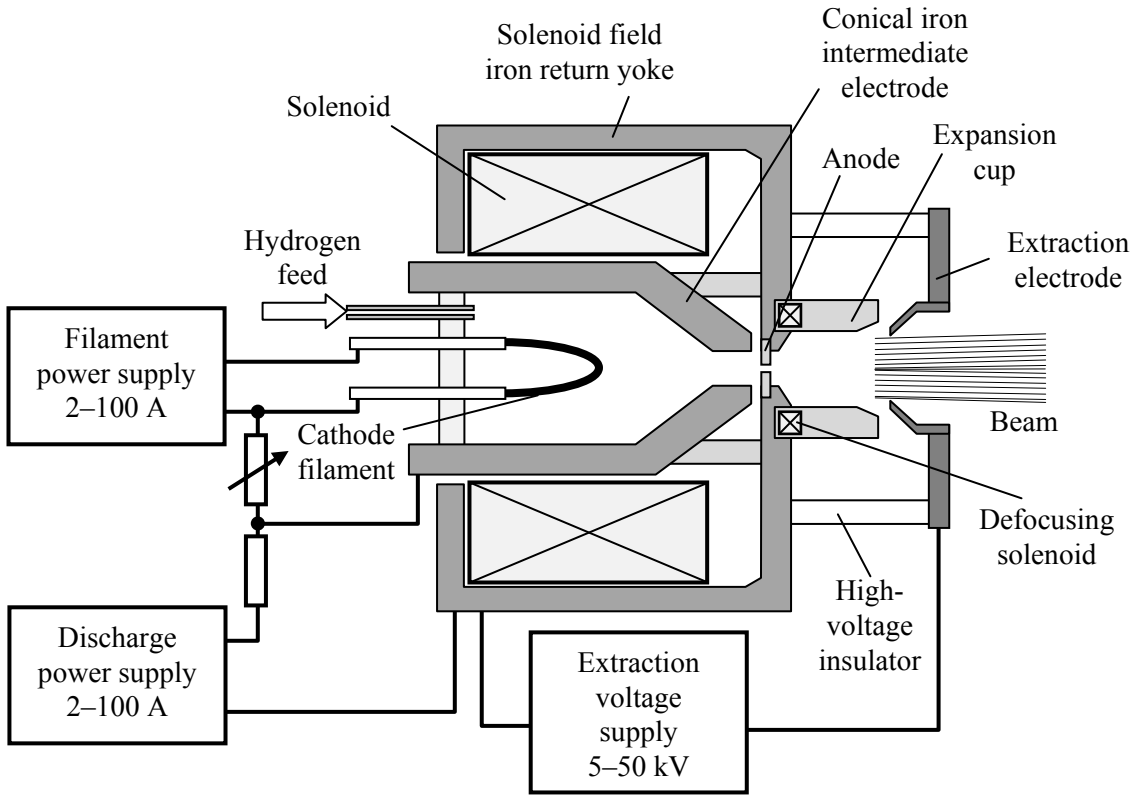


Fig. 9: A schematic of a duoplasmatron source

4.4 Microwave ion sources

4.4.1 Introduction

Microwave ion sources use alternating electric fields in the gigahertz (GHz) frequency range to generate the plasma. Instead of using electrodes the microwave energy is coupled to the discharge via a waveguide. With no electrodes to erode away microwave ion sources can have lifetimes in excess of 1 year. The plasma chamber has similar dimensions to the wavelengths of the microwaves and is surrounded by DC solenoids that produce an axial magnetic field.

Microwave ion sources can be separated into two families: “on resonance” and “off resonance”.

4.4.2 On resonance (ECR sources)

On resonance sources are called ECR sources. The electrons are cyclotron accelerated by the combination of microwave frequency electric fields and static magnetic fields. The magnetic field makes electrons gyrate around at a frequency that matches the frequency of the microwave electric field. The solenoidal magnetic field also acts to confine the positive ions.

ECR ion sources were first developed in the late 1960s by Richard Geller and his group at CEA. ECR ion sources are very good at producing multiply charged positive ions. ECR sources work by step-wise ionization: the accelerated electrons progressively remove the outer orbital electrons of the

ions by impact ionization. The comparatively slow moving positive ions are confined by the magnetic field to be ionized again by the re-accelerated electrons. High-charge-state positive ions can be produced by this technique. This is particularly useful for making high-charge-state beams of heavy elements such as uranium.

The ECR ion source is based on plasma heating at the electron cyclotron frequency (f_{ECR}) in a magnetic field, given by

$$\omega_{ECR} = 2\pi f_{ECR} = \frac{eB}{m} \tag{13}$$

For a 2.45 GHz frequency the electron ECR field is 875 G.

For electrons in a magnetic field in the range 0.05–1 T, this corresponds to a frequency range of 1.4 GHz to 28 GHz. The availability of commercial magnetrons and klystrons results in most sources working at 2.45, 10, 14.5, 18, 28 and 37.5 GHz.

The 2.45 GHz frequency is used because of this is also the frequency used in microwave ovens, so cheap reliable magnetron tubes are readily available. Also the waveguides are of manageable size (35 mm × 73 mm). Lower frequencies yield lower emittance beams and require lower magnetic fields.

Since 1994 CERN have used a 14.5 GHz ECR ion source to produce 100 eμA of Pb^{27+} ions. Daniela Leitner and her team at Lawrence Berkeley National Laboratory have recently produced 200 eμA beams of U^{34+} ions and 4.9 eμA beams of U^{47+} ions with the 28 GHz superconducting VENUS ion source [3].

4.4.3 Off resonance (Microwave discharge sources)

Off resonance sources are called microwave discharge ion sources. They also use microwaves to produce a discharge, but the magnetic field is above the ECR field for the applied microwave frequency. Microwave discharge ion sources produce high currents of singly charged ions with low emittance. Higher plasma density is obtained by the higher magnetic fields rather than using higher gas pressure. Microwave discharge ion sources were first developed by Noriyuki Sakudo's team at Hitachi [4] and Junzo Ishikawa's team at Kyoto University [5] in the late 1970s and early 1980s. The basic design of all modern microwave discharge ion sources are based on the proton source developed by Terence Taylor and Jozef Mouris at Chalk River National Laboratory in the early 1990s [6].

4.4.4 Basic design

Figure 10 shows a schematic of a microwave ion source. The key aspects of the design are a small compact plasma chamber with two solenoids at the front and back. A stepped matching section is used to allow smooth transition between the waveguide and plasma impedances. An extraction system with a suppressor electrode is employed to limit the back-streaming electrons.

The main difference between ECR and microwave discharge sources is that ECR sources have an additional hexapole field surrounding the plasma chamber. This helps to further confine the ions so that high charge states can be obtained. The hexapole field can either be produced by permanent magnets or with coil windings.

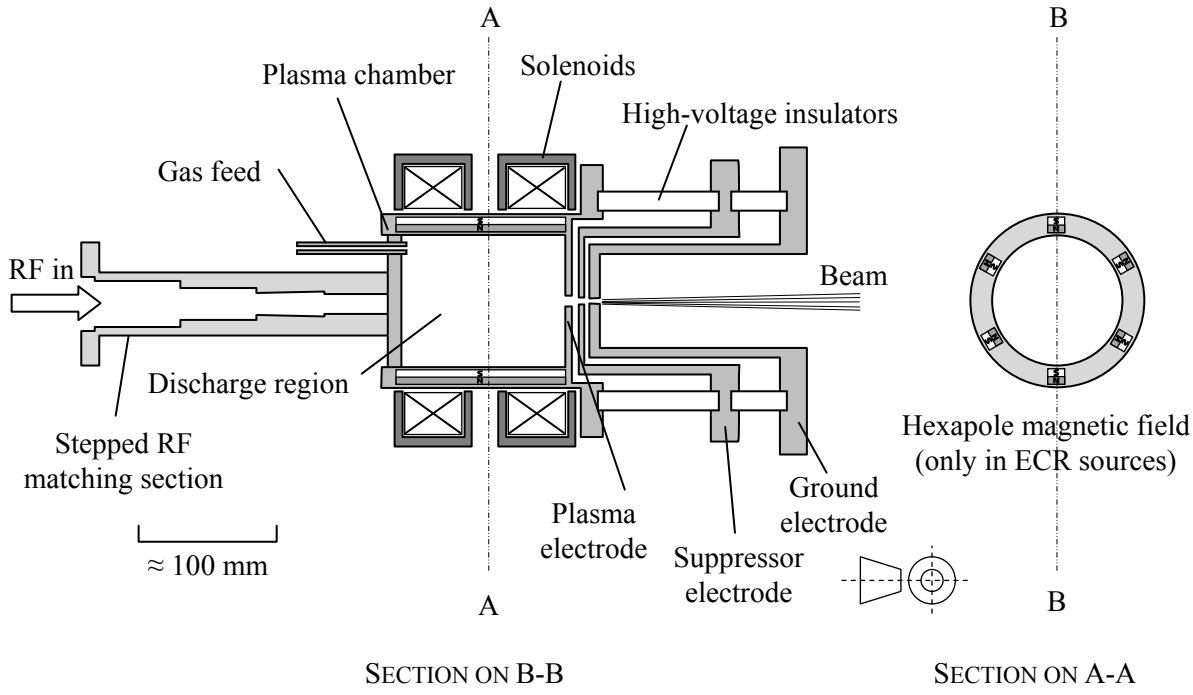


Fig. 10: Sectional schematic of a microwave ion source

4.4.5 Further developments

4.4.5.1 Low Energy Demonstration Accelerator

The design was further improved by Joe Sherman and his team at LANL in the mid 1990s [7]. They optimized the extraction system and the LEBT to maximize transmission to the radio-frequency quadrupole (RFQ) of the Low Energy Demonstration Accelerator (LEDA) project. They also developed a pulsed mode operation with a rise/fall time of the order of tens of microseconds. The LEDA project demanded a decrease in beam emittance and a high reliability. The LEDA source could reliably deliver enough current to produce a 100 mA, 7 MeV DC beam of protons at the exit of the RFQ.

4.4.5.2 Source d´Ions Légers à Haute Intensité

During the second half of the 1990s, Raphael Gobin and his team at CEA Saclay developed the Source d´Ions Légers à Haute Intensité (SILHI) source [8, 9]. They obtained greater brightness and even higher reliability. DC proton beam currents of 140 mA with 0.2 π·mm·mrad normalized emittance were achieved. Lifetimes of around 1 year were demonstrated.

4.5 Electron beam ion sources

4.5.1 Introduction

EBISs use a high current density electron beam to ionize the particles. The EBIS was first developed in the late 1960s by E.D. Donets and his team at JINR, Dubna. Reinard Becker and his team at Frankfurt demonstrated that DC beams are possible but only with very low beam currents. EBISs are complex, expensive and can only produce relatively short pulse lengths of high currents. Nevertheless, they are capable of reliably producing beams of positive ions with very high charge state. Heavy elements can be completely stripped of their electrons leaving bare nuclei.

4.5.2 Basic operation

Figure 11 shows a schematic of an EBIS. A high-current electron gun produces a 1 keV to 20 keV electron beam that is compressed to a current density of the order of 1000 A cm^{-2}. The electron beam passes though through a set of drift tubes in a 1–5 T solenoidal field. The strong solenoidal field compresses the electron beam. Electrical damping components on the drift tubes help maintain the beam stability.

The material to be ionized is either pulsed into the middle of the ionization chamber or injected as a low-energy, low-charge-state beam from another ion source. The strong space charge of the negative electron beam creates a potential well that traps the injected positive ions. The amount of charge that can be trapped is limited by the size of the potential well created by the electron beam. Once trapped the ions undergo successive ionizations by the electron beam. During the trapping and ionization phase greater positive voltages are applied to the end drift tubes to longitudinally confine the ions as shown in Fig. 11. Once the required charge state has been reached the extraction phase begins by modifying the potential distribution on the drift tubes as shown in Fig. 11. The EBIS has been developed by many researchers, most recently by Jim Alessi and his team at BNL to produce a 1.7 emA, 10 μs, 5 Hz beam of Ag^{32+} ions [10].

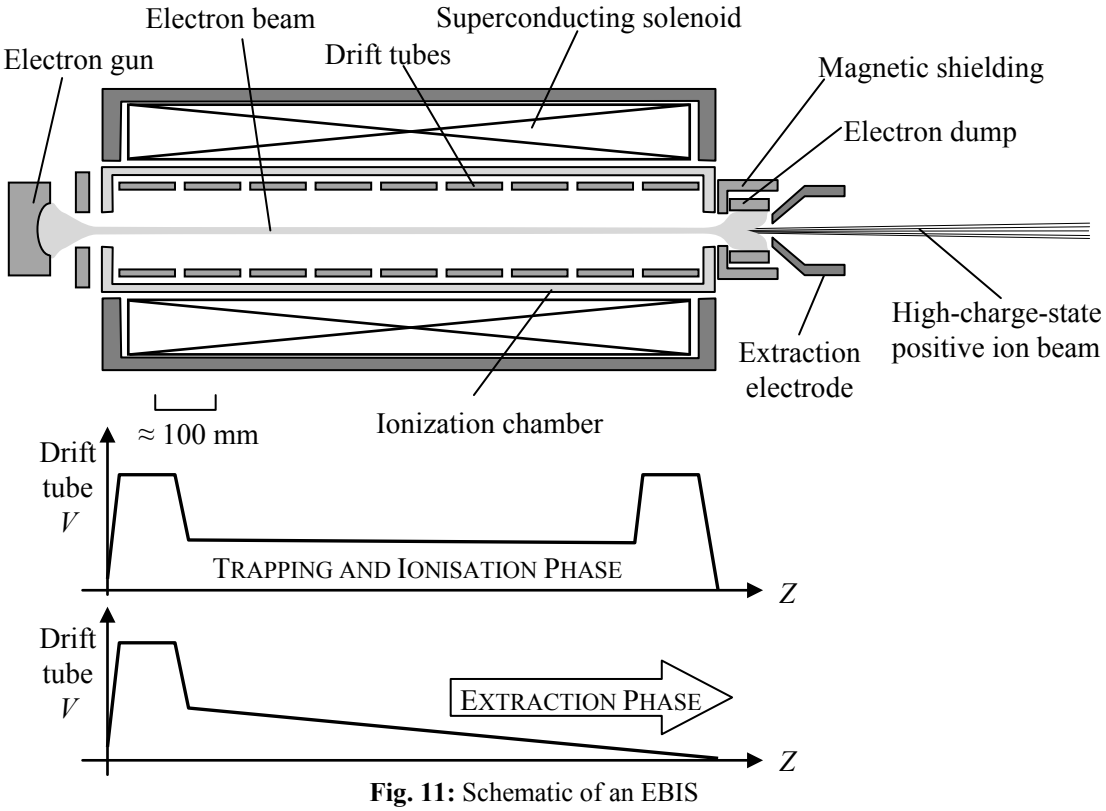


Fig. 11: Schematic of an EBIS

4.6 Laser ion sources

4.6.1 Introduction

Laser ion sources use a powerful laser to vaporize and ionize target material. They can produce high-current and high-charge-state beams of almost every element. The idea for laser ion sources was first proposed independently in 1969 by Peacock and Pease at UKAEA Culham and by Byckovsky and colleagues in Russia.

Laser ion sources are limited to short pulse lengths and low repletion rates. The particles are ablated from the target material so a fresh area of the target surface must be exposed for each pulse.

4.6.2 Basic operation

The beam from a pulsed high-power laser is focused at a target through a KCl salt window in the target chamber as shown in Fig. 12. When the laser beam hits the solid target it first vaporizes the material then ionizes it into a plasma. The electrons are accelerated by inverse bremsstrahlung to several hundred electronvolts. The electrons stepwise ionize the target atoms to higher and higher charge states. The dense plasma rapidly expands into a plasma plume which propagates along the expansion region until it reaches the extraction aperture. The target chamber sits on a high-voltage platform to allow a beam to be extracted by a grounded electrode. A suppressor electrode is also used to prevent back-streaming electrons.

The beam current produced depends on the amount of target material ionised which depends on the amount of energy delivered by the laser, this can range from 0.1 J to a few tens of Joules per pulse. The highest charge state obtained depends on the power density on the target surface. Power densities employed range between 10^9 W/cm^2 and 10^{16} W/cm^2. The pulse length depends on the length of the expansion region.

The laser ion source for the TWAC at ITEP Moscow produces 7 mA, 10 μs pulses of C^{4+} at 50 keV/u [11]. Recently, Masahiro Okamura at BNL has successfully matched a 35 mA, 2.1 μs pulsed beam of C^{4+} from a laser ion source directly into a RFQ.

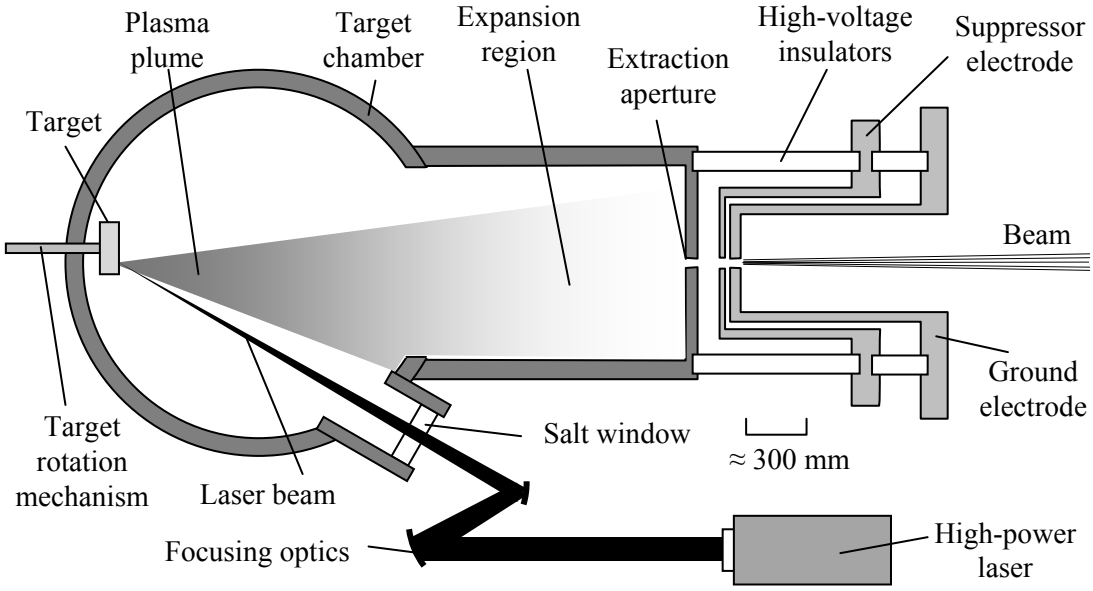


Fig. 12: Schematic of a laser ion source

4.7 Vacuum arc ion sources

4.7.1 Introduction

When an arc occurs in a vacuum the current-carrying particles are created by vaporizing the cathode material. Vacuum arc ion sources exploit this to produce a beam of particles made of the cathode material. Vacuum arc sources are often called metal vapour vacuum arc (MEVVA) sources. The first reliable sources were developed in the 1980s in the USA by Ian Brown, S. Humphries, Jr and

S. Picraux. MEVVA sources can produce high currents of medium-charge-state metal ions but the beam can be quite noisy. Cathode lifetimes are limited to about 1 day or less depending on the duty cycle.

4.7.2 Basic operation

The arc is triggered by applying a short (≈ 10 μs), high-voltage (≈ 10 kV) pulse to the trigger electrode (shown in Fig. 13). This initiates an arc between hot spots on the cathode and the anode. Microscopic irregularities on the cathode surface emit large quantities of electrons which cause localised heating. Material is vaporized from these cathode hot spots which feeds into the arc discharge. Each tiny (1–10 μm) cathode hot spot carries about 10 A and is only active for a few tens of nanoseconds before it explodes. In a typical 100–300 A arc discharge, dozens of hot spots are active at any one moment and the overall behaviour is dynamic and extremely complex.

The arc plasma expands through the expansion region which sometimes has a solenoidal field to confine the ions. The dynamic cathode hot spots make the source intrinsically noisy. Some sources use electrostatic grids [12] to stop the flow of plasma electrons so that only ions are allowed to continue to the extraction aperture. The space charge of the ions helps to smooth out the plasma density variations caused by the hot spot explosions before the ions reach the extraction aperture. The source sits on a high-voltage platform (up to 50 kV) to allow a beam to be extracted by a grounded electrode. A suppressor electrode is also used to prevent back-streaming electrons.

The MEVVA ion source for the High Current Injector at GSI in Germany can produce 15 mA of U^{4+} ions [13].

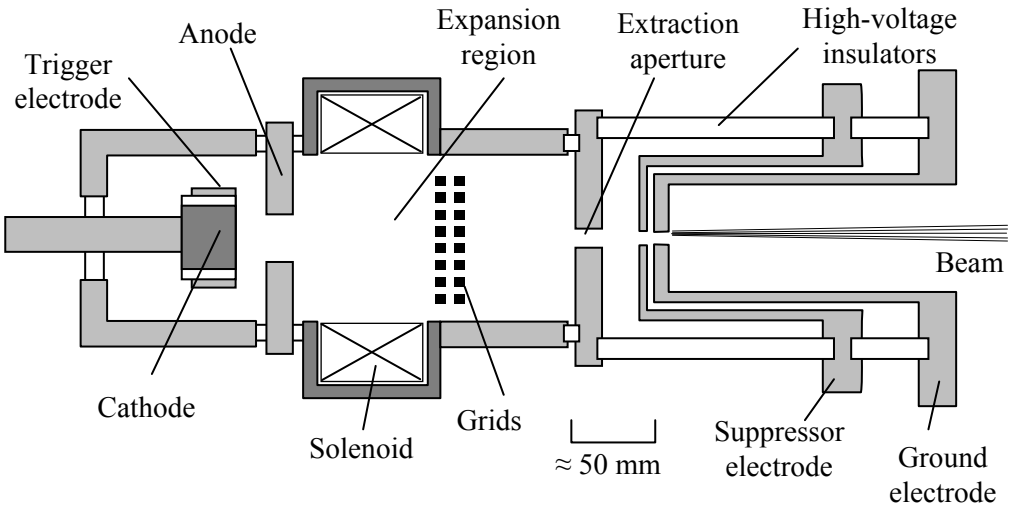


Fig. 13: Schematic of a vacuum arc ion source

5 Negative ion sources

5.1 Introduction

5.1.1 The negative ion

Negative ion sources produce beams of atoms with an additional electron. The binding energy of the additional electron to an atom is termed the electron affinity. Some elements have a negative electron affinity (such as beryllium, nitrogen or the noble elements) which means they cannot form stable

negative ions. H^- is the most commonly produced negative ion. Hydrogen has an electron affinity of 0.7542 eV. Considering that the electron binding energy of neutral hydrogen is 13.6 eV, the extra electron on an H^- ion is very loosely held on. All of the ion sources in this section have been used to produce D^- ions as well as other heavy negative ions, such as O^-, B^-, C^-, etc.

5.1.2 Uses

Negative ion sources were first developed to allow electrostatic accelerators to effectively double their output beam energy. In a tandem generator H^- ions are first accelerated from ground to terminal volts, they are then stripped of their two electrons when they pass through a thin foil. The resulting protons are then accelerated from terminal volts back to ground, at which point they have an energy of twice the terminal volts.

Cyclotrons use negative ions and stripping foils to extract the beam from the cyclotron. The stripping foil is positioned near the perimeter of the cyclotron poles. As the negative ion beam is accelerated it circulates on larger and larger radii until it passes through the stripping foil, which converts the beam from being negative to positive. The Lorenz force on the beam is reversed and instead of the force pointing into the centre of the cyclotron it points outwards and the beam is cleanly extracted.

In high-power proton accelerators H^- ions are used to allow charge accumulation via multiturn injection. An H^- beam from a linear accelerator is fed through a stripping foil into a circular ring (a storage, accumulator or synchrotron ring) leaving protons circulating in the ring. The H^- beam from the linear accelerator continues to enter the ring whilst the circulating beam repeatedly passes through the stripping foil unaffected. The incoming beam curves in one way through a dipole as an H^- beam then curves out of the dipole in the opposite direction as a proton beam on top of the circulating beam. This allows accelerator designers to beat Liouville's theorem and build up charges in phase space. Without this negative ion stripping trick only one turn could be accumulated in the ring.

5.2 Physics of negative ion production

5.2.1 Mechanisms and challenges

The physical processes involved with the production of negative ions are still not fully understood, but they can be generally described as: charge exchange, surface and volume production processes. In different types of source one production process may dominate, however all three processes might contribute to the overall extracted negative ion current.

5.2.2 Charge exchange

The first H^- ion sources were charge exchange devices. There are two ways of doing this: with foils or gases. With foils a proton beam, with an energy of about 10 keV, is passed though a negatively biased foil and by electron capture an H^- beam is produced. For gases the proton beam is passed though a region filled with a gas. The H^- beam is produced by sequential electron capture; first protons are converted to neutral H^0, then to H^-. The gas acts as an electron donor. Only about 2% of the protons are converted into H^- ions. Until the 1960s this was the main technique used to make H^- beams. Beams of up to 200 μA were produced using this method. In 1967 Bailey Donnally [14] discovered that the yield of He^- ions can be increased by using caesium vapour as an electron donor. This lead to the development of a series of negative ion sources using alkali vapour.

Resonant charge exchange between fast H^- ions and slow neutral hydrogen atoms (H^0) is essential to the operation of a Penning source (see Section 5.3.3):

$$H^- (\approx 60 \text{ eV}) + H^0 (< 1 \text{ eV}) \rightarrow H^- (< 1 \text{ eV}) + H^0 (\approx 60 \text{ eV})$$

5.2.3 Early Negative Ion Sources

For several decades numerous researchers [15, 16] had been experimenting with sources originally designed to produce positive ions, but by reversing the polarity of the extraction they were able to extract negative ions. Nevertheless, the co-extracted electron current was always at least an order of magnitude higher than the negative ion current.

In the early 1960s George Lawrence and his team at Los Alamos [17] were using a duoplasmatron to produce H^- ions when they first noticed that substantially higher beam currents and lower electron currents could be extracted when the extraction was actually off-centred from the intermediate electrode (Fig. 14). They concluded that the extracted H^- ions must be produced near the edge of the plasma. (This was also discovered independently by a team at the UK Atomic Weapons Establishment [18].)

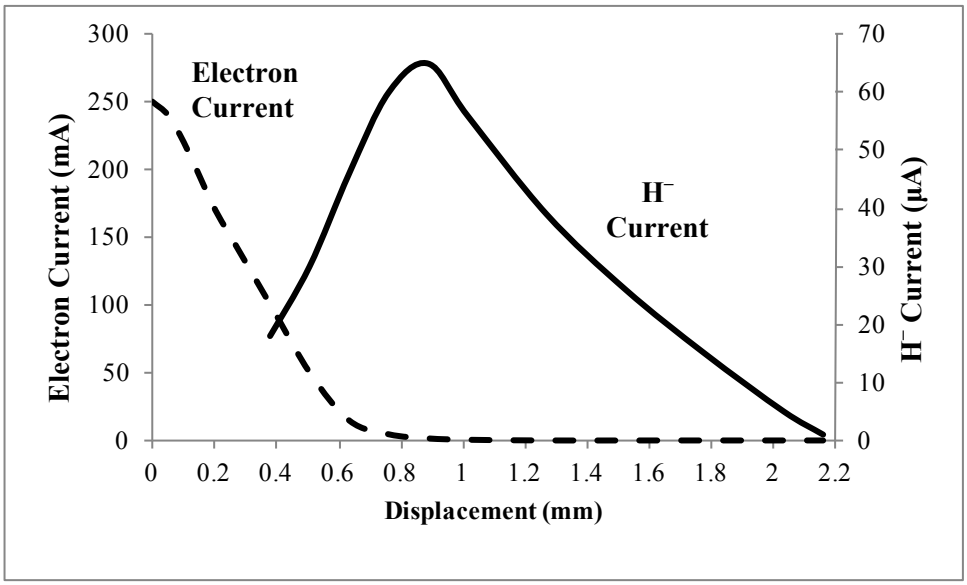


Fig. 14: H^- and electron currents as a function of extraction offset in a duoplasmatron measured at Los Alamos

During the 1960s various sources originally designed to produce positive ions were adapted and modified to produce H^- ions and beam currents up to a few milliamps were produced.

5.2.4 Caesium and surface production

In the early 1960s Victor Krohn, Jr and his team [19] at Space Technology Laboratories, Inc. California were experimenting with surface sputter ion sources. Surface sputter ion sources are mainly used to produce beams of heaver ions (such as metals) for coating and etching applications. Krohn noticed that when Cs^+ ions were used to sputter a metal target the yield of sputtered negative ions increased by an order of magnitude.

In the early 1970s Gennadii Dimov, Yuri Belchenko and Vadim Dudnikov at the Budker Institute of Nuclear Physics started experimenting with caesium in ion sources. Using a magnetron ion source (see Section 5.3.2), Vadim Dudnikov added Cs vapour to the discharge for the first time. A dramatic increase in H^- current was observed along with a decrease in co-extracted electrons. The Dimov team went on to extract a colossal 880 mA pulsed H^- beam from an experimental magnetron ion source [21]. This success led them to develop a Penning type ion source (see Section 5.3.3) that could produce 150 mA of H^- beam current with only 250 mA of extracted electrons. The H^- currents produced were orders of magnitude higher than anything seen previously. When these revolutionary

results were published interest in caesiated ion sources took off. Researchers all over the world started using caesium in their ion sources and a large number of new ion source designs were developed.

A very different type of H^- source that relies on surface production of H^- ions is the surface converter source (see Section 5.5). Developed in the 1980s by Ehlers and Leung at the Lawrence Berkeley Laboratory, it also relies on a caesiated surface. The caesiated surface sits in the middle of the plasma and is curved with a radius centred on the extraction region. H^- ions produced on this surface are "focused" towards the extraction hole because they are repelled by the negative potential on the converter surface, this is why this type of source is also called "self-extracting".

In addition to aiding H^- surface production, caesium also helps to stabilize the plasma by readily ionizing to produce additional electrons for the discharge. This reduces the amount of noise in the discharge and extracted beam current.

5.2.5 Surface physics processes

The types of particles arriving at a surface could be protons, ionized hydrogen molecules, ionized caesium, energetic neutral atoms or molecules. When a particle interacts with a surface many complex and competing processes can occur:

- reflection;
- adsorption;
- sputtering;
- desorption;
- recombination;
- dissociation;
- ionization;
- secondary electron emission;
- photo emission;
- excitation.

Particles ejected from the surface could be of a different charge state from the incoming particle, be excited, be in a molecule, or some combination of all three. The surface may also be altered. Complex interactions can take place at the surfaces.

The most important factor affecting H^- production at a surface is the work function, φ. To make H^- ions the surface must provide electrons, so a low work function surface is essential. The work function of a surface obviously depends on what it is made of. If different atoms of a different element are adsorbed on that surface (such as caesium) then the work function can be altered. The "thickness" of the adsorbed layer will also have an effect on the surface's work function.

The thickness of the adsorbed layer is usually defined in terms of the number of "monolayers" of the adsorbed atoms:

$$\text{Thickness (number of monolayers)} = \frac{\text{Number of adsorbate atoms per unit area}}{\text{Number of adsorbate atoms for a monolayer per unit area}} \quad (14)$$

When talking about negative ion production, the surface is usually the cathode and is typically made of a high melting point metal such as tungsten φ = 4.55 eV or molybdenum φ = 4.6 eV. Caesium has the lowest work function of all elements: φ = 2.14 eV. The work function of a caesium-coated molybdenum surface is actually lower than that of bulk caesium. As caesium covers the molybdenum surface the work function decreases to 1.5 eV at 0.6 of a monolayer and then rises to about 2 eV for one monolayer or greater of caesium as shown in Fig. 15. This minimum at 0.6 monolayers is caused by atomic interactions increasing the Fermi level at the surface, thus decreasing the amount of energy required to liberate the electrons.

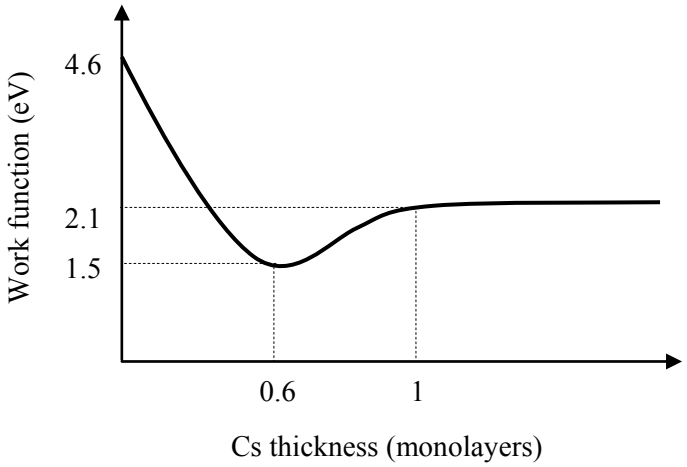


Fig. 15: Surface work function versus caesium thickness on a molybdenum surface

The thickness of the caesium coating on electrode surfaces will rarely ever come close to one full monolayer because thermal emission and plasma sputtering removes excess caesium. This is because the Cs–Cs bond is actually weaker than the Cs–Mo or Cs–Ta bonds, so Cs atoms adsorbed on Cs atoms are rapidly sputtered away by the plasma or thermally emitted from hot surfaces. (It is possible to build up multiple layers but only on cold surfaces that are shielded from the plasma.)

5.2.6 Maintaining caesium coverage

To minimize the work function and maximize the H^- production an optimum layer of caesium must be maintained on the surface. The surface is a dynamic place, caesium atoms are constantly being desorbed by plasma bombardment. To maintain optimum caesium coverage a constant flux of caesium is required. This is often provided by an oven containing pure elemental caesium but it can also be provided using caesium chromate cartridges that release Cs when heated.

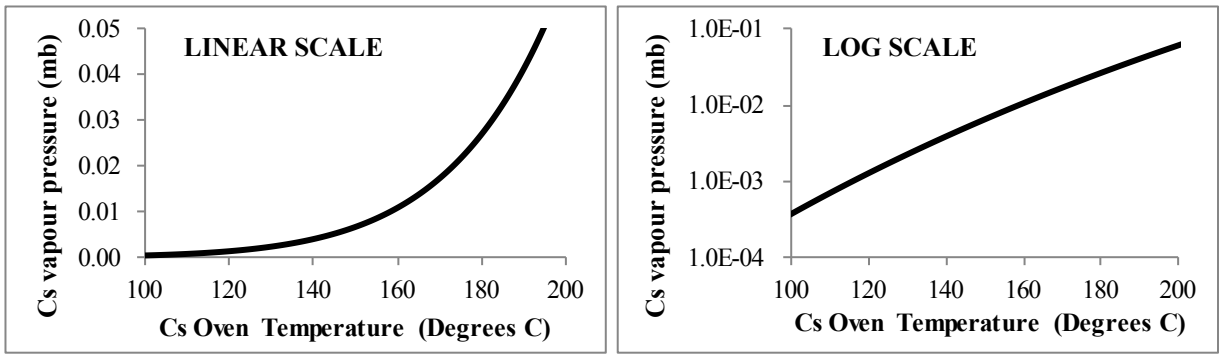


Fig. 16: Caesium vapour pressure versus caesium oven temperature

The flux of caesium into the plasma can be precisely controlled by setting the temperature of the caesium oven. Figure 16 shows how the vapour pressure of caesium varies with temperature [20]. All sources that use elemental caesium in an oven operating between 100 °C and 190 °C. This covers a large range of vapour pressures. Most sources operate in the 130 °C to 180 °C range but some sources such as the RF driven volume multicusp source only require very small Cs fluxes and operate in the 105 °C to 110 °C range.

5.2.7 Volume production

Also in the 1970s in parallel to the discovery of caesium-enhanced H^- surface production, Marthe Bacal and her team at Ecole Polytechnique developed a completely new type of source that relied on H^- production in the plasma volume itself. Initially people were sceptical because H^- ions are so fragile: only 0.7542 eV is required to detach the extra electron. The plasma in the discharge was thought to be too energetic for any H^- ions produced in the volume to survive long enough to make it to the extraction region. The breakthrough that made volume production possible was separation by a magnetic filter field of the plasma production region from the extraction region. In the 1980s, Leung, Ehlers and Bacal used a filament-driven multicusp ion source with a magnetic dipole filter field positioned near the extraction region (see Section 5.6). The filter field blocked high-energy electrons from entering the extraction region, whereas ions and cold electrons could diffuse across the filter field. This effectively separated the discharge into two regions: a high-temperature driver plasma on the filament side of the filter field, and a low-temperature H^- production plasma on the extraction region side. Magnetically filtered multicusp ion sources are sometimes called "tandem" sources because of these two regions of different plasma temperatures (not to be confused with tandem accelerators).

The volume production process relies on the dissociative attachment of low energy electrons to rovibrationally excited H_2 molecules:

$$H_2^* + e\ (\leq 1\ eV) \rightarrow H^- + H^0$$

If the H_2 molecule is vibrationally cold the dissociative attachment cross section is extremely low (10^{-21} cm^2). When the H_2 molecule is rovibrationally excited, however, the cross section increases by five orders of magnitude. Thus, low-energy electrons can be very effective in generating H^- ions by dissociative attachment to highly vibrationally excited molecules. The rovibrationally excited molecules are produced not only in the plasma but also on the walls of the chamber and electrode surfaces.

5.2.8 H^- destruction

In both volume and surface sources there are many processes that can destroy H^- ions. The most common are:

$H^- + H^+ \rightarrow H^0 + H^0$	mutual neutralization;
$H^- + e \rightarrow H^0 + 2e$	electron detachment;
$H^- + H^0 \rightarrow H_2^* + e$	associative detachment.

Another important factor is the cross section of the H^- ion in comparison to the H^0 atom. The H^- ion cross section is 30 times larger than the neutral H^0 atom for collisions with electrons and 100 times larger for collisions with H^+ ions. As well as being fragile and easily destroyed it is much more likely to be hit.

The aim of the source designer is to minimize the H^- destruction processes by controlling the geometry, temperature, pressure and fields in the source. The following sections describe the design of some of the most successful H^- ion sources.

5.3 Surface plasma cold cathode ion sources

5.3.1 Introduction

The term cold cathode refers to the fact that the cathode is not independently heated, however the name can be misleading because the cathode can still operate at elevated temperature due to heating by the discharge itself. They are called surface plasma sources because H^- ions are produced on the surface of the cathode (see Section 5.2.5). Both of the sources discussed in this section were used for many years as positive ion sources before they were employed to make H^- ions. The discharge is in

direct contact with the anode and cathode, so sputtering processes will eventually erode the electrode surfaces. This puts a fundamental limit on their lifetime.

5.3.2 Magnetron

The magnetron (also called a planotron) was the first ion source where the H^- current was first significantly increased by adding caesium vapour. This work was done by Gennadii Dimov, Yuri Belchenko and Vadim Dudnikov at the Budker Institute of Nuclear Physics in the early 1970s [21]. Chuck Schmidt developed the Fermilab version of Magnetron ion source in the late 1970s, a design which was adopted and further developed by Jens Peters at DESY, and Jim Alessi at Brookhaven.

The magnetron source has a racetrack-shaped discharge bounded on the inside by the cathode and the outside by the anode as shown in Fig. 17. The anode and cathode are only about 1 mm apart so the discharge is in the shape of a ribbon wrapped around the cathode. A magnetic field of between 0.1 T and 0.2 T is applied perpendicular to the plane of the racetrack, this causes the plasma to drift around the racetrack. On one of the long sides of the racetrack discharge the anode has a hole through which the beam is extracted. Pulsed hydrogen is fed into discharge on the opposite side to the extraction hole. Caesium vapour is introduced via an inlet on one side. H^- ions are produced on the cathode surface and are accelerated away by the cathode sheath potential. A concave region on the cathode surface opposite the extraction hole can give an initial focus to the extracted H^- beam. Some of the cathode sheath accelerated H^- ions undergo resonant charge exchange with slow thermal H^0 on the way to extraction, resulting in a beam energy distribution with two peaks.

The magnetron can give high H^- currents up to 80 mA and can have very long lifetimes of over 6 months, but it will only operate at very low duty factors of up to 0.5 %. This because it is not possible to maintain optimum caesium coverage on the cathode surfaces during the time the discharge is on.

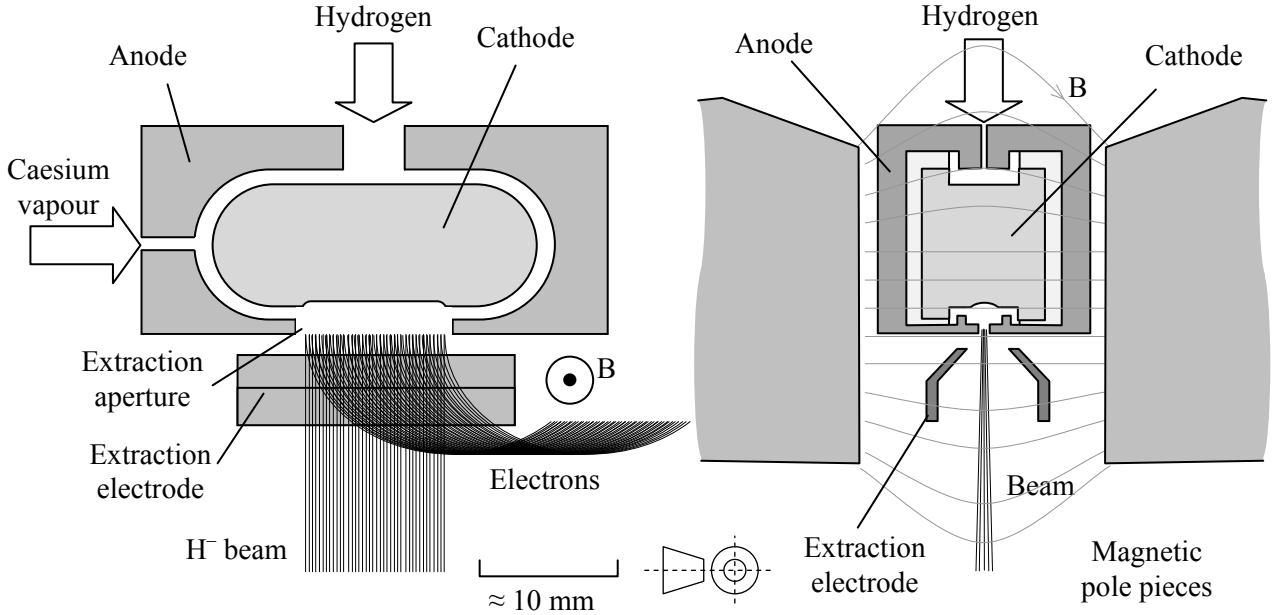


Fig. 17: Sectional schematic of a magnetron with slit extraction

5.3.3 Penning

The H^- Penning ion source was invented by Vadim Dudnikov, and developed with Gennadii Dimov and Yuri Belchenko at the Budker Institute of Nuclear Physics in the early 1970s at the same time as the magnetron source. Vadim Dudnikov reported high H^- currents up to 150 mA and duty cycles all of the way up to DC [22]. Vernon Smith, Paul Allison and Joe Sherman at Los Alamos developed a scaled up Penning source [23] that gave similarly high currents with low emittances.

A Penning source (Fig. 18) has a small (10 mm × 5 mm × 5 mm) rectangular discharge region with a transverse magnetic field. The long sides of discharge are bounded by two cathodes, with the other four walls at anode potential, creating a 'quadrupole like' electric field arrangement. The magnetic field is orientated orthogonally to the cathode surfaces so that electrons emitted from the cathode are confined by the magnetic field lines and reflex back and forth between the parallel cathode surfaces. The primary anode is hollow and has holes through which hydrogen and caesium vapour are fed into the discharge. The extraction aperture plate is also at anode potential. The beam is extracted from the plasma through a slit in the extraction aperture plate by a high-voltage applied to an extraction electrode. The electrodes are made of molybdenum.

Like with the magnetron, H^- ions are produced on the cathode surfaces and accelerated by the plasma sheath potential that exists next to the cathode. The plasma sheath potential is about 60 V. Unlike the magnetron, however, the cathode surface is not directly opposite the extraction aperture. The addition of ribs on the inside of the extraction aperture plate mean there is no direct line of sight between the cathode surface and the extraction aperture. Thus, it is impossible for the fast cathode produced H^- to be extracted directly. Only H^- ions that have undergone resonant charge exchange with slow H^- ions (see Section 5.2.2) are extracted resulting in a beam with a lower energy spread than from the magnetron.

The Penning is the brightest H^- ion source with current densities at extraction above 1 A cm^{-2} possible. The lifetime of the Penning source is limited to a few weeks because of cathode sputtering by caesium ions. This type of source will not operate without caesium vapour and requires the electrode surfaces to be between 400 °C and 600 °C.

This source is currently under development at the Rutherford Appleton Laboratory by the present author and his team; 60 mA, 1 ms, 50 Hz pulses can be routinely produced [24].

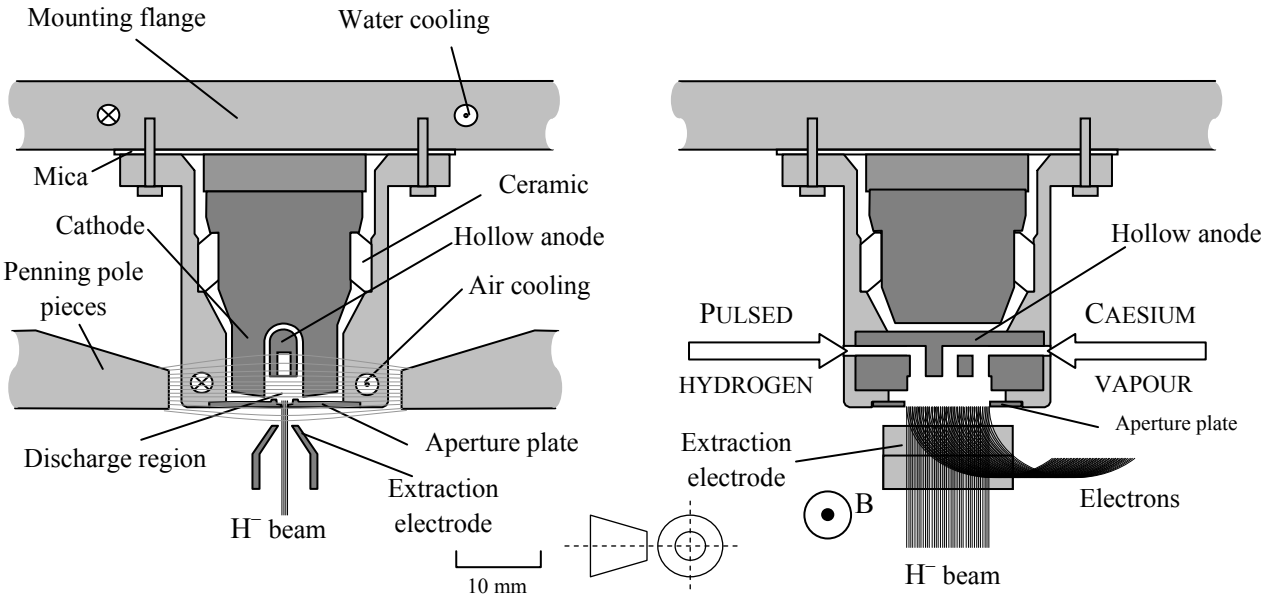


Fig. 18: Sectional schematic of a Penning H^- ion source

5.4 Multicusp ion sources

5.4.1 Introduction

Multicusp ion sources have permanent magnets positioned around the perimeter of the plasma chamber with alternating north and south poles. This alternating arrangement creates magnetic cusps

around the chamber walls which serve to confine the plasma and keep it away from the walls. The containment of the plasma by the multipole field arrangement is why multicusp sources are also referred to as "bucket" sources. Originally developed in the 1970s for fusion research these sources truly were giant buckets of plasma with dimensions of the order of 0.5 m to 1 m. They were designed to produce several amps of current extracted through multiple apertures in an extraction grid.

For high-current multicusp sources the means of plasma production is either in a hot cathode discharge or in an inductively coupled discharge with a RF solenoid antenna. Researchers have investigated both microwave and ECR discharge plasma production in multicusp sources, but only low negative ion beam currents in the microamp range have been produced, so they are not used for high-power accelerator applications.

5.4.2 Hot-cathode-driven plasma production

Most multicusp sources use a hot tungsten filament cathode bent into the shape of a hairpin. Using a filament heater power supply, several hundred amps of DC are passed through the filament to heat it up so that it thermionically emits electrons. The main plasma discharge is then created with a second power supply, between 10 A and 500 A DC flows from the filament cathode to an anode. The anode can either be the metallic walls of the plasma chamber or a specific electrode in the plasma chamber.

The temperature of the filament should not be too high otherwise the electron emission will become space charge limited. This causes the main discharge voltage to increase and become very noisy. The shape and position of the hot cathode filament is important. Coil filaments can be used but the hairpin shape is more common. A notable exception is the JPARC source that uses a coil filament made from LaB_6 as a cathode [25]. At about 1500 °C LaB_6 produces large amounts of electrons.

Wide hairpins (U-shaped) have been shown to produce more stable plasmas over a large range of operating conditions. The position of the filament relative to the multicusp field is important because the high current required to heat the filament loop creates a magnetic field itself. The filament should therefore be positioned so that the filament produced magnetic field is in the same direction as the local cusp field.

5.5 Filament cathode multicusp surface converter source

The multicusp surface converter source is a multicusp filament-driven discharge with the H^- ions produced on a molybdenum converter surface opposite the extraction hole. It was first developed by Ehlers and Leung and the team at Lawrence Berkeley University in the late 1970s and early 1980s. This type of source has also been used to generate other negative ions.

A discharge of several hundred amps is created between the filament cathode and the anode walls. The multicusp field as shown in Fig. 19 confines the plasma. Caesium vapour is fed into the discharge chamber and it coats the converter electrode surface. This type of source is sometimes called a self-extracting source because: (1) the converter is negatively biased so H^- ions produced on its caesium-coated surface are repelled toward the extraction hole; and (2) the radius of curvature of the converter surface is centred on the extraction hole to focus the H^- ions towards the extraction aperture. The multicusp magnets on either side of the extraction region act as a filter field to allow some volume production of H^- ions to supplement the ions produced by the surface converter. They also provide the magnetic field to dump the co-extracted electrons.

Ehlers and Leung developed this source for neutral beam injectors for fusion research to produce a 1 A DC H^- beam by adding more cathode filaments (up to six) and increasing the discharge current to 1000 A (see Ref. [26]).

Filament-driven surface converter sources tend to consume large amounts of Cs to negate the effect of the sputtered filament atoms adsorbing on the Cs layer on the surface converter electrode.

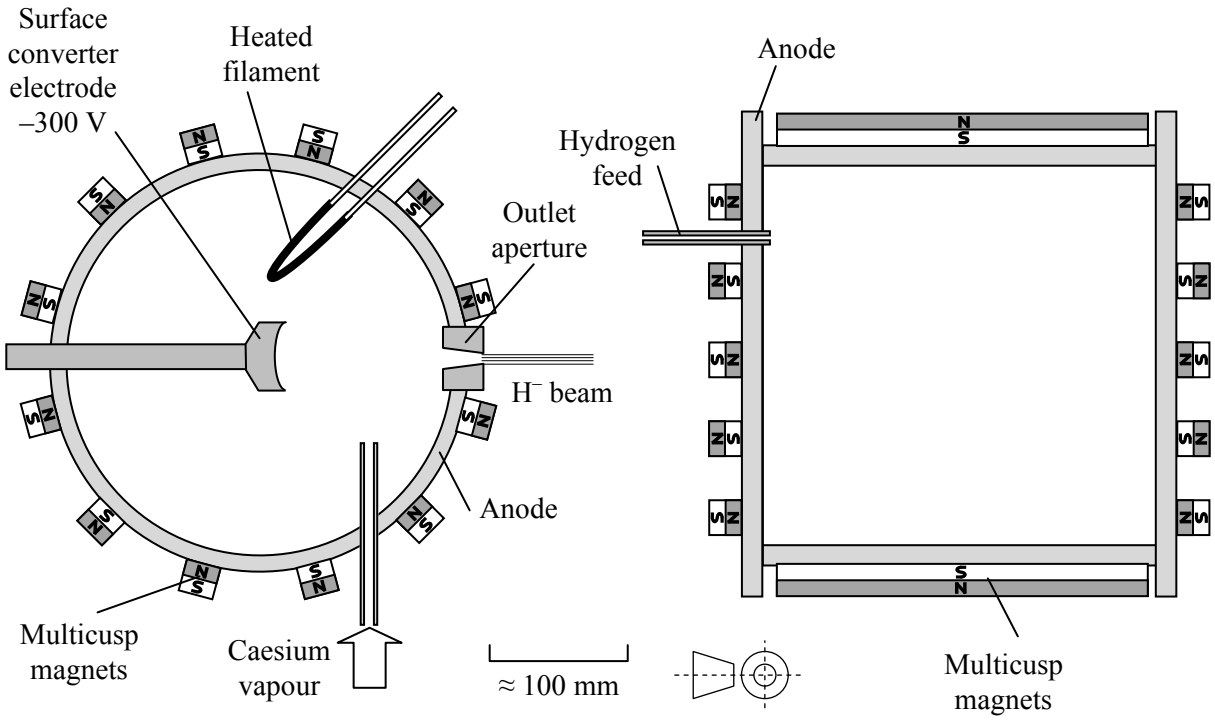


Fig. 19: Sectional schematic of a filament cathode multicusp surface converter source.

Rod Keller and Gary Rouleau at Los Alamos National Laboratory use this type of source operations on the LANSCE machine routinely producing a 16 mA, 60 Hz H^- beam with a lifetime of 35 days [27]. Like all filament driven discharge sources it suffers from lifetime limitations due to filament erosion. This source takes about 10 hours to start up and stabilize its output. It takes this long to develop the equilibrium coverage of caesium on the surface converter. This can impose operational restrictions on the rest of the machine.

5.6 Filament cathode multicusp volume source

The filament cathode multicusp volume source with filter field was first developed by Ehlers, Leung and Marthe Bacal and the teams at LBNL and Ecole Polytechnique in the 1980s [28]. A schematic of the source is shown in Fig. 20. The filter field creates a low-electron-temperature plasma region that is conducive to H^- production in the volume just in front of the extraction region (see Section 5.2.7 for an explanation of the physics).

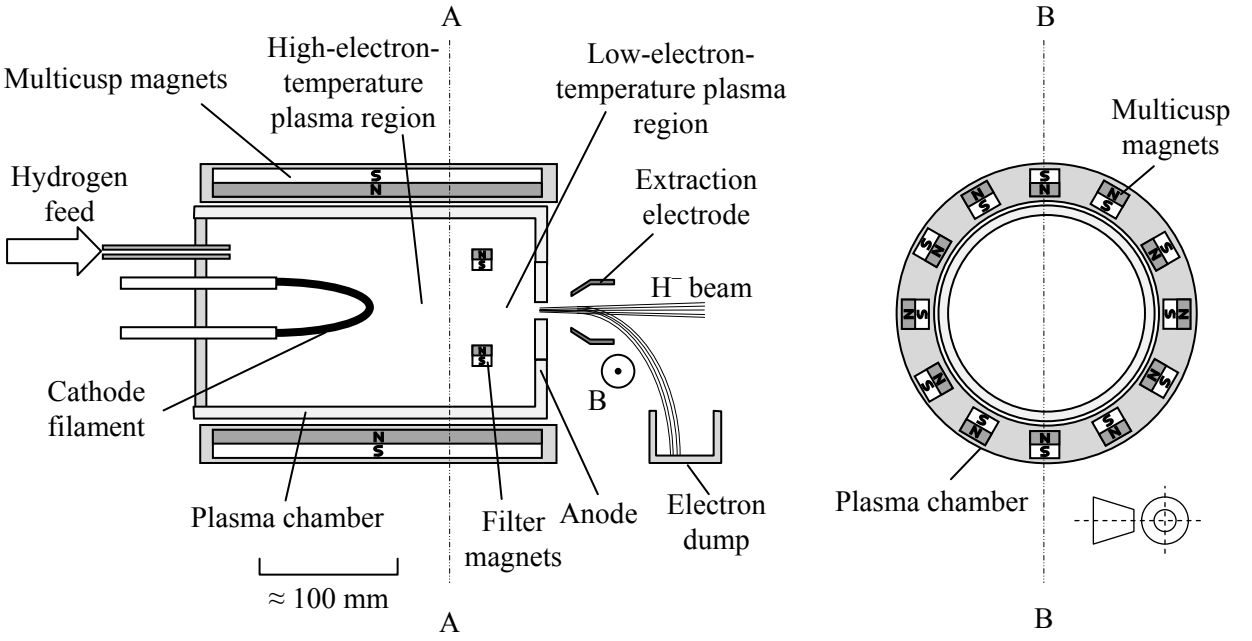


Fig. 20: Schematic of a filament cathode multicusp volume source

This type of source has been successfully developed all over the world and a few companies now sell them commercially. They are reliable and although they only have a lifetime of a few weeks they are very low maintenance, only requiring a very simple filament change. They are used mainly on cyclotrons.

Andrew Holmes and his team at Culham found the maximum current that can be extracted from this type of source is about 40 mA DC; it is limited by the maximum H^- density obtainable at the extraction region. The H^- current does not increase above a discharge current of about 200 A because H^- destruction processes start to dominate [29].

If caesium is added to this type of source the H^- current can be doubled to 80 mA. The caesium does not increase the volume production of H^- instead it actually facilitates surface production making this source a combined volume and surface source.

Frankfurt University created a source with three cathode filaments and using caesium they produced pulsed H^- currents of 120 mA [30], however the emittance and persistence of this beam was never measured and it is unlikely that anywhere near that current could actually be transported to the next stage of an operational accelerator. The lifetime of the Frankfurt University source was also never fully evaluated.

5.7 Internal RF antenna multicusp volume source

Instead of producing a discharge between a hot cathode filament and an anode the discharge can be generated by inductively coupled RF heating. An alternating magnetic field is produced by solenoid antenna fed with a RF power supply at a frequency of between 1 MHz and 10 MHz. Figure 21 shows a schematic of the source which also includes a filer field to screen the H^- volume production region from fast electrons.

In the early 1990s Ka-Ngo Leung and his team at Lawrence Berkeley Laboratory first developed this type of source with a two and a half turn antenna inside the plasma chamber. The antenna was made of 4.7 mm diameter copper tubing and coated with porcelain. It was powered with a

2 MHz, 50 kW RF power supply. The antenna was water-cooled. They obtained a H^- current of about 40 mA which could be increased to about 90 mA by adding caesium. The lifetime is limited to a few weeks due to antenna erosion.

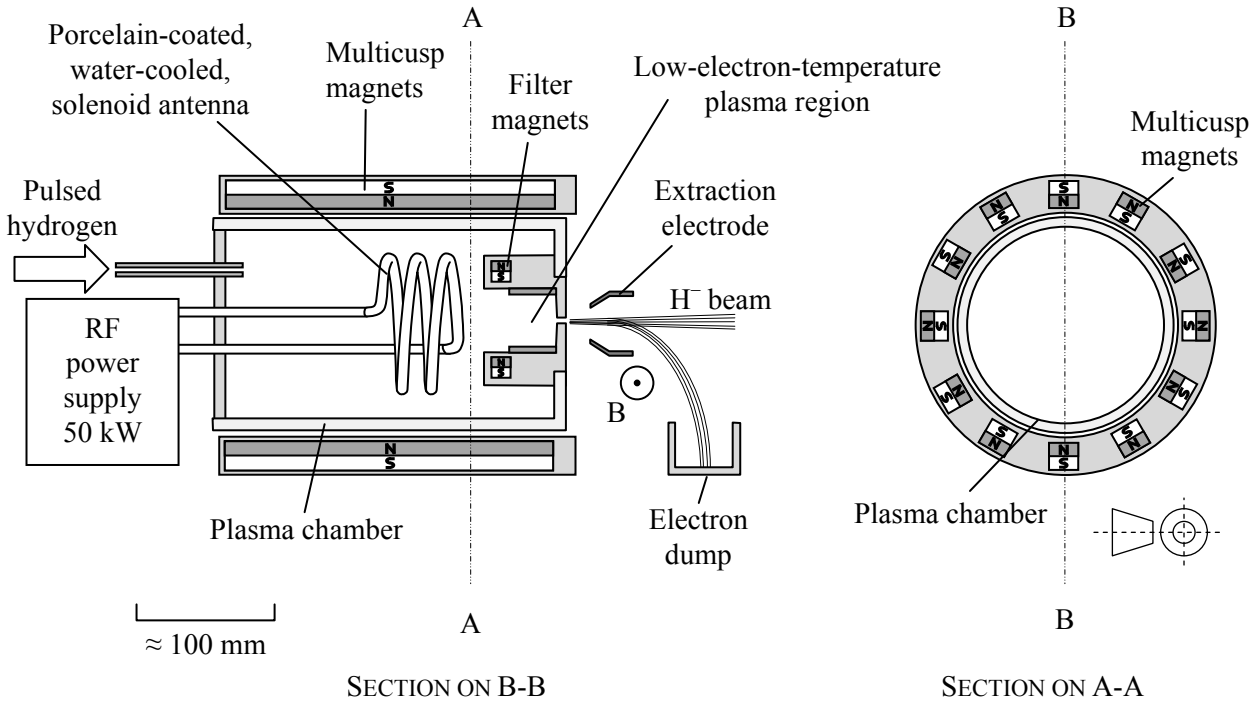


Fig. 21: Schematic of an internal RF antenna multicusp volume source.

In the 2000s Martin Stockli and his team at Oak Ridge National Laboratory developed this source for SNS operations. This source now routinely injects 50 mA, 1 ms, H^- pulses at 60 Hz into the RFQ, of which 38 mA are accelerated by the LINAC [31]. Source lifetimes are of the order of 4–5 weeks. Failures are eventually caused by hot spots occurring on the antenna which melts the 0.6 mm porcelain coating.

During start up, approximately 3 mg of caesium is introduced into the discharge by heating caesium chromate cartridges. No more caesium is added for the lifetime of the source [32].

5.8 External RF antenna multicusp volume source

The problem of antenna failure in multicusp sources can be avoided by putting the antenna outside the plasma chamber. Jens Peters and his team at DESY [33] were the first to successfully try this for negative ions in the late 1990s. They fed 50 kW RF into a three-turn solenoid antenna outside an Al_2O_3 ceramic chamber and obtained a 40 mA H^- beam with a duty factor of 0.05% and a pulse length of 100 μs without caesium. The source ran for a record breaking 300 days. They also experimented with different number of turns on the antenna and different frequencies ranging from 1.65 MHz to 9 MHz. The optimum appeared to be about five turns and 2 MHz.

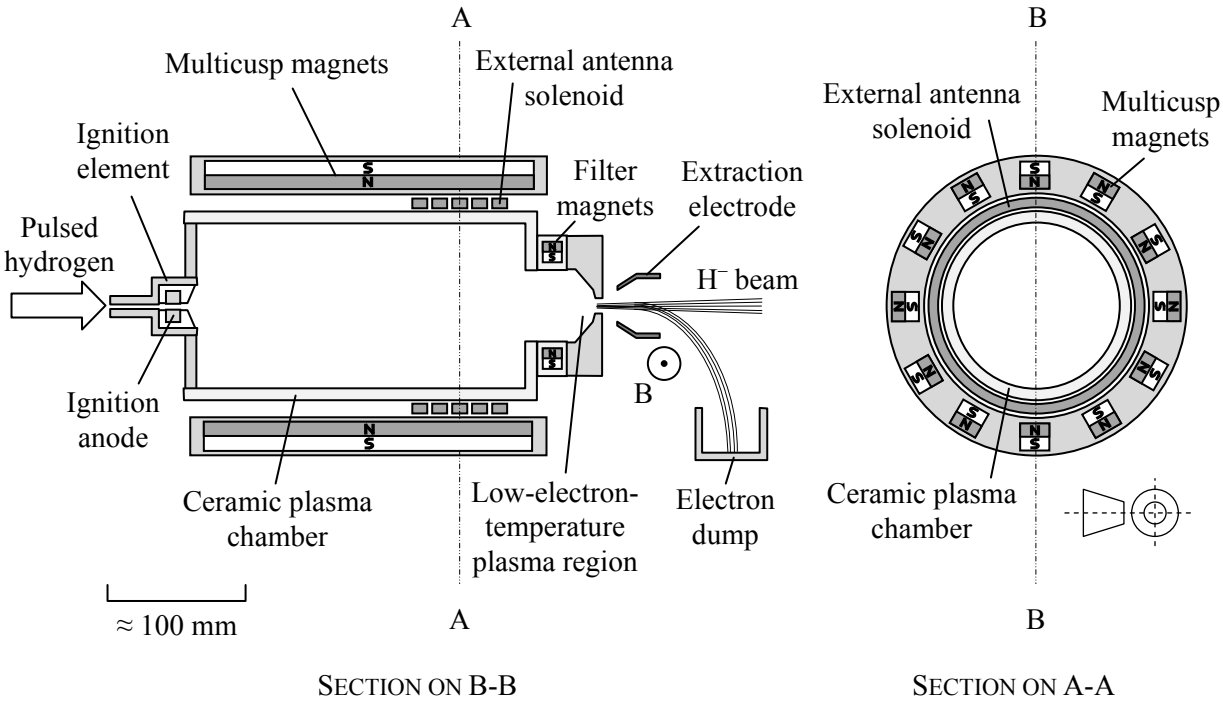


Fig. 22: Sectional schematic of an external antenna RF multicusp volume ion source

This epic lifetime inspired Oak Ridge National Laboratory to start developing an external antenna source for the SNS in the mid 2000s. The DESY source ran at very low repetition rate and duty cycle. SNS requires 1 ms beam pulses at 60 Hz: two orders of magnitude greater. When they tried to scale up the duty cycle they found that they could not extract high enough beam currents. Rob Welton and the SNS ion source group are currently developing this source. Experiential caesiated sources on a test rig have demonstrated unanalysed currents up to 100 mA (see Ref. [34]), but lacked persistence at the required duty factor. External helicoil and saddle antennas have been tested in an attempt to increase the plasma density near the extraction region.

In the late 2000s Jacques Lettry and his team at CERN also started developing an external solenoid antenna multicusp plasma generator. As of 2011 a stable 2 MHz, 1.2 ms, 50 Hz plasma has been produced, but extraction has yet to be demonstrated [35].

Figure 22 shows a five-turn external solenoid antenna multicusp volume source. The pulsed hydrogen is pre-ionized by a spark gap on injection into the plasma chamber. A filter field is positioned in front of the extraction aperture to provide a volume production region.

6 Running and developing sources

6.1 Which source?

The type of source an accelerator uses obviously depends on what types of ions are required and what beam current is needed. The reason why one type of source in a family is used and not another is usually historic. It depends on when the accelerator was built, what facility was there previously and what expertise is available. Cost can sometimes be an issue as well.

CERN use a duoplasmatron because they need a 300 mA proton current. Fermilab use a magnetron because it was the best source that met their needs when it was developed in the late 1970s.

Similarly RAL uses a Penning source because it was the best source that met their needs in the early 1980s.

If you were building a new machine today which source would you choose? For proton sources up to 100 mA the microwave discharge ion source is the obvious choice because it offers exceptional lifetime and reliability. For proton sources up to 500 mA, the only option is the duoplasmatron. For heavy or multiply charged ions the ECR ion source is the best option. For high-charge-state ions, the EBIS is the best option.

The question is much more interesting when considering negative ion sources. Intense development work continues at all of the major labs operating H^- ion sources.

External RF antenna multicusp volume sources offer the promise of great reliability and long lifetimes (>1 year), but have only been shown to work on test stands at very low duty factors (0.05 %) at 40 mA. Much development work is going into external antenna RF sources at SNS and CERN, only time will tell what performance is ultimately achievable.

Internal RF antenna volume sources have demonstrated high currents (>100 mA) at low duty factors (0.1 %) and 50 mA at 6 % duty factors, however in both cases lifetime is limited to a few weeks due to antenna wear.

Surface converter multicusp sources have demonstrated DC beams at reasonable H^- beam currents of 20 mA but only with lifetimes of a few weeks and very long setup times (10 hours). Development sources have demonstrated 120 mA but with even shorter lifetimes.

Magnetrons have demonstrated high currents (80 mA) and very long lifetimes (>6 months) but only at very low duty factors (<0.5 %).

Penning sources offer high currents up to 60 mA (170 mA on experimental sources) and long duty cycles up to DC, however lifetime is limited to a few weeks.

Hot cathode multicusp volume sources have experimentally delivered DC currents up to 40 mA DC, and this current can be doubled with the addition of caesium. Filament lifetimes are short for high currents, but maintenance is very easy.

6.2 Power supplies

Ion sources present particular challenges for power supplies: they must deliver stable high currents to a plasma load that is often unstable. They must deliver stable high voltages to extraction electrodes that often breakdown. All of the electrodes that the power supplies are connected to are often in close proximity and often coated with caesium that increases the probability of sparking especially in an environment full of charge carriers in the presence of strong magnetic fields. All of these factors mean that the power supplies must not only be very robust and resistant to breakdown, but also very stable. Digital control electronics is susceptible to errors in the event of inevitable breakdowns, so analogue control circuitry is often a better choice.

6.3 Control systems

All of the ion sources discussed in this paper have to operate floating on a high-voltage platform, so it is essential to have some form of isolated control and measurement system to allow source tuning during operation and for the provision of timing signals. This is usually done over fibre optics, but some much older sources use insulated mechanical linkage to change power supply settings. The control system is invariably microprocessor based and must be very well isolated from all of the power supplies and housed in a well-screened chassis.

6.4 Developing a source

Most of the sources discussed in this paper are still being developed by labs all over the world. Beam currents are being increased, duty cycles extended, emittances reduced and reliabilities improved. Modern finite-element modelling and computational fluid dynamics allow the electrical, magnetic and thermal operation of the source to be investigated to a detail never before possible. Beam transport and plasma codes allow extraction and beam formation to be studied. All of these computer modelling tools allow ion source development to progress without having to go through as many prototype iterations.

It must be remembered, however, that ion sources and plasmas are incredibly complex: they exhibit numerous emergent behaviours that could never be predicted by simulation alone. The only way to find out how a new source design will behave is to actually test it. This is why development rigs or test stands are essential to designing new sources. These test rigs should replicate the actual environment where the source will run; ideally it should also include a LEBT to test exactly what beam can be transported.

A development test rig must be equipped with as many diagnostics as possible to try to understand how the source is performing.

These could include:

- beam current, e.g. toroids, Faraday cups;
- emittance, e.g. slit–grid, pepperpot, slit–slit, Alison electric sweep scanner;
- profile, e.g. scintillator, wire scanner, laser wire scanner;
- energy spread, e.g. retarding potential energy analyser;
- optical spectroscopy;
- Langmuir probes.

Sources must also run 24 hours a day if lifetimes are to be tested. Even then the true performance of source will not be known until it runs on an operational machine for several years, exposed to variable conditions and the inevitable human error.

7 Summary and conclusions

Meeting the beam current, pulse length and emittance required by the accelerator is only part of the job of an ion source. Operational sources must be reliable and they must have a lifetime that is compatible with the operating schedule of the accelerator. They must be easy to maintain so that in the event of a failure they can be easily fixed. If they have to be replaced they should be easy to dismantle. The start-up procedure should be made as easy and quick as possible.

Ion sources are a very interesting and stimulating area to work in. Sources cover a huge range of different disciplines, requiring skills in both engineering and physics. It can take a lifetime to become an expert in just one type of source. This paper has provided an introduction to some of the most common sources in use today in high-power hadron accelerators.

For further reading see the bibliography given in the Appendix. Huashun Zhang's book, *Ion Sources*, contains comprehensive references for every type of ion source discussed in this paper.

Acknowledgements

I wish to thank Marthe Bacal, Jim Alessi, Martin Stockli, Rod Keller, Reinard Becker, Jacques Lettry, Raphael Gobin, Andrew Holmes, Joe Sherman, Alan Letchford, Scott Lawrie and Melisa Akdoğan for their comments and assistance during the writing of this paper.

References

[1] F. Paschen Wiedemann, *Ann. Phys. Chem.* **37** (1889) 69–96.

[2] A.J. Dempster, *Phys. Rev.* **8** (1916) 651–662.

[3] D. Leitner, *et al.*, *Rev. Sci. Instrum.* **79** (2008) 02C710.

[4] N. Sakudo, *Rev. Sci. Instrum.* **49** (1978) 940.

[5] J. Ishikawa, Y. Takeiri and T. Takagi, *Rev. Sci. Instrum.* **55** (1984) 449.

[6] T. Taylor and J.F. Mouris, *Nucl. Instrum. Methods* **336** (1993). 1-5.

[7] J. Sherman *et al.*, *Rev. Sci. Instrum.* **69** (1998) 1003.

[8] R. Gobin, *et al., Rev. Sci. Instrum.* **73** (2002) 922–924.

[9] O. Tuske, *et al., Rev. Sci. Instrum.* **79** (2008) 02B710.

[10] A. Pikin, *et al., JINST* **5** (2010) C09003.

[11] N. Alekseev, *et al.,* TUPSA020, Proc. of RuPAC2010.

[12] S. Humphries, *et al., J. Appl. Phys.* **59** (1986) 1790.

[13] U. Ratzinger, TUZF204, Proc. of EPAC 2000.

[14] B.L. Donnally, *Phys. Rev.* **159** (1967) 87.

[15] K.W. Ehlers, *et al.*, *Nucl. Instrum. Methods* **22** (1963) 87-92.

[16] K.W. Ehlers, *Nucl. Instrum. Methods* **32** (1965) 309-316.

[17] G.P. Lawrence, *et al.*, *Nucl. Instrum. Methods* **32** (1965) 357-359.

[18] L.E. Collins and R.H. Gobbett, *Nucl. Instrum. Methods* **35** (1965) 277-282.

[19] V.E. Krohn, Jr, *J. Appl. Phys.* **33** (1962) 3523-3525.

[20] J.B. Taylor and I. Langmuir, *Phys. Rev.* **51** (1937) 753-760.

[21] Yu.I. Belchenko, G.I. Dimov and V.G. Dudnikov, *Nucl. Fusion* **14** (1974) 113.

[22] V.G. Dudnikov, Proc. 4th All-Union Conf. on Charged Part. Accel., Moscow, 1974, (Nauka, Moscow, 1975), Vol. 1. pp. 323-325.

[23] H. Vernon Smith, *et al., Rev. Sci. Instrum.* 65, 123 (1994).

[24] D.C. Faircloth *et al., AIP Conf. Proc.* **1390** (2011) 205-215.

[25] A. Ueno, *et al., Rev. Sci. Instrum.* **75** (2004) 1714.

[26] K.N. Leung and K.W. Ehlers, *Rev. Sci. Instrum.* **53** (1982) 803.

[27] K.F. Johnson, *et al.,* THP114, Proc. of LINAC2010.

[28] K.N. Leung, K.W. Ehlers and M. Bacal, *Rev. Sci. Instrum.* **54** (1983) 56-61.

[29] A.J.T. Holmes *et al., Rev. Sci. Instrum.* **65** (1994) 1153.

[30] K. Volk, *et al.*, TH4057*,* Proc. of LINAC 1998.

[31] B.X. Han, *et al., AIP Conf. Proc.* **1390** (2011) 216-225.

[32] M.P. Stockli, *et al*., *Rev. Sci. Instrum.* **81** (2010) 02A729.

[33] J. Peters, TUP061, Proc. of LINAC 2006.

[34] R.F. Welton, *et al., AIP Conf. Proc.* **1390** (2011) 226-234.

[35] J. Lettry, *et al., Rev. Sci. Instrum.* **81** (2010) 02A723.

Appendix: Bibliography

A.1. Papers

S. Nikiforov *et al.*, *Ion Sources for use in Research and Applied High Voltage Accelerators,* Proceedings of PAC95, vol. 2, pp. 1004-1006 (1995).

C.E. Hill, *Ion and Electron Sources*, Proc. CERN Accelerator School, La Hulpe, Belgium, CERN-94-36 (1994).

R. Scrivens, *Electron and Ion Sources for Particle Accelerators,* Proc. CERN Accelerator School, Zeuthen, Germany, pp. 494-504 (2003).

R. Scrivens, *Proton and Ion Sources for High Intensity Accelerators*, Proceedings of EPAC04, Lucerne, Switzerland, pp. 103-107 (2004).

L. Celona *et al., Microwave to plasma Coupling in Electron Cyclotron Resonance and Microwave Ion Sources,* Rev. Sci. Instrum. 81, 02A333 (2010).

S. Gammino *et al., Review on High Current 2.45 GHz Electron Cyclotron Resonance Sources,* Rev. Sci. Instrum. 81, 02B313 (2010).

R. Becker and O. Kester, *Electron Beam Ion Source And Electron Beam Ion Trap*, Rev. Sci. Instrum. 81, 02A513 (2010).

C. W. Schmidt, *Historical Perspective of the H^- Ion Source Symposia*, AIP Conf. Proc. Volume 439, 254-258 (1998).

R. Keller, *High-intensity Ion Sources for Accelerators with Emphasis on H^- Beam Formation and Transport*, Rev. Sci. Instrum. 81, 02B311 (2010).

D.P. Moehs *et al.*, *Negative Hydrogen Ion Sources for Accelerators,* IEEE Trans. Plasma Sci. 33(6), 1786-1798 (2005).

C.W. Schmidt, *Review of Negative Hydrogen Ion Sources,* Proceedings of LINAC 90, Albuquerque, NM (1990).

M. Bacal, A. Hatayama, J. Peters, *Volume Production Negative Ion Sources*, IEEE Trans. Plasma Sci., 33(6), 1845-1871 (2005).

M. Bacal, *Physics Basis and Future Trends for Negative Ion Sources*, Rev. Sci. Instrum. 79, 02A516 (2008).

M. Bacal, *Physics Aspects of Negative Ion Sources,* Nucl. Fusion 46, S250–S259 (2006).

J. Peters, *Review of High Intensity H^- Sources And Matching to High Power RFQ´S,* Proceedings of EPAC 2000, Vienna, Austria, pp. 113-117.

J. Peters, *New Developments in RF and Filament Volume H^- Ion Sources for Accelerators,* Rev. Sci. Instrum. 75(5), 1709-1713 (2004).

J. Peters, *Negative Ion Sources for High Energy Accelerators,* Rev. Sci. Instrum. 71(2), 1069-1074 (2000).

J. Peters, *New Developments in Multicusp H^- Ion Sources for High Energy Accelerators,* Rev. Sci. Instrum. 79, 02A515 (2008).

A.2. Internet

M.P. Stockli, *Ion Source 101*, Internet, 2001.

Wikipedia.

A.3. Books

Huashun Zhang, *Ion Sources* (Science Press, 1999).

Ian G Brown, *The Physics and Technology of Ion Sources* (Wiley-VCH, 2004).

Bernhard Wolf, *Handbook of Ion Sources* (CRC Press, 1995).

Collimators

Slawomir Wronka
National Centre for Nuclear Research, Otwock, Poland

Abstract
The collimator system of a particle accelerator must efficiently remove stray particles and provide protection against uncontrolled losses. In this article, the basic design concepts of collimators and some realizations are presented.

1 Introduction

In all types of linear and circular accelerators, collimators are required to narrow the beam of particles. Owing to differences in the construction of the various types of accelerators, there are various approaches to beam collimation. In linacs, it is important to collimate the beam before the target or before or within the transfer line. In this type of machine, the beam interacts with the collimating system only once. In contrast, in synchrotrons and accumulator rings, the collimating system affects the beam parameters continuously and the proper selection of collimator locations is a more complicated problem.

Historically, collimators have been used in *hadron machines* to reduce the radiation background at the experimental sites. However, new machines, owing to the high energy and high luminosity of the beam and also the use of superconducting technologies, require sophisticated collimation systems for beam cleaning and machine protection. In modern accelerators, high intensities occur not only in the beam core but also in the beam halo created by particle migration due to collisions, beam–gas interactions, or nonlinearities in the magnetic fields. Lost particles originating from beam tails create uncontrolled losses, emittance growth, activation of accelerator components, heat deposition, and potential quenches in superconducting magnets. Properly designed collimator sections allow controlled, cumulative deposition of the losses in well-known, prepared locations, and thus minimize the impact of radiation on equipment.

In general, and regardless of the type of machine (synchrotron or linac), the objectives of collimators can be defined as follows:

- to obtain low uncontrolled beam loss;
- to minimize the halo around the proton or ion beam;
- to minimize the activation of downstream beam line components;
- to allow faster access to the machine and to experimental sites;
- to protect the machine itself against damage.

2 Types of collimators

The typical components of the collimation system can be categorized by their function, as follows.

2.1 Jaws

These typically consist of two solid blocks and are used for efficient beam cleaning. The thickness is high enough to stop impacting particles, although the production of secondary particles is possible (see Fig. 1). The correct selection of the material is important to avoid damage in the case of high power

deposition. For ultrasmall beams (as in the LHC; see Fig. 2), high precision and stringent tolerances are required. Primary jaws are often followed by secondary and tertiary collimators.

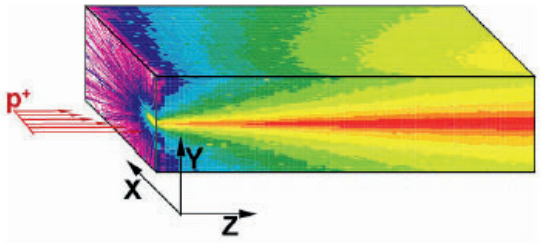


Fig. 1: Schematic drawing of a collimator jaw [1]

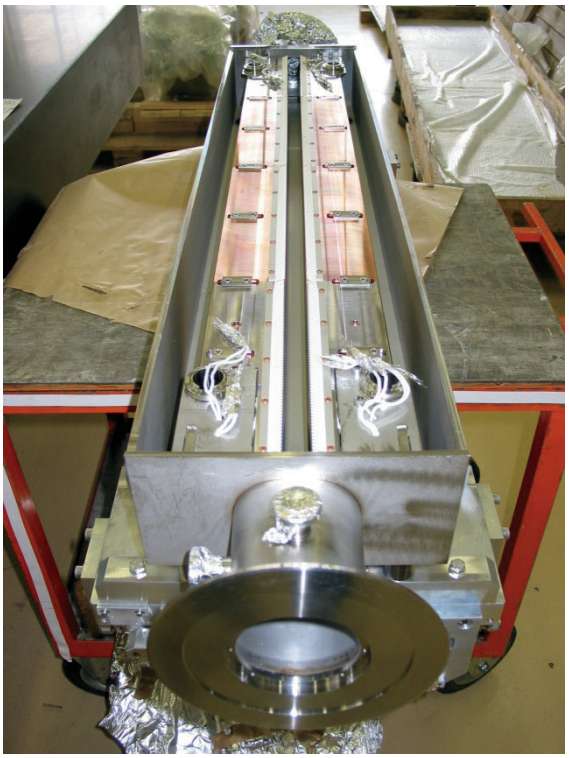


Fig. 2: Jaws of the LHC accelerator [2]

A heavy charged particle traversing matter loses energy primarily through the ionization and excitation of atoms [3]. The energy transferred may be sufficient to knock an electron out of an atom and thus ionize it, or it may leave the atom in an excited, non-ionized state. Because the mass of a heavy charged particle is thousands of times larger than the electron mass, the particle can transfer only a small fraction of its energy in a single electronic collision. The deflection of the particle in the collision is negligible. Thus, a heavy charged particle travels in an almost straight path through matter, losing energy almost continuously in small amounts through collisions with atomic electrons, leaving ionized and excited atoms in its wake. The average linear rate of energy loss of a heavy charged particle in a medium (expressed, for example, in $\mathrm{MeV \cdot cm^{-1}}$) is of fundamental importance in radiation physics and dosimetry. This quantity, denoted by $-\mathrm{d}E/\mathrm{d}x$, is called the stopping power of the medium for the particle.

The range of a charged particle is the distance it travels before coming to rest. The reciprocal of the stopping power gives the distance travelled per unit energy loss. Therefore, the range $R(T)$ of a particle of kinetic energy T is the integral of this quantity down to zero energy:

$$R(T) = \int_0^T \left(-\frac{\mathrm{d}E}{\mathrm{d}x}\right)^{-1} \mathrm{d}E. \quad (1)$$

In practice, Monte Carlo codes are used to calculate the minimum length required for a set of jaws, for example MCNP [4], Fluka [5], and Geant4 [6].

2.2 Scrapers

Thin objects (e.g. foils) are typically used for beam shaping and diagnostics. Scrapers can be used together with magnets located in the beam line. For example, in the case of H^- particles, scrapers change the charge and the magnets then naturally bend the particles out of the beam line [7]. An example of a scraper is presented in Fig. 3.

Fig. 3: Scraper at the Spallation Neutron Source (SNS) [1]

2.3 Absorbers

Movable absorbers can be quite similar in design to jaws. The main goal of these elements is to absorb miskicked beam or products of particle-induced showers. In comparison with jaws, larger gaps and relaxed tolerances are used.

2.4 Additional equipment for precise set-up and alignment

Beam loss monitors (BLMs) are used for precise alignment of collimators. When a jaw or scraper 'touches' the beam, a stronger secondary shower is produced, which is detected by the BLMs. Such a procedure can be repeated for each element and each side of the beam.

3 Practical realization of collimator systems

Typically, the whole of a collimation system must be carefully calculated and designed. It is important to define the length of the collimating tube; the shape, location, and number of collimators (along the entire lattice and in each set); and the aperture size of the primary, secondary, and tertiary collimators. The material of the collimators must effectively stop the particles, survive the resulting heat load, and receive as low an activation as possible. Systematic material studies must be performed to verify the ability of materials to withstand thermomechanical shock and to measure other parameters, such as

thermal expansion coefficients and radiation resistance for long exposures. Finally, the requirements for a stationary or movable collimator imposed by the design of the accelerator or detector must be taken into account; for example, one must analyse if any movement is required and, if so, what algorithm will be used to control the movement.

In such calculations and engineering design work, one has to take the following into account:

- the beam power;
- all potential losses and beam halo;
- showers in the collimators and other equipment, and electron clouds;
- material behaviour and beam-induced damage, and elastic and inelastic deformation;
- the possible use of coatings on the base material to modify its parameters;
- the need for precise mechanical movement and highly efficient cooling;
- the radioactivity levels in the collimator regions (which affect both the materials and personnel);
- tolerance requirements;
- a risk analysis of the potential failure scenarios (for example, the particle trajectories may change as a result of a klystron misfire or a magnet failure).

Such projects are almost always very complicated and require tight co-operation among the people responsible for different accelerator and detector sections. For example, the activation of the collimators themselves and of downstream elements, together with the shielding requirements for each collimation section, must be taken into account (see [8, 9] for a description of an example). It is necessary to study a 'distributed collimation' approach, where small collimators are located in many places sandwiched between other elements, and compare it with a 'bulk collimation' philosophy, in which beam collimation is done at only two or three locations. The output of such complex studies should define the collimator geometry, the collimation material, and the cooling requirements for the various levels of intercepted power that will be encountered.

The future development of collimators will take account of the possibility of using new materials to achieve longer lifetimes in an ultrahigh-radiation environment and under high thermal and mechanical stresses. Some new concepts are based on multiuse collimators, an example of which is presented in Fig. 4.

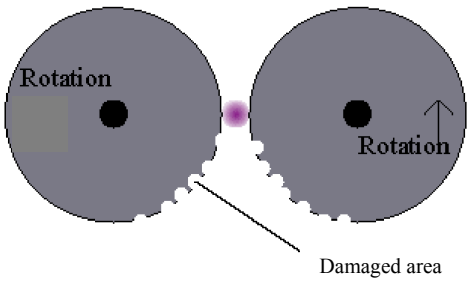


Fig. 4: Concept of a multiuse rotating-wheel collimator. This is an example of a 'consumable spoiler' to be used at the SLAC NLC [9, 10].

4 Summary

A general introduction to collimators, which are very important elements of accelerators, has been presented in this paper. A short description of different types of collimators and their roles was given, with examples from the LHC and SNS. In modern high-power hadron machines, collimators are of principal importance for ensuring the safety of the machine itself and of the detectors downstream of the beam, as well as being an important part of the radiation safety system.

References

[1] N. Simos *et al*., Experimental studies of targets and collimators for high intensity beams, Proc. HB2006, Tsukuba, Japan.

[2] LHC Collimation Project, http://lhc-collimation-project.web.cern.ch/lhc-collimation-project/images/IMG_3377.JPG

[3] J. Turner, *Atoms, Radiation and Radiation Protection* (Wiley, New York, 1995), Chapter 5.

[4] Los Alamos National Laboratory, https://laws.lanl.gov/vhosts/mcnp.lanl.gov/

[5] FLUKA, http://www.fluka.org/fluka.php

[6] S. Agostinelli *et al*., Geant4 – a simulation toolkit, *Nucl. Instrum. Methods A* **506** (2003) 250–303.

[7] H. Ludewig *et al*., Integration of the beam scraper and primary collimator in the SNS ring, Proc. 2003 Particle Accelerator Conference, Portland, Oregon, USA.

[8] I. Popova and F.X. Gallmeier, Full-scale radiation dose analyses for the SNS accelerator system, Supercomputing in Nuclear Applications (SNA-2003), Paris, 2003.

[9] I. Popova et al., Residual dose rate analyses for the SNS accelerator facility, Proc. Hadron Beam 2008, Nashville, TN.

[10] A. Seryi, Seminar at CERN, 3 May 2005, http://clic-meeting.web.cern.ch/clic-meeting/2005/05_03as.pdf

Radiation protection at CERN

Doris Forkel-Wirth, Stefan Roesler, Marco Silari, Marilena Streit-Bianchi, Christian Theis, Heinz Vincke, and Helmut Vincke
CERN, Geneva, Switzerland

Abstract
This paper gives a brief overview of the general principles of radiation protection legislation; explains radiological quantities and units, including some basic facts about radioactivity and the biological effects of radiation; and gives an overview of the classification of radiological areas at CERN, radiation fields at high-energy accelerators, and the radiation monitoring system used at CERN. A short section addresses the ALARA approach used at CERN.

1 Introduction

CERN's radiation protection policy stipulates that the exposure of persons to radiation and the radiological impact on the environment should be as low as reasonably achievable (the ALARA principle), and should comply with the regulations in force in the Host States and with the recommendations of competent international bodies. This paper gives a brief overview of the general principles of radiation protection legislation; explains radiological quantities and units, including some basic facts about radioactivity and the biological effects of radiation; and gives an overview of the classification of radiological areas at CERN, radiation fields at high-energy accelerators, and the radiation monitoring system used at CERN. Finally, a short section addresses the ALARA approach used at CERN.

2 General principles of radiation protection legislation

The International Commission on Radiological Protection (ICRP) has specified in its Recommendation 60 [1] that any exposure of persons to ionizing radiation should be controlled and should be based on three main principles, namely:

- *justification:* any exposure of persons to ionizing radiation has to be justified;
- *limitation:* personal doses have to be kept below legal limits;
- *optimization:* personal and collective doses have to be kept as low as reasonably achievable (ALARA).

These recommendations have been fully incorporated into CERN's radiation safety code [2].

3 Radiological quantities and units [3]

It would be desirable if the legal protection limits could be expressed in directly measurable physical quantities. However, this does not allow the biological effects of exposure of the human body to ionizing radiation to be quantified. For this reason, protection limits are expressed in terms of so-called protection quantities, which, although calculable, are not measurable. Protection quantities quantify the extent of exposure of the human body to ionizing radiation from both whole-body and partial-body external irradiation and from the intake of radionuclides. In order to demonstrate

compliance with dose limits, so-called operational quantities are typically used, which are aimed at providing conservative estimates of protection quantities. The radiation protection detectors used for individual and area monitoring are often calibrated in terms of operational quantities.

3.1 Physical quantities

The *fluence* Φ (measured in units of $1/m^2$) is the quotient of dN by da, where dN is the number of particles incident upon a small sphere of cross-sectional area da:

$$\Phi = \frac{\mathrm{d}N}{\mathrm{d}a}. \tag{1}$$

In dosimetric calculations, the fluence is frequently expressed in terms of the lengths l of particle trajectories. It can be shown that the fluence is also given by

$$\Phi = \frac{\mathrm{d}l}{\mathrm{d}V}, \tag{2}$$

where dl is the sum of the particle trajectory lengths in the volume dV.

The *absorbed dose D* (measured in units of grays; 1 Gy = 1 J/kg = 100 rad) is the energy imparted by ionizing radiation to a volume element of a specified material divided by the mass of that volume element.

The *kerma K* (measured in units of grays) is the sum of the initial kinetic energies of all charged particles liberated by indirectly ionizing radiation in a volume element of a specified material divided by the mass of that volume element.

The *linear energy transfer L* or LET (measured in units of J/m, but often given in keV/μm) is the mean energy dE lost by a charged particle owing to collisions with electrons in traversing a distance dl in matter. Low-LET radiation (L < 10 keV/μm) comprises X-rays and gamma rays (accompanied by charged particles due to interactions with the surrounding medium), and light charged particles such as electrons that produce sparse ionizing events far apart on a molecular scale. High-LET radiation (L > 10 keV/μm) comprises neutrons and heavy charged particles that produce ionizing events densely spaced on a molecular scale.

The *activity A* (measured in units of becquerels; 1 Bq = 1/s = 27 pCi) is the expectation value of the number of nuclear decays in a given quantity of material per unit time.

3.2 Protection quantities

The *organ absorbed dose* D_T (measured in units of grays) in an organ or tissue T of mass m_T is defined by

$$D_T = \frac{1}{m_T} \int_{m_T} D \, \mathrm{d}m. \tag{3}$$

The *equivalent dose* H_T (measured in units of sieverts; 1 Sv = 100 rem) in an organ or tissue T is equal to the sum of the absorbed doses $D_{T,R}$ in an organ or tissue caused by different radiation types R weighted by so-called radiation weighting factors w_R:

$$H_T = \sum_R w_R * D_{T,R}. \tag{4}$$

This expresses the long-term risks (primarily cancer and leukaemia) from low-level chronic exposure. The values of w_R recommended by the ICRP [4] are unity for photons, electrons, and muons, 2.0 for protons and charged pions, 20.0 for ions, and a function of energy for neutrons (of energy E_n):

$$w_R = \begin{cases} 2.5 + 18.2 * \mathrm{e}^{-\left[\frac{\ln(E_\mathrm{n})^2}{6}\right]} & \text{if } E_\mathrm{n} < 1\ MeV, \\ 5.0 + 17.0 * \mathrm{e}^{-\left[\frac{\ln(2*E_\mathrm{n})^2}{6}\right]} & \text{if } 1\ \mathrm{MeV} < E_\mathrm{n} < 50\ MeV, \\ 2.5 + 3.25 * \mathrm{e}^{-\left[\frac{\ln(0.04*E_\mathrm{n})^2}{6}\right]} & \text{if } E_\mathrm{n} < 50\ MeV. \end{cases} \quad (5)$$

The *effective dose E* (measured in units of sieverts) is the sum of the equivalent doses, weighted by the tissue weighting factors w_T (where $\sum_T w_T = 1$), for several organs and tissues T of the body that are considered to be the most sensitive [4]:

$$E = \sum_T W_T * H_T. \quad (6)$$

3.3 Operational quantities

The *ambient dose equivalent H**(10) (measured in units of sieverts) is the dose equivalent at a point in a radiation field that would be produced by a corresponding expanded and aligned field in a 30 cm diameter sphere of tissue of unit density at a depth of 10 mm, on the radius vector opposite to the direction of the aligned field. The ambient dose equivalent is the operational quantity for area monitoring.

The *personal dose equivalent* $H_\mathrm{p}(d)$ (measured in units of sieverts) is the dose equivalent in standard tissue at an appropriate depth d below a specified point on the human body. The specified point is normally taken to be where an individual dosimeter is worn. The personal dose equivalent $H_\mathrm{p}(10)$, with a depth d = 10 mm, is used for the assessment of the effective dose, and $H_\mathrm{p}(0.07)$, with d = 0.07 mm, is used for the assessment of doses to the skin and to the hands and feet. The personal dose equivalent is the operational quantity for monitoring of individuals.

3.4 Dose conversion coefficients

Dose conversion coefficients allow the direct calculation of protection or operational quantities from the particle fluence and are functions of the particle type, energy, and irradiation configuration. The most commonly used coefficients are those for the effective dose and ambient dose equivalent. The former are based on simulations in which the dose to organs of anthropomorphic phantoms is calculated for approximate actual conditions of exposure, such as irradiation of the front of the body (antero-posterior irradiation) or isotropic irradiation. Dose conversion coefficients from fluence to effective dose for antero-posterior irradiation are shown in Fig. 1.

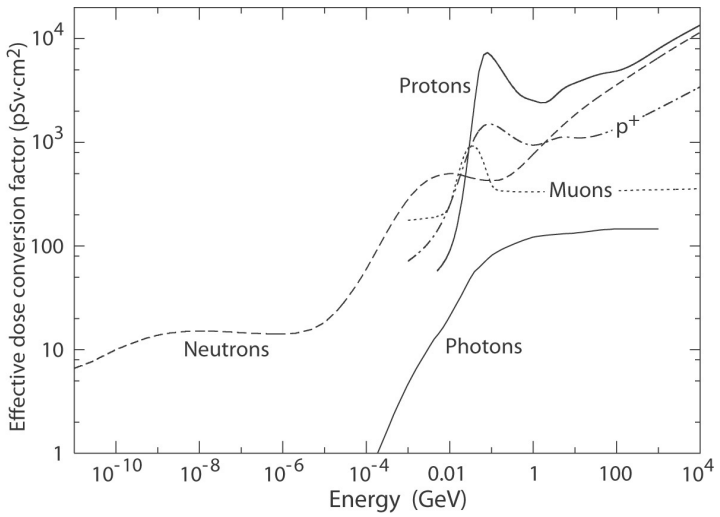


Fig. 1: Conversion coefficients from fluence to effective dose for antero-posterior irradiation

4 Health effects of ionizing radiation

Radiation can cause two types of health effects, deterministic and stochastic.

Deterministic effects are tissue reactions which cause injury to a population of cells if a given threshold of absorbed dose is exceeded. The severity of the reaction increases with dose. The quantity used for tissue reactions is the absorbed dose *D*. When particles other than photons and electrons (low-LET radiation) are involved, a dose weighted by the Relative Biological Effectiveness (RBE) may be used. The RBE of a given radiation is the reciprocal of the ratio of the absorbed dose of that radiation to the absorbed dose of a reference radiation (usually X-rays) required to produce the same degree of biological effect. It is a complex quantity that depends on many factors such as cell type, dose rate, and fractionation.

Stochastic effects are malignant diseases and inheritable effects for which the probability of an effect occurring, but not its severity, is a function of dose without a threshold.

4.1 Biological effects

The biological effect of radiation depends on the type and energy of the radiation (photons, neutrons, protons, heavy nuclei, etc.), on whether the irradiation is external or internal, on whether it is from radionuclides inhaled or ingested, and on the dose and dose rate received. Furthermore, the type of organ irradiated (for example, the bone marrow is much more sensitive than the liver) and whether local or total body irradiation has occurred will strongly affect the severity and outcome of the damage produced. All this explains the need for and use of various weighting factors to derive equivalent and effective doses in radiation protection.

The cascade of reactions and interactions that occurs when radiation hits a biological system is a mixture of direct and indirect effects, each of them occurring on a different time-scale. The damage starts with the direct ionization and excitation of biological molecules or the creation of free radicals, which gives rise to peroxides, and the interaction of these with DNA molecules produces both repairable and non-repairable damage. Breaks in DNA single strands are highly repairable, but the problem is to know how much misrepair will occur for various doses and types of radiation. In fact, misrepair can either induce programmed cell death, called apoptosis, or produce non-lethal mutations. The damage will result either in deterministic effects (cell death, necrosis, or damage to tissues, organs, or the body, etc.) or in stochastic effects. The latter, resulting from non-lethal mutations, may become visible only many years after irradiation as a cancer or, if the germ cells have been affected, it may be transmitted to future generations in the form of inheritable damage.

The dose for which 50% of individuals will die within 30 days after acute irradiation exposure ($LD_{50/30}$) is 2.5 to 4.5 Gy. More recently, the doses for which 10% and 90% of the population may die from acute irradiation have been estimated; these values are 1–2 Gy for LD_{10} and ~5–7 Gy for LD_{90}, respectively.

For each type of deterministic effect (erythraemia, depletion of bone marrow and blood cells, necrosis, vomiting, etc.), there is a dose threshold for the damage to become assessable or visible. The various types of damage observable after acute irradiation, and their dose equivalents are listed in Table 1.

In spite of the long controversy about the presence or absence of damage at extremely low doses less than 0.2 Gy, the absence of a threshold for the stochastic effects is generally accepted. Based on such an assumption, the probability of risk at extremely low doses has been calculated and applied to set occupational and public dose limits for radiation protection. More detailed information about the biological effects of ionizing radiation is given in Ref. [6].

Table 1: Radiation damage to the human body [5]

Dose (whole-body irradiation)	Effects
<0.25 Gy	No clinically recognizable damage
0.25 Gy	Decrease in white blood cells
0.5 Gy	Increasing destruction of leukocyte-forming organs (causing decreased resistance to infections)
1 Gy	Marked changes in the blood (decrease in the numbers of leukocytes and neutrophils)
2 Gy	Nausea and other symptoms
5 Gy	Damage to the gastrointestinal tract causing bleeding and ~50% death
10 Gy	Destruction of the neurological system and ~100% death within 24 h

5 Radiation levels [3]

- *Natural background radiation*. On average, worldwide, the annual whole-body dose equivalent due to all sources of natural background radiation ranges from 1.0 to 13 mSv, with an average of 2.4 mSv [7]. In certain areas, values up to 50 mSv have been measured. A large fraction (typically more than 50%) originates from inhaled natural radioactivity, mostly radon and radon decay products. The dose equivalent due to radon can vary by more than one order of magnitude: it is 0.1–0.2 mSv per year in open areas, 2 mSv per year on average in houses, and more than 20 mSv per year in poorly ventilated mines.
- *Cosmic ray background radiation*. At sea level, the whole-body dose equivalent due to cosmic ray background radiation is dominated by muons; at higher altitudes, nucleons also contribute. The dose equivalent rates range from less than 0.1 μSv/h at sea level to a few μSv/h at aircraft altitudes.
- *Cancer induction*. The cancer induction probability is about 5% per sievert on average for the entire population [4].
- *Lethal dose*. The whole-body dose from penetrating ionizing radiation resulting in 50% mortality in 30 days, assuming no medical treatment, is 2.5–4.5 Gy (RBE-weighted when necessary), as measured internally on the longitudinal centre line of the body. The surface dose varies because of variable body attenuation and may be a strong function of energy.
- *Recommended dose limits*. The ICRP recommends a limit for radiation workers of 20 mSv effective dose per year averaged over five years, with the provision that the dose should not exceed 50 mSv in any single year [4]. The limit in the EU countries and Switzerland is 20 mSv per year; in the US, it is 50 mSv per year (or 5 rem per year). Many physics laboratories in the US and elsewhere set lower limits. The dose limit for the general public is typically 1 mSv per year.

5.1 Radiation levels in Switzerland

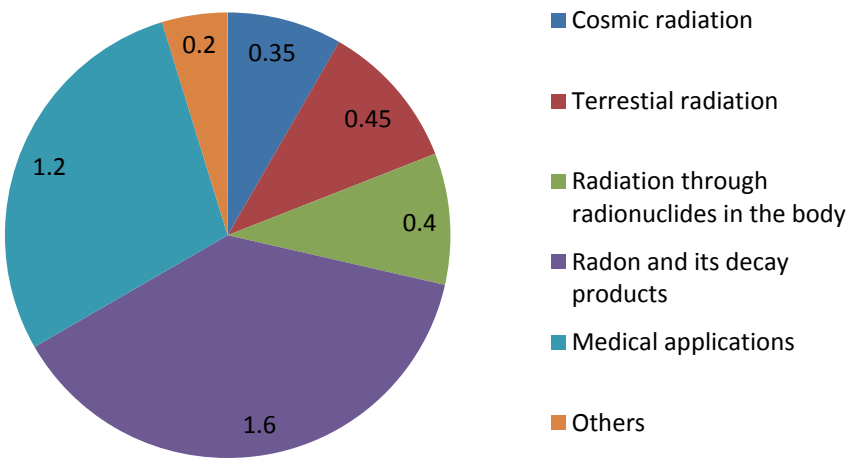


Fig. 2: Mean radiation exposure in Switzerland per year (in mSv) [8]

The contributions to the mean radiation exposure in Switzerland [8] are given in Fig. 2. However, the contribution of radon to the total radiation exposure varies strongly in Switzerland. This is determined mainly by the amount of natural radon (which is a product of the natural decay of uranium and thorium) in the soil. A radon map of Switzerland is shown in Fig. 3.

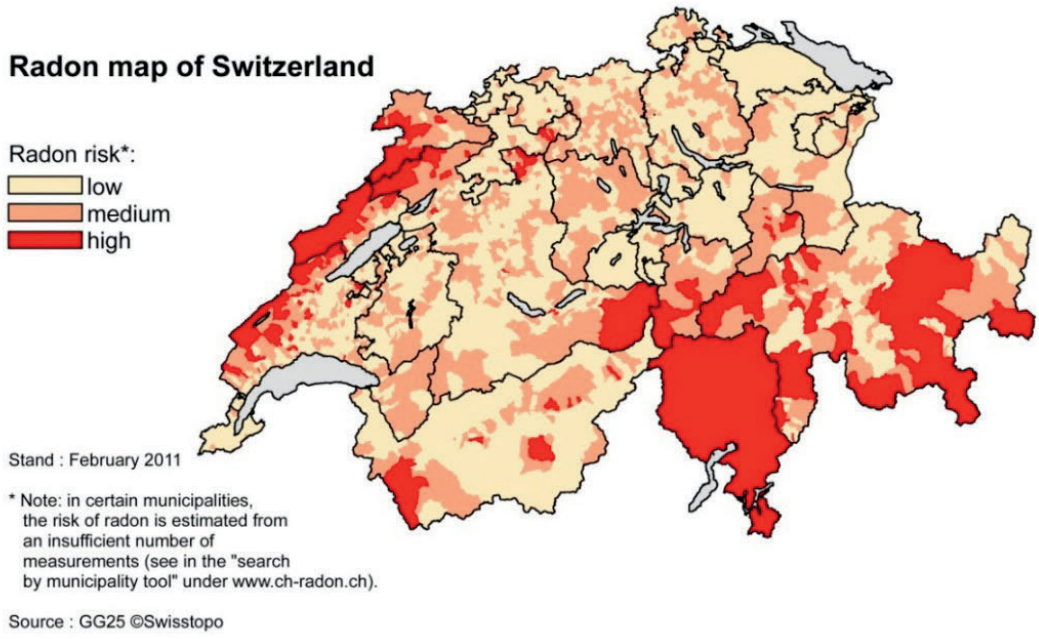


Fig. 3: Radon risk in Switzerland. The map is based on measurements performed in buildings (occupied rooms) [9].

6 Radiological classification of CERN's areas, and dose limits

6.1 Radiological classification at CERN

The areas inside CERN's perimeter are classified as a function of the effective dose a person is liable to receive during his stay in the area under normal working conditions during routine operation. In line with Safety Code F (2006) [2], three types of areas are distinguished:

- Non-designated Areas;
- Supervised Radiation Areas;
- Controlled Radiation Areas.

The latter two are jointly termed Radiation Areas.

The potential external and internal exposures have to be taken into account when assessing the effective dose that persons may receive when working in an area under consideration. Limitation of exposure in terms of effective dose is ensured by limiting an operational quantity, the ambient dose equivalent rate $H^*(10)$ for exposure from external radiation, and by setting action levels for airborne radioactivity and specific surface contamination at the workplace for exposure from incorporated radionuclides. The radiological classification used at CERN is shown in Table 2.

Table 2: Synopsis of the classification of Non-designated Areas and Radiation Areas at CERN

	Area	Dose limit [year]	Ambient dose equivalent rate		Sign
			Work place	Low occupancy	
	Non-designated	1 mSv	0.5 µSv/h	2.5 µSv/h	
Radiation Area	Supervised	6 mSv	3 µSv/h	15 µSv/h	Dosimeter obligatory / Dosimètre obligatoire
Radiation Area	Simple	20 mSv	10 µSv/h	50 µSv/h	Dosimeter obligatory / Dosimètre obligatoire
Radiation Area	Limited Stay	20 mSv		2 mSv/h	LIMITED STAY / SÉJOUR LIMITÉ; Dosimeters obligatory / Dosimètres obligatoires
Radiation Area	High Radiation	20 mSv		100 mSv/h	HIGH RADIATION / HAUTE RADIATION; Dosimeters obligatory / Dosimètres obligatoires
Radiation Area	Prohibited	20 mSv		> 100 mSv/h	PROHIBITED AREA / ZONE INTERDITE; No Entry / Défense d'entrer
Controlled Area					

6.2 Dose limits and classification of workers

All occupationally exposed persons at CERN are classified into one of two categories:

Category A: persons who may be exposed in the exercise of their profession to more than 3/10 of the limit in terms of effective dose in 12 consecutive months;

Category B: persons who may be exposed in the exercise of their profession to less than 3/10 of the limit in terms of effective dose in 12 consecutive months.

The CERN dose limits are compliant with those of most European countries or even more restrictive. Examples of the dose limits in some European countries are given in Table 3.

Table 3: Dose limits at CERN and in some European countries

	Dose limits for 12 consecutive months (mSv)		
	Non-occupationally exposed persons	Occupationally exposed persons	
		Category B	Category A
EURATOM members	1	6	20
Germany and France	1	6	20
CERN	1	6	20
Switzerland	1	20	

6.3 CERN's limits for radionuclides of artificial and natural origin

At CERN, material is considered as radioactive if one or more of the following three criteria are fulfilled.

6.3.1 Specific activity and total activity

CERN's Safety Code F [2] applies to any practice involving material containing radionuclides for which

- the specific activity exceeds the CERN exemption limits [10]; and
- the total activity exceeds the CERN exemption limits [10].

For material containing a mixture of radionuclides of artificial origin, the following sum rule is applied to exempt it from any further regulatory control:

$$\sum_{i=1}^{n} \frac{a_i}{LE_i} < 1, \tag{7}$$

where a_i is the specific activity (Bq/kg) or the total activity (Bq) of the i-th radionuclide of artificial origin in the material, LE_i is the CERN exemption limit for that radionuclide, and n is the number of radionuclides present.

6.3.2 Dose rate

CERN's Safety Code F [2] applies to all materials for which the ambient dose equivalent rate measured at a distance of 10 cm from the item exceeds 0.1 μSv/h after subtraction of the background.

6.3.3 Surface contamination

CERN's Safety Code F [2] applies to all materials for which the surface contamination exceeds 1 Bq/cm^2 in the case of unidentified beta and gamma emitters and 0.1 Bq/cm^2 in the case of unidentified alpha emitters. Once a radionuclide has been identified, specific CERN CS-values [10] can be used, and the following sum rule should be applied:

$$\sum_{i=1}^{n} \frac{c_i}{CS_i} < 1, \tag{8}$$

where c_i is the value of the surface contamination (Bq/cm^2) of the i-th radionuclide, CS_i is its CS-value, and n is the number of identified radionuclides.

7 Induced radioactivity [11]

Neutrons are not affected by the Coulomb barrier of nuclei, and can thus react at any energy and produce radioactive nuclides. Neutron capture dominates for thermal neutrons, whereas reactions of type (n, p), (n, α), (n, 2n), etc. occur with increasing energy. High-energy neutrons cause spallation reactions that can produce any nuclide lighter than the target nucleus.

Charged particles with energies lower than the Coulomb barrier (a few MeV) do not react effectively with nuclei. As soon as the energy exceeds the Coulomb barrier, compound nuclei may be formed, which de-excite by the emission of photons, nucleons, or light nuclei (e.g., in the case of protons, reactions of type (p, n), (p, d), (p, α), etc. can occur). Similarly to neutrons, high-energy charged particles interact by spallation reactions.

Electromagnetic particles may also cause activation through photonuclear interactions, although with a much lower cross-section than for hadronic reactions (at high energy, lower by the fine structure constant). Thus, activation by electrons and photons is typically not a concern at hadron accelerators, whereas it might be important at electron accelerators. The threshold energies for photonuclear reactions are a few MeV, depending on the target material. Just above threshold, so-called giant dipole resonance reactions dominate, in which the nucleus de-excites by the emission of neutrons, protons, and light nuclei.

7.1 Fundamental principles

Radioactive decay is a random process characterized by a decay constant λ. If a total number $N_{\mathrm{tot}}(t)$ atoms of a radionuclide are present at time t, the total activity $A_{\mathrm{tot}}(t)$ is determined by

$$A_{\mathrm{tot}}(t) = \frac{\mathrm{d}N_{\mathrm{tot}}(t)}{\mathrm{d}t} = \lambda N_{\mathrm{tot}}(t), \tag{9}$$

for which the solution at $t = T$ is

$$A_{\mathrm{tot}}(T) = A_{\mathrm{tot}}(0)\mathrm{e}^{-\lambda T}. \tag{10}$$

Often, the time required to decay to half of the original activity, the half-life $t_{1/2}$, is given; this is related to the decay constant by

$$t_{1/2} = \frac{\ln 2}{\lambda}. \tag{11}$$

If we assume steady irradiation of a material with a spatially uniform fluence rate Φ (cm$^{-2}\cdot$s^{-1}), the density of atoms $n(t)$ of the radionuclide of interest per unit volume at time t (cm^{-3}) during the irradiation is governed by

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} = -\lambda n(t) + N\sigma\Phi, \tag{12}$$

where σ is the production cross-section (cm^2) and N is the density of target atoms (cm^{-3}). This equation has the solution

$$n(t) = \frac{N\sigma\Phi}{\lambda}(1 - \mathrm{e}^{-\lambda t}), \tag{13}$$

where the specific activity during irradiation is given by $A(t) = \lambda n(t)$. For $t >> t_{1/2}$, Eq. (13) yields $A(t) = A_{\mathrm{sat}} = N\sigma\Phi$, i.e. the saturation activity equals the production rate.

The activity after an irradiation period t and a cool-down time t_{cool} can be written as

$$A_{tot}(T) = A_{sat}(1 - e^{-t/\tau})e^{-t_{cool}/\tau} \quad , \tag{14}$$

where $\tau = 1/\lambda$.

7.2 Radionuclides in solid materials

The most important medium- and long-lived radionuclides produced in typical accelerator materials are given in Table 4. As can be seen, the heavier the elements in the material are, the greater the number of radionuclides that can be created. Thus, light materials should be preferred if possible in the construction of accelerator components. For example, aluminium supports have better radiological characteristics than steel supports owing to the significantly lower number of nuclides produced.

Reactions with trace elements in materials give rise to additional nuclides which might also be important, especially if they are long-lived. A typical example is ^{60}Co, produced by thermal-neutron capture reactions with traces of cobalt in aluminium or iron components. This nuclide can dominate the activity in a component many years after irradiation, when most other nuclides have already decayed.

The activation properties of the materials used in accelerator construction must be considered during the design process as they may have a direct impact on later handling (maintenance and repair) and waste disposal. Gamma-emitting nuclides dominate the residual dose rates at longer decay times (more than one day), whereas at short decay times β^+ emitters are also important (as a result of photons produced by β^+ annihilation). Owing to their short range, β^- emitters are usually relevant only to doses to the skin and eyes and doses due to inhalation or ingestion.

Figures 4 and 5 show the contributions of gamma and β^+ emitters, respectively, to the total dose rate close to an activated copper sample [12]. Typically, the dose rates at a given decay time are determined mainly by radionuclides with half-lives of the order of the decay time. Extended irradiation periods might be an exception to this general rule, as in this case the activity of long-lived nuclides can build up sufficiently that it dominates over that of short-lived nuclides even at short cooling times.

Activation in concrete is dominated by ^{24}Na (at short decay times) and ^{22}Na (at long decay times). Both of these nuclides can be produced either by low-energy neutron reactions with the sodium component in the concrete or by spallation reactions with silicon and calcium. At long decay times, the nuclides of radiological interest in activated concrete can also include ^{60}Co, ^{152}Eu, ^{154}Eu, and ^{134}Cs, all of which are produced by (n, γ) reactions with traces of natural cobalt, europium, and caesium. Thus, such trace elements might be important even if their content in the concrete is only a few parts per million or less by weight.

Explicit simulation of radionuclide production with general-purpose Monte Carlo codes has become the method most commonly applied to calculate induced radioactivity and its radiological consequences. Nevertheless, other more approximate approaches, such as the use of 'ω-factors' [13], can still be useful for fast order-of-magnitude estimates. These ω-factors give the dose rate per unit star density (the density of inelastic reactions above a certain energy threshold, e.g. 50 MeV) in contact with an extended, uniformly activated object after 30 days of irradiation and one day of decay. The ω-factor for steel or iron is approximately 3×10^{-12} Sv cm^3/star. This does not include possible contributions from thermal-neutron activation.

Table 4: Nuclides of radiological importance in the elements of typical accelerator materials. The last column indicates the half-life.

Element or material	Nuclide	$t_{1/2}$
Carbon	^{3}H	12.3 y
	^{7}Be	53.29 d
	^{11}C	20.38 min
Aluminium	All of the above plus	
	^{22}Na	2.6 y
	^{24}Na	15.0 h
Iron	^{m44}Sc	2.44 d
	^{46}Sc	83.8 d
	^{48}Sc	1.81 d
	^{48}V	16.0 d
	^{51}Cr	27.7 d
	^{52}Mn	5.6 d
	^{54}Mn	312.1 d
	^{55}Fe	2.73 y
	^{59}Fe	44.5 d
	^{55}Co	17.54 h
	^{56}Co	77.3 d
	^{57}Co	271.8 d
	^{58}Co	70.82 d
Stainless steel	All of the above plus	
	^{60}Co	5.27 y
	^{57}Ni	35.6 h
Copper	All of the above plus	
	^{63}Ni	100 y
	^{61}Cu	3.4 h
	^{64}Cu	12.7 h
	^{65}Zn	244.3 d

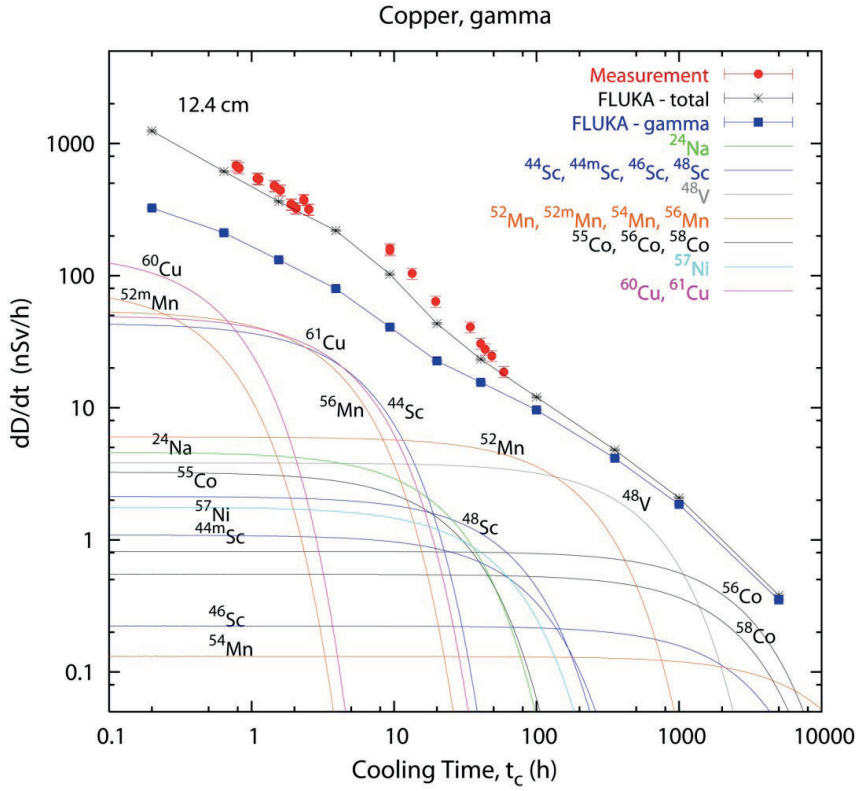


Fig. 4: Contribution of individual gamma-emitting nuclides to the total dose rate at 12.4 cm from an activated copper sample [12]

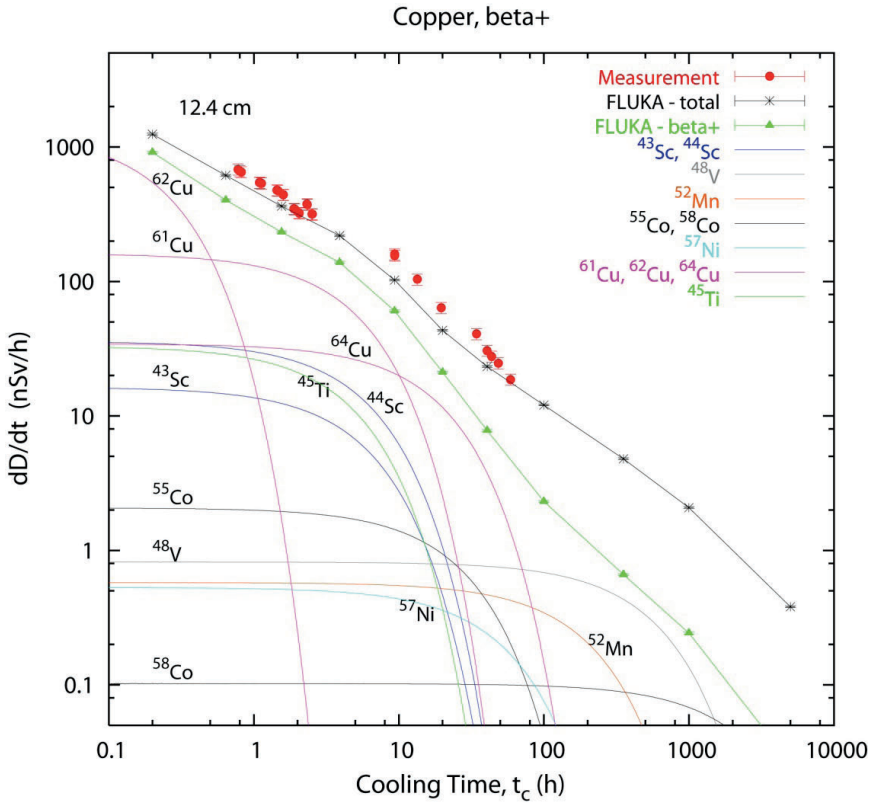


Fig. 5: Contribution of individual positron-emitting nuclides to the total dose rate at 12.4 cm from an activated copper sample [12]

7.3 Radionuclides in liquids

At accelerators, liquids are used mainly for cooling purposes (e.g. demineralized water and liquid helium), but liquid targets also exist (e.g. mercury).

Spallation reactions of secondary particle showers with oxygen in demineralized water can create tritium ($t_{1/2}$ = 12.3 y), ^{7}Be ($t_{1/2}$ = 53.29 d), and a number of short-lived β^{+}-emitters (^{11}C, ^{13}N, and ^{15}O). The production of tritium by thermal-neutron capture in natural hydrogen can be neglected in most application owing to the low abundance of deuterons and the small cross-section. Sometimes cooling-water circuits also contain nuclides from corrosion products (e.g. cobalt nuclides); however, a large fraction of these is collected, together with ^{7}Be, in the resin of ion exchanger cartridges. In natural water, radionuclides can also be produced in reactions with trace elements (i.e. minerals).

During accelerator design, the activation of cooling liquids is most conveniently assessed by folding fluence spectra with energy-dependent nuclide production cross-sections. Direct calculation is also possible using Monte Carlo codes for nuclides produced from oxygen, but this direct method would fail for nuclides produced from trace elements owing to a lack of statistical significance.

Activated cooling liquids pose contamination hazards during interventions in accelerator components and may also cause external irradiation close to pipes and cartridges. Although the decay of tritium proceeds only via the emission of a low-energy electron, its concentration in water, especially if released off-site, has become a critical parameter as it may attract the attention of the public.

7.4 Radionuclides in air

Airborne radionuclides are produced mainly by the interaction of beam particles or associated showers of secondary particles with air molecules. Other sources include activated dust and outgassing of nuclides from activated accelerator components. The latter two sources, however, are typically of lower importance and can only be assessed by measurement.

Table 5 gives the nuclides of highest radiological importance. At hadron and ion accelerators, most of them are created by spallation reactions with air molecules. Only ^{41}Ar results from thermal-neutron capture reactions with argon (σ_{th} = 660 mb). At electron accelerators, photonuclear interactions of type (γ, n) contribute to the production of ^{13}N and ^{15}O. Although the radiological impact of ^{3}H in air is small, it easily becomes attached to humidity and can reach waste water circuits, especially via condensation in air conditioning units.

Table 5: Airborne nuclides of radiological importance (the second column indicates the half-life)

Nuclide	$t_{1/2}$
^{3}He	12.3 y
^{7}Be	53.29 d
^{11}C	20.38 min
^{13}N	9.96 min
^{15}O	2.03 min
^{41}Ar	1.83 h

Apart from the list in Table 5, specific situations and exposure pathways may require the consideration of further nuclides, such as ^{32}P ($t_{1/2}$ = 14.26 d), which is produced by spallation reactions with argon. This nuclide can reach milk consumed by infants through ground deposition on grazing land and thus dominate the committed dose due to ingestion.

The low density of air usually renders a direct calculation of the activation of air by Monte Carlo models highly inefficient. Instead, particle fluence spectra are multiplied by energy-dependent nuclide production cross-sections, which are obtained from Monte Carlo models, experimental data, or both (the latter are called evaluated cross-sections). This yields nuclide production rates per unit volume or, after application of Eq. (13), the specific activity.

The results of air activation studies play a crucial role in the design of the ventilation system of an accelerator. Closed circuits that are flushed with fresh air prior to access but otherwise remain closed have the advantage of reducing the total annual release of short-lived nuclides. However, the concentration of long-lived nuclides may build up and lead to undue exposure if the nuclides are released at once over a period of time too short for there to be any benefit from changing wind conditions. In addition, tritium can build up, attach to water, and accumulate, for example in sumps. On the other hand, constant venting with fresh air causes an increased annual release of short-lived nuclides, although there is a benefit from natural dilution of long-lived nuclides. Apart from the environmental aspects, ventilation systems have safety functions in ensuring the containment of radioactive gases and should follow international standards [14].

Adjustments for the presence of ventilation can be made by introducing an effective decay constant λ' that includes the physical decay constant along with a ventilation term:

$$\lambda' = \lambda + \frac{D}{V}, \tag{15}$$

where D is the ventilation rate (volume of air exchanged per unit time) and V is the enclosure volume. Thus, with ventilation, the saturation activity A'_{sat} becomes

$$A'_{\mathrm{sat}} = \frac{\lambda A_{\mathrm{sat}}}{\lambda + D/V}. \tag{16}$$

8 Radiation fields around high-energy accelerators

8.1 Prompt stray radiation fields

Stray radiation fields are created at high-energy particle accelerators by the intentional interaction of the accelerated beam with targets, beam dumps, and collimators and by unintentional beam losses on structural components of the machine.

At electron accelerators, the most important secondary radiation is bremsstrahlung photons and high-energy electrons produced in electromagnetic cascades. An electromagnetic cascade is initiated when either a high-energy electron or a high-energy photon enters a material. At high energy, photons interact with matter mainly via pair production, whereas electrons and positrons lose their energy in a medium primarily by emitting bremsstrahlung photons. These two processes continue alternately, leading first to an exponential increase in the number of particles present in the cascade, which then starts to decline when removal processes (the photoelectric effect, ranging-out of electrons, and Coulomb and Compton scattering) dominate over the processes that generate new particles. Finally, low-energy electrons lose their residual energy by ionization and excitation processes.

At high-energy electron accelerators, neutrons are also present, released by photon-induced reactions rather than by electrons directly. High-energy neutrons are often the dominant secondary radiation outside a thick shield, which usually absorbs most of the bremsstrahlung photons.

At proton accelerators, interaction of the beam with materials generates a hadron cascade containing neutrons, charged hadrons, muons, photons, and electrons, with energy spectra extending over a wide range. The number of secondary particles produced per primary proton (the multiplicity)

increases as the proton energy increases. The average energy of these secondary particles also increases with the energy of the primary proton, making them capable of producing further inelastic interactions. The dominant radiation at workplaces outside accelerator shielding is the neutron field, with minor contributions from other particles. The neutron spectrum at the source, for example a beam loss point, is modified by transport through the shield, so that the energy distribution of neutrons at a workplace may be significantly different from the source spectrum. The shape of the spectrum also depends on the thickness of the shielding: the various components of the spectrum are attenuated differently, and only after a certain depth in the shield does the neutron spectrum reach equilibrium. This can be seen in Figs. 6 and 7, which show the neutron energy distributions in the transverse direction generated by 250 MeV protons impinging on an iron target thicker than the proton range. These figures show the energy distribution of the source neutrons and that behind a thin (20 cm to 1 m) and a thick (1–5 m) concrete shield. The distributions have been normalized to unit area in order to show better the change in the shape of the spectrum with increasing shield thickness.

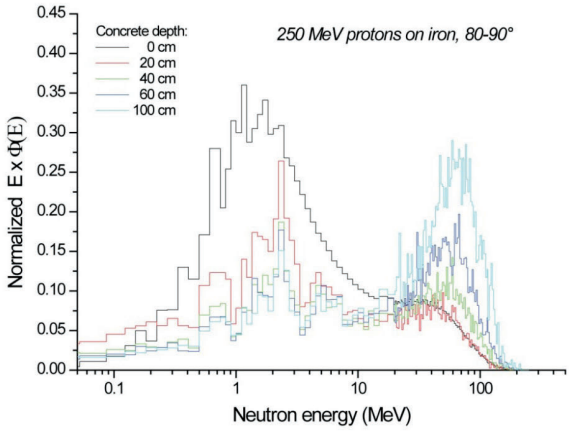


Fig. 6: Neutron energy distributions $E\Phi(E)$ in the transverse direction generated by 250 MeV protons impinging on an iron target thicker than the proton range. The distributions are for source neutrons and behind concrete shields of thicknesses ranging from 20 cm to 1 m. The distributions have been normalized to unit area in order to show better the change in the shape of the spectrum with increasing shield thickness.

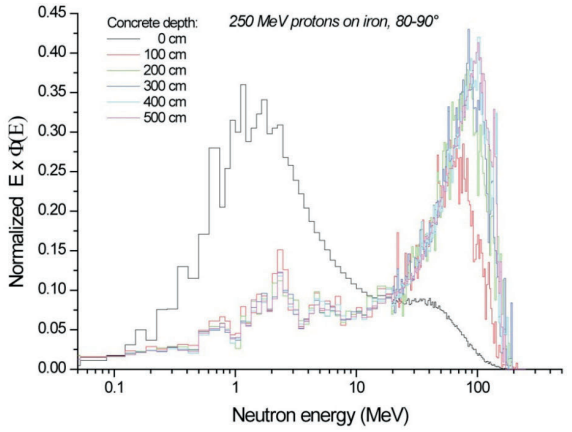


Fig. 7: Neutron energy distributions $E\Phi(E)$ in the transverse direction generated by 250 MeV protons impinging on an iron target thicker than the proton range. The distributions are for source neutrons and behind concrete shields of thicknesses ranging from 1 m to 5 m. The distributions have been normalized to unit area in order to show better the change in the shape of the spectrum with increasing shield thickness.

Figure 8 shows typical neutron energy distributions outside two types of shield at a multi-GeV proton accelerator [15]. The difference between the shapes of the two spectra outside the concrete shields is because in one case the neutrons emerging from the shield are scattered further by an additional surrounding concrete structure which softens the spectrum, a situation commonly found at accelerators.

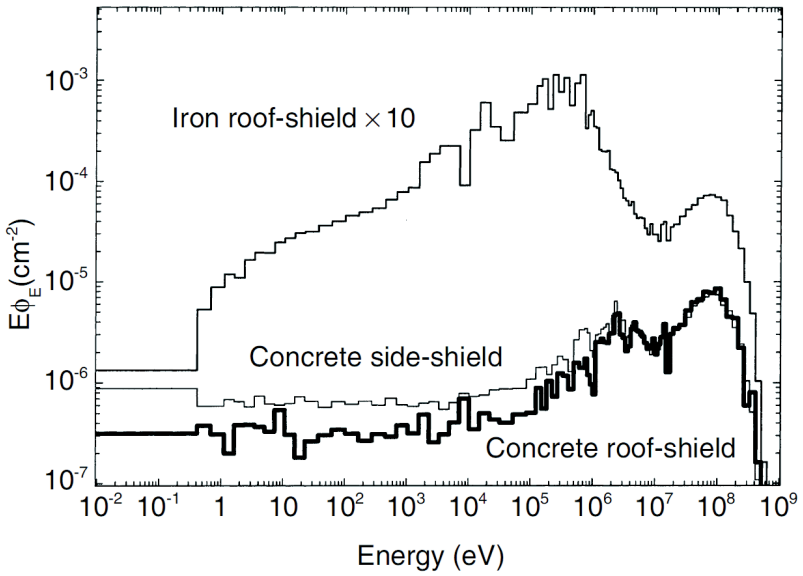


Fig. 8: Neutron spectral fluences $E\Phi(E)$ outside a concrete roof shield (80 cm thickness of concrete), an iron roof shield (40 cm thickness of iron), and an 80 cm thick concrete side shield (80 cm thickness of concrete, but the neutrons are scattered further by surrounding concrete) at the CERF facility at CERN (neutrons per primary beam particle incident on a copper target) [15]

As an example of the contribution of particles other than neutrons to $H^*(10)$, Figs. 9 and 10 plot the ratio of the values of $H^*(10)$ due to protons, photons, and electrons at various depths in a concrete shield to the total, in the forward and transverse directions, for 250 MeV protons impinging on a thick iron target. One sees that in the forward direction, protons contribute more than photons, while in the transverse direction, the opposite is the case.

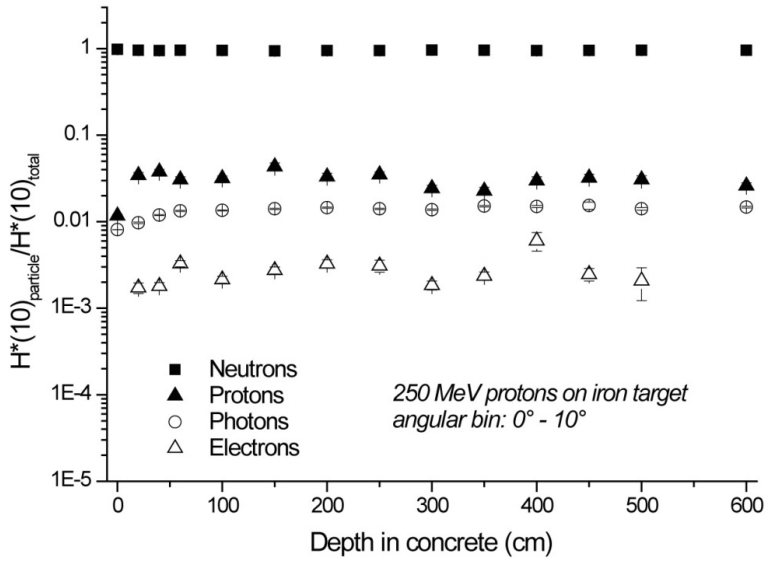


Fig. 9: Ratio of $H^*(10)$ due to secondary particles at various depth in a concrete shield to the total, in the forward direction, for 250 MeV protons impinging on an iron target thicker than the proton range.

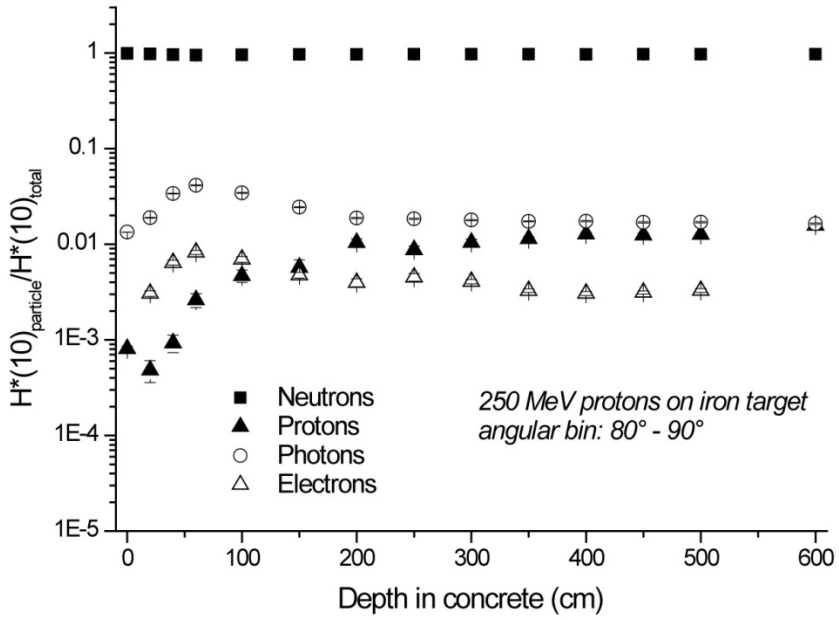


Fig. 10: Ratio of $H^*(10)$ due to secondary particles at various depth in a concrete shield to the total, in the transverse direction, for 250 MeV protons impinging on an iron target thicker than the proton range

Above about 10 GeV, muon-shielding requirements dominate in the forward direction for high-intensity proton beams, meaning that a residual muon beam is often present behind a shield thick enough to attenuate the hadron component of the field [16]. Muons arise from the decay of pions and kaons, either in the particle beam or in cascades induced by high-energy hadrons [17]. They can also be produced in high-energy hadron–nucleus interactions. The decay lengths for pions and kaons are 55.9 m and 7.51 m, respectively, times the momentum (in GeV/c) of the parent particle. Muons are weakly interacting particles and can only be stopped by 'ranging them out'. Muons lose energy mainly by ionization, as their cross-section for nuclear interaction is very low.

Muons from pion decay have a momentum spectrum that extends from 57% of the momentum of the parent pion to the pion momentum itself. Secondary pion beams generally have dumps containing a longitudinal depth of 1–2 m of Fe, and thus decay muons will penetrate these dumps for pion beams with a momentum larger than a few GeV/c.

To give an example [17], a beam of 10^7 pions per pulse with a momentum of 20 GeV/c travelling over a distance of 50 m will generate about 5×10^5 muons per pulse (5% of the parent beam). For a pulse repetition period of 2 s (a typical order of magnitude for a high-energy synchrotron), taking an approximate fluence-to-dose-equivalent conversion factor equal to 40 fSv·m^2 [18] and assuming that the muon beam is averaged over a typical area for a human torso of 700 cm^2, this fluence translates into a non-negligible dose equivalent rate of 500 μSv/h. Thus, under some circumstances (e.g. if the area downstream of the beam line is not interlocked for access), a muon component can be present in a mixed workplace field and contribute substantially to personnel exposure.

Radiation protection quantities such as the dose rate at workplaces and shielding thickness are generally not simple functions of energy. The parameters which most directly affect radiological safety are the particle type, the particle energy, the average beam power, and the number of lost particles per unit time at a given energy.

Some accelerators operate in pulsed mode, which means that the beam is present in the machine (or lost somewhere) during only a fraction of the time. With a synchrotron, the relevant parameters are the repetition rate (the number of cycles per unit time) and the flat-top duration (the time during which the beam is extracted from the accelerator to be transported somewhere else), whereas a cyclotron produces a virtually continuous beam. With a linear accelerator, an important parameter is the duty factor (DF), which is the fraction of the operating time during which the linac is actually producing radiation:

$$\mathrm{DF} = p * T_{\mathrm{p}}, \tag{17}$$

where p is the pulse repetition rate (in Hz) and T_{p} is the pulse length (in seconds).

At a given energy E, the dose rate generated by the interaction of the beam with a material is directly proportional to the average beam power P (i.e. to the number of 'lost' particles).

8.2 Stray radiation from residual radioactivity

Residual radioactivity is mainly a problem at proton accelerators, as dose rates at electron machines from induced radioactivity in accelerator structures are typically two orders of magnitude lower. At most accelerator facilities, the largest contribution to personnel dose actually arises from maintenance work near dumps, targets, septa, and collimators and generally near any object hit directly by the primary beam or located close to a beam loss point, rather than from exposure during machine operation. External (and sometimes internal) exposure to radiation from induced radioactivity can also occur in connection with the handling, transport, machining, welding, chemical treatment, and storage of irradiated items. A place where personnel can be exposed to such types of radiation can also be a workplace.

In spite of the fact that this radiation source is actually responsible for most of the individual and collective doses at accelerator laboratories, the associated radiation field is much simpler than that of the prompt radiation generated during accelerator operation, and the personnel exposure is due only to beta- and gamma-emitting radionuclides (and whole-body exposure is due essentially only to gamma emitters). The most common radionuclides with sufficiently long half-lives found in accelerator components are ^{7}Be, ^{22}Na, ^{54}Mn, ^{65}Zn, and the cobalt isotopes ^{56}Co, ^{57}Co, ^{59}Co, and ^{60}Co; in activated shielding structures, ^{133}Ba, ^{134}Cs, and ^{137}Cs are found; and in earth used as a shielding material, ^{152}Eu and ^{154}Eu are found.

The monitoring of such workplaces thus requires only beta/gamma monitors, such as ion chambers.

9 Instrumentation for area monitoring

CERN has a legal obligation to protect the public and persons working on its site from any unjustified exposure to ionizing radiation. For this purpose, CERN's Occupational Health & Safety and Environmental Protection (HSE) Unit monitors ambient dose equivalent rates inside and outside CERN's perimeter and releases of radioactivity in air and water. The results of the measurements allow the preventive assessment of radiological risks and the minimization of individual and collective doses. CERN's HSE Unit currently operates two radiation monitoring systems:

- ARCON (ARea CONtroller), which was developed at CERN for LEP and has been in use since 1988;
- RAMSES (RAdiation Monitoring System for the Environment and Safety), which was designed for the LHC based on current industry standards and has been in use since 2007.

About 800 monitors are employed in ARCON and RAMSES, about 400 for each system. Both installations comprise data acquisition, data storage, and the triggering of radiation alarms and beam interlocks. The most recent CERN facilities (the LHC, CNGS, and CTF3) are equipped with RAMSES, whereas the entire LHC injector chain, the remaining facilities (e.g. ISOLDE, n-TOF, and AD), and all experimental areas are still equipped with ARCON. In the long run, it is envisaged that ARCON will be replaced by the more recent RAMSES technology.

9.1 Radiation monitors

Both ARCON and RAMSES use the same or at least very similar types of radiation detectors. Environmental radiation protection monitors record stray radiation and the releases of radioactivity into air and water. Recording of other measured values such as wind speed, wind direction, and flow rates is required to obtain relevant input parameters for calculating doses to members of the public. An environmental stray-radiation monitoring station consists of one high-pressure ionization chamber filled with argon (from Centronics) for photons and penetrating charged particles such as muons, one REM counter (from Berthold) for neutrons, and a locally installed unit for data acquisition, alarm generation, and data transfer. The radiation protection part of a CERN water monitoring station consists of an NaI detector for in-situ measurements of gamma-emitting radionuclides and a device to collect water samples for laboratory analyses such as measurements of tritium and for cross-checks of the on-line results. The ventilation monitoring system is based on silicon surface detectors to measure the total activity of beta emitters released. In addition, removable filters are installed to allow laboratory analysis of radionuclides attached to aerosols using gamma spectroscopy. The active parts of the air and water monitoring stations (the Si and NaI detectors) are equipped with alarm functions.

The Radiation Protection Group uses three different types of monitors to measure ambient dose equivalent rates at CERN and in the close neighbourhood of CERN's facilities. The radiation monitors employed to protect workers against prompt ionizing radiation [19] during beam operation are special REM counters (from WENDI/Thermo) and hydrogen-filled, high-pressure ionization chambers (from Centronics). Both are optimized to measure high-energy neutrons with energies up to the GeV range; the hydrogen chamber responds to all particles contributing to the high-energy mixed radiation fields [20, 21].

The ambient dose equivalent rates which can be monitored inside the machine tunnel and the experimental caverns after the beam has been stopped are due to radiation emitted by the decay of radionuclides induced during operation of the beam. The energies of the emitted photons do not exceed 2.7 MeV (emitted by ^{24}Na) [19]. The induced radioactivity is measured with air-filled plastic ionization chambers (from PMI) in order to assess risks during maintenance and repair work [22]. The radiation monitoring system is completed by hand and foot monitors at the exits from the accelerator and experimental areas and by gate monitors at the exits of the CERN sites (Site Gate Monitors, SGMs). The RAMSES system provides an option to connect the SGMs to the access system; that is, when there is an alarm, the barriers can remain closed.

Outside the shield of an accelerator facility, the ambient dose equivalent rates during operation range from a few hundreds of microsieverts per year to a few millisieverts per year. To measure such rates, one needs detectors that are of high sensitivity or capable of integrating over long periods.

9.2 Dosimetry at CERN

Exposure to ionizing radiation (gamma, beta, and particle radiation) accompanies all work at a particle accelerator and in the associated experimental facilities. Legal dose limits assure the safety of personnel working under these conditions. The dose received by individuals working with ionizing radiation at CERN is monitored with personal dosimeters. Every person working at CERN in Radiation Areas or with sources of ionizing radiation must wear a CERN dosimeter. The CERN dosimeter registers the personal dose from sources of ionizing radiation around particle accelerators. It

combines an active detector for gamma and beta radiation based on the Direct-Ion Storage (DIS) technology and a passive detector for quantifying neutron doses.

The gamma/beta dose registered by a CERN dosimeter can be read out as frequently as deemed necessary, but it must be read at least once per month on one of the approximately 50 reader stations which are installed CERN-wide The monitoring period for the neutron dosimeter is, in principle, one year. It must then be returned to the supplier for evaluation.

For work in Controlled Radiation Areas, where the radiological risk and the dose rate are above 50 μSv/h, the additional use of an operational dosimeter is required. CERN provides all staff who may work in Limited Stay Radiation Areas or High Radiation Areas with a system for active dosimetry with an alarm, in the form of an dosimeter, model DMC-2000 from MPG instruments.

10 ALARA at CERN

CERN introduced a formalized approach to ALARA [23–25] at the end of 2006, as a result of collaboration between the former Accelerator and Beams department and the Radiation Protection Group. This approach was applied first to the SPS and LHC complex, and since 2009 has been applied to all CERN facilities. The goal was to optimize work coordination, work procedures, handling tools, and even the design of entire facilities. Consequently, all work in Radiation Areas has to be optimized. In particular, all work in Controlled Radiation Areas must be planned and optimized, including an estimate of the collective and individual effective doses to the workers participating in the completion of a task.

Five different criteria were established in 2006 and are used for the determination of the so-called ALARA level of an intervention. These five criteria are shown in Table 6. Depending on the level of the intervention, different means of optimization have to be applied. For example, level 3 interventions need formal approval from the ALARA Committee, which is chaired by the Director of Accelerators.

Table 6: ALARA criteria at CERN

Criteria: Ambient dose equivalent		
	50 μSv/h	2 mSv/h
Level I	Level II	Level III

Criteria: Individual dose		
	100 μSv	1 mSv
Level I	Level II	Level III

Criteria: Collective dose		
	500 μSv	10 mSv
Level I	Level II	Level III

Criteria: Airborne activity in CA values according to [26]		
	5 CA	200 CA
Level I	Level II	Level III

Criteria: Surface contamination in CS values according to [26]		
	10 CS	100 CS
Level I	Level II	Level III

Acknowledgements

Figures 6, 7, 9, and 10 are the results of Monte Carlo simulations performed at CERN by Alessio Mereghetti for his diploma thesis at the Polytechnic of Milan. The authors wish to thank Alessio for kindly providing the figures for this paper.

References

[1] ICRP, *1990 Recommendations of the International Commission on Radiological Protection*, ICRP Publication 60 (Pergamon, Oxford, 1991) [*Ann. ICRP* **21** (1991) 1].

[2] CERN, Safety Code F 2006, EDMS 335729.

[3] Particle Data Group, *Review of Particle Physics* (2010), http://pdg.lbl.gov.

[4] ICRP, *The 2007 Recommendations of the International Commission on Radiological Protection*, ICRP Publication 103 (Elsevier, Amsterdam, 2007) [*Ann. ICRP* **37**(2–4) (2007)].

[5] K.H. Lieser, *Nuclear and Radiochemistry: Fundamentals and Applications* (VCH/Wiley, Weinheim, 1997).

[6] M. Streit-Bianchi, http://videolectures.net/marilena_streit_bianchi/

[7] United Nations, Report of the United Nations Scientific Committee on the Effect of Atomic Radiation, General Assembly, Official Records A/63/46 (2008).

[8] Bundesamt fuer Gesundheit, Radioaktivitaet und Strahlenschutz, BAG Schweiz (2007).

[9] Federal Office of Public Health, Switzerland, http://www.bag.admin.ch/themen/strahlung/00046/11952/index.html?lang=en

[10] S. Roesler and C. Theis, CERN operational radiation protection rule, exemption and clearance of material at CERN, 02.12.2009, EDMS 942170.

[11] A.W. Chao and M. Tigner, *Handbook of Accelerator Physics and Engineering*, 4th edn (in press), World Scientific Publishing.

[12] S. Roesler *et al*., Proc. Sixth International Meeting on Nuclear Applications of Accelerator Technology, 2003, p. 655, San Diego.USA.

[13] R.H. Thomas and G.R. Stevenson, *Radiological Safety Aspects of the Operation of Proton Accelerators,* IAEA Technical Report No. 283 (1988).

[14] International Organization for Standardization, International Standard ISO 17873 (2004), http://www.iso.org

[15] A. Mitaroff and M. Silari, *Radiat. Prot. Dosim*. **102** (2002) 7–22.

[16] M. Silari and G.R. Stevenson, *Radiat. Prot. Dosim*. **96** (2002) 311–321.

[17] R.H. Thomas and G.R. Stevenson, Radiological safety aspects of the operation of proton accelerators, Technical Report No. 288, IAEA, Vienna (1988).

[18] M. Pelliccioni, *Radiat. Prot. Dosim*. **77** (1998) 159–170.

[19] D. Forkel-Wirth, F. Corsanego, S. Roesler, C. Theis, Heinz Vincke, Helmut Vincke, P. Vojtyla, L. Ulrici, M. Brugger, and F. Cerutti, How radiation will change (y)our life, Proc. Chamonix Workshop 2010.

[20] C. Theis, D. Forkel-Wirth, D. Perrin, S. Roesler, and H. Vincke, *Radiat. Prot. Dosim*. **116** (2005) 170–174.

[21] C. Theis, D. Forkel-Wirth, M. Fuerstner, S. Mayer, T. Otto, S. Roesler, and H. Vincke, Field calibration studies for ionisation chambers in mixed high-energy radiation fields, *Radiat. Prot. Dosim*. 2007, doi:10.1093/rpd/ncm062 (2007).

[22] H. Vincke, D. Forkel-Wirth, D. Perrin, and C. Theis, *Radiat. Prot. Dosim*. **116** (2005) 380–386.

[23] P. Bonnal and D. Forkel-Wirth, Instruction générale de sécurité CERN, Règles générales d'exploitation, critères et exigences ALARA, 20.12.2006, EDMS 810176.

[24] P. Bonnal, Instruction générale de sécurité CERN, Règles générales d'exploitation, constitution et convocation du comité ALARA du CERN, 20.12.2006, EDMS 810178.

[25] P. Bonnal and D. Forkel-Wirth, Instruction générale de sécurité CERN, Règles générales d'exploitation, démarche ALARA applicable aux interventions, 06.03.2007, EDMS 825353.

[26] Schweizer Strahlenschutzverordung (StSV), SR-Nummer 814.501.

Activation and radiation damage in the environment of hadron accelerators

Daniela Kiselev
Paul Scherrer Institute, Villigen, Switzerland

Abstract
A component which suffers radiation damage usually also becomes radioactive, since the source of activation and radiation damage is the interaction of the material with particles from an accelerator or with reaction products. However, the underlying mechanisms of the two phenomena are different. These mechanisms are described here. Activation and radiation damage can have far-reaching consequences. Components such as targets, collimators, and beam dumps are the first candidates for failure as a result of radiation damage. This means that they have to be replaced or repaired. This takes time, during which personnel accumulate dose. If the dose to personnel at work would exceed permitted limits, remote handling becomes necessary. The remaining material has to be disposed of as radioactive waste, for which an elaborate procedure acceptable to the authorities is required. One of the requirements of the authorities is a complete nuclide inventory. The methods used for calculation of such inventories are presented, and the results are compared with measured data. In the second part of the paper, the effect of radiation damage on material properties is described. The mechanism of damage to a material due to irradiation is described. The amount of radiation damage is quantified in terms of displacements per atom. Its calculation and deficiencies in explaining and predicting the changes in mechanical and thermal material properties are discussed, and examples are given.

1 Activation on a microscopic level

Nuclear reactions of particles with the nuclei of the chemical elements of a component may cause an initially non-radioactive component to become radioactive. In contrast to chemical reactions, where the elemental composition is of interest, the isotopic composition of the component has to be known. An isotope is characterized by its number of protons, Z, and its number of neutrons, N. The total number of protons and neutrons is called the mass number, A. Before exposure to irradiation, the distribution of the isotopes of each element is given by the natural abundance. (Cosmic radiation also produces radioactive isotopes in materials but at a very low level. This becomes relevant, for example, when a material is used as shielding for measurements sensitive to the natural background. In this case ancient lead, which can be found in ancient Roman ships buried in deep water, can be used. The water provides effective shielding against cosmic irradiation, and the radioactive isotopes induced by cosmic rays in ancient times have mostly decayed.) Bombarding isotopes with particles leads to a change in their proton and neutron numbers, i.e. transmutation. When we look at a nuclear chart of isotopes, we see that most of the stable isotopes are directly surrounded by radioactive ones. For interactions with particles of energies larger than 100 MeV, spallation is the dominant process. Part of the energy of the primary particle is transferred to several nucleons in independent reactions. The nucleons inside the nucleus subsequently collide with each other, distributing the energy almost equally. This is called an intranuclear cascade and leads finally to a highly excited nucleus. This process takes about 10^{-22} s. Some of the nucleons may leave the nucleus if their energy is higher than their binding energy. Nucleons at the surface, i.e. in less strongly bound states, have binding energies of a few MeV, whereas nucleons in the deeper-lying shells in a medium-heavy nucleus need about 50 MeV to remove

them from the nucleus. Charged particles such as protons or pions (produced mainly in the decay of the Delta resonance) have to overcome the Coulomb barrier (the potential energy between the charged nucleus and the particle) in addition. (The Delta resonance is the first excited state of a nucleon, about 300 MeV above the mass of the nucleon. At this energy, a clear enhancement (bump) appears in the cross-section.) Therefore, in the first stage of the spallation process, i.e. the cascade/pre-equilibrium stage, high-energy secondary particles such as protons, pions, and neutrons are emitted. Their energy usually exceeds 20 MeV and most of them move in the direction of the primary particle, which has transferred a high momentum in the first collision to the secondary particles now leaving. In a thick target or component, it is likely that the secondary particles will interact with other nuclei. This is called an internuclear cascade.

In the second stage of the spallation process, the excited nucleus releases its energy by emitting particles. This is called the evaporation phase and takes about 10^{-18} s. Mostly neutrons, with a Maxwell–Boltzmann distribution peaking around 1–2 MeV, are emitted. Charged particles such as protons, pions, and light ions are suppressed owing to the Coulomb potential, particularly for nuclei with a large charge number. If the energy left falls below the threshold for particle emission, the remaining energy is released by photon emission. Owing to the deficit of neutrons, the remaining nucleus is likely to be radioactive. For heavy nuclei, high-energy fission competes with evaporation during the evaporation phase and is the preferred channel for very heavy nuclei such as lead, tungsten, and mercury. At high excitation energy, the nucleus breaks into two halves, i.e. symmetric fission is preferred. Some lighter particles such as neutrons and protons are lost during the fission process.

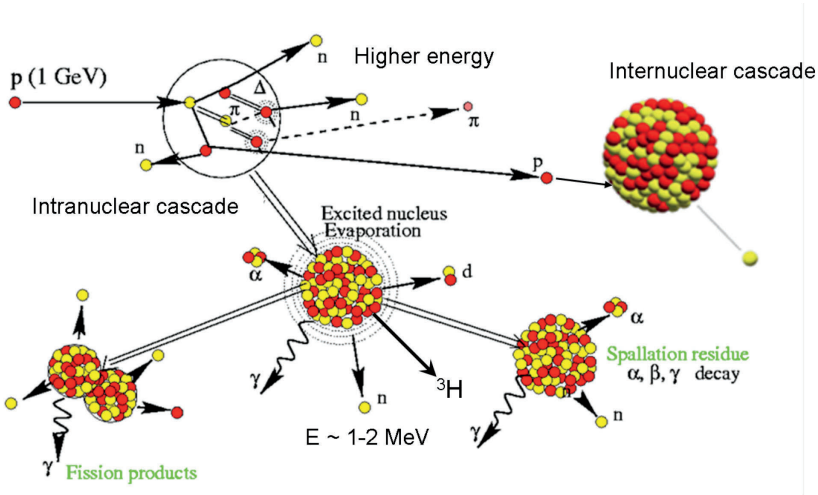


Fig. 1: Illustration of the spallation process (modified from an image on www.cea.fr)

The spallation process is illustrated in Fig. 1 for a 1 GeV primary proton. At higher energies, the process is similar. More particles and other species such as exotic mesons can be produced. More and larger fragments can be emitted. In general, the number of neutrons increases with energy. For 1 GeV protons, 20 neutrons per primary particle are produced in lead.

Starting from a target nucleus (Z, A), all nuclei with mass numbers less than A down to 1 can be produced by spallation. For a heavy nucleus, the mass region around $A/2$ is filled by fission residues. How many isotopes of a given type remain after bombardment of a target nucleus with high-energy particles depends on the energetics of the process and, in more detail, on the structure of the initial and final nuclei. The production rate of an isotope obviously depends on the number of target nuclei N_A and on the number of primary particles n_i incident per second onto the target material. The quantity that describes the transformation of a target nucleus A into a given isotope Y is called the (production) cross-section. Since the energy spectrum ϕ_x of particles of type x (e.g. secondary neutrons) is given per incident primary particle and per unit area (1/area), the cross-section has units of area. Owing to the

small value of the cross-section, the barn (b), equal to 10^{-24} cm^2, is often used as a unit. The production rate P_Y for isotope Y is given by

$$P_Y = N_A n_i \int \frac{\mathrm{d}\phi_x(E)}{\mathrm{d}E} \cdot \sigma_{xA\to Y}(E) \cdot \mathrm{d}E \tag{1}$$

Since the production cross-section $\sigma_{xA\to Y}$ depends on the energy, the spectrum of the particles as a function of energy is needed as well.

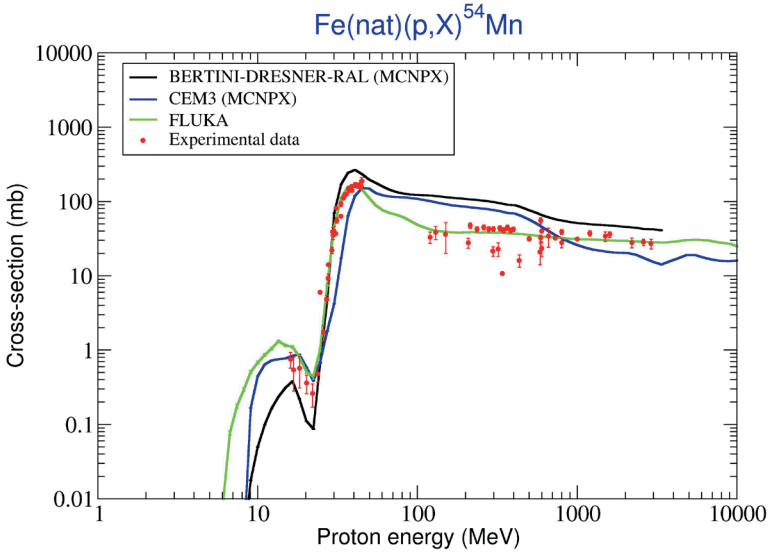


Fig. 2: Production cross-section for ^{54}Mn using proton irradiation of natural Fe. Results obtained from several models [1–6] implemented in MCNPX [7] and experimental data [8] are shown. FLUKA [9] simulation: courtesy of S. Roesler, CERN.

A typical production cross-section for ^{54}Mn using protons on natural iron is shown in Fig. 2. For protons with energies larger than a few hundred MeV, the cross-section becomes almost constant. This behaviour continues with a slight slope up to the TeV range. A similar behaviour of the cross-section is observed for high-energy neutrons, except at the threshold. Owing to the charge of the proton, a minimum energy is needed to interact with the charged nucleus. This threshold energy depends on the nuclear structure and Z. For medium-heavy nuclei, the threshold energy is a few MeV. The bump between 10 and 20 MeV in Fig. 2 is caused by a large number of close-lying resonances, which correspond to nuclear levels that are free to capture a proton.

The curves in Fig. 2 are the predictions of various models implemented in the particle transport codes MCNPX [7] and FLUKA [9]. The models CEM3.02 [5, 6] and BERTINI-DRESNER-RAL [1– 4] were used in the form of versions implemented in MCNPX2.7.c. Compared with the experimental data [8] shown by the points in the figure, all models are in reasonable agreement. This is not always the case, however. In addition, there may be little no data available for elements less common than iron and, particularly, for targets consisting of a single isotope. Models need to predict not only the production cross-sections but also the production of secondaries and their angular distribution. Owing to the interest in low-energy neutrons for spallation sources and the high penetration capability of high-energy neutrons through biological shielding, the prediction of neutron production is the most important requirement in many cases. In addition, the main contribution to the activity is often caused by low-energy neutrons, owing to their large capture cross-sections (see below).

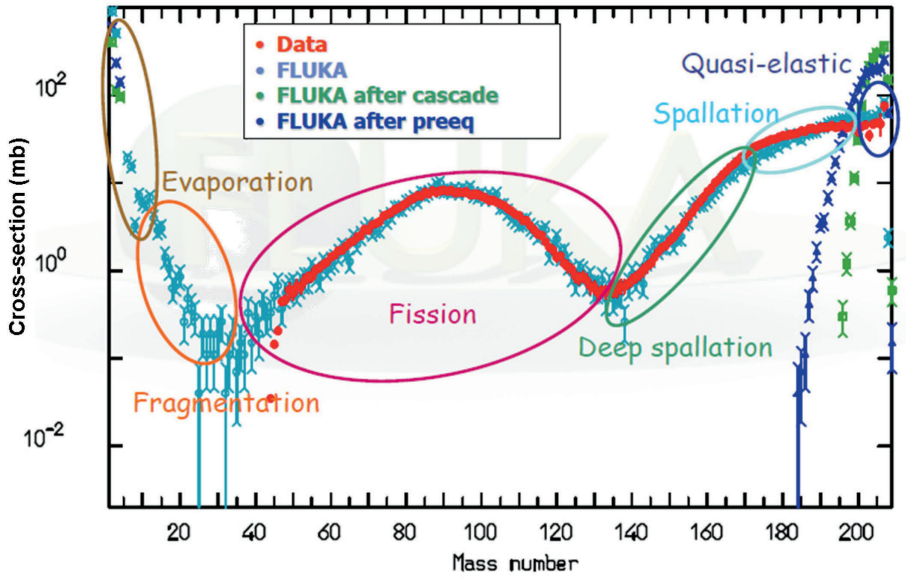


Fig. 3: Mass distribution predicted by FLUKA after bombardment of Pb with 1 GeV protons. The data are taken from Ref. [10]. Image courtesy of A. Ferrari, CERN.

Figure 3 shows the mass distribution for 1 GeV protons on ^{208}Pb, calculated with FLUKA; the inverse reaction was measured at GSI [10]. The FLUKA results are in very good agreement with the data shown as dots. Up to 15 mass units below the target nucleus, the isotopes produced arise from spallation reactions. The production cross-section for these isotopes is quite high. Isotopes which lose more nucleons receive a larger energy transfer from the projectile, and therefore this type of reaction is called 'deep spallation'. Afterwards, the fission products are broadly distributed around half of the mass number of the target nucleus, since symmetric fission is favoured for heavy nuclei at higher energy, with a few nucleons being lost during the break-up. At very low mass numbers, evaporation products up to light ions have a high probability of being produced, since they accompany almost every spallation process. Products with larger mass numbers are produced by fragmentation, where the target nucleus disintegrates into large clusters or one cluster is emitted. This region is particularly difficult to reproduce by models.

Further away from the loss points of the primary particles, i.e. in the biological shielding around these loss points, the charged particles are slowed down by ionization of atoms along their paths. Depending on the type of particle, they may decay, be absorbed, or be changed by reactions. An almost stopped proton, for example, can capture an electron and remain as hydrogen in a component. After a certain distance in the shielding material, which depends on the stopping range of the charged particles and therefore on their energy, only neutral particles are left. Since neutrons have a larger lifetime than other neutral particles, and because they are produced in vast amounts, the reactions occurring in the shielding are driven by neutrons. The neutrons lose energy in many (elastic) collisions, i.e. they are moderated. This process is particularly effective for nuclei with small mass numbers, for kinematical reasons. For example, a ball thrown against a wall bounces back with the same velocity, i.e. it does not lose any energy but just changes momentum. The wall represents a collision partner of large mass number, and the ball represents a neutron.

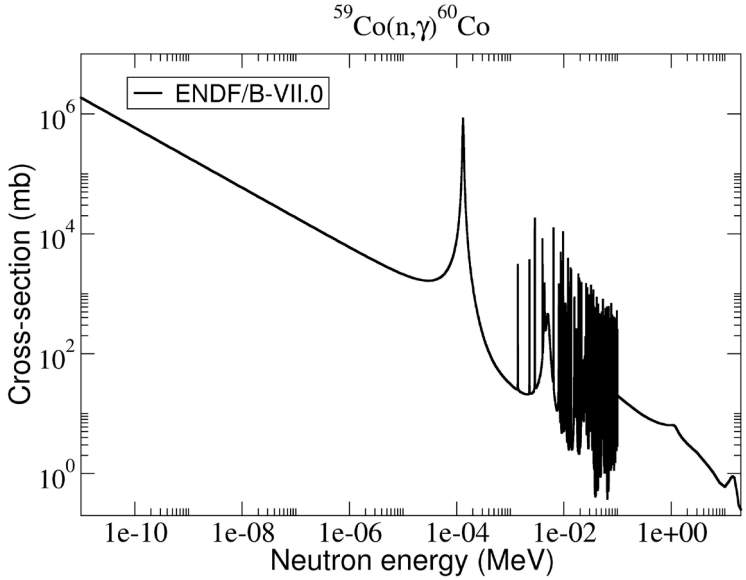


Fig. 4: Cross-section for neutron capture by Co for energies up to 20 MeV, from the ENDF/B-VII library [11]

There are still higher-energy neutrons left, but these are in equilibrium with the low-energy neutrons. This means that low-energy neutrons which are lost by absorption are replaced by moderated, formerly higher-energy neutrons. In this case, all energy groups of neutrons are in equilibrium with each other. As a consequence, the shape of the neutron spectrum does not change any more (or at least not significantly); only the amplitude changes, owing to the continuous loss of neutrons. The main reaction process for this is neutron capture, where a neutron is absorbed by a nucleus. As an example, the neutron capture cross-section of ^{59}Co is shown in Fig. 4. For neutron energies from a few keV onwards, large spikes in the cross-section can be seen. The probability of capturing a neutron with an energy just corresponding to a level of the nucleus that is free to capture a neutron is especially high. This situation is called a resonance. The energy width of the levels is around 1 eV. At higher energies, the density of states increases and the resonances overlap. The neutron is usually captured into an excited state. The surplus energy, including the binding energy, is often released as a photon. For energies less than 1 eV, the cross-section increases linearly on a logarithmic plot, with the inverse of the square root of the energy or the inverse of the velocity. As the velocity decreases, the time for interaction between the neutron and the nucleus increases, and therefore the probability of capture of the neutron increases.

2 Direct calculation of the activation

Spallation is a complicated process with many open reaction channels. The particles have to be tracked, particularly in thick targets where secondary particles produce particles again and so on. This is a kind of chain reaction, for which Monte Carlo simulations are useful. The code for the simulation has to provide cross-sections for all relevant reaction channels and particles involved. Usually, the cross-sections for low-energy neutrons up to 20 MeV are taken from libraries, where experimental cross-sections have been cross-checked against each other and combined with models where necessary. A well-known and often used library is one from the USA and Canada compiled by the Cross Section Evaluation Working Group (CSEWG). The newest version available is ENDF-B-VII [11] (ENDF is an abbreviation for ‘Evaluated Nuclear Data Files’). The result of the European effort in this field is the Joint Evaluated Fission and Fusion (JEFF) file [12], compiled by the NEA Data

Bank member countries. Only recently have libraries for neutrons, protons, and deuterons up to 200 MeV, such as EAF-2010 [13], become available. Usually, models are used in simulation codes for neutrons above 20 MeV and for all other particles. At present, most of these models are based on the mechanism of the internuclear cascade (INC) described above. Their origin can often be traced back to the 1960s. The INC models are coupled to models for fission and evaporation on a microscopic basis.

The user of the Monte Carlo code has to provide the geometry of the system and the characteristics of the particle beam, which define the primary particles. If the component of interest is far away from the first point of interaction and the space between is filled with material with a complex geometry, the geometric model that must be provided is elaborate and the calculation is CPU-consuming. Often tricks (biasing) have to be applied to get sufficient statistics at the point of interest. In some restricted cases, other methods are more efficient than a Monte Carlo simulation. One method will be presented below.

Besides the geometry, the material composition of all components of interest has to be known. The most important constituents are elements with large production cross-sections, such as Co for neutrons. Only a few thousand parts per million of Co is contained in steel, and a few parts per million in aluminium (depending on the grade). Owing to its large neutron capture cross-section, ^{60}Co becomes dose-relevant after a cooling time of about a month. In aluminium, the dose is determined by ^{22}Na initially, and by ^{60}Co after a few years of cooling. Therefore it is important to know precisely the Co content. In practice, however, the Co content is not precisely known. The definition of the material is one large source of uncertainty in the prediction of the dose. Some examples will be given below.

A pure Monte Carlo simulation follows each particle through the material, including all interactions between the material and the particle. At the end of the simulation, the number of reaction products is counted for each isotope per primary particle, which leads to the production rate for a specific isotope (see Eq. (1)). Alternatively, particle fluxes can be provided, which are later folded with the corresponding cross-sections. To obtain the activity of a component, the time periods during which the component was irradiated have to be known. Cooling times, i.e. periods with no irradiation, are also important, because of the decay of the radioactive nuclei. First, the time evolution of an ensemble of ^{60}Co nuclei will be considered for simplicity. At the beginning ($t_0 = 0$), there are N_0 nuclei in the ensemble; after a time t, the number of nuclei that have not yet decayed follows the law of radioactive decay:

$$N(t) = N_0 \exp(-\lambda t) , \tag{2}$$

where λ is the decay constant and is related to the half-life $T_{1/2}$ via

$$\lambda = \frac{\ln 2}{T_{1/2}} . \tag{3}$$

The activity is defined as the number of nuclei that decay per second, i.e.

$$A(t) = -\frac{\mathrm{d}N}{\mathrm{d}t} = \lambda N(t) . \tag{4}$$

The expression on the right-hand side is derived from Eq. (2). During the irradiation time, the number of nuclei of an isotope can change not only by decay but also due to the production rate P of that isotope. The rate equation is then

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = P - \lambda N(t). \tag{5}$$

The expression for the activity follows from the solution of the differential equation in Eq. (5):

$$A(t) = P(1 - \exp(-\lambda t)). \tag{6}$$

For very long irradiation times, the activity is equal to the production rate P and becomes independent of the irradiation time. It is important to realize that at a given particle flux, the activity cannot exceed a certain value, which is called the saturation activity A_{sat}. An example of the production and decay of ^{60}Co is shown in Fig. 5. ^{60}Co has a half-life of 5.3 years; 90% of the saturation activity is reached after an irradiation time of about three times the half-life, and half of the saturation activity is produced after an irradiation time of the order of the half-life. Since ^{60}Co decays into stable ^{60}Ni, the decay chain ends at that point.

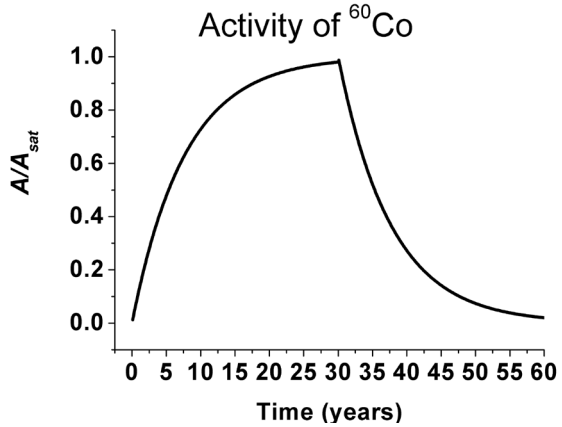


Fig. 5: Build-up and decay of ^{60}Co.

In the general case, many isotopes are produced, each with different production rates. The number of nuclei of an isotope of type m per unit time is given by the Bateman equation,

$$\frac{\mathrm{d}N_m(t)}{\mathrm{d}t} = -N_m(t)\left(\lambda_m + \varphi_x \sigma_{m+x}^{\mathrm{abs}}\right) + \sum_{k \neq m} N_k(t)\varphi_x \sigma_{k \to m} \,. \tag{7}$$

Here the first term describes the decrease in the number of nuclei of an isotope due to radioactive decay and transmutation to other isotopes. The last term contains the production of isotope m. The sum runs over all production channels.

There are Monte Carlo codes such as MCNPX and PHITS [14] which have to be coupled to external build-up and decay codes to obtain the activation after a specified irradiation history. “External” means that these codes are not provided by the Monte Carlo distributor. PHITS is often used in connection with DCHAIN [15], and for MCNPX a script [16] exists to transfer the information about the neutron flux (for neutron energies less than 20 MeV) and the residual production rates to Cinder’90 [17] or FISPACT [18]. These programs use a built-in or external library for neutron energies of less than 20 MeV (sometimes reaching to higher energies). They also provide all decay properties of the isotopes. MARS [19] comes with its own build-up and decay code, which is based on DCHAIN. The calculation of the activation for user-specified irradiation conditions is done during runtime. FLUKA provides its own build-up and decay codes. It offers two possibilities, to perform the calculation of the activation during runtime or to do the operation of coupling to the build-up and decay code separately afterwards. The later option needs more effort, but has the advantage that a request for results at more cooling times or different irradiation times after the simulation has been run can be quickly fulfilled.

Owing to the dependence of the production and decay rates on the half-life, the question of which radioisotope dominates the activity in a specified material has a time-dependent answer. It is clear that short-lived isotopes decay rapidly and their presence is negligible after long cooling times. On the other hand, the final nuclide inventory depends not only on the total number of bombarding particles but also on the particle rate (or flux, or current). For example, consider two isotopes with

very different half-lives but the same production cross-section. This condition leads to the same saturation activity for both isotopes. The difference is that during a short intense irradiation time, half of the saturation activity is reached for the short-lived isotope, whereas the activity of the long-lived isotope is still negligible. ('Short' means that the irradiation period is of the order of the half-life of the shorter-lived isotope.) When the same number of particles is kept but at a lower flux, the consequence is a lower saturation activity for both isotopes. Compared with the case of the short irradiation time, the activity of the long-lived isotope will be larger. For very long irradiation times, the same activity will be reached for both types of isotopes. This does not mean, however, that the same numbers of nuclei of the isotopes are produced. Since

$$A = \lambda N, \tag{8}$$

the number of nuclei of the short-lived isotope will be greater by a factor of $T_{1/2}$ (long)/$T_{1/2}$ (short).

3 Example: activation of a directly irradiated component

'Directly irradiated' means that the component is hit by the primary proton beam. At such locations, a major fraction of the beam is often lost. Such dedicated beam loss points include targets, collimators, and beam dumps. The particular activated component considered in this section is the first version of the target for the neutron spallation source SINQ at the Paul Scherrer Institute (PSI). The 57 cm long target contains about 456 Zircaloy tubes, each of them 10 cm long. The beam hits the tubes, known as 'cannelloni', laterally, i.e. the tubes are installed perpendicular to the beam. Nowadays, these tubes are filled with lead, but in the trial version of the target, solid Zircaloy sticks were installed instead of tubes. From the beginning of the development of the target, the containment around the target served as a 'safety hull'. This containment was made from AlMg3, an aluminium alloy containing 3 weight per cent of magnesium. The target was mounted vertically and the beam hit it from below. Every two years, the target, together with the safety hull and the shielding behind the target, was changed. Samples from the front window of the safety hull, the shielding, and a Zircaloy screw in the target were taken in a hot cell using remote-controlled manipulators. The gamma spectra of the samples were measured with a high-purity germanium detector, and the isotopes present were identified by their characteristic gamma energies. Owing to their high self-absorption, the beta emitters needed a chemical treatment, after which they were dissolved in a scintillating fluid and measured. Isotopes with a very long half-life would need a very long measurement time to achieve the required statistics; these isotopes were instead counted one by one using the technique of Accelerator Mass Spectrometry (AMS). Chemical preparation was needed for this purpose also. The samples were not examined for alpha emitters. The measurement of alpha emitters would need a windowless solid state detector, and the preparation of the samples would be difficult and elaborate.

The geometrical model prepared for the simulation that was performed using MCNPX is shown in Fig. 6, which also gives a good overview of the SINQ facility. Around the target, water (shown in dark grey) is used to decelerate the neutrons to thermal energies. This is similar to the set-up of a reactor. Instead of normal water, the tank is filled with deuterated water, D_2O. The reason for this is that hydrogen has a large capture cross-section for thermal neutrons and would therefore reduce the neutron flux significantly. A smaller tank, shown in light grey in Fig. 6, contains D_2 at 25 K. Scattering of neutrons by cold deuterons leads to energy transfer from the faster neutrons to the deuterons, resulting in cold neutrons. The intermediate grey areas are made of iron and serve as biological shielding. The concrete shielding outside the iron shielding was not included in the model. The samples were taken from SINQ target 3, which was irradiated for almost two years from 1998 to 1999 with protons equivalent to 6.77 A h. In the MCNPX simulation, the ENDF-B-VI cross-section library was used for neutrons of energy less than 20 MeV. The resulting neutron flux was folded with the cross-sections provided by the Cinder'90 library, version 7.4. The same program also calculated the time evolution of the radioisotopes according to the irradiation history.

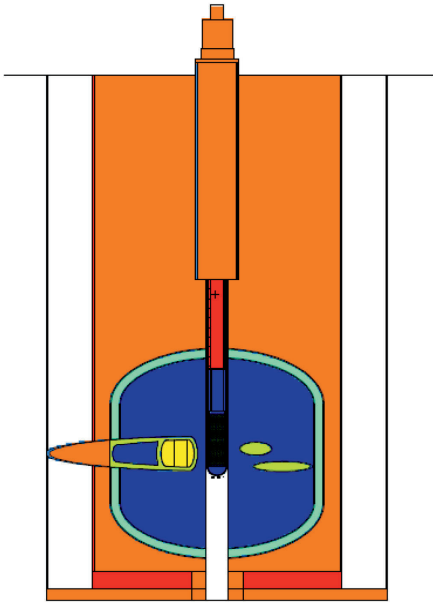


Fig. 6: Geometrical model of the SINQ facility at the PSI used for calculations with MCNPX

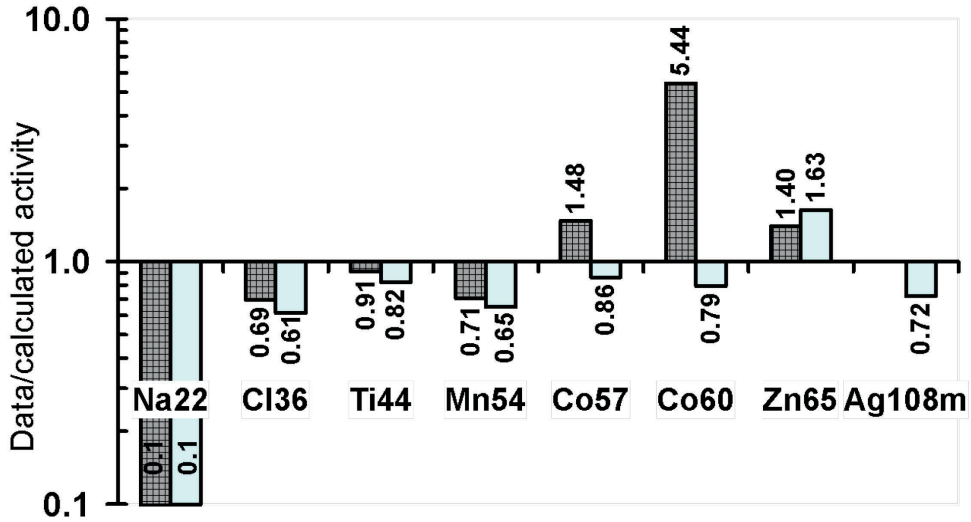


Fig. 7: Comparison of the measured and calculated activities of several isotopes produced in the safety hull of SINQ target 3.

The ratio of the measured to the calculated activity is shown in Fig. 7 for several isotopes in the sample from the safety hull. The two shades of grey represent two different material compositions that were used for the calculation. The main differences in these compositions were that the first composition (shown in the darker shade) had a lower Co content and no Ag. As usual, the exact material composition was not known, since an analysis of the material of this particular component was not performed. The concentration of Ag assumed was only 40 parts per million, but this produced a measurable activity of 72 Bq/g ^{108m}Ag, which is a factor of 10 above the free-exemption limit. This isotope is produced by neutron capture by elementary silver. Therefore no ^{108m}Ag was obtained with the first material composition, where silver was absent. This example reveals that the material

composition is an important input and source of uncertainty when one is calculating nuclide inventories. When the calculated activity is compared with the ^{60}Co data, the higher Co content in the second material composition fits the measured activity better. Except for ^{22}Na, good agreement is achieved between the measured and calculated activities. Using a later version of the Cinder'90 library, the measured value of Na-22 could be reproduced almost exactly. The new value was similar to the activity obtained with SP-FISPACT, which leads to a ratio of 0.72.

4 Radioactive waste from accelerator facilities

In every country, there are regulations that determine when a component that has been used in a radiologically controlled area can be freely released and when it has to be disposed of as radioactive waste in a final repository. In Switzerland, a (solid) component can be freely released when all of the following three conditions are fulfilled:

1. The dose rate at a distance of 10 cm is less than 0.1 μSv/h. The dose rate is equal to the absorbed energy in 1 kg of human tissue times a biological factor, which takes the damage done to human tissue into account.
2. The sum rule $\sum_i (A_i/LE_i) < 1$ is obeyed, where the LE_i are the exemption limits given in the radioprotection regulations. The sum runs over all radioisotopes.
3. The surface of the component is free of radioactive particles (no contamination) according to the limits defined in the regulations.

Radioactive waste has to be kept for 30 years if there is a chance that the material might be freely released at the end of this storage period or at some earlier time. If this is not the case, components have to be disposed of as radioactive waste. For this purpose, it is required that the documentation accompanying a container of waste contains, among other things, a nuclide inventory of the components in the container. The nuclide inventory should be as complete as possible and should not only provide the main isotopes. However, in practice it is almost impossible to determine all isotopes experimentally. As already mentioned, isotopes which cannot be identified by their gamma spectra are difficult to measure. To maintain confidence in the codes used for calculating nuclide inventories and maintain the acceptability of those codes, comparisons of the calculated results with experimental data are periodically performed. The code used most often for this purpose is called PWWMBS and was developed at the PSI [13]; it will be described in more detail below.

In the last few years, the authorities in Switzerland have also required an estimate of the total amount of future waste, including that arising from decommissioning, when operating approval is being sought for an installation or facility. These calculations usually require a huge effort, since particle transport Monte Carlo calculations are necessary for large regions, sometimes up to the outer biological shielding. This requires biasing techniques, since otherwise only very few particles reach the outer shielding, not enough to make statistically conclusive statements about the nuclide inventory. When the flux consists mainly of low-energy neutrons, it is possible to obtain the activities by folding the flux with known cross-sections using an external program.

The radioactive waste at the PSI is produced mainly at the proton accelerator. The amount of waste from the electron accelerator at the Swiss Light Source (SLS) is negligible. The reason for this is that electrons produce mainly bremsstrahlung and undergo almost no nuclear reactions, although a small probability exists that hard gamma photons will produce neutrons via photonuclear reactions. Compared with a hadron facility, however, the residual activity remains small. In contrast to nuclear power plants, where high-level waste is regularly produced when the spent fuel elements are exchanged, most of the operating waste at the PSI is low-level waste. The majority of components have dose rates below 100 μSv/h. In a few hundred years, the activity of these components will decay to a level where they can be freely released. In France, there are repositories on the surface already in

use for waste which needs a cooling time of less than 300 years to be declared as conventional waste. In Switzerland, however, such a storage facility is not foreseen. Here, waste is only separated into constituents that do and do not generate a considerable dose. Fuel elements belong to the category HAA (High Activity Waste), and waste from accelerators to the category SMA (which refers to waste with low and intermediate levels of activity). Alpha-toxic waste, i.e. waste containing an amount of alpha-emitters larger than 20 000 Bq/g, is distributed between repositories for both types of waste according to its dose.

The regular operating waste at the PSI is 90% normal and stainless steel. Typical components are vacuum chambers, beam diagnostic elements, magnets, and shielding, and to a lesser amount targets, collimators, and beam dumps. The large amount of biological shielding made out of steel and concrete will be disposed of at the time of decommissioning of the facility. Owing to the higher energies of neutrons produced at accelerators compared with nuclear power plants, the inner shielding consists of a material denser than concrete; normal steel is usually chosen.

At the PSI, the solid radioactive waste is placed in concrete containers. Finally, the waste components in the container are fixed in place with concrete and the container is closed with a cover. The resulting package is suitable for a final repository. Because the combination of aluminium and concrete is a source of hydrogen, the surface area of aluminium components has to be reduced by melting them down; the molten metal is the cast into coquilles. On average, a container with an opening of size 1.26 m × 1.26 m and a height of 1.88 m can be filled with 4.5 t of waste; the last 30 containers achieved a better filling factor and held an average of 6 t of waste. In the case of aluminium, 36 coquilles can be fitted into a container, which gives a total of 1.8 t. The activity per container is about 10^{10}–10^{12} Bq. If there is no major reconstruction work at or in the close environment of the proton accelerator, only one container per year is filled.

5 Determination of the nuclide inventory for radioactive waste

As already mentioned, a nuclide inventory is required by the authorities before radioactive waste can be accepted for disposal in a repository. To perform a particle transport Monte Carlo simulation for every component in a container would require a huge effort, particularly if a complex geometry needed to be modelled. This is often the case when a component is far away from the loss point(s), i.e. the points where the primary beam interacts with a material to produce a shower of a variety of particles. If the particles have no charge, they can travel large distances and thus activate components far from the source (i.e. the loss point). (Although the muon is charged, its mean free path is large. In contrast to the electron, the main mechanism of energy loss, due to bremsstrahlung, is suppressed owing to its 206 times larger mass; the bremsstrahlung cross-section is proportional to $1/m^2$. As the muon is a lepton, its cross-section for hadronic reactions is small.) Because these loss points are surrounded by numerous components (some of which are listed above) and large blocks of shielding material, most of the radioactive waste in a container has not been directly irradiated by the primary beam. Among the neutral particles, only neutrons and π^0 particles can be produced from the 590 MeV proton beam available at the PSI. The π^0 has a lifetime of 8.4×10^{-17} s, much too short to cover a considerable distance. Therefore the only significant source of activation far from the loss points is the neutron flux. It should be mentioned that photons also reach far beyond the loss points; these are produced via bremsstrahlung from charged particles. However, they do not play a role, since their cross-section for nuclear reactions is much smaller than that for neutrons.

The basic idea is to use Eq. (1) to calculate the production rate for each isotope by folding the neutron flux spectrum with the corresponding cross-section. This is a fast and simple method, provided that the neutron flux spectra are known. The key to avoiding calculating the neutron spectra at every location by a complex Monte Carlo simulation is the observation that the shapes of the neutron spectra do not vary much within a large region, i.e. the energy dependence is almost constant

even though the amplitude varies. The reason is that high-energy neutrons lose energy while, at the same time, low-energy neutrons are captured or stopped. In the ideal case, the high- and low-energy parts are in energy balance.

This method is applied in the code PWWMBS [20], developed at the PSI. The name is an abbreviation of 'PSI West Waste Management Bookkeeping System'. The code contains its own cross-section library for neutrons from 2 MeV upwards, extracted from older models based on INC. This library is called PSIMECX [21]. For neutrons with energies less than 20 MeV, evaluated cross-sections are used. The neutron spectra are calculated under simplified assumptions about the geometrical layout, omitting many of the details. A crucial input is the material composition. As was shown above, small amounts of impurities can produce significant amounts of radioisotopes if the production cross-section is large. The material compositions used are averages over several material analyses performed on different types of materials used at the PSI. Several different methods can be applied to determine the material composition. Gaseous products are more difficult than other elements to determine. Furthermore, detection at the level of a few parts per million is more difficult to achieve for heavy elements.

To determine the nuclide inventory of a component, its irradiation period, location, and weight have to be known. PWWMBS contains a data bank which stores the charge delivered by the proton accelerator to the target stations integrated over one year, as well as the main shutdown periods. The corresponding neutron flux spectrum is chosen depending on the location. Since there is no normalization of the spectrum to the actual neutron flux, the nuclide inventory is calculated in an initial step with an arbitrarily assumed flux. The surface dose rate is then calculated from the resulting inventory and compared with the measured value, and the calculated dose and the nuclide inventory are scaled to the measured dose rate.

PWWMBS is applicable to more than 90% of the waste generated at the PSI. In the case of targets, collimators, and beam dumps, MCNPX or FLUKA is used to determine the nuclide inventory.

At CERN, a code based on the same principles has been developed recently, called Jeremy. Owing to the higher energies of the protons used at CERN, spectra for pions, protons, and photons are included as well as neutrons [22].

6 Example: activation of an indirectly irradiated component

In 2004, several samples were taken from the μE4 beam line. This beam line is one of the six beam lines around Target E, which is a graphite wheel rotating at 1 Hz. As a result of reaction of the 590 MeV protons with the graphite, pions are produced, in addition to neutrons. The pions decay into highly polarized muons. Both types of particles are used for experiments in the fields of basic physics and materials science. The μE4 beam line was reconstructed in 2004 with the aim of improving the efficiency with which muons were collected and transported to the experimental area. Many components were removed on this occasion and became radioactive waste. Samples were taken from the bending magnet ASK61, the shutter behind ASK61, and the shielding around the shutter. The location of ASK61 relative to Target E and the main beam line is shown in Fig. 8. Since the components in the μE4 beam line are not directly irradiated, it is justifiable to calculate their nuclide inventory with PWWMBS. We shall present a comparison between the measured and calculated values of the activity of several isotopes for a sample taken from the beam tube in front of ASK61.

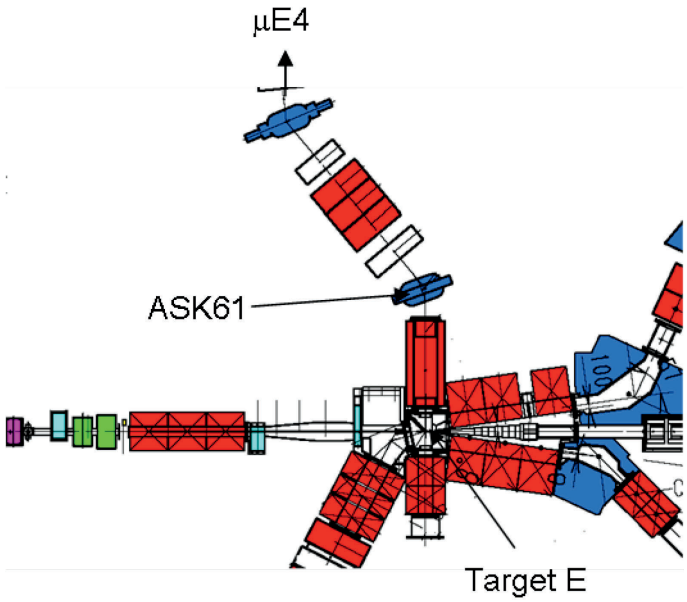


Fig. 8: Area around Target E with the μE4 beam line at the top

The beam tube was made from stainless steel and had been in place since 1991, when the Target E area was rebuilt. The sample was obtained by drilling a hole into the beam tube and therefore consisted of swarf. The operation was performed with a manipulator in a hot cell at the PSI. As usual, the activities of several isotopes were measured via gamma and beta spectroscopy. For long-lived isotopes such as ^{36}Cl and ^{26}Al, the activities were determined via AMS. Since the gamma measurement was performed as soon as possible after the last irradiation period, isotopes with half-lives of the order of a few months could be detected.

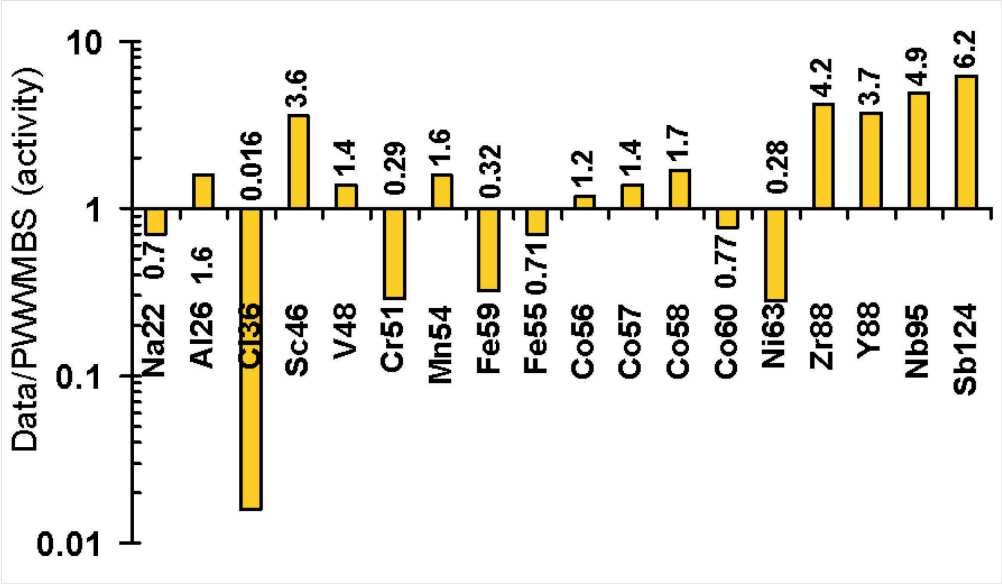


Fig. 9: Ratio of the measured to the calculated activity for several isotopes in a sample from the beam tube at the entry to ASK61

The experimentally determined activities of several isotopes are compared with the values calculated with PWWMBS in Fig. 9. The neutron flux spectrum for the region perpendicular to Target E was chosen. The dose rate at the location where the sample was taken was measured to be 70 mSv/h at the surface, two months after end of beam. (In fact, it was not possible to measure the surface dose

rate, but only the dose rate at some distance from the surface. This distance was 3 cm for the Geiger–Müller counter used for the measurements. Since the components were large compared with the sensitive area of the counter, the dose rate did not vary significantly with distance.) The results in Fig. 9 show fair agreement between the measured and calculated results, within a factor of 10; this factor is anticipated owing to the uncertainty of the method. An exception is the overestimation of ^{36}Cl in the calculation by almost a factor of 100. A likely reason for this is that the chloride content in the material composition used for the stainless steel was too large. The standard method of material analysis (inductively coupled plasma optical emission spectrometry) is not sensitive to gaseous elements, and therefore the Cl content was conservatively estimated. This example shows again the sensitivity to the material composition. The latter is an important source of uncertainty.

7 Radiation damage in materials

In this second part of the paper, another implication of particle irradiation of materials will be examined. This is the change in material properties due to damage to the lattice structure, which sometimes leads to the failure of components. This effect is called radiation damage. It is a threat particularly to components at loss points in high-power accelerators. These components include targets, beam dumps, and highly exposed collimators. There is renewed interest in the topic of radiation damage owing to new projects and initiatives which require high-power accelerators, and therefore materials which will withstand the high power for sufficiently long. One such project is the European Spallation Source (ESS), which will be built in Lund, Sweden [23]. A rotating wheel made of tungsten with a tantalum cladding is proposed for the target, which will be irradiated with 5 MW of 2.5 GeV protons. Some key values which are important when the behaviour of components under radiation is considered will be given below. The Facility for Rare Ion Beams (FRIB) will be built at the National Superconducting Cyclotron Laboratory (NSCL) at Michigan State University. This will deliver heavy ions with an extremely high power density of 20–60 MW/cm^3 [24]. The Daedalus project is an initiative at MIT with the aim of studying CP violation [25]. For this purpose, a neutrino beam is produced by three cyclotrons, each delivering a proton beam with an energy of about 800 MeV. The beam power on the target will be 1, 2, and 5 MW for the first, second, and third cyclotron, respectively. An upgrade to higher beam power is already foreseen. At the PSI, a 1.3 MW proton beam is routinely available, which constitutes the most intense proton source in the world at present. An upgrade to a higher beam power is planned (up to 1.8 MW) in the future.

For these projects, it is essential to know how long the heavily irradiated components will survive. In addition, improvement of the lifetime of components needs knowledge about the underlying mechanism of radiation damage and its relation to the changes in material properties. One problem is that the components cannot be tested under the same conditions as the ones they will be exposed to when the facility is in operation. Therefore the correlations between data obtained under different conditions need to be understood.

The macroscopic effects on structural materials caused by radiation damage are the following:

- hardening, which leads to a loss of ductility;
- embrittlement, which leads to fast crack propagation;
- growth and swelling, which lead to dimensional changes of components, and can also induce additional mechanical stress;
- increased corrosion rates, in particular in contact with fluids;
- irradiation creep, which leads to deformation of components;
- phase transformations in the material or segregation of alloying elements, which leads to changes in several mechanical and physical properties.

A component heavily irradiated by charged particles often needs cooling, since the charged particles deposit additional energy. Predictions of the temperature distribution in a component rely on a knowledge of the thermal conductivity of the material. Unfortunately, the thermal conductivity is subject to change. Its precise behaviour is not known for all conditions and materials. Owing to the damage to the lattice structure, one can expect that the thermal conductivity of a high-conductance material will decrease. This has been confirmed for most materials. The consequence is that the component might reach higher temperatures than foreseen, which could lead to the failure of the component. Since the thermal conductivity is difficult to measure, the electrical resistance can be determined with a nanovoltmeter. The two quantities are related to each other by the Franz–Wiedemann law, which contains a phenomenological constant, called the Lorenz factor; this varies for different materials and depends slightly on the temperature.

With pulsed sources of charged particles, components suffer from thermal cycles, which might lead to fatigue. Cracks may occur, which could lead to failure of the component. This phenomenon is also influenced by radiation. Thermal shock is best absorbed in materials with a low thermal expansion. Therefore the thermal expansion serves as a key parameter when one is examining materials after irradiation. Moreover, a drastic change in the thermal expansion with temperature is a clear sign of a phase transformation.

In the following, some examples of observations of radiation damage will be given. In preparation for the above-mentioned FRIB, several objects were studied at the NSCL with respect to radiation damage due to heavy ions. For this purpose, a 580 mg/cm^2 tungsten target, which corresponds to a thickness of 0.03 cm, was irradiated with $^{76}Ge^{30+}$ ions (which means that only 30 electrons were stripped off the Ge atoms) at an energy of 130 MeV/nucleon. The total energy of the ions was 130 MeV × 76, which leads to 9880 MeV. After irradiation of the tungsten foil with 5.77×10^{16} Ge ions on a beam spot with a diameter of 0.6–0.8 mm, a crack was observed. In Fig. 10, it can be seen that this crack is centred on the beam spot, where it has a small bow. Further investigations [24] revealed that the crack was caused by swelling and embrittlement, which induced stress in the foil. The stress might have been increased by thermal stress due to a decrease in the thermal conductivity.

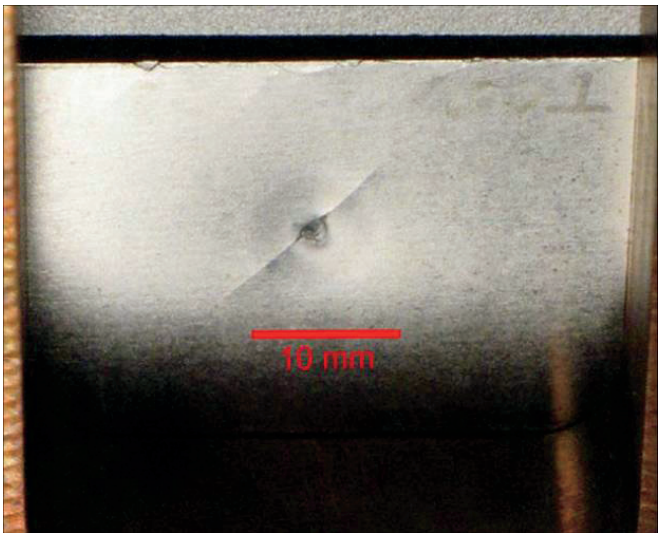


Fig. 10: Tungsten foil irradiated with $^{76}Ge^{30+}$ ions at the NSCL. Image taken from [24]

At the Los Alamos National Laboratory, tungsten was considered as a material for spallation targets. To investigate its suitability, hardness and compression tests were performed at room temperature and at 475°C on irradiated and unirradiated specimens. Tungsten rods were irradiated for up to 6 months with 800 MeV protons at a current of 1 mA, which corresponds to a dose of 23

Displacements Per Atom (DPA). The temperature during irradiation was kept constant for each sample; it varied between 50 and 270°C for different samples. In the compression tests, the samples were compressed to a strain of about 20%. The irradiated samples suffered from a loss of ductility, which showed up in the compression tests as a longitudinal crack, i.e. in the direction of the force. The compressive yield stress and the hardness increased linearly with the dose after a strong increase at small values. Optical micrographs of the tungsten compression specimens were also taken [26].

The pyrolytic graphite target at TRIUMF was cooled on the edges with water. After irradiation with 500 MeV protons at a current of 120 μA, the graphite delaminated, i.e. segmented into slices perpendicular to the beam. It is interesting that the target survived currents below 100 μA but never above that value. For details and a picture of the target after irradiation, see [27].

8 Underlying mechanism of radiation damage

To understand the mechanism of radiation damage, we have to take a look at the various interactions of particles with the atoms of a material. When particles penetrate into a material, they lose energy by several different mechanisms. These are

- electronic excitations;
- elastic interactions;
- inelastic reactions.

The first of these types of interaction is due to the Coulomb interaction and is therefore possible only for charged particles. Here, energy is used to shift electrons from the atomic core to an outer shell. This is called excitation and can also lead to the removal of an electron, i.e. ionization of the atom. The excess energy is dissipated as heat. In some cases this heat might also cause damage to a structural material, although this kind of damage has nothing to do with radiation damage. In elastic and inelastic interactions, energy and momentum are transferred from the particle to the nucleus. In the case of an elastic interaction, the nucleus is not changed but remains the same isotope. In all cases, the atom gains a recoil momentum. If the recoil energy exceeds a threshold (see below), the atom can leave its lattice place. Since the energy and momentum are transferred to the nucleus and not to the electrons, the atom moves in a partly ionized state through the lattice. The recoil energy is lost mainly by Coulomb interactions (ionization and excitation) and is again dissipated as heat. If the energy is large enough, the primary atom can knock on another atom, which again leaves its site. As a result, many atoms can be moved from their original lattice sites. This is called a displacement cascade. The third type of interaction, inelastic reactions, was described earlier in this paper in the part dealing with activation. This leads to transmutation of the nucleus, which can become radioactive. The transmuted nucleus, referred to as an impurity in the following, does not fit ideally into the lattice structure and therefore changes the mechanical properties of the material. Furthermore, in high-conductivity materials such as very pure copper, impurities are known to reduce the conductivity, i.e. they also have an influence on the physical properties. Usually, the damage done to the lattice by the recoils is much larger than that due to the impurities. An exception is when large amounts of helium and hydrogen are produced in highly energetic reactions. In particular, helium can have a considerable influence on the change in material properties, which is not yet fully understood. This will be discussed below.

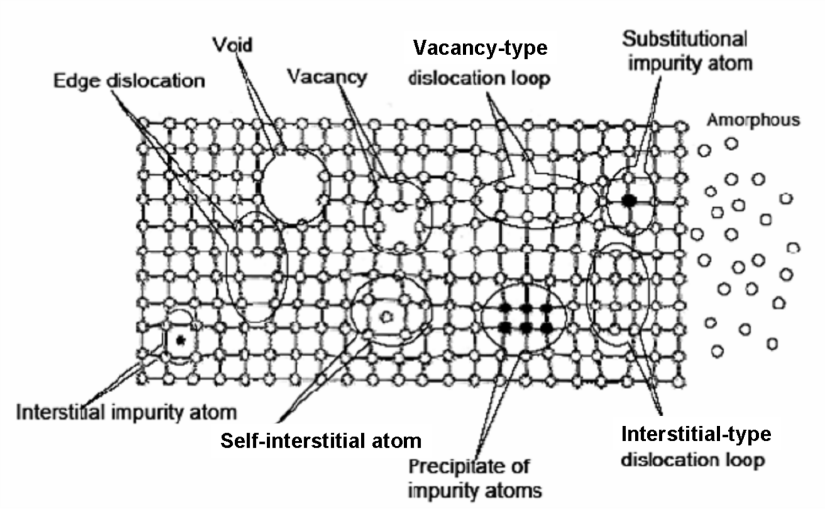


Fig. 11: The most important defects in a lattice structure (modified version of an image from Professor H. Föll, University of Kiel).

The most important defects in a lattice are shown in Fig. 11. The open dots belong to the original crystal, and the black dots indicate impurities. The simplest defects are the point defects, also known as zero-dimensional defects. The most prominent representatives of the point defects are self-interstitials and vacancies. Self-interstitials are atoms from the lattice which have left their lattice site for a site not provided in the lattice. The influence of a self-interstitial on its surroundings is a shift of neighbouring atoms away from the self-interstitial to make space for it. A vacancy is just the opposite of a self-interstitial. Here, a lattice atom is missing. These defects also exist in unirradiated materials. If a defect of this type is caused by irradiation, a vacancy and a self-interstitial appear in a pair. This is called a Frenkel pair. Also, an atom on an interstitial site may have been transmuted by an inelastic reaction to an impurity. It is then called an interstitial impurity atom or extrinsic interstitial.

The dislocation loop belongs to the class of one-dimensional defects. Here, part of a lattice plane is missing or has been added. There are two types of dislocation loop: the vacancy-type dislocation loop and the interstitial-type dislocation loop. In the vacancy type, part of a plane of lattice sites is missing. In the interstitial type, part of a plane of additional atoms has been incorporated into the lattice structure. Dislocations move under the influence of external forces, which cause internal stresses in a crystal. In the ideal case, dislocations move out of the lattice. If more than one plane is involved, a cluster is formed. If several planes are partly missing, one has an agglomeration of vacancies. This is called a void. An agglomeration of impurity atoms replacing neighbouring lattice sites on more than one plane is called a precipitate. Owing to their different sizes and properties, the neighbouring atoms are slightly shifted from their original positions. All of these defects make the lattice less flexible against strain, which manifests itself in a loss of ductility and an increase in hardness. In addition, the material becomes brittle. Small cracks can develop, which may grow further, and this can lead to the failure of a component.

Usually, the interaction with a particle with an energy of more than a few MeV does not cause single defects of the kind described above; instead, a large region containing millions of atoms is affected. For example, a nucleus in gold with a recoil energy of only 10 keV destroys the lattice structure in its surroundings within a radius of about 5 nm. This is called a displacement spike and happens within 1 ps. Since a huge number of atoms is involved in the process, a simulation via a Monte Carlo technique needs considerable effort and a large amount of computer power. Such a simulation has to solve the equation of motion for all atoms at the same time, since each atom can interact with and be influenced by all the other atoms. This is a multibody problem, and the computer time needed grows with the square of the recoil energy of the first knock-on atom. The higher the

recoil energy, the greater the number of atoms that are involved. Therefore such calculations are limited to recoil energies less than 100 keV for practical reasons. In addition, the simulation has to be repeated for every recoil energy. This kind of calculation is called Molecular Dynamics Simulation (MDS). The advantage is that the results are quite realistic, and the various kinds of defects produced can be studied in the simulation. The MDS method is the only way to evaluate how many defects disappear as a result of recombination with other defects (see below). Unfortunately, the MDS method can follow the process for only a few picoseconds, whereas the healing process can last for months (again, see below).

A faster but less accurate method is the Binary Collision Approximation (BCA), where only collisions between two (hence the name 'binary') atoms are considered. The other atoms are considered as spectators. The particles are followed via trajectories as in a Monte Carlo particle transport program. This calculational method is much faster than the MDS method and also works well at higher energies. However, when such approximations are made, much less information about the process and the state of the lattice is available compared with the MDS method. For example, no statements about the healing of defects can be made.

To estimate and quantify the severity of the damage, a phenomenological approach was developed by Norgett, Robinson, and Torrens, which dates back to the 1970s [28], known as the NRT model after the authors' initials. To quantify the radiation damage, a value is chosen which indicates how often each atom is displaced on average during the irradiation. This quantity is called the DPA, and is obtained by convolution of the energy-dependent particle fluence $\phi(E)$ (in units of cm^{-2}) with the displacement cross-section $\sigma_{\mathrm{disp}}(E)$:

$$DPA = \int \sigma_{\mathrm{disp}}(E)\ \frac{\mathrm{d}\phi(E)}{\mathrm{d}E}\ \mathrm{dE} \tag{9}$$

The displacement cross-section gives the number of displacements per primary (bombarding) particle or secondary particle (neutrons, protons, etc.). It is a function of the energy of the particle responsible for the damage. The displacement cross-section is obtained from the damage cross-section $\sigma_{\mathrm{dam}}(E, E_{\mathrm{R}})$. This damage cross-section is, in addition, a function of the recoil energy of the Primary Knock-on Atom (PKA), i.e. the atom which was first knocked on by the particle. The PKA displaces other atoms if its recoil energy is large enough (generating a displacement cascade). How many atoms are displaced is given by the damage function $\nu(E_{\mathrm{R}})$, which is the ratio of the energy available to the energy required for displacing atoms. The energy available for the displacement cascade is called the damage energy, T_{dam}. This is equal to the recoil energy minus the energy E_{e} dissipated in ionization and excitation of the atom. For recoil energies larger than 10 keV, most of the energy is lost by ionization. The fraction of the recoil energy left for the damage energy is called the partition function or, sometimes, the damage efficiency. To displace an atom, energy is required to break bonds. The amount of energy required is roughly twice the sublimation energy because, at the surface, only half of the bonding needs to be broken. In Cu, the energy needed ranges from 18 to 43 eV, depending on the crystal orientation [29]. In most calculations, the effective threshold energy E_{D} for copper is taken equal to 30 eV, but sometimes 40 eV is used. When the recoil energy E_{R} is larger than E_{D} but less than $2E_{\mathrm{D}}$, just one atom can be displaced. The PKA may be captured on the lattice site of the second atom. Since for $E_{\mathrm{R}} = 2E_{\mathrm{D}}$ only one atom is effectively displaced, the damage function is given by

$$\nu(E_{\mathrm{R}}) = \frac{\kappa T_{\mathrm{dam}}}{2E_{\mathrm{D}}} \tag{10}$$

The factor κ is set to 0.8. It compensates for forward scattering, which will not lead to a displaced atom, owing to its low energy transfer. For $E_{\mathrm{R}} > 2E_{\mathrm{D}}$, a cascade of collisions and displacements will take place.

The displacement cross-section is obtained by folding the damage cross section with the damage function:

$$\sigma_{\mathrm{disp}}(E) = \int_{E_{\mathrm{D}}}^{E\,\mathrm{max}} \frac{\mathrm{d}\sigma_{\mathrm{dam}}(E, E_{\mathrm{R}})}{\mathrm{d}E_{\mathrm{R}}}\, \nu(E_{\mathrm{R}})\, \mathrm{d}E_{\mathrm{R}} \tag{11}$$

The integration runs over all recoil energies from the threshold, i.e. E_{D}, to the maximum possible recoil energy. The damage cross-section is not a reaction cross-section but is related to the recoil energy spectrum $w(E_{\mathrm{R}})$ of the nuclei. It states how many nuclei can be found with a certain recoil energy. It is obtained from

$$\frac{\mathrm{d}\sigma_{\mathrm{dam}}(E, E_{\mathrm{R}})}{\mathrm{d}E_{\mathrm{R}}} = \frac{\mathrm{d}w(E, E_{\mathrm{R}})/\mathrm{d}E_{\mathrm{R}}}{xN_V}\,, \tag{12}$$

where x is the thickness of the sample and N_V is the atom density in atoms/cm^3. To obtain the recoil spectrum, the cross-sections of all reactions occurring in the material have to be known. Since Monte Carlo particle transport programs contain models for all nuclear reaction cross-sections over a wide energy range, a popular application of these programs is to use them to obtain the recoil spectrum. Here, it is important to use a thin target to avoid significant energy loss of the primary particle in the sample. If the object of interest has larger dimensions, the recoil spectrum has to be calculated for different energies of the primary particle to account for the energy loss of that particle. This requires several Monte Carlo runs. Furthermore, the fluence of the particles in the object of interest has to be obtained, which is a standard option in such codes. The DPA value is calculated from these two quantities via Eq. (9). Nowadays most Monte Carlo particle transport programs, such as FLUKA, PHITS, and MARS, already have a built-in option to obtain the DPA in one run. This is very convenient and avoids a larger effort.

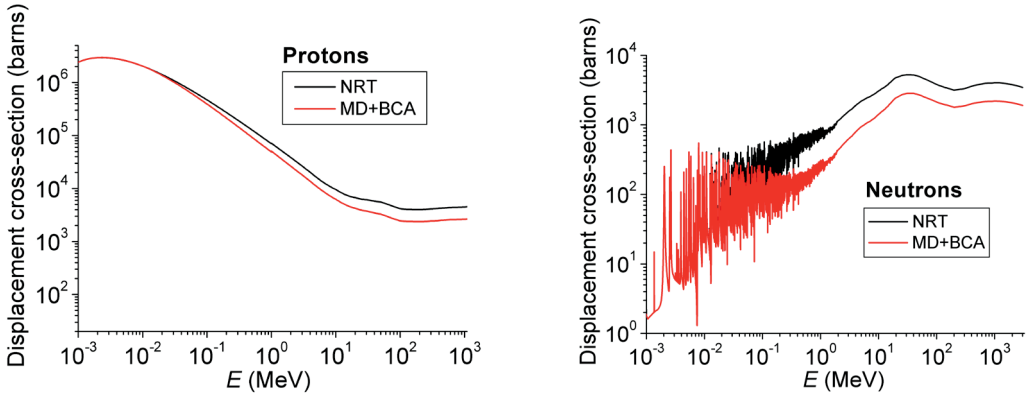


Fig. 12: Displacement cross-sections of protons (left) and neutrons (right) in copper obtained by two different approaches (see legend).

As an example, displacement cross-sections in copper are shown in Fig. 12 as a function of energy for protons and neutrons. These cross-sections were extracted by Konobeyev *et al.* [30] using two different approaches: the NRT model, and a combination of the MDS and BCA methods. Owing to the large computer power needed, the MDS method could only be used for energies less than 28 keV. For larger energies, the computation was continued using the BCA. Having these cross-sections, the DPA can be obtained simply via Eq. (9). For thin objects, where the primary energy is roughly constant and the fluence of secondary particles is small, the integration over the energy can be dropped. The displacement cross-sections for protons and neutrons are quite similar in size and shape except at small energies. There, the Coulomb interaction of the proton and the large capture cross-section of the neutron are the main drivers. Above the pion threshold at about 150 MeV, the cross-section is almost constant. This is due to the total inelastic reaction cross-section, which shows the

same behaviour for the two particles. The increase in the cross-section at smaller energies is due to the large elastic cross-section, which means that the particle uses its energy most efficiently to displace atoms and not for the release of particles as in the inelastic case.

It should be emphasized that the NRT approach is a simplified method. It completely neglects the details of the process of the displacement cascade. No interactions of the struck atom with the remaining lattice atoms are taken into account. Parameters of the crystal lattice such as the atomic bonding energy and the properties of the solid are completely absent. Instead, all this is condensed into the displacement threshold energy E_D. In the NRT model, it is implicitly assumed that the defect concentration is equal to the calculated number of displacements. Moreover, the displacements formed are taken to be stable. Molecular dynamics simulations have shown that the defects are not isolated Frenkel pairs as assumed in the NRT model, but are concentrated in a small region and influence each other. A high density of displaced atoms is produced in the first few tenths of a picosecond. This is called the collisional phase. In this phase, the number of displaced atoms is in fact much larger than that predicted by the NRT model. A few picoseconds later, most of the displaced atoms have recombined with vacancies. This is called 'healing'. The interstitial–vacancy annihilation process is completely omitted in the NRT model. This process is especially important at large PKA energies (>5 keV), where cascades of displaced atoms are produced in the initial state and defects are produced close to each other. At higher PKA energies (>20 keV), subcascades are formed and the number of recombination events decreases. Such a high-energy atom shakes the whole lattice and also deposits thermal energy, localized in the defect region. This makes the defects more mobile and facilitates recombination. The assumption of the NRT model that it is sufficient to count the initially produced Frenkel pairs cannot be justified at energies larger than 0.5 keV, where high-energy cascades start to develop. An example of the effect of healing in copper is shown in Fig. 13. The effective healing is just $1 - \eta$, where η is the defect or cascade efficiency. This is given as a function of the recoil energy of the PKA. The defect efficiency η is defined as the ratio of the number of Frenkel pairs at the end of phase 1, i.e. at the end of the collision cascade, obtained by an MDS, to the number obtained from the NRT model [31]. The MDS calculation here was performed for a temperature of 4 K. The results confirm that the NRT approach is only justified at small recoil energies. From 5 keV onwards, the recombination of defects dominates. The number of Frenkel pairs is five times lower than that predicted by the NRT model. It is interesting to note that the defect efficiencies in other materials such as W, Fe, and Al show very similar values, even though the final distribution of the defects is different.

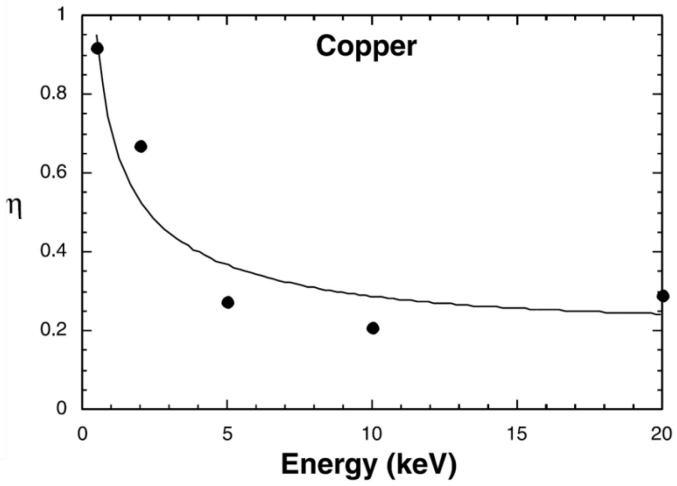


Fig. 13: Defect efficiency as a function of the recoil energy in copper [31] at 4 K.

The reduction of the defect efficiency at high PKA energy is important when one is comparing the damage produced by low- and high-energy particles. The materials that suffer radiation damage at

today's accelerators are irradiated by high-energy particles, whereas most of the material studies were done in reactors. Figure 13 suggests that one has just to multiply the recoil spectrum of the PKA by the defect efficiency to compare the results and to profit from the large data set that has been taken with reactor neutrons. A comparison of experimental data taken at BNL, where the electrical resistance was measured in irradiated samples of Cu and W, shows that the truth is somewhere in between the predictions obtained by the MDS and NRT methods [31]. The good news is that the NRT–DPA method provides a conservative estimate for predicting the lifetime of a component.

This was also observed in the case of a Cu collimator, called KHE2, in the 590 MeV proton beam line at the PSI. This collimator is located about 5 m behind a graphite target wheel 4 cm thick, which enlarges the beam spot size significantly owing to multiple scattering of protons in the target. The collimator is needed to reduce the beam size, and this also reduces the losses along the beam line. With a 2 mA proton beam on the graphite target, about 150 kW is deposited in KHE2, which has to be cooled with water. KHE2 is 30 cm long, with an outer diameter of 26 cm, and consists of six segments with an elliptical aperture (the half-axes are approximately 4 cm and 8 cm). In the first segment, within a radius of 10 cm, the DPA value obtained with MCNPX was 17.0. Using the NRT cross-sections for neutrons and protons from Konobeyev *et al.*'s work [30] (see Fig. 12), the DPA result was 13.2. At the inner sides of the collimator close to the beam, the DPA value was much higher. The radial distribution of the DPA was calculated with MARS15; the average value found for the DPA was 31.4. The value closest to the beam was about 150 DPA, i.e. five times higher. Although there was a problem that different calculations, all based on the NRT model, led to different results, the DPA at the inner side of the collimator was expected to be at least 80 DPA. In material studies in fast and thermal reactors, a linear volume increase of 0.5% per DPA was measured in several experiments at a temperature of about 400°C, which corresponds to the temperature at the inner side of the collimator at a current of 2 mA. The volume increase, also known as swelling, was measured up to a DPA value of 100, where the volume continued to increase linearly and no saturation was observed (see Ref. [32] and references therein). If one assumes 80 DPA in a volume of 1 cm^3, then the opening of the collimator is expected to have shrunk by 1.2 mm on each side. To clarify the situation, the collimator was taken out of the beam line during a long shutdown period of the accelerator. Owing to the high dose rate from the collimator, of up to 300 Sv/h, the handling had to be completely remote-controlled, and additional shielding was provided by an exchange flask (see M. Wohlmuther's contribution to this volume). The aperture was measured with two calibrated laser distance meters. The deviation between the expected and measured values was less than 0.2 mm. The accuracy of the measurement was estimated as 0.5 mm. Images taken with a high-resolution camera revealed no obvious or serious damage to the inner and outer sides of the collimator. Details of the measurement and results can be found in Ref. [33].

The DPA value calculated with the MD/BCA displacement cross-sections shown in Fig. 12 is 7.1. This would lead to 35 DPA at the inner side of the collimator, or a shrinkage of the opening by about 1 mm. This would still be measurable with the laser system used for the distance measurement.

The number of stable defects is reduced at higher temperatures. Two effects contribute to the healing of defects. First, the defect efficiency shown in Fig. 13 was obtained at 4 K, but at higher temperatures it is reduced. According to Ref. [34], the number of defects that survive the collisional phase at room temperature and higher is reduced to about 30% of the value at zero temperature. The reason is that the defects are more mobile at higher temperature. This also helps to reduce the number of defects in the period following the collisional phase; this is the second effect. At room temperature, interstitial clusters migrate to dislocations or recombine with vacancies. This reduces the total number of defects. Vice versa, vacancies also migrate and recombine with interstitials. Large vacancy clusters become mobile at about 400 K. The migration of defects takes time. At room temperature, 25% of the defects surviving the collisional phase disappear after 1 s. After a few months, 50% of the defects have been healed [31]. This effect has been confirmed experimentally with samples which were stored for longer times at room temperature. At higher temperatures, the healing effect would be even larger.

This might be an explanation for why so little damage was observed in KHE2. First, the operating temperature is about 350–400°C at the inner surface. Second, after eight or nine months every year, there is a shutdown period in which the defects have time to recombine.

Many experiments have been done to study the effect of heating samples after irradiation. This heating process is called annealing, and is used for treatment of materials before machining. In these experiments, physical and mechanical properties of samples are measured before and after irradiation, with and without heat treatment. It turns out that part of the damage can be restored, i.e., after the additional heat treatment, the measured values are closer to the original values.

Despite the uncertainty in the prediction of the production of defects in KHE2 and about the possible healing effects during and after the irradiation period, it is questionable whether the swelling rate obtained for irradiation with fast neutrons is the same as that for 590 MeV protons. Irradiation studies at high-energy accelerators are scarce. Some studies on aluminium can be found, because in the 1990s it was a candidate material for the inner wall of ITER. Fortunately, Al has the same lattice structure as Cu (faced-centred cubic, fcc), and therefore the results regarding swelling under irradiation with 600 MeV protons will be summarized here. Details and further references can be found in Ref. [35]. The experiments were carried out at the PSI, which had at that time a station (PIREX) dedicated to irradiation studies. Al foils were irradiated at 120°C with a damage rate of 3.5×10^{-6} DPA/s. Samples with doses between 0.2 and 5 DPA were produced. Their microstructure was examined by Transmission Electron Microscopy (TEM), which revealed helium bubbles, as well as voids and other defects. The helium production rate per DPA was about 200 appm (atomic parts per million), which is mentioned here because helium influences the swelling rate (see below). A surprising result [35] was that void formation was observed only at the grain boundaries, and occurred in a specific heterogeneous fashion. No voids were visible in the grain interiors. The original (unirradiated) grain size of the Al samples used was about 200 μm. The observations were completely different from observations on samples irradiated with electrons and neutrons, where there was no lack of voids in the grain interiors. In the TEM images of the proton-irradiated samples, large voids about 50 nm in size could be distinguished from small voids (~20 nm). It is thought that the large voids were nucleated by residual gas atoms, whereas the small voids were produced around helium atoms. Even the large voids in the proton-irradiated samples were much smaller (by a factor of about 3) than the voids produced by fast neutrons. The most important results for solving the 'KHE2 puzzle' came from a comparison of the swelling rates obtained for 600 MeV protons and fast neutrons. The swelling rate as a function of DPA is shown in Fig. 14. For small DPA values of the order of a few tenths, the swelling rates are the same for neutrons and protons. The proton data shown were taken only at the grain boundaries, i.e. in the peak zone (PZ), where the amount of void formation was largest. As already mentioned, no voids were observed in the grain interiors (GI), and therefore the swelling there was zero. At 1 DPA, the difference in the swelling rates in the peak zone for materials irradiated with protons and with neutrons is a factor of 10. Since there were no voids in the grain interiors, the macroscopic swelling rate during 600 MeV proton irradiation was almost negligible compared with that for neutron irradiation. If these findings can be confirmed for Cu also, this would be the key to understanding the good condition of KHE2 after 20 years of heavy proton irradiation.

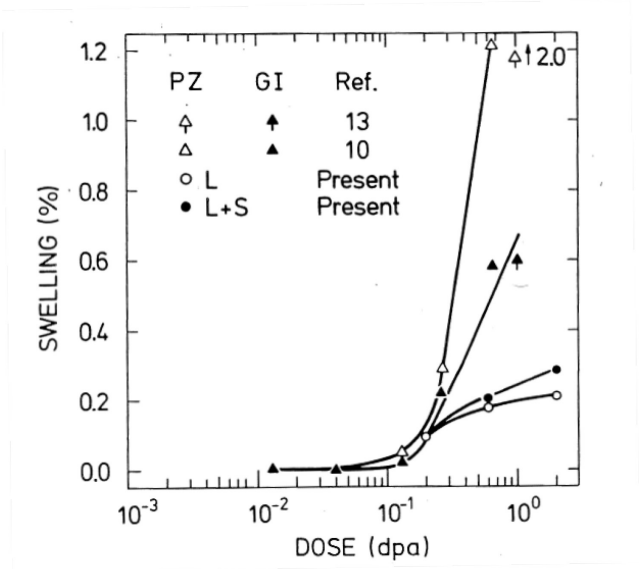


Fig. 14: Swelling of Al for irradiation with 600 MeV protons (circles) and fast neutrons (triangles) in the peak zone (PZ) and the grain interiors (GI). The proton data are for the peak zone only. L, large voids; S, small voids. Image reproduced from Ref. [35]. (The column headed 'Ref.' is not relevant to the present paper.)

Another approximation in the NRT model not yet mentioned is obviously the factor κ, which is taken to be constant in the original NRT model. In MARS, the displacement efficiency κ depends on the recoil energy. The values are chosen according to Ref. [36]. For example, at 0.1 keV, κ is 1.4. For higher recoil energies, κ drops to 0.3 at 100 keV. This modification effectively takes the defect survival into account. The value above 1 for κ is due to the fact that the NRT model results used for comparison were calculated with $T_d = 40$ eV instead of 30 eV. In [36], a very weak dependence of the defect efficiency on the temperature and the material was found. These studies were performed with the MDS method.

In their newest versions, the DPA calculations done with MARS and PHITS agree quite well in several cases [37, 38]. According to the authors of those studies, the reason for the agreement and the discrepancy with other codes is the correct treatment of the Coulomb interaction. For highly ionized heavy atoms, the Coulomb interaction is the main driving force up to a few GeV/nucleon. For protons, this is true only for energies up to 20 MeV. Above this energy, inelastic reactions are possible and a lot of secondary particles are produced. Then, mainly secondary particles interact with the atoms and transfer their recoil energy. Since the secondary particles have lower energies than the primary particle, the Coulomb interaction might again be important (depending on the set-up). However, the deviation of the MARS result for the DPA by a factor of two from the result obtained with Konobeyev *et al*.'s displacement cross-sections of cannot be understood in this way. These displacement cross-sections were derived with an explicit treatment of a screened Coulomb interaction. The screening is due to the electron cloud around the nucleus, which effectively reduces the charge of the nucleus for particles approaching the atom. This is only relevant for particles with energies of less than about 2 MeV, since faster particles come close to the nucleus. Moreover, the displacement cross-section for protons incident on Cu given by Konobeyev *et al*. agrees well with the evaluation of Jung [39] in the energy range from 0.5 keV to 20 MeV, which is critical with respect to the Coulomb interaction. Jung deduced cross-sections from measurements of the electrical resistivity at liquid helium temperature under different particle irradiations. Since the change in the electrical resistivity per Frenkel pair defect is known from X-ray scattering, measurements of the electrical resistivity can be related to the number of defects (displacements). More irradiation experiments, for example with neutrons and high-energy protons and a thin-plate target, where no secondaries are produced, are needed to resolve the discrepancy.

Regardless of the details of the calculation and the different results, however, the DPA value predicted by the NRT model is obviously too high and does not reflect the actual number of defects in the material. Unfortunately, the DPA value calculated in the NRT model cannot be measured directly. The reason is that only a small fraction of the displaced atoms lead to permanent lattice defects, as already explained. The state of defects estimated by the NRT model is that after a few picoseconds. The advantage of the NRT–DPA method is that calculations can be done more easily and much faster than MDS calculations. It serves the purpose of comparing and quantifying radiation damage induced in materials irradiated under slightly different conditions. If the irradiation conditions are significantly different, i.e. if the mechanism of defect production changes, the NRT–DPA values are not useful for comparison any more. This is the case, for example, if the lattice type of the material is different or the particles are changed in such a way that the interactions are different.

Heavier ions have a much shorter stopping range in materials than do light ions or protons. This means that the concentration of defects is much higher for the same kinetic energy, and the defects are more localized. At lower energies, it is possible that only the surface of the material might be damaged. Owing to the larger mass of the ions, their energy can also be transferred more efficiently to the target nuclei. At higher energies, the heavy ion rather than its secondary particles is the particle that transfers energy to the target nuclei, i.e. inelastic reactions are negligible. Up to 1 GeV/nucleon, the Coulomb interaction is the most important interaction. Owing to the dominant Coulomb interaction and the high charge, the damage cross-section for a U ion is four orders of magnitude higher than that for a proton at the same energy.

So far, the difficulty of predicting the number of stable defects has been discussed. However, the final number of stable defects may be interesting to scientists who wish to learn more about the underlying mechanisms. In practical cases, the macroscopic effects on the material properties matter. The main changes in materials which can occur under irradiation were listed at the beginning of this paper. The prediction of the behaviour of material properties under irradiation is itself difficult. Much more challenging is the prediction of when this will lead to the failure of a component. The influence of radiation on materials depends on many parameters, such as the rate of irradiation and the type of particle causing the damage. Another category of relevant parameters deals with the structure of the material. These include the grain size of the material and the presence of boundaries, and impurities which have an influence on the lattice. These impurities might be present in the material from the beginning or might be produced during irradiation by inelastic interactions. In addition, the temperature plays an important role. On the one hand, it determines the rate of defect recombination, i.e. healing, after irradiation. On the other hand, the temperature during irradiation also determines which material properties are subject to change. To make the following statements more general, the temperature will be expressed as a fraction of the melting temperature T_m in kelvin. Hardening, i.e. the loss of ductility, occurs after a few tenths of a DPA at temperatures less than $0.4T_m$, and then saturates. From $0.2T_m$ to $0.4T_m$, the material may creep and become deformed. This requires some mobility of the defects, which increases with temperature. Swelling due to voids or vacancies is expected at around $0.5T_m$. Among the impurities produced during the irradiation period, helium is the most important. Since helium is mobile, it migrates into vacancies, which leads to their stabilization. Helium can also accumulate easily in metals, since it is not soluble. It can then form bubbles, which grow with increasing temperature. The result is that the metal becomes brittle and cracks can develop. Only a few atomic parts per million of helium is sufficient to lead to drastic changes in the material properties. Many experiments have performed, some using helium implantation, to study these effects.

Since spallation reactions are often accompanied by helium emission, the helium produced per DPA is much larger than in reactors. In fission reactors, about 0.5 appm is produced. This is negligible compared with the European Spallation Source, which is planned to operate with 2.5 GeV protons at a power of 5 MW. In this case there will be approximately 100 appm of helium per DPA produced in the steel window of the target [40]. In Ref. [40], a beam energy of 1.5 GeV was assumed. Since the inelastic cross-section is almost constant, the proton flux needed to achieve 5 MW will be reduced by

roughly a factor of two (corresponding to the ratio of the beam energies). This means that only 50 instead of 100 appm/DPA of helium will be produced, which is still a very large number, as we will see below. To calculate the DPA per year, one needs the displacement cross-section, which is approximately 3000 barn. The proton flux integrated over one year is roughly 10^{22} protons/cm^2. Multiplying the number of protons by the displacement cross-section according to Eq. (9) leads to 30 DPA per year. After three years, about 2% of the material will be helium. This calculation is not realistic, however, since the steel window will break before it reaches a helium concentration of 2%. A limit of 10 DPA was estimated for the lifetime of the MEGAPIE target, which consisted of a steel vessel (T91 martensitic steel) filled with liquid lead–bismuth eutectic [41]. This was used in the 590 MeV continuous proton beam at the PSI, and therefore the conditions of the irradiation at the ESS will be different. As has already been explained, such predictions can depend on many parameters.

It should be mentioned that about 10 times more hydrogen than helium is formed during the irradiation. Hydrogen is even more mobile than helium but it usually leaves the material, except in the case of metals which form hydrides. In this case hydrogen embrittlement is an issue, and occurs even at temperatures from $0.1T_\mathrm{m}$ to $0.4T_\mathrm{m}$. Above this temperature range, the bonding in the hydride becomes unstable and the hydrogen can leave the metal.

Owing to the dependence on several parameters, a large number of experiments have to be performed to make predictions for any specific case. In particular, not much data for high-energy accelerator particles is available. At present, it is not at all clear how to transfer the database obtained from irradiation with low-energy neutrons to high-energy particle beams. The solution of this very complex problem cannot come from theoretical considerations alone; irradiation test experiments are needed. The good news is that the displacement cross-section becomes constant at higher energies, and therefore experiments at energies of about 1 GeV are sufficient. The results can then be used over a wider range of energies.

9 Summary

In this paper, two subjects were presented, the activation of and damage to materials, which are both caused by reactions between incident particles and target nuclei. The predictive power of Monte Carlo simulations for activation is usually limited by the knowledge of the production cross-sections. Most of the reaction cross-sections have to be provided by theories, owing to the lack of experimental data. However, more information is available for common dose-relevant isotopes. Therefore the prediction of the dose for these is usually much more accurate than that of the amount of activity for an exotic isotope. In practical applications, another large source of uncertainty is the material composition. The activity of isotopes which are produced by direct reactions, for example neutron capture, is proportional to the abundance of the element from which they are produced. A knowledge of the nuclide inventory is required for the disposal of radioactive waste and also for the transport of radioactive material. The radioactive waste from accelerators consists mainly of steel with a low to intermediate level of radioactivity. This is different from the situation for nuclear power plants, where the amount of steel is much lower than the total volume of waste.

Often, radiation damage leads to the failure of a component which is being used in a harsh environment. As accelerators become increasingly powerful, the understanding of radiation damage and the prediction of the lifetime of components becomes increasingly important. Unfortunately, the mechanisms leading to radiation damage are difficult to predict from theory alone. In addition, their effect depends on many parameters, such as the temperature and the type and energy of the particles. Higher temperatures and higher energies lead to a higher rate of recombination of defects, i.e. healing. The DPA is widely used as a measure of irradiation doses leading to radiation damage. Unfortunately, its value can depend strongly on the code used for calculation. The relation between irradiation conditions and macroscopic effects such as hardening and loss of thermal conductivity is not yet

understood. Since experimental data taken at accelerator energies are scarce, more irradiation stations are needed to perform experiments under different conditions, the result of which can then be compared.

Acknowledgement

I would like to thank Sabine Teichmann and Yong Dai for reading the manuscript.

References

[1] H.W. Bertini, *Phys. Rev.* **131** (1963) 1801.

[2] H.W. Bertini, *Phys. Rev.* **188** (1969) 1711.

[3] L. Dresner, EVAP – A Fortran program for calculating the evaporation of various particles from excited compound nuclei, Oak Ridge National Laboratory, Report ORNL-TM-7882 (1981).

[4] F. Atchison, *Nucl. Instrum. Methods B* **259** (2007) 909.

[5] S. Mashnik, M. Baznat, K. Gudima, A. Sierk, and R. Prael, *J. Nucl. Radiochem. Sci.* **6**(2) (2005) A1.

[6] S.G. Mashnik, K.K. Gudima, R.E. Prael, A.J. Sierk, M.I. Baznat, and N.V. Mokhov, CEM03.03 and LAQGSM03.03 event generators for the MCNP6, MCNPX, and MARS15 transport codes, Joint ICTPIAEA Advanced Workshop on Model Codes for Spallation Reactions, ICTP, Trieste, Italy, e-print arXiv:0805.0751v2 (2008).

[7] D.B. Pelowitz *et al.*, MCNPX 2.7.0 extensions, Los Alamos National Laboratory, Report LA-UR-11-02295 (2011).

[8] A.S. Iljinov, V.G. Semenov, M.P. Semenova, N.M. Sobolevsky, L.V. Udovenko, *Production of Radionuclides at Intermediate Energies*, Landolt-Börnstein I/13 (Springer, Berlin, 1999).

[9] A. Fasso, A. Ferrari, J. Ranft, and P.R. Sala, FLUKA: a multi-particle transport code, CERN-2005-10 INFN/TC_05/11, SLAC-R-773 (2005).

[10] T. Enquist *et al.*, *Nucl. Phys. A* **686** (2001) 481.

[11] M.B. Chadwick, P. Oblozinsky, M. Herman *et al.*, *Nucl. Data Sheets* **107** (2006) 2931.

[12] A.J. Koning, *et al.*, The JEFF evaluated data project, Proc. International Conference on Nuclear Data for Science and Technology, Nice, 2007, Eds. O. Bersillon, F. Gunsing, E. Bauge, R. Jacqmin, and S. Leray (EDP Sciences, Les Ulis, France, 2008), p. 194.

[13] J.-C. Sublet, L.W. Packer, J. Kopecky, R.A. Forrest, A.J. Koning, and D.A. Rochman, The European activation file: EAF-2010 neutron-induced cross section library, CCFE Report R (10) 05 (2010).

[14] K. Niita, N. Matsuda, Y. Iwamoto, H. Iwase, T. Sato, H. Nakashima, Y. Sakamoto, and L. Sihver, PHITS: Particle and Heavy Ion Transport code System, Version 2.23, JAEA, Data/Code 2010-022 (2010).

[15] T. Kai *et al*., DCHAIN-SP 2001: High energy particle induced radioactivity calculation code, JAERI, Data/Code 2001-016 (2001).

[16] F.X. Gallmeier, W.L. Wilson, M. Wohlmuther, B.J. Micklich, E.B. Iveron, E. Pitcher, W. Lu, H.R. Trellue, C. Kelly, G. Muhrer, I.I. Popova, and P. Ferguson, Proc. 8th Int. Topical Meeting on Nuclear Applications and Utilization of Accelerators, Pocatello, ID (2007), p. 207.

[17] W. Wilson, S. Cowell, T. England, A. Hayes, and P. Möller, A manual for Cinder'90 version 07.4 codes and data, Los Alamos National Laboratory, Report LA-UR07-8412 (2007).

[18] R. Forrest, FISPACT-2007: user manual, UKAEA, Report FUS 534 (2007).

[19] N.V. Mokhov, Recent Mars15 developments: nuclide inventory, DPA and gas production, Fermilab, Report Conf-10-518-APC (2010).

[20] F. Atchison, PWWMBS: a computer based book-keeping system for radioactive waste from PSI-West accelerator complex, PSI, Report AN-96-01-20 (2001).

[21] F. Atchison, The PSIMECX medium-energy neutron activation cross section library, Parts I, II, III, IV, PSI, Reports 98-09, 98-10, 98-11, 98-12 (1998).

[22] R. Fröschl and M. Magistris, personal communication.

[23] European Spallation Source (ESS), http://ess-scandinavia.eu.

[24] R. Ronningen, Proc. 46th ICFA Advanced Beam Dynamics Workshop on High-Intensity and High-Brightness Hadron Beams (HB2010), Morschach, Switzerland, 2010, p. 662.

[25] J.M. Conrad *et al.*, arXiv:1012.4853v1 [hep-ex] (2010).

[26] S.A. Maloy *et al.*, *J. Nucl. Mater.* **343** (2005) 219.

[27] E.W. Blackmore *et al.*, Proc. PAC 2005, p. 1919.

[28] M. Norgett, M. Robinson, and I. Torrens, *Nucl. Eng. Des.* **33** (1975) 50.

[29] H.B. Huntington, *Phys. Rev.* **93** (1954) 1414.

[30] A. Konobeyev, U. Fischer, C.H.M. Broeders, and L. Zanini, Displacement cross section files for structural materials irradiated with neutrons and protons, IAEA, Report NDS-214 (2009), http://www-nds.iaea.org/displacement/iaea-nds-0214.pdf.

[31] M.J. Cartula *et al.*, *J. Nucl. Mater.* **296** (2001) 90.

[32] S. Zinkle, Proc. 15th Int. Symposium on Effects of Radiation on Materials, Eds. R.E. Stoller, A.S. Kumar, and D.S. Gelles (STP 1125, ASTM, Philadelphia, 1992), p. 813.

[33] A. Strinning, S.R.A. Adam, P. Baumann, V. Gandel, D.C. Kiselev, and Y. Lee, Proc. 46th ICFA Advanced Beam Dynamics Workshop on High-Intensity and High-Brightness Hadron Beams (HB2010), Morschach, Switzerland, 2010, p. 245.

[34] B.N. Singh, S. Zinkle *et al.*, *J. Nucl. Mater.* **206** (1993) 212.

[35] M. Victoria, W.V. Green, B.N. Singh, and T. Leffers, Proc. 13th Int. Symposium on Radiation Induced Changes in Microstructure, Eds. F.A. Garner, N.H. Packan, and A.S Kumar (STP 955, ASTM, Philadelphia, 1987), p. 233.

[36] R.E. Stoller, *J. Nucl. Mater.* **276** (2000) 22.

[37] Y. Iwamoto, K. Niita, T. Sawai, R.M. Ronningen, T. Baumann, DPA calculation for proton and heavy ion incident reactions in wide-energy region using PHITS code, Proc. 4th High Power Targetry Workshop, 2011, Malmö, Sweden, forthcoming.

[38] N. Mokhov and S. Striganov, Radiation Damage: Accelerator Surprises, Proc. 4th High Power Targetry Workshop, 2011, Malmö, Sweden, forthcoming.

[39] P. Jung, *J. Nucl. Mater.* **117** (1983) 70.

[40] K.N. Clausen, R. Eccleston, P. Fabi, T. Gutberlet, F. Mezei, and H. Tietze-Jaensch (Eds.), *The ESS Project Volume III Update: Technical Report*, ISBN3-893336-304-1 (2003), http://neutron.neutron-eu.net.

[41] Y. Dai *et al.*, *J. Nucl. Mater.* **356** (2006) 308.

Commissioning strategies and methods

John Galambos
Spallation Neutron Source, Oak Ridge National Laboratory, USA

Abstract
Accelerator beam commissioning is a challenging and exciting period. It is generally the first integrated operation of the many systems in an accelerator and, most importantly, of the beam. First, general preparation is discussed. Then general methods for initial beam commissioning are described, including methods for transverse and longitudinal beam set-up. The particular emphasis here is on tuning methods for linear accelerators.

1 Introduction

Beam commissioning is one of the most exciting times in the accelerator life cycle—the birth of the beam. As with the delivery of a child, things can be not only exciting but also stressful. In order for the commissioning experience to go as smoothly as possible, proper preparation is important, along with following certain guiding principles. Tools and principles for commissioning are discussed here. In keeping with the thrust of the other lectures in this school, emphasis is given to the commissioning of high-power proton linear accelerators, but many of the concepts discussed here are general.

High-power, high-intensity accelerators tend to be large, expensive facilities. Often these construction projects tend to run behind the anticipated schedule, with added pressure and reduced time available for beam commissioning. Nevertheless, several recently completed facilities have had successful beam commissioning despite minimal available time. These include the Spallation Neutron Source (SNS) [1–3], the Japan Proton Accelerator Research Complex (J-PARC) [4], and more recently the Large Hadron Collider [5].

Beam commissioning refers to the initial transport and tuning of the beam through an accelerator. However, much of the effort to accomplish this is put in beforehand. The preparation activities will be discussed first. Next, the methods used to ensure proper magnet and RF set-up are reviewed. Finally, some overall commissioning strategies are discussed.

2 Commissioning preparations

In general, it is much easier to shake out software and equipment before the actual commissioning, in a controlled environment, rather than after commissioning starts. Having the eyes of a commissioning team looking at you while you try to mend a system that is holding everyone up is a situation to be avoided. Generally, system engineers will test their individual components with acceptance tests, but additional verification is useful. For example, magnets may appear to be working properly, but there are ample opportunities for polarity swaps. A useful check is to carry a Hall probe along the beam line with the magnets powered and verify that the quadrupoles and dipoles all have the correct polarity.

One important commissioning tool is an ‘on-line’ model that can be used in predictive ways to help understand and correct the state of the machine. One of the primary inputs to such a beam model is the field strength of the magnets in the accelerator. Engineers usually provide magnet control via prescribed-current control, but the physicist needs to read or prescribe the magnetic field strength or focusing strength. A careful magnetic-field–current mapping should be done for the magnets prior to commissioning, and the magnetic-field ‘physics units’ must made available to the model in real time.

When measuring the magnetic-field–current mapping in a magnet-measuring set-up before magnet installation, one can also perform hysteresis tests on magnets containing iron, to determine whether a magnet-cycling procedure is needed for reliable magnetic-field setting (e.g. one can determine how fast the cycling can be performed).

In addition to providing a physically meaningful interface to the magnets, providing a physically meaningful interface to the RF accelerating structures is also useful. This, however, is more difficult. For the RF structures, there are two important parameters: the phase with respect to the beam, and the amplitude. The precise setting of these parameters using beam-based methods will be discussed below, but a rough calibration of the RF amplitude can be made a priori, using RF power measurements and knowledge of the cavity shunt impedance [6].

Beyond equipment checks, software reliability is always a concern. Recently, 'virtual accelerators' have been prepared before beam commissioning to simulate the behaviour of the beam [7, 8]. These virtual accelerators are model-based representations of the beam. They receive input for magnet and RF settings through the accelerator control system, initialize a model appropriately, and perform a beam simulation calculation. Typically, the models are simple linear matrix representations of the beam. Space charge is often treated as a defocusing (linear) correction term. But an important metric is computational speed; in the control room, one cannot usually afford to wait minutes for feedback from a model. From the model output, the responses of the beam-line instruments (e.g. beam position monitor signals) are then generated and sent out through the control system. These virtual accelerator set-ups certainly cannot catch many equipment-related issues, but are useful for debugging high-level applications used to analyse and control beam behaviour.

Finally, beam simulations are a critical part of the preparation for commissioning. Most beam simulations during the accelerator design stage tend to focus on the final, high-power, operational scenarios. But understanding the behaviour of the beam during the initial low-intensity commissioning stage is also important. These simulations form the basis of the beam-commissioning plan, and are used to define the requirements for beam diagnostics. Having a notion of what to expect for off-normal settings is useful, as this will likely be the case during the initial set-up. For example, low-energy proton beams are easily debunched, and if an RF structure initially has a synchronous phase that debunches the beam, downstream strip-line pickups may not be excited. There may be nothing wrong with the pickup, even though there is no signal coming from it. The beam distribution for an off-normal RF setting is simply different from the design expectation for which the pickup may have been designed.

Most beam-commissioning methods involve varying an external input to the beam (e.g. a magnetic field, RF phase, or RF amplitude) and observing the downstream effect on the beam with an instrument. Having a priori simulations of what to expect is important. Many of the commissioning techniques discussed below involve comparing measured beam changes with calculations of expected changes, using parametric variations in the changes applied. By comparing the observed changes with model calculations, one can calibrate the control of the external input and set it to a design value.

3 Transverse beam set-up

3.1 Trajectory correction

One the first beam-commissioning tasks is sending the beam down the centre of a section of beam pipe. Trajectory correction (or orbit correction in a circular machine) is one of the most common types of tuning. It involves (1) generating a response matrix for the change in the beam position at a measurement point (i.e. the position of a Beam Position Monitor, BPM) with the strength of each dipole corrector element to be used, (2) measuring the beam position at each BPM, (3) solving for the corrector strengths required to best return the beam to the quadrupole centres (or other desired

positions) along the beam line, and (4) applying the new corrector magnet settings and seeing if they work.

The response matrix can be generated in two ways: model-based or empirically. The advantage of the model-based approach is that it is fast. The model must be initialized with the magnet strengths in the beam line (correctors, focusing and bending magnets, etc.), and with the RF cavity set-up. Note that the latter requires knowledge of the synchronous phase of the beam relative to the RF (as RF cavities apply a transverse kick to off-axis proton beams), which is generally not known at the start of commissioning. So one can turn RF cavities off until they are properly set up, when doing initial trajectory correction, and then repeat the operation after the RF is set up. A disadvantage of the model-based approach is that it does not account for any BPM or quadrupole misalignments (at least initially—they can be added in to the model later). The empirical approach to response matrix generation has the disadvantage of being slower: one must apply a kick with each corrector element individually and measure the response of all downstream BPMs (or all BPMs in the case of a circular orbit). It also requires applying a large enough kick to generate a wave with a displacement sufficiently greater than the noise level of the BPMs (typically, a wave of displacements of a few millimetres is required to generate a clean response matrix), and this can sometimes be a high-loss exercise. The empirical approach has the advantage of including the effects of any quadrupole misalignments. Also, it accounts for any possible polarity swaps of dipole correctors and BPMs (if one is trying to centre the beam trajectory).

Regarding finding the solution for the best dipole corrector strengths in step 3 above, there are many approaches, ranging from simple least squares minimization of the beam displacements to sophisticated constrained optimization methods. The advantage of constrained optimization techniques is the straightforward inclusion of the limitations of the dipole corrector power supplies. With a sparse corrector deployment, the ability to minimize the beam trajectory displacement will likely be limited by the corrector power supplies. Another consideration in the solution is the weighting between the benefit gained in reducing the trajectory displacement versus the corrector strength applied. Sometimes only a small additional benefit in trajectory correction is gained from a large increase in a corrector power supply setting. These effects can be handled by appropriate weighting of the figure of merit used in the minimization procedure.

3.2 Orbit difference

During commissioning, many issues can cloud the beam-centring procedure, including polarity swaps of dipole correctors and BPMs. Trajectory difference techniques (or orbit difference techniques in the case of a circular device) provide a powerful technique for identifying these issues. The method is simple: apply a kick to the beam and compare the measured change in the beam trajectory with a model-predicted change. Although it is difficult to know the initial absolute position and angle of the beam at the start of a beam-line section, the prediction of the change in the trajectory is independent of the initial conditions of the beam (for a linear transport system). The concept is shown schematically in Fig. 1. The red line in this figure indicates the difference between the original beam trajectory (black) and the perturbed trajectory (blue). Note that the difference is zero upstream of the kick (for a non-ring application).

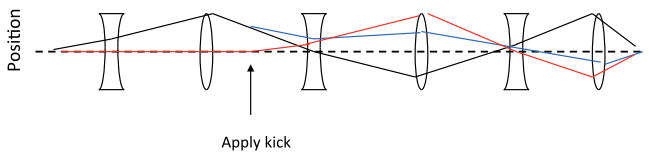


Fig. 1: Schematic illustration of the application of the trajectory difference technique. The original trajectory is shown in black, the trajectory after a kick has been applied is shown in blue, and the difference between the original and kicked trajectories is shown in red.

Figure 2 shows an example of an application of the trajectory difference technique. In this case one measured change at a BPM is opposite to the predicted change—a likely suspect for a BPM polarity issue. If, instead, all the measured beam position changes are opposite in sign to the expected changes, and only one corrector element shows this behaviour, there is likely to be a corrector polarity issue. It is also possible that the model-predicted and measured differences agree with each other, but both are incorrect. To determine if this is the case, one can vary the beam position at the location of an insertable device (such as a retractable-wire profile measurement device) and compare the measured sign change of the beam position at the insertable device with the expected sign change. This technique is useful for addressing systematic sign issues.

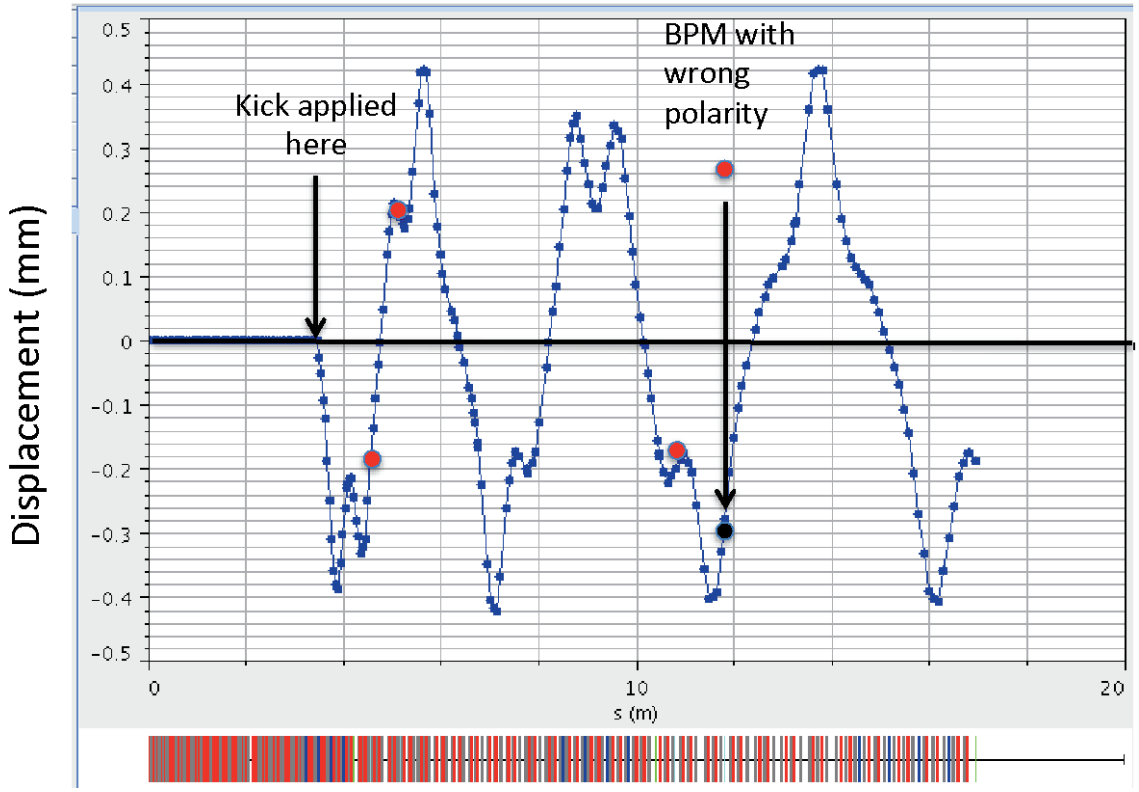


Fig. 2: Example of a trajectory difference. The line shows a model-predicted response of the trajectory to a dipole kick, and the large points show the measured BPM responses.

3.3 Transverse matching

Understanding the transverse beam size throughout a beam line and correcting it to the desired design lattice is a fundamental tuning process. A transverse beam mismatch increases the maximum beam size in the beam line, resulting in a closer approach to the aperture. Transverse matching optimizes the effective use of the aperture with respect to the core beam, and is a good starting step for beam loss reduction. Also, beam mismatch can cause halo growth [9].

Matching techniques require beam size measurements and independently adjustable quadrupoles. Typically, profile measurement stations are situated at the start of lattice transitions, and adjustable quadrupoles (often referred to as matching quads) are provided upstream of the beam measurement stations to facilitate matching. There are two primary approaches to transverse matching: (1) model-based methods and (2) a beam response matrix method.

3.3.1 Beam size measurement

For any transverse-matching algorithm, beam sizes along a lattice sequence must be measured. The most common method is to use wire scanner profile measurements [10, 11], which typically provide profiles of the beam distribution in the horizontal and vertical directions. Characteristic beam sizes must be obtained from these profiles, with the most common techniques being (1) calculation of the r.m.s. deviation from the centre of the beam, and (2) fitting the beam distribution with a Gaussian profile. The former technique is more general, as beam distributions are not always Gaussian. However, it is not easy to produce a robust generalization of this technique, and the results tend to be sensitive to the noise floor cut-off of the data. Usually, the Gaussian approximation works well. Figure 3 shows example linac beam profiles: one case is well approximated by a Gaussian profile but the other has a non-Gaussian beam halo. Additional pitfalls in the interpretation of profile data for use in transverse matching are discussed in Ref. [12].

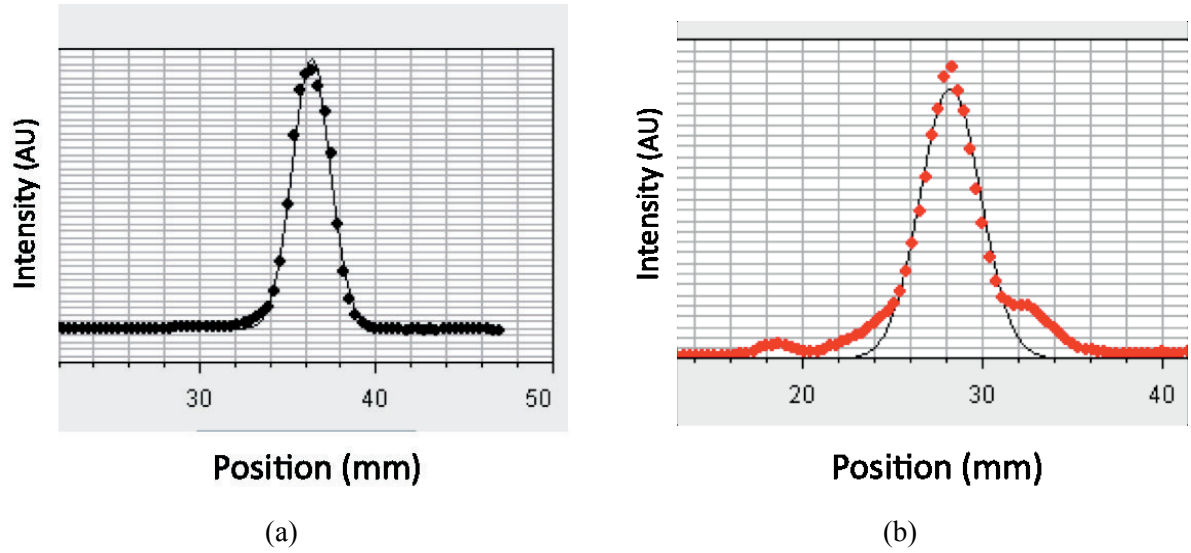


Fig. 3: Example of measurements of beam size from profile data. The dots show the wire-scanner-measured intensity, and the black lines show the best-fit Gaussian profile. Case (a) is an example at the beginning of the SNS linac (drift-tube linac section) where the beam is well represented by a Gaussian profile, and case (b) is an example at the end of the SNS linac where the beam has tails and is not well fitted by a Gaussian profile.

3.3.2 Model-based matching

An envelope model, given an initial set of Twiss parameters, can predict the r.m.s. beam size throughout a beam-line section. Envelope models (e.g. [13]) provide solutions for linear transport, often with corrections for space charge effects, which are important for high-intensity, low-energy beams. Given a set of beam size measurements in a beam line and knowledge of the lattice focusing strengths at the time of the measurements, one can solve for the initial Twiss parameters α, β, and ε [14] upstream of the measurements, so that the model-predicted beam sizes at the measurement points match the measured sizes best. The lattice region between the upstream Twiss solution point and the profile measurements should contain quadrupoles to be used for subsequent matching. At least three separate beam size measurements are needed to determine the three independent Twiss parameters. The use of three independent size measurement locations is easier, but it is possible to use a single beam size measurement station with different upstream focusing conditions. If only a single profile station is available, the upstream focusing should be varied to span a waist at the profile station to ease the fitting of the model.

Figure 4 shows an example of transverse matching. The dots represent measured beam sizes from wire profile devices and the lines show results from an envelope model. The blue lines represent the vertical beam size, and the red lines the horizontal beam size. The upstream Twiss parameters of the beam (at the point marked 'Initial model point' in Fig. 4(a)) were adjusted so that the model beam sizes best matched the measured sizes at the profile measurement stations. Note that in this case the problem is overconstrained, with five profile measurements and three Twiss variables per plane, so the model does not exactly match the measurements. The transport line section shown in Fig. 4(a) includes the end of the superconducting linac (from 150 to 171 m) and the start of a transport line (HEBT) from 171 m to 220 m. The superconducting linac portion of the beam line contains matching quadrupoles, and the HEBT portion is designed to have a FODO lattice structure. The initially measured beam sizes indicate that the beam in the HEBT is not well matched (the FODO lattice structure has a regular periodic beam size variation). Using the envelope model, the quadrupole strengths in the 'matching quadrupoles' region shown in Fig. 4(a) were varied to produce the design Twiss parameters α and β at the start of the HEBT. In this step, it is important to include any magnet limits in the solution for the quadrupole settings. Sometimes additional magnets must be employed in the matching process if some quadrupoles are being run near power supply limits. Figure 4(b) shows the measured beam sizes when the new strengths were applied. Now the proper FODO structure is apparent in the HEBT portion of the beam line, as seen by comparing the post-matching results in Fig. 4(b) with Fig. 4(c), which displays the design beam sizes at the beginning of the HEBT.

(a)

(b)

(c)

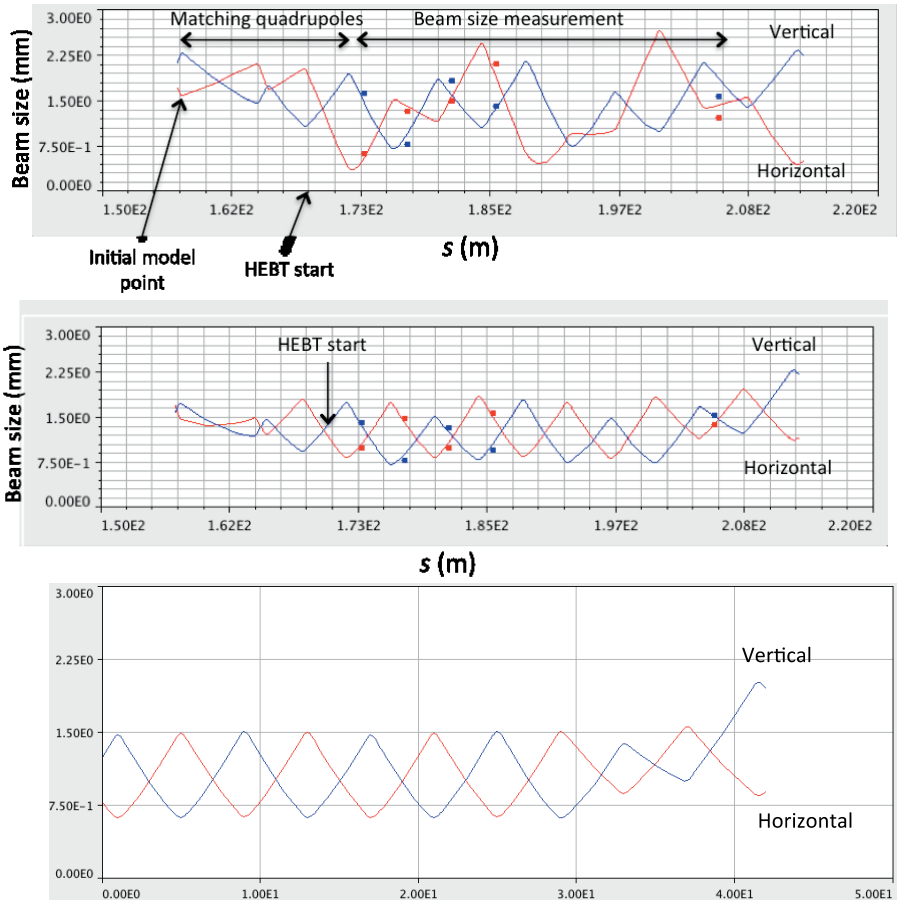


Fig. 4: Beam size along the beginning of the SNS HEBT transport line. The blue lines represent the vertical size and the red lines the horizontal size. The dots show measured values, and the lines show envelope model results. (a) Initial Twiss solution, (b) after applying quadrupole corrections, (c) the design lattice.

3.3.3 Beam response methods

The advantage of the model-based matching approach described above is that it is faster, as there is a minimal need for beam size measurements. However, it does require a machine-configured model, which may not available. If there is not a machine-configured model available, a beam response technique is sometimes possible. To generate the response matrix, beam sizes are measured while each matching quadrupole is varied independently. From this data ensemble, a response matrix can be constructed and used to provide quadrupole settings that produce the desired beam sizes at the measurement locations. For example, in a FODO structure, with profile measurements arranged to be adjacent to the quadrupoles, equal beam sizes should be expected at each profile station. When one is simply solving for the quadrupole strengths required to produce equal beam sizes at the measurement stations, a beam model is not needed. This approach is used in Refs. [15, 16].

4 Longitudinal beam set-up

4.1 Longitudinal RF set-up in linacs

In order to properly accelerate a beam in a linac, the phase of the RF field must be properly synchronized with the arrival of the beam at the entrance to the accelerating structure (referred to as a cavity here). For most copper accelerating structures, the amplitude must also be set to a prescribed level. Typical requirements for the RF set-up are a few degrees of RF phase, and about 1% in amplitude. Accurate setting of the accelerating structure is needed to minimize the loss of beam from the accelerating acceptance.

The primary goal is to determine (1) the calibration of the RF drive and the cavity voltage (setting the amplitude) and (2) the timing offset of the RF drive phase with respect to the arrival of the beam (setting the phase). The phase requirement typically translates to a few picoseconds, which requires beam-based solutions. Although a rough calibration of the cavity voltage with the RF drive can be done in advance, beam-based techniques offer resolutions within ~1%. These methods involve varying the RF phase or amplitude or both of a cavity, observing the effect on the beam downstream, and comparing the observed behaviour of the beam with a model prediction.

4.1.1 Energy degrader approach

An early approach to the RF set-up was the ‘energy degrader’ approach [17, 18], in which an intercepting material (degrader) of known thickness is inserted into the beam downstream of a cavity, followed by a charge-measuring device (e.g. a Faraday cup), as shown schematically in Fig. 5. The degrader thickness is chosen to stop an incoming beam with an energy slightly below the design output of the cavity. Thus, only when the output beam is close to being fully accelerated is charge detected. The cavity phase is scanned, and the width of an ‘acceptance’ can be determined by examining the width of the detected beam signal in the Faraday cup. This process is repeated for several amplitudes, as indicated in Fig. 6(a). The width of the acceptance is plotted versus the RF amplitude, as shown in Fig. 6(b), and the calibration between the RF drive and the actual cavity voltage can be determined by matching this curve with the expected trend based on model predictions for the given degrader thickness. The cavity phase setting is determined relative to the acceptance boundaries at the nominal amplitude setting, again by comparison with model-predicted expectations for the given degrader thickness. Finally, we note that the bunch width of the beam can also be determined from the width of the rise of the curves in the phase scans shown in Fig. 6(a). A drawback of this method is the need for an intercepting device. Also, at higher beam energies (above about 100 MeV for protons), the stopping distance becomes large and this approach becomes impractical.

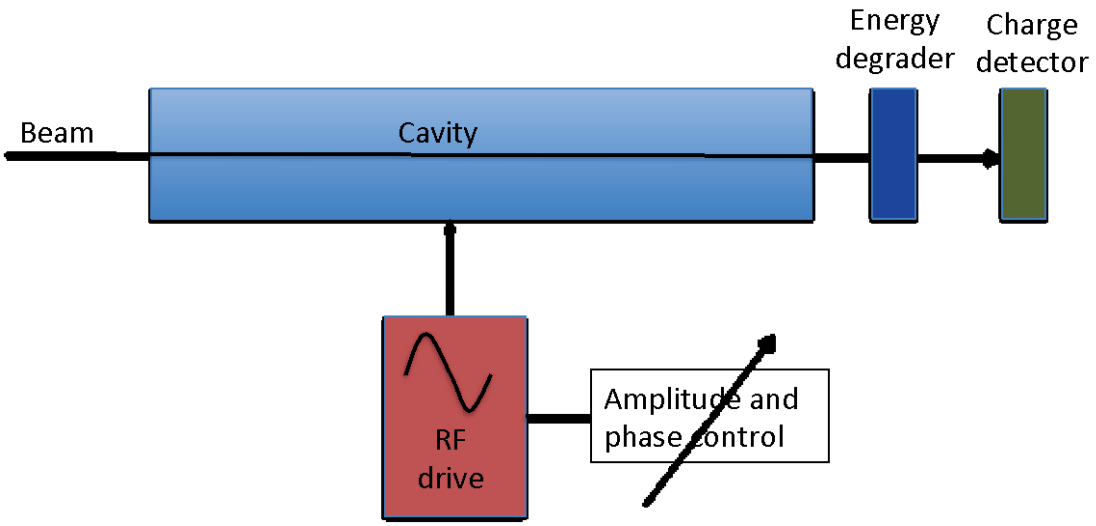
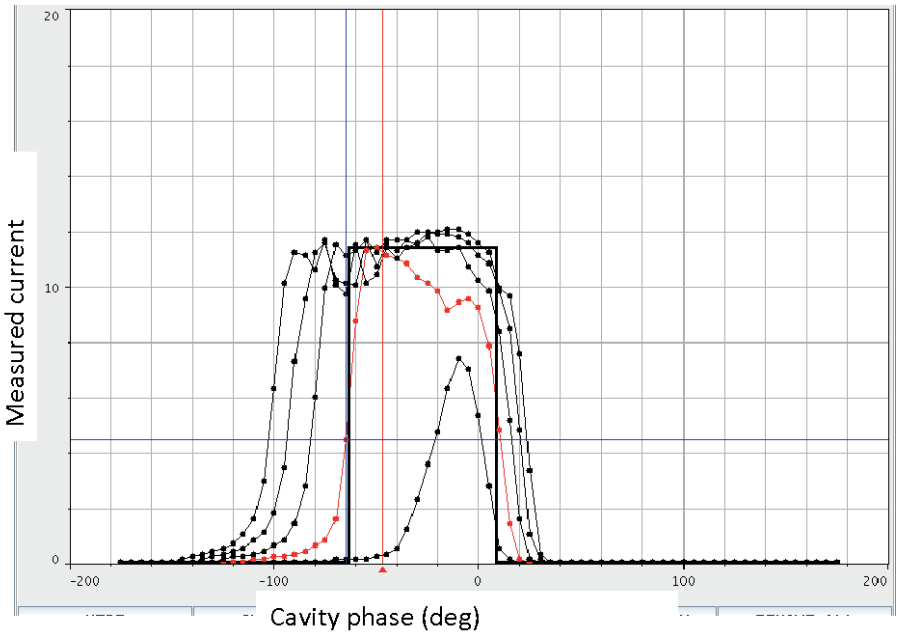


Fig. 5. Schematic illustration of the experimental set-up for the energy degrader method

(a)



(b)

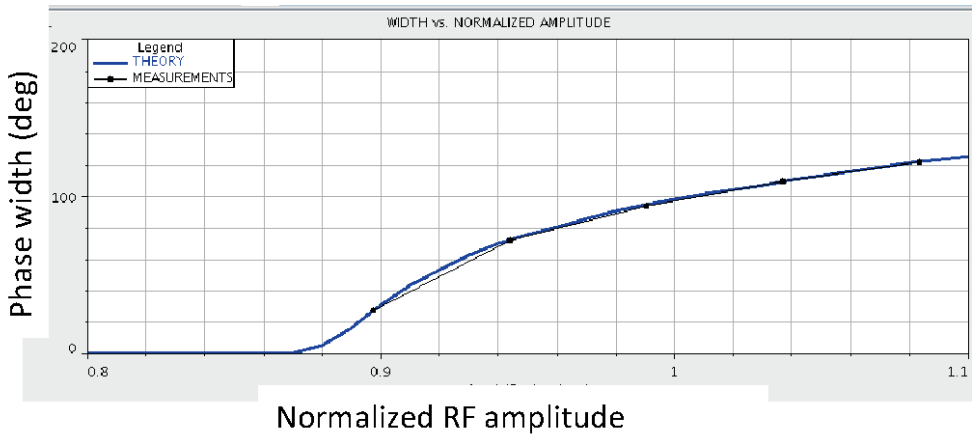


Fig. 6: Example of the application of the energy degrader method to a drift tube linac tank in the SNS linac. (a) Measured beam downstream of the degrader vs. RF phase for several amplitudes; (b) width of the detected acceptance windows in the scans vs. the amplitude setting of the RF drive.

4.1.2 Time-of-flight methods

A more common approach to setting the RF phases and amplitudes in a linac is a technique in which changes in the beam arrival time downstream of the accelerating structure are measured over a range of RF amplitude and phase settings, and the results are compared with model predictions. By adjusting the model RF settings to best match the observed beam behaviour, one can calibrate the phase and amplitude of the RF hardware. To reduce the sensitivity of the measurements to uncertainties, such as those in the distances between the cavity and the beam pickups, difference techniques are often employed. That is, changes in the Time of Flight (TOF) between two detectors are compared rather than the comparing changes in the beam arrival time at a single detector. The basic concept is shown schematically in Fig. 7. This class of RF set-up is referred to as the TOF method, and includes the Delta-T and signature-matching techniques. All of these techniques require accurate relative-phase measurements between the two beam bunch detectors for varying RF drive conditions. These TOF measurements are typically performed using dedicated diagnostics, for example Fast Current Transformers (FCTs) in the J-PARC linac [19] and specialized treatment of the BPM strip-line signals in the SNS linac [20, 21]. The typical accuracy of the measured changes in the beam TOF is ~1° of RF phase.

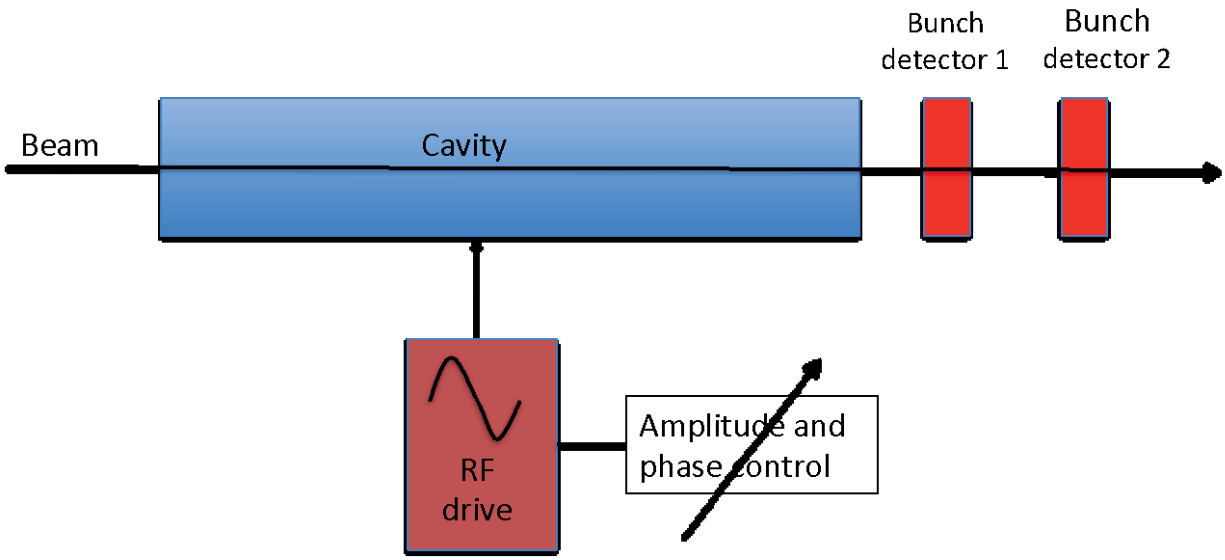


Fig. 7: Schematic set-up for beam-based time-of-flight measurements.

In all these approaches, there are three unknowns in the model, which are used to match the observed changes in the TOF with the predictions: (1) the RF amplitude calibration, (2) the RF phase offset, and (3) the energy of the incoming beam in the cavity being tuned. The beam models are typically simple: a beam bunch is treated as a single particle, and the Panofsky equation is solved for the synchronous particle throughout the gaps in the accelerating structure and the subsequent drifts through the TOF detectors.

A concern in the application of these schemes is maintaining the integrity of the bunch as it drifts between the RF structure and the phase detectors. Some pitfalls are beam debunching for low-energy beams (less than a few MeV), which can limit the range of useful RF phase variation. Also, any intervening RF structures between the cavity being tuned and the detectors must be turned off. In the case of superconducting cavities, excitation of intervening cavities by the drifting beam itself can impact on the measured arrival time, and care must be taken to use a beam of low enough intensity or to detune the cavities in some way so that they do not affect the drifting beam.

4.1.2.1 Delta-T method

The earliest of these TOF approaches was the Delta-T method [22–25], pioneered at Los Alamos National Laboratory. In this procedure, the predicted variations in arrival time at two downstream detectors B and C are calculated a priori, as difference values:

$$\Delta t_B = (t_{B\,off} - t_{B\,on}) - (t_{B\,off,\,design} - t_{B\,on,\,design}),$$

$$\Delta t_C = (t_{C\,off} - t_{C\,on}) - (t_{C\,off,\,design} - t_{C\,on,\,design}).$$

The subscripts 'off' and 'on' refer to the RF being on and off. The first term on the right-hand side is evaluated for small variations from the design beam energy, the design RF amplitude, and the design RF phase. The second term on the right-hand side is evaluated for the design conditions. An example of such a calculation is shown in Fig. 8. In this figure, there are three curve clusters, each cluster representing different incoming beam energies (nominal and a range of ±50 keV). Within a cluster, each curve represents a different RF amplitude input (±5%, with 1% increments). Along each curve, the RF phase is varied, with each dot representing a 2.5° step.

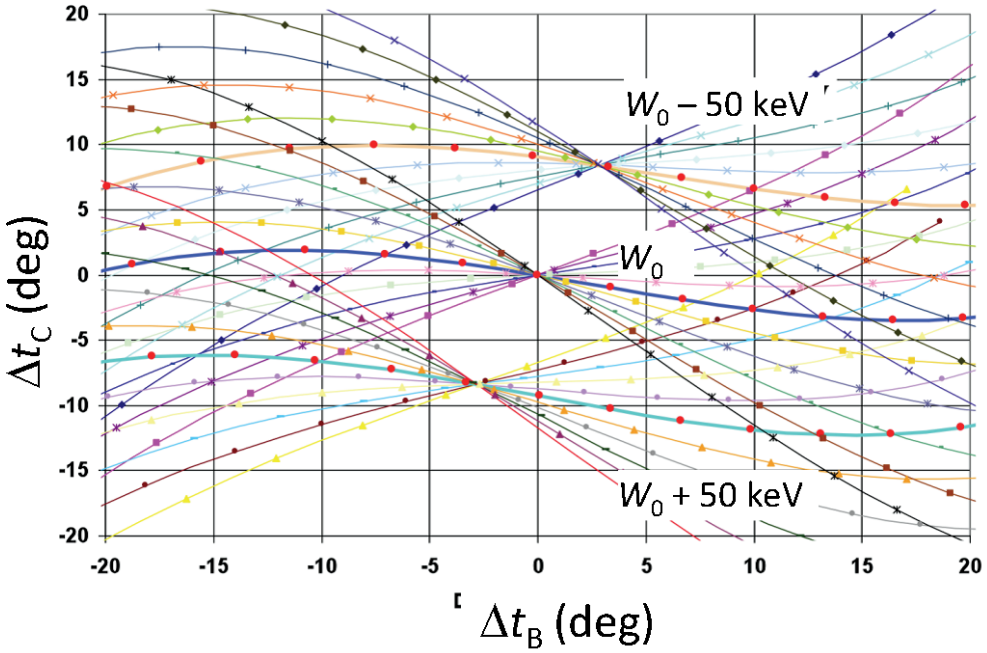


Fig. 8: Arrival time difference terms calculated for use in the Delta-T procedure for the SNS Coupled Cavity Linac module 1, with a design input energy of 86 MeV. Note that the time differences here are expressed in units of the detector frequency of 402.5 MHz.

Next, a measured curve of Δt_C versus Δt_B is obtained by scanning the RF phase and measuring arrival times with detectors B and C, with the RF turned on and turned off. An example measurement is shown in Fig. 9, including a linear fit of the measured values. For small deviations of the incoming beam energy, RF amplitude, and RF phase from their design values, the responses of Δt_B and Δt_C are linear in these changes. With this assumption, linear algebra techniques can be applied to determine where the measured relation between Δt_B and Δt_C best lies with respect to the calculated curves, which will have been tabulated. Knowing the RF amplitude and phase offsets, it is straightforward to determine how to change the RF phase and amplitude to reach the design values. The rising straight line in Fig. 9 passes through the locus of points defined by the RF design, but with energy deviations (the centres of the clusters in Fig. 8). The blue line is a linear fit to the measured differences. The intersection of these two lines defines the deviation the beam energy from the design.

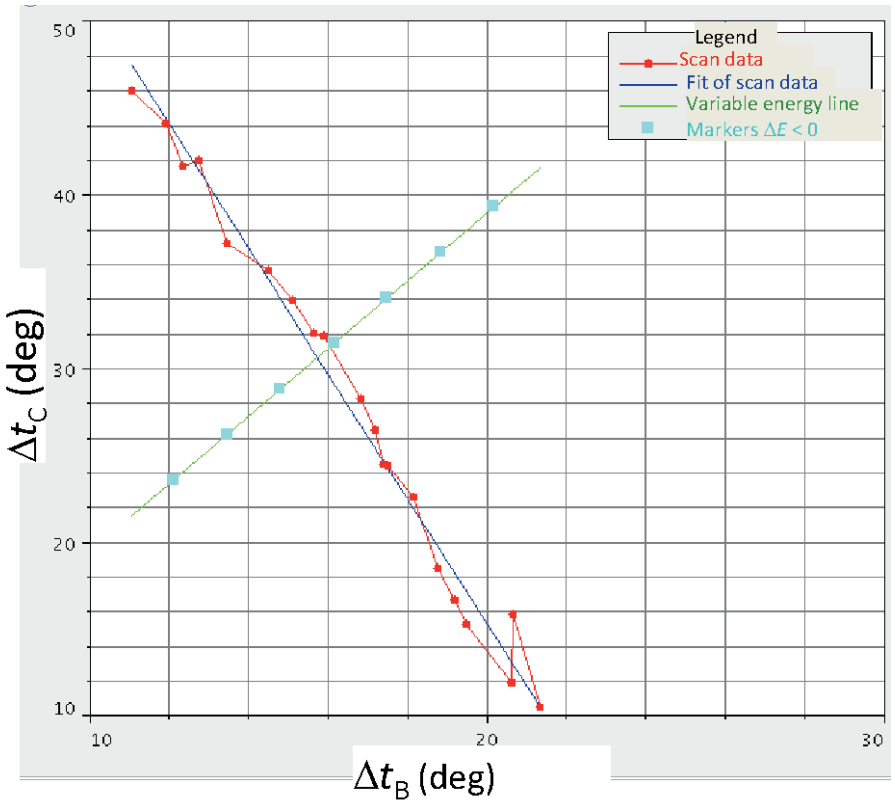


Fig. 9: Example of a Delta-T scan of beam phase detector differences for a coupled cavity module in the SNS linac (curve) and a linear fit (descending line).

The processing time for this method is extremely fast, which was an important consideration when the Delta-T method was first deployed, years ago. However, a limitation of this method is the assumption of a linear response of the beam arrival times at detectors B and C to changes in the RF amplitudes and phases. Even in Fig. 8, it is evident that for deviations beyond ±5° or ±10° the response is no longer linear. The method works well when the RF settings are close to the design values, but convergence can be difficult if the initial attempt is far from the design set-up.

4.1.2.2 Signature matching

More recently, a generalized approach to the problem of the cavity phase and amplitude has been developed, which is often referred to as a signature-matching scheme. The concept is simple: perform scans across a broad range of RF phase settings, amplitudes, or both, measure the variation in the TOF between two downstream detectors B and C, and use a model to match the measured TOF variation. The method was first demonstrated at the Fermi National Accelerator Laboratory [26], and has subsequently been used at the SNS [27] and J-PARC [28, 29]. This technique requires more computational resources than the Delta-T method, but with modern computers this is not a problem. No linearization of the response of the beam to RF changes is assumed, so this approach is more general and can be applied over wider ranges of RF conditions.

The usual approach is to measure the TOF variation between detectors B and C over a range of RF phases and amplitudes. As in the Delta-T method, measurements are taken with the RF off when possible, and the difference in TOF with the RF on and with it off is tracked:

$$\Delta_{\mathrm{TOF}} = (t_{\mathrm{C}} - t_{\mathrm{C\,off}}) - (t_{\mathrm{B}} - t_{\mathrm{B\,off}}),$$

where the t's are the measured arrival times with the RF on, and the subscripts 'off' represent the measured arrival times with the RF off. Parametric scans of this quantity at different RF phase and

amplitude settings are usually done, producing sets of nonlinear TOF responses. As in the Delta-T method, the model-predicted variations in this quantity are compared with the measured curves. Using optimization methods, the input beam energy, the RF amplitude calibration, and the RF phase offset with respect to the beam are varied to achieve the best match between the predicted and measured curves.

Examples of the application of this technique are shown in Fig. 10. The solid curves show the measured values of the TOF differences, and the dots represent model calculations after the solution was found. The case shown in Fig. 10(a) is for a Drift Tube Linac (DTL) tank at the SNS, with an input energy of about 40 MeV, a longitudinal phase advance of about 2π, 28 accelerating cells, and an energy gain of 17 MeV. RF scans were performed over about 40°, which is considerably more than the range used in the Delta-T method, and the two curves in Fig. 10(a) represent RF amplitude settings which differ by 3.5%. Often, at low energies, the phase scan range is limited by significant beam debunching at incorrect RF settings, which makes arrival time detection problematic. There is a strong nonlinear dependence of Δ_{TOF} on the input RF amplitude and phase, and only when the model has the proper calibration and the proper input beam energy do the model and measured values agree. For structures such as this with many accelerating cells, a large change in the beam β, and a fairly large phase advance, the dependence of Δ_{TOF} on the RF settings is quite strong and unique—hence the name 'signature matching' for this method.

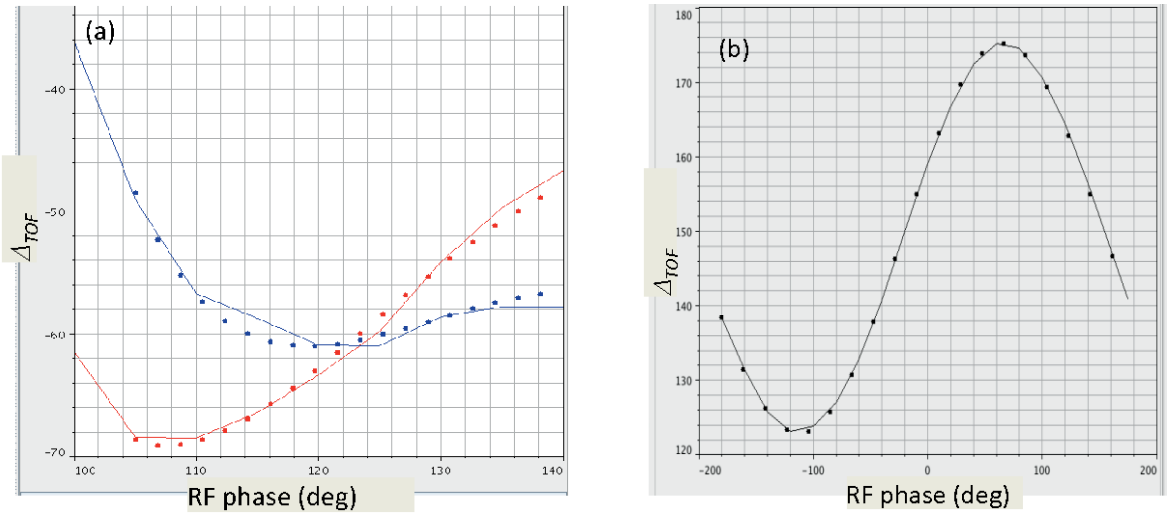


Fig. 10: Examples of the phase scan signature-matching technique for (a) the fourth drift tube linac tank at the SNS, and (b) a typical six-cell superconducting cavity at the SNS.

At high enough beam energies and for short enough accelerating structures, the beam remains bunched well enough for detection over a full 360° variation of the RF phase. An example of this is shown in Fig. 10(b), which is for a superconducting cavity with six cells, an energy gain of about 10 MeV, and a beam energy of about 350 MeV. In this case there is a small change in the beam β, and a small longitudinal phase advance through the cavity. The resultant variation of Δ_{TOF} is nearly sinusoidal, as would be expected from an ideal RF gap kick. The determination of the RF phase and amplitude calibrations and of the incoming beam energy is trivial in this case.

4.1.2.3 RF shaking

All of the above techniques are used to determine the proper RF set-up for a single accelerating structure. Often it is useful to check how well a group of accelerating structures is tuned collectively. A simple technique that is useful for this purpose is a beam-based difference technique, analogous to

the transverse-orbit-difference technique described in Section 3.2, but in this case in longitudinal space. Typically, a perturbation is applied to the RF drive in an upstream accelerating structure, and the change in beam arrival time is observed throughout a downstream set of accelerating structures. For example, the phase setting of the upstream RF cavity may be changed by 5°. The change in the arrival times of the beam in the downstream phase detectors in the range of cavities being examined is predicted with a model and compared with the measured differences in arrival time. This technique is a useful way of identifying an incorrectly tuned cavity, by observing where the deviation between the measured and predicted differences is initiated. The difference quantity at each detector i measured is

$$\Delta t_{\text{shake } i} = t_{\text{initial } i} - t_{\text{pert } i},$$

where the first term is the beam arrival time at detector i before the RF perturbation and the second term is the beam arrival time after the perturbation. Examples of this procedure are shown in Fig. 11. The first example (Fig. 11(a)) is for the SNS warm linac (a drift tube and coupled cavity structures) and includes an incorrectly set cavity, evidenced by the deviation between the model-predicted change and the observed change. The second example (Fig. 11(b)) shows a case related to the SNS superconducting linac, in which the cavities have properly set RF structures, evidenced by the agreement between the predicted and observed changes.

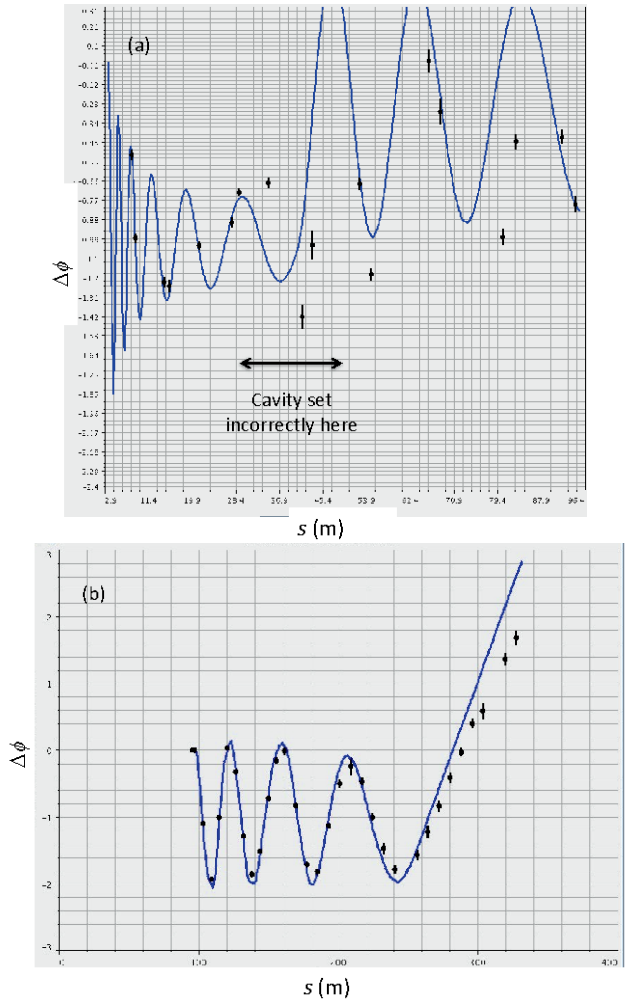


Fig. 11: The 'RF shaker' difference technique applied to SNS warm linac structures. (a) A case with an incorrectly set DTL tank; (b) a case of application to the SNS superconducting linac with cavities set correctly.

5 Overall commissioning strategies

The above sections have described the preparation for beam commissioning and some fundamental commissioning techniques. There are also some considerations about the commissioning strategy with respect to organizing commissioning schedules. Generally, it is quite difficult to turn on the equipment for a large accelerator complex, so the equipment is kept switched on for long periods. Given that the equipment is on, there is a natural tendency to use it as effectively as possible. At a large institution (e.g. [5]), there are sufficient personnel resources to run it 24 hours a day, seven days a week ('24/7'), with adequate support for the requisite subsystems (e.g. the physics, instrumentation, control, RF, and mechanical subsystems).

At smaller institutions, it is not possible to schedule complete coverage across many groups for extended periods. There are several different ways to handle this situation. At the SNS, the strategy was to have 24/7 coverage of beam physicists, and call in experts as needed [30]. This approach maximizes the utilization of potential beam time, but there is often downtime associated with the time needed for experts to come in when needed. Another approach was taken by J-PARC [30], using 12 h commissioning shifts, but with a broader contingent of experts available. This method allows issues to be addressed faster, but does not fully utilize the potential beam time (note that the machine hardware is still left on all the time). Commissioning has been successfully demonstrated using both approaches.

Another important overall commissioning strategy is to begin commissioning beam line as early as possible. Beam commissioning is the first opportunity for systems to work together in an integrated fashion, and no matter how careful the planning is, there are always surprises. Beam commissioning is the first real test of the ability of systems such as controls, RF systems, timing systems, machine protection systems, and diagnostics to work together, and system interface issues are often identified only when a beam is present. Staging commissioning in such a way that upstream accelerator components can be tested early allows systems (and interfaces) to be shaken out before they are fully deployed throughout the accelerator. Problems identified in a short early commissioning period can be addressed during subsequent downstream equipment installation, and greatly facilitate longer-term progress.

Finally, the most important advice about beam commissioning is to enjoy it. Although it may seem troublesome at the time, it passes quickly and is one of the most exciting periods in the lifetime of an accelerator.

Acknowledgement

ORNL is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the US Department of Energy.

References

[1] S. Henderson *et al*., SNS beam commissioning status, Proc. EPAC 2004, Lucerne, Switzerland, http://accelconf.web.cern.ch/AccelConf/e04/PAPERS/TUPLT168.PDF.

[2] S. Henderson *et al*., Proc. 2005 Particle Accelerator Conference, Knoxville, TN, pp. 3423–3425, http://accelconf.web.cern.ch/AccelConf/p05/PAPERS/FPAE058.PDF

[3] M. Plum, Proc. 2007 Particle Accelerator Conference, Albuquerque, NM, 2007, p. 2603.

[4] K. Hasegawa, Commissioning of the J-PARC linac, Proc. PAC07, Albuquerque, NM, 2007.

[5] M. Lamont, Proc. IPAC2011, San Sebastián, Spain, pp. 11–15, http://accelconf.web.cern.ch/AccelConf/IPAC2011/papers/moyaa01.pdf

[6] I. Campisi, Testing of the SBS superconducting cavities and cryomodules, Proc. 2005 Particle Accelerator Conference, Knoxville, TN, http://accelconf.web.cern.ch/AccelConf/p05/PAPERS/ROAC001.PDF

[7] H. Harada, K. Shigaki, F. Noda, H. Hotchi, H. Sako, H. Suzuki, Y. Irie, K. Furukawa, and S. Machida, Current status of virtual accelerator at J-PARC 3 GeV RCS, Proc. PAC07, Albuquerque, NM, 2007.

[8] A. Shishlo, P. Shu, J. Galambos, and T. Pelaia, The EPICS based virtual accelerator—concept and implementation, Proc. PAC, 2003.

[9] R. Duperrier, Review of instability mechanisms in ion linacs, Proc. HB2010, Morschach, Switzerland, 2010, http://accelconf.web.cern.ch/AccelConf/HB2010/papers/tho1b01.pdf

[10] M. Plum, W. Christensen, R. Meyer, and C. Rose, Proc. LINAC 2002, Gyeongju, Korea, pp. 172–174, http://accelconf.web.cern.ch/AccelConf/l02/PAPERS/MO458.PDF

[11] H. Akikawa, Z. Igarashi, M. Ikegami, S. Lee, Y. Kondo, S. Sato, T. Tomisawa, and A. Ueno, Proc. LINAC 2006, Knoxville, TN, pp. 293–295, http://accelconf.web.cern.ch/AccelConf/l06/PAPERS/TUP021.PDF

[12] C. Allen, W. Blokland, and J. Galambos, Beam profile measurements and matching at SNS: practical considerations and accommodations, Proc. XXV Linear Accelerator Conference, Tsukuba, Japan, 2010, forthcoming.

[13] C.K. Allen, K. Furukawa, M. Ikegami, Proc. LINAC 2006, Knoxville, TN, pp. 397–399, http://accelconf.web.cern.ch/AccelConf/l06/PAPERS/TUP064.PDF

[14] H. Wiedemann, *Particle Accelerator Physics, Basic Principles and Linear Beam Dynamics* (Springer, Berlin, 1993).

[15] H. Sako, A. Ueno, T. Ohkawa, Y. Kondo, T. Morishita, M. Ikegami, and H. Akikawa, Proceedings of LINAC08, Victoria, BC, Canada, pp. 260–262, http://accelconf.web.cern.ch/AccelConf/LINAC08/papers/mop078.pdf

[16] M. Ikegami, Transverse tuning scheme for J-PARC linac, Proc. 2005 Particle Accelerator Conference, Knoxville, TN.

[17] D. Jeon *et al*., *Nucl. Instrum. Methods A* **570** (2007) 187.

[18] D. Jeon, J. Stovall, and K. Crandall, Proc. LINAC2002, Gyeongju, Korea, pp. 368–370, http://accelconf.web.cern.ch/AccelConf/l02/PAPERS/TU427.PDF

[19] S. Lee, Z. Igarashi, M. Tanaka, S. Sato, H. Akikawa, F. Hiroki, T. Tomisawa, H. Yoshikawa, J. Kishiro, and T. Toyama, Proc. LINAC 2004, Lübeck, Germany, pp. 441–443, http://tdweb.fnal.gov/HINS/Library/BPM/TUP74_JPARC_stripline.pdf

[20] C. Deibele, S. Kurennoy, Matching BPM stripline electrodes to cables and electronics, Proc. 2005 Particle Accelerator Conference, Knoxville, TN.

[21] J.F. Power, M.W. Stettler, A.V. Aleksandrov, S. Assadi, W. Blokland, P. Chu, C. Deibele, J. Galambos, C.D. Long, J. Pogge, and A. Webster, Proc. LINAC 2006, Knoxville, TN, pp. 247–249, http://accelconf.web.cern.ch/AccelConf/l06/PAPERS/TUP003.PDF

[22] K.R. Crandall and D.A. Swenson, Side-coupled cavity turn-on problem, Los Alamos National Laboratory Internal Report MP-3- 98 (February 9, 1970).

[23] K.R. Crandall, The delta-t tuneup procedure for the LAMPF 805-MHz linac, Los Alamos National Laboratory Report LA-6374-MS (UC-28) (May 1976).

[24] S.V. Dvortsov, A.V. Feschenko, S.J. Jarylkapov, and P.N. Ostroumov, Proc. European Particle Accelerator Conf., Berlin, 1992, pp. 1209–1211, http://accelconf.web.cern.ch/AccelConf/e92/PDF/EPAC1992_1209.PDF

[25] A. Feschenko, S. Bragin, Yu. Kiselev, L. Kravchuk, O. Volodkevich, A. Aleksandrov, J. Galambos, S. Henderson, and A. Shishlo, Proc. 2005 Particle Accelerator Conference, Knoxville, TN, pp. 3065–3067.

[26] T.L. Owens, M.B. Popovic, E.S. McCrory, C.W. Schmidt, and L.J. Allen, *Part. Accel*. **98** (1994) 169.

[27] J. Galambos, A. Aleksandrov, C. Deibele, and S. Henderson, Proc. 2005 Particle Accelerator Conference, Knoxville, TN, pp. 1491–1493, http://accelconf.web.cern.ch/AccelConf/p05/PAPERS/FPAT016.PDF

[28] G. Shen, H. Sako, and S. Sato, Proc. PAC07, Albuquerque, NM, pp. 1529–1531, http://accelconf.web.cern.ch/AccelConf/p07/PAPERS/TUPAN062.PDF

[29] M. Ikegami, Y. Kondo, and A. Ueno, Proc. LINAC 2004, Lübeck, Germany, pp. 414–416, http://accelconf.web.cern.ch/AccelConf/l04/PAPERS/TUP65.PDF

[30] J. Galambos *et al*., Commissioning strategies, operations and performance, beam loss management, activation machine protection, Proc. Hadron Beam 2008, Nashville, TN, http://accelconf.web.cern.ch/AccelConf/HB2008/papers/cpl04.pdf

Reliability and fault tolerance in the European ADS project

Jean-Luc Biarrotte
CNRS/IN2P3, IPN Orsay, France

Abstract
After an introduction to the theory of reliability, this paper focuses on a description of the linear proton accelerator proposed for the European ADS demonstration project. Design issues are discussed and examples of cases of fault tolerance are given.

1 Introduction

The aim of an Accelerator-Driven System (ADS) is to transmute long-lived radioactive waste in a subcritical reactor. This typically requires a continuous proton beam with an energy of 600 MeV to 1 GeV, and a current of a few milliamps for demonstrators and a few tens of milliamps for large industrial systems. Such machines belong to the category of high-power proton accelerators, with an additional requirement of unprecedented reliability levels: because of the thermal stress induced in the subcritical core, the number of unwanted beam trips should not exceed a few per year, a specification that is far above usual performance and turns the issue of reliability into the main challenge and a constant consideration in all research and development activities pertaining to this type of accelerator.

This paper describes, basically, a reference solution adopted for such a machine, a superconducting linac with a combination of redundant and fault-tolerant schemes. The focus is primarily on the MYRRHA project, led by SCK•CEN in Belgium. A short presentation of the theory of reliability is also given, inspired mainly by Refs. [1, 2].

2 The basic concepts of reliability

Reliability deals with the analysis of failures, their causes, and their consequences. A commonly used definition of reliability is the following: 'reliability is the probability that a system will perform its intended function under a specified working condition—i.e. without failure—for a specified period of time'. This definition makes it clear that two important aspects have to be taken into account when speaking about reliability:

- a functional definition of failure is needed: in the case of an accelerator, a failure will typically be the absence of a beam on the target, or a beam on the target with the wrong parameters;
- a period of time, or 'mission time', is needed to define the reliability level of a system: unlike the availability, which measures the mean system uptime (see Section 2.3), the reliability is very time-dependent.

2.1 Reliability function, failure rate, mean time to failure

Mathematically, the reliability function $R(t)$ of a system can be defined as the probability that the system experiences no failures during the time interval 0 to t, given that it was operating at time zero. The reliability function therefore ranges between 0 and 1, by definition. In the simple but unrealistic case where a system systematically experiences a failure at time t_{fail}, the reliability of the system would be $R(t < t_{\text{fail}}) = 1$ and $R(t \geq t_{\text{fail}}) = 0$. In the real world, things are of course more complex, and statistical tools need to be used to correctly model how a system experiences failures.

All of the functions commonly used in reliability engineering can be derived directly from and described by a probability density function, namely the failure density distribution. The failure density distribution $f(t)$ of a system is the probability that the system experiences its first failure at time t, given that the system was operating at time zero.

Once this statistical distribution is defined, it is easy to derive the failure probability $F(t)$ of the system (see Fig. 1), which is the probability that the system experiences a failure between time zero and time t:

$$F(t) = \int_0^t f(x)\,dx\,. \tag{1}$$

The reliability function $R(t)$ can then be expressed simply as

$$R(t) = 1 - F(t) = \int_t^\infty f(x)\,dx\,. \tag{2}$$

Note also that, conversely,

$$f(t) = \frac{dF(t)}{dt} = -\frac{dR(t)}{dt}\,. \tag{3}$$

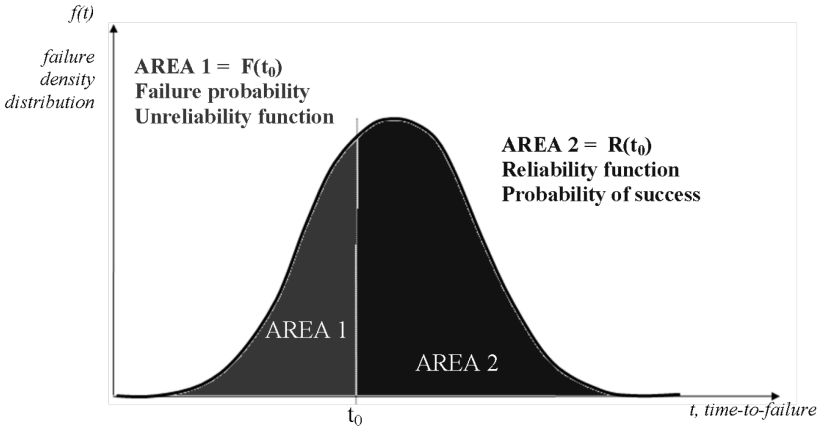


Fig. 1: Failure probability and reliability for a Gaussian distribution of failure density

The failure rate function $\lambda(t)$ is also an important concept; it enables the determination of the number of times the system will fail per unit time. Mathematically, it is given by

$$\lambda(t) = \frac{1}{R(t)}\frac{dF(t)}{dt} = \frac{f(t)}{R(t)}\,. \tag{4}$$

This instantaneous value is also known as the hazard function. It is useful for characterizing the failure behaviour of a product, determining the allocation of maintenance crews, and planning the provision of spares. Note that Eq. (4) also leads to

$$\ln(R(t)) = -\int_0^t \lambda(t)\,dt\,. \tag{5}$$

Finally, the mean time to failure (MTTF) is defined as the average time of operation of the system before a failure occurs. This value is widely used, and is often the main value of interest in characterizing the reliability of equipment. It can be computed from:

$$\mathrm{MTTF} = \int_0^\infty t f(t)\, dt\,. \tag{6}$$

The MTTF is used for non-repairable systems. When dealing with repairable systems, it is more meaningful to speak about the mean time between failures (MTBF). These two metrics are identical if the failure rate of the system is constant.

2.2 Commonly used distributions

The reliability function, failure rate function, and mean time functions can be determined directly from the failure density distribution $f(t)$. Several distributions exist, such as the normal (Gaussian), exponential, and Weibull distributions. The simplest of these and the most commonly used—even in cases to which it does not really apply—is the exponential distribution, which can be expressed as

$$f(t) = \lambda \mathrm{e}^{-\lambda t}, \tag{7}$$

where λ is a constant. In this case, one can derive the following relationships:

$$F(t) = 1 - \mathrm{e}^{-\lambda t}, \tag{8}$$

$$R(t) = \mathrm{e}^{-\lambda t}, \tag{9}$$

$$\lambda(t) = \lambda, \tag{10}$$

$$\mathrm{MTTF} = \frac{1}{\lambda}. \tag{11}$$

With this exponential distribution, the failure rate function $\lambda(t)$ is a constant. This means that in this case, the system does not have an ageing property. This assumption allows one to calculate the MTBF by dividing the total operating time of the system by the total number of failures encountered. Practically, this is usually valid for software systems, but most of the time, for hardware systems, the failure rate can have other, more complex shapes. A remarkable aspect of the exponential distribution is the fact that once the MTTF is known, the distribution is fully specified.

Another convenient distribution is the well-known normal distribution, given by

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}, \tag{12}$$

where μ is the mean value and σ is the standard deviation of the distribution. In this case, the failure rate function is always increasing, and the MTTF is equal to the mean value μ.

Two more powerful distributions are the lognormal and Weibull distributions. These can be applied to describe various failure processes correctly. They are given by

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}}\, \mathrm{e}^{-\frac{(\ln t-\mu)^2}{2\sigma^2}}, \tag{13}$$

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \mathrm{e}^{-\left(\frac{t}{\eta}\right)^{\beta}}, \tag{14}$$

respectively. For the lognormal distribution, μ is the mean value and σ is the standard deviation; for the Weibull distribution, the two parameters are the shape parameter β and the scale parameter η. Note that the Weibull distribution is very commonly used because it allows one to describe the ageing property of a system easily, by mixing different Weibull distributions for different stages of the life of the system, as illustrated in Fig. 2.

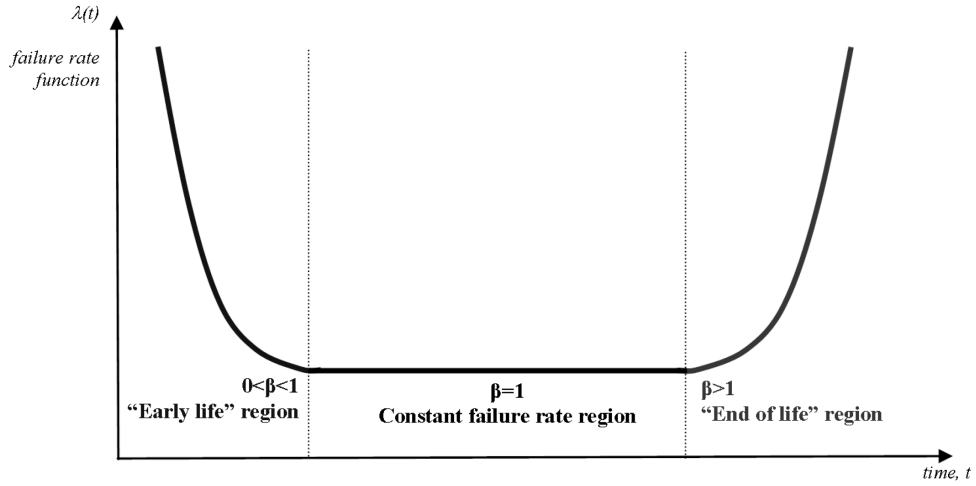


Fig. 2: Shape of the classical failure rate function, reconstructed using Weibull distributions, for which $\lambda(t) = (\beta/\eta)(t/\eta)^{\beta-1}$

2.3 Maintainability and availability

When a system fails to perform satisfactorily, repair work is normally carried out to locate and correct the fault. The system is restored to its functioning state by making an adjustment or by replacing a component, according to prescribed procedures and resources.

The maintainability is defined as the probability of isolating and repairing a fault in a system within a given period of time. Generally, exactly the same formalism as that used for reliability is used to describe maintainability. The random variable here is the time to repair, in the same way as the time to failure is the random variable in the case of reliability. One can therefore describe various aspects of maintainability using a repair density distribution and by defining a maintainability function, a repair rate function, and the mean time to repair, usually denoted by MTTR, which is the expected value of the repair time.

If one considers both the reliability (the probability that an item will not fail) and the maintainability (the probability that the item will be successfully restored after failure), then an additional metric is needed for the probability that the system is operational at a given time t (i.e. either it has not failed or it has been restored after failure). This metric is called the availability and is usually denoted by $A(t)$. The availability is therefore defined as the probability that the system is operating properly when it is required for use. The availability function, which is a complicated function of time, has a simple steady-state or asymptotic expression A, given by

$$A = \lim_{t\to\infty} A(t) = \frac{\text{system uptime}}{\text{system uptime} + \text{system downtime}} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}. \tag{15}$$

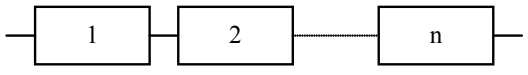
Equation (15) is very simple and is widely used, because we are usually concerned mainly with systems running for a long time. However, one should keep in mind that until steady state is reached, the MTBF may be a function of time and the above formulation should be used cautiously.

2.4 Common techniques in reliability analysis

There are many techniques in reliability analysis. The most common of these are reliability block diagrams, fault tree analysis, and Monte Carlo simulations.

The reliability block diagram (RBD) is one of the conventional tools for system reliability analysis and one of the most commonly used. A major advantage of using the RBD approach is the ease of expression and evaluation of the reliability. An RBD shows the structure of the reliability of the system. It is made up of individual blocks, each block corresponding to a module or function of the system. These blocks are connected to each other by basic relationships that represent the operational configuration of the modules. The most usual connections are the following.

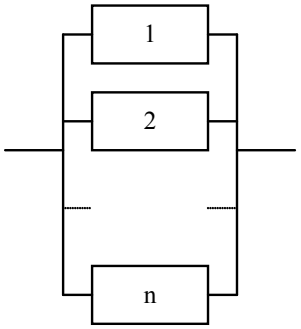
– *Series connection* (Fig. 3): when any module fails, the whole system fails. For a pure series system, the reliability of the system is equal to the product of the reliabilities of its constituent components.



$$R_{tot} = \prod_{i=1}^{n} R_i$$

Fig. 3: RBD for series connection of modules with reliability R_i

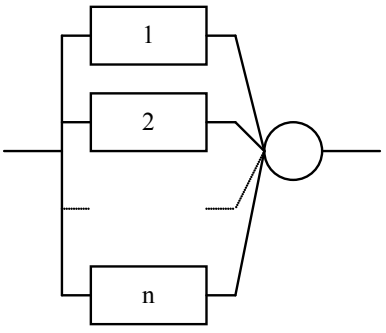
– *Simple parallel connection* (Fig. 4): the modules are redundant, so that the system requires only one module to be operational.



$$R_{tot} = 1 - \prod_{i=1}^{n} (1 - R_i)$$

Fig. 4: RBD for parallel connection of modules with reliability R_i

– *k-out-of-n parallel connection* (Fig. 5): the system requires at least k modules out of n to be operational. In this case, the redundancy is only partial.



$$R_{tot} = \sum_{i=k}^{n} \frac{n!}{i!(n-i)!} R^i (1-R)^{n-i}$$

Fig. 5: RBD for k-out-of-n connection of modules with identical reliability R

Figure 6 shows a simple practical example of an RBD model for an RF system, with associated results for the reliability and availability. This example is taken from Ref. [3]. Note that this model

assumes exponential failure density distributions, leading to simple derivations for the various metrics, in particular the MTBF of a series connection system:

$$\frac{1}{\mathrm{MTBF}} = \sum_i \frac{1}{\mathrm{MTBF}_i} . \tag{16}$$

Reliability and availability results for RF system (for a standard mission time of 168 h)

Component	MTBF (1/h)	MTTR (1/h)	Failure rate (1/h)	Reliability	Availability
Transmitters	10000	4	1.0E-4	0.98	0.99
High-voltage power system	30000	4	3.3E-5	0.99	0.99
Low-level radio frequency	100000	4	1.0E-5	0.99	0.99
Power amplifiers	50000	4	2.0E-5	0.99	0.99
Power components	100000	4	1.0E-5	0.99	0.99
1 comp./system	5769	4		0.97	0.99
60 comp./system	96	4		0.17	0.96

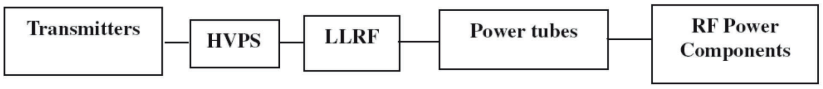


Fig. 6: Example of RBD analysis applied to an RF system [3] (courtesy of P. Pierini & L. Burgazzi)

Fault tree analysis, adapted from system safety analysis, is also a common tool in the application of the concepts of reliability. Whereas the reliability block diagram is mission-success-oriented, a fault tree diagram shows which combinations of component failures will result in a system failure. The fault tree diagram represents logical relationships of 'AND' and 'OR' among the various failure events, as depicted in Fig. 7. Since any logical relationship can be transformed into a combination of 'AND' and 'OR' relationships, the status of the output or top event can be derived from the status of the input events and the connections between the logical gates. A fault tree diagram can therefore describe fault propagation in a system. However, it is not always easy to describe complex systems using a fault tree formulation, especially when one wants to include aspects of repair and maintenance.

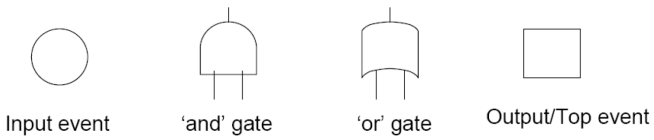


Fig. 7: Basic shapes used in a fault tree diagram [2]

Monte Carlo simulations are also very useful for reliability analysis and can be very powerful. In a Monte Carlo simulation, a reliability model is evaluated repeatedly using random parameter values drawn from a specific distribution. Monte Carlo simulations are often used to evaluate the MTBF for complex systems. However, they usually require the development of a customized program, and also lengthy computer runs if accurate, converging computations are desired.

3 The reference ADS-type accelerator

3.1 MYRRHA, the European ADS demonstrator project

The basic purpose of an ADS is to reduce by orders of magnitude the radiotoxicity, volume, and heat load of nuclear waste before underground storage in deep geological depositories [4]. In this context, a

new research reactor, named MYRRHA (Multipurpose hYbrid Research Reactor for High-tech Applications) is being planned. It will be located at SCK●CEN, Mol, Belgium, and it is hoped that construction will start in 2015 [5]. It is designed to be able to operate in both subcritical and critical modes with the following general objectives: first, to be an experimental device to serve as a test bed for transmutation by demonstrating the ADS technology and the efficient transmutation of high-level waste; second, to be operated as a flexible, multipurpose, high-flux, fast-spectrum irradiation facility ($\Phi_{>0.75\ MeV} = 10^{15}\ n \cdot cm^{-2} \cdot s^{-1}$); and third, to contribute to the demonstration of the Lead Fast Reactor technology, as underlined in the European Roadmap for Sustainable Nuclear Energy [6], without jeopardizing the first two objectives.

MYRRHA is composed of a proton accelerator, a spallation target, and a core with a power of ~70 MW_{th} cooled by liquid lead–bismuth eutectic (LBE). To feed its subcritical core with an external neutron source, the MYRRHA facility requires a powerful proton accelerator, featuring above all a very limited number of unforeseen beam interruptions, i.e. an extremely high reliability level. The present general specifications for the proton beam are the following:

- beam energy: 600 MeV; beam energy stability: better than ±1%;
- beam pulse current: 2.5 mA, and up to 4 mA for core burn-up compensation; beam current stability: better than ±2%;
- beam time structure: continuous, with low-frequency 200 μs zero-current interruptions for on-line subcriticality monitoring of the core;
- beam footprint on the spallation target window: 'doughnut-shaped', 85 mm diameter; beam footprint stability: better than ±10%;
- beam reliability: fewer than 10 beam interruptions longer than 3 s during a three-month operation period.

Extrapolation to a 0.5 GW_{th} industrial 'transmuter' leads to the following figures: 800 MeV, 20 mA proton continuous beam (total beam power 16 MW), and fewer than three beam trips per year.

3.2 The reliability requirement

Until now, the reliability goal for accelerators has been 'we do the best we can'. In the case of an ADS, however, reliability is a constraint for the first time. The stringent reliability requirement arises from the fact that frequently repeated beam interruptions can induce high thermal stresses and fatigue in the reactor structures, the target, and the fuel elements, with the possibility of significant damage, especially to the fuel cladding. Moreover, these beam interruptions can dramatically decrease the availability of the plant, implying plant shutdowns of tens of hours, which could quickly become unsustainable, especially for industrial transmuters.

In the case of MYRRHA, the present tentative limit for the number of allowable beam trips, 10 transients longer than 3 s per three-month operation cycle, comes from the conclusions of the EUROTRANS project [7]. This specification has been slightly relaxed compared with the initial requirements inspired by an analysis of the operation of the PHENIX reactor plant, because the MYRRHA core exhibits a fairly large thermal inertia due to the large LBE pool, and because higher margins seem to exist concerning the behaviour of the fuel and cladding during transients. This beam trip frequency nevertheless remains very significantly lower than today's reported achievements for comparable accelerators (see Fig. 8), and therefore the issue of reliability is considered as the main challenge and will be a constant consideration in all the design and R&D activities pertaining to the MYRRHA accelerator.

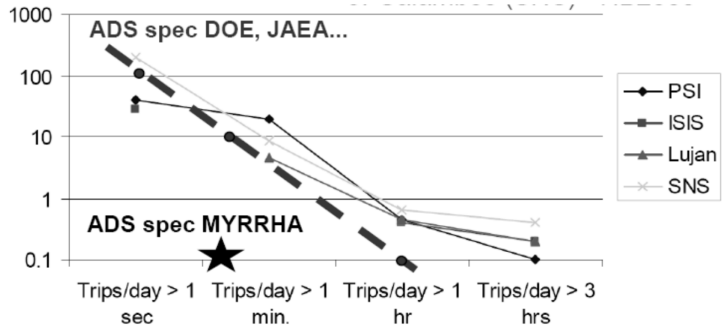


Fig. 8: Trip frequency vs. trip duration for high-power proton accelerators, from [8], and ADS specifications

Nevertheless, it is worth noting that several other ADS studies in Japan and the US claim beam trip limits two orders of magnitude less stringent than the MYRRHA requirements, almost compatible with the present state of the art shown in Fig. 8. In particular, a recent US Department of Energy white paper on ADS technology stated [9]: "Finding #6: Recent detailed analyses of thermal transients in the subcritical core lead to beam trip requirements that are much less stringent than previously thought; while allowed trip rates for commercial power production remain at a few long interruptions per year, relevant permissible trip rates for the transmutation mission lie in the range of many thousands of trips per year with duration greater than one second."

3.3 Reliability-oriented conceptual design of an ADS accelerator

The MYRRHA reliability constraint may be reformulated in the following way: the mean time between failures of the beam delivery system must be longer than ~250 h, a failure being defined as a beam trip longer than 3 s. This MTBF value is one to two orders of magnitude more demanding than what is typically achieved at present in facilities such as PSI, with a MTBF of about 1 h in 2009 [10], and the ESRF, where the operators are very much concerned about reliability and which is improving year after year, with an MTBF of more than 60 h in 2006 [11]. These figures underline the fact that reliability-oriented design practices need to be followed in the early design stage of an ADS accelerator if one wants to be able to achieve the goal.

In the accelerator context, the beam MTBF is a combination of the failure behaviour of many subsystems and sub-subsystems, all contributing fundamentally to successful beam generation. It has been shown that with such a machine configuration, an important increase in the beam MTBF may be obtained only if a single failing element does not automatically imply a global failure [3]. The key to implementing this concept of 'fault tolerance' is redundancy. Parallel redundancy can, of course, be used. It is common to use two elements for one function, as described in Section 2, but for clear economic reasons this parallel scenario has to be minimized. Serial redundancy, in contrast, replaces a missing element's functionality by retuning adjacent elements with nearly identical functionalities. This concept of serial redundancy, which is closely linked to modularity, is to be preferred when applicable.

The concept of the MYRRHA ADS machine requires a 2.4 MW proton accelerator operating in continuous mode. In principle, both cyclotrons and linear accelerators are candidates for providing such a beam. But in order to be able to implement fault tolerance and enhance the reliability figures, a modular machine, i.e. a linear accelerator ('linac'), is to be preferred. Moreover, such a solution also allows the same machine concept to be used both for the demonstrator (MYRRHA) and for long-term industrial machines, as pointed out in Ref. [12].

Thus, basically, the MYRRHA accelerator is a high-power proton accelerator with strongly enhanced reliability, but also with state-of-the-art availability (about 85%, since every beam failure will imply a rather long machine shutdown). The technical solution adopted is that of a

superconducting linac, in agreement with most of the high-power accelerator projects that are in operation or to be built. The continuous operation of this type of accelerator strengthens this choice, as it ensures optimized operation costs. To implement a reliability scheme (or, equivalently, a redundancy scheme), the linac will consist of two clearly distinct sections, as illustrated in Fig. 9.

- A medium- and high-energy section (the main linac, i.e. an independently phased superconducting section), highly modular, based on individual, independently controlled accelerating cavities. In this section, serial redundancy may be applied successfully so as to yield strong fault tolerance. The function of a faulty cavity may typically be taken over by four adjacent cavities. The same concept can be applied to a faulty focusing magnet.
- A low-energy section (the injector, or linac front end), in which modularity and fault tolerance are not applicable, since the beam velocity is too low. Here the number of elements is minimized using multicell cavities, and redundancy is applied in parallel form, so that two complete injectors with fast switching capabilities are foreseen. The transition energy between the two sections has been set at 17 MeV.

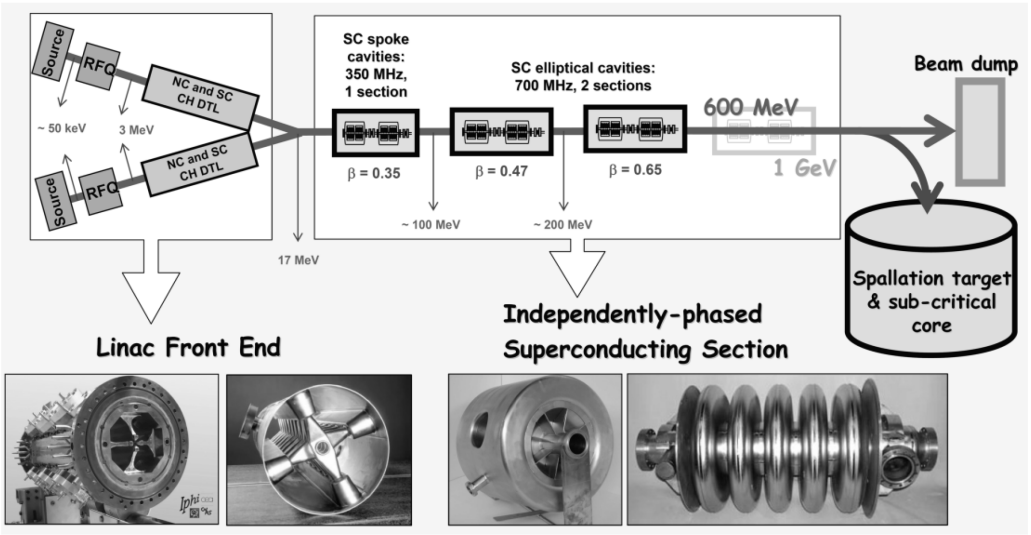


Fig. 9: Conceptual scheme of the MYRRHA accelerator [12] (RFQ: Radio-Frequency Quadrupole; NC: Normal Conducting; SC: SuperConducting; CH: Cross-bar H-type; DTL: Drift Tube Linac)

Throughout the design phase of the MYRRHA accelerator, which is presently ongoing with support from the FP7 European MAX project (2011–2014) [13], the following three principles therefore have been and are being followed regarding the goal of reliability:

- strong design: make it simple, avoid ‘not-so-useful’ complicated elements, use components far from their technological limits, and ensure ‘no beam loss’ operation;
- fault tolerance, and hence redundancy, with the maximum amount of serial redundancy, as already underlined, coupled with realistic fast fault recovery scenarios;
- repairability (on line where possible), coupled with a short enough MTTR and efficient maintenance schemes.

Finally, the reliability of the MYRRHA accelerator that is aimed at will only be realized if these principles are applied in every machine subsystem, including all ancillary equipment. This reliability issue will deserve continuous attention during the engineering design of all components, and during the commissioning of the components and machine.

4 Fault tolerance cases for MYRRHA and expected impact on reliability

4.1 Hot spare injector

The injector part of MYRRHA (0–17 MeV) is based on some rather unconventional solutions. These have been chosen in view of the optimal efficiency and minimized number of components that they provide, given the fact that serial redundancy cannot be applied in this section.

The present design of the injector is described in Ref. [14]; more details can be found in Ref. [13]. It is about 13 m long from the ion source exit to the entrance of the Medium Energy Beam Transport (MEBT), and is composed of four subsections:

- a 30 kV ECR (Electron Cyclotron Resonance) proton source and a 2 m long Low Energy Beam Transport (LEBT);
- a 176 MHz four-rod RFQ, 4 m long, accelerating the beam to 1.5 MeV and operating with a very conservative intervane voltage (30 kV) and Kilpatrick factor (1.0);
- two copper multicell CH-DTL structures for acceleration to 3.5 MeV;
- four superconducting multicell CH-DTL structures [15], combined in one single cryomodule, for acceleration to 17 MeV.

To increase the reliability, the philosophy here was to duplicate this 17 MeV section, providing a hot standby spare injector able to quickly resume beam operation in the case of any failure in the main injector. The fault recovery procedure is based on the use of a switching dipole magnet with a laminated steel yoke connecting the two injectors through a 'double-branch' MEBT. The injector reconfiguration process should last not more than 3 s, and is defined as follows (see Fig. 10).

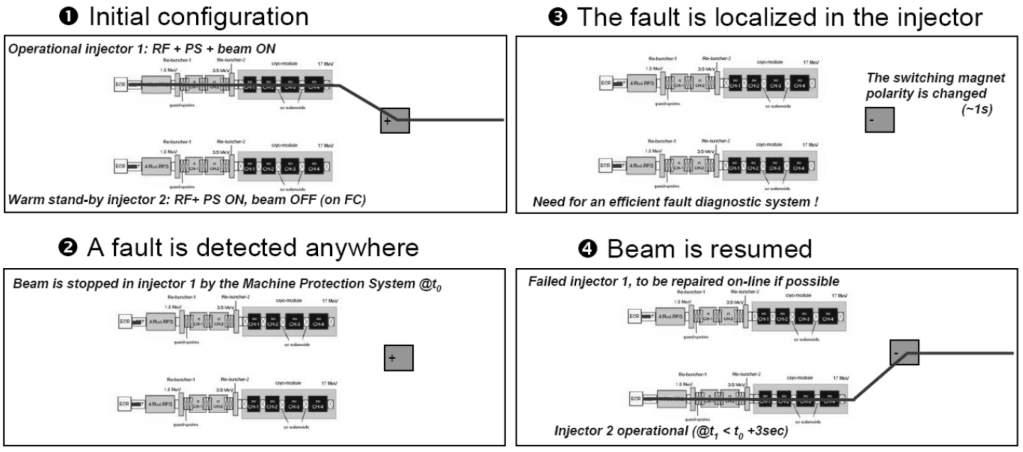


Fig. 10: Fault recovery scenario for the MYRRHA injector

1. In the initial configuration, one of the injectors (e.g. Injector 1) provides a beam to the main linac; the hot spare injector is also fully operational (RF on, power supplies on…), but the beam is intercepted at the source exit with a Faraday cup.

2. A fault is detected somewhere in the linac, and the beam is immediately and automatically stopped at the source exit of Injector 1 by the machine protection system.

3. The fault is localized in Injector 1 by the fault diagnostic system. The polarity of the power supply feeding the MEBT switching magnet is therefore changed.

4. Once steady state is reached, the beam is resumed using Injector 2. It is of course supposed that beam tuning of Injector 2 has been previously performed. The failed Injector 1 should then be repaired as soon as possible, during operation if possible.

4.2 Fault-tolerant main linac

From 17 MeV, a fully modular superconducting linac accelerates the proton beam to the final energy of 600 MeV, over a total length of about 220 m. The linac is composed of an array of independently powered 350 MHz spoke cavities and 700 MHz elliptical cavities. The design ensures a high energy acceptance and moderate energy gain per cavity, using a low number of cells and conservative accelerating gradients (with a nominal operation point of around 50 mT and 25 MV/m peak field), in order to increase the tuning flexibility as much as possible and provide a sufficient margin (about 30%) to allow the implementation of fault-recovery scenarios, as detailed below. A regular focusing-lattice scheme is used, with not-too-long cryostats and room-temperature quadrupole doublets in between. Such a scheme provides several advantages: easy maintenance and fast replacement if required, easier magnet alignment at room temperature and no fringe field issues, the possibility of providing easily accessible diagnostic ports at each lattice location, and, last but not least, nearly perfect regularity of the optical lattice (no specific beam matching is required from cryostat to cryostat). This main linac is then followed by a final transport line, composed of an achromatic triple deviation able to inject the proton beam with the specified footprint into the spallation target. More details of this design can be found in Refs. [7, 13, 16].

Because we are dealing with a non-relativistic proton beam, any fault in an RF cavity, which implies beam energy loss, will also lead to a phase slip along the linac. This will increase with distance, and thus push the beam out of the stability region: the beam will be completely lost. To recover from such RF fault conditions, the philosophy is to re-adjust the accelerating fields and phases of some of the non-faulty RF cavities to recover the nominal beam characteristics at the end of the linac, in particular the transmission, phase, and energy. A simple way to achieve this is to act on the accelerating cavities next to the failed one, as shown in Fig. 11. This 'local compensation method' has the advantage of involving only a small number of elements, and therefore of being able to compensate for multiple RF faults in different sections of the machine at the same time. A similar system based on the updating of the phase set points of all downstream superconducting cavities is being developed very successfully at the SNS [17].

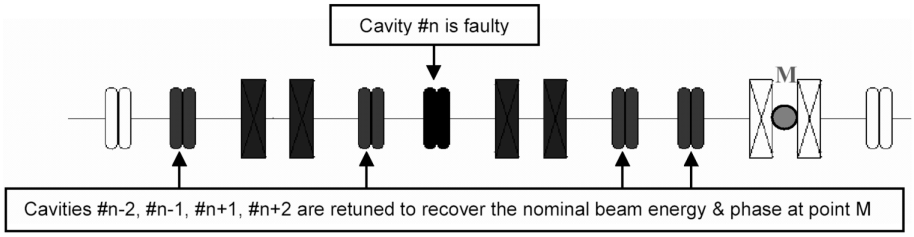


Fig. 11: Principle of the local compensation method

Beam dynamics simulations show that by using such a retuning method in the MYRRHA design, the nominal beam parameters at the target can always be restored with very limited additional emittance growth, given the condition that a 20–30% rise in accelerating field and RF power can be sustained in the few (four to eight) retuned elements [18]. This method is, of course, rather demanding in terms of linac length—it requires an increase of about 20%—and installed RF power budget, but is on the other hand totally in line with the ADS overdesign criterion, and is in any case required if one is to try to reach the required reliability level. This method can also be successfully applied to the failure of a focusing element. In the case of a quadrupole-based focusing scheme, it is preferable either to switch off the whole doublet in the case of a quadrupole failure—this limits the beam mismatch induced by the fault and eases the retuning of the surrounding doublets—or to use triplets.

In any case, it has to be underlined that this local compensation method can be applied successfully only if the beam velocity is high enough to avoid unsustainable defocusing effects while

the beam is travelling across the drift length corresponding to the failed element. The typical limit has been found to be around $\beta = 0.15$, leading to the choice of a 17 MeV injector for MYRRHA.

Simulations of transient beam dynamics need to be performed to analyse accurately what happens to the beam during the above retuning procedures, keeping in mind that the retuning has to be performed in less than a few seconds. A new simulation tool has been developed, based on the TraceWin CEA code [19], that allows the effect of time-dependent perturbations on the beam optics to be analysed by modelling of the RF control loop. From such work [20], a reference scenario for fast on-line recovery from failures of the accelerating system has been settled on. This scenario uses the following sequence, to be performed in less than a few seconds.

1. A fault is detected somewhere in the linac, and the beam is immediately and automatically stopped at the source exit by the machine protection system.

2. The fault is localized in the RF loop of a superconducting cavity by the fault diagnostic system.

3. The field and phase set-points are updated in some RF cavities adjacent to the failed one. These set-points need to be determined previously during the commissioning phase, and possibly stored directly in digital chips in the Low Level RF (LLRF) systems.

4. To avoid any beam-loading effect, the failed cavity is detuned by a few hundred bandwidths using a suitable cold tuning system, possibly piezo-based. This sequence is the most demanding in terms of duration.

5. Once steady state is reached, the beam is resumed. The failed RF system should then be repaired as soon as possible, during operation if possible, and put back on line using a similar opposite procedure.

Details of related simulations and developments can be found in Ref. [21].

4.3 Expected impact on reliability

When one looks at the literature on accelerator reliability, it appears that injectors and RF systems usually represent a significant proportion of the faults that generate beam trips in operating facilities. This is the reason why the two fault cases described above have been taken into account in the early conceptual design stage of the ADS accelerator.

Based on this conceptual design, two independent integrated reliability analyses have been performed to try to estimate the number of malfunctions of the MYRRHA accelerator that could cause beam or plant shutdowns in a three-month operation cycle, and to analyse the influence of the MTBFs, the MTTRs, and the whole system architecture on the results. These studies were performed by means of a reliability block diagram analysis using the Relex© software package [3] and by means of Monte Carlo simulations using home-made software with slight differences in the hypotheses between simulations.

In both cases, the results show that a linear accelerator has high potential for reliability improvement if the system is properly designed with this objective: from about 100 unexpected beam shutdowns per three-month operation period for a classical ‘all-in-series’ linac, this figure falls to around three to five beam interruptions in the MYRRHA case, where a second redundant injector stage with fast switching capabilities is used, and where fault tolerance is included in the independently phased linac via fast fault recovery scenarios. Nevertheless, the absolute figures obtained remain rather questionable at present, because of the still somewhat crude modelling used for such a complex system, and because of the lack of a well-established database of component reliability figures. The development of a more accurate reliability model of the MYRRHA accelerator is therefore very much required for guidance of the engineering design. This work is presently ongoing [13], using the methodology applied in current nuclear power plants, and trying to make

efficient use of existing data and models developed in the accelerator community, especially in machines rather similar to MYRRHA such as the SNS [22].

It is clear in any case that in order to reach the extremely ambitious reliability level of the MYRRHA accelerator that is desired, failure cases will have to be anticipated for all systems and subsystems (e.g. power supplies, the cryogenic system, controls, cooling systems, and vacuum systems), and suitable engineering design solutions implemented. Here again, it is extremely probable that redundancy will be a key issue, using serial redundancy as much as possible, much more often than classical duplication. Some particular fields in which very promising progress is being made in this respect are modular DC power supplies and solid-state-based RF amplifiers.

Solid-state RF amplifiers [23], for example, are totally suited to application in an ADS, with an expected MTBF of more than 50 000 h, which is clearly higher than that of classical RF tubes. These amplifiers are based on a combination of elementary modules of a few hundred watts each, providing extreme modularity and therefore inherent redundancy and flexibility towards failures. In fact, interruption of the source is not required if a failure happens in one or a few amplifier modules: operation can be still sustained with fewer modules, given that the available power remaining is sufficient. Compared with IOTs (Inductive Output Tubes) and klystrons, they also have some operational advantages such as low voltages and easy maintenance. Moreover, the continuous nature of the MYRRHA beam, and hence the absence of a peak in power demand, and the relatively low operational RF frequencies, below 1 GHz, are very compatible with such a solution.

5 Conclusion

In situations where accelerators are applied, one usually cares about beam availability. But with an ADS, for the first time, reliability is an additional constraint to be taken into account. Even in the case of a demonstrator, i.e. the MYRRHA machine, which requires an MTBF of about 250 h, the reliability requirement is extremely ambitious compared with the present state of the art.

Reliability models show that to be able to achieve this goal, redundancy needs to be included in all stages of the machine design in order to provide a strong level of tolerance to faults. At the level of the first conceptual design, this is achieved using a redundant injector followed by a fully modular superconducting linac with fault tolerance capabilities. At the level of the subsystems, this strategy should also be applied widely, with implementation of redundancy where possible, robust designs, and efficient maintenance strategies. Finally, once the machine is constructed, a few years of commissioning and training will be necessary to identify, repair, and optimize the weak elements so as to maximize the reliability level of the operation of the machine.

6 Acknowledgements

The author would like to thank all of his colleagues involved so far in these developments linked to the MYRRHA accelerator design, especially Tomas Junquera (ACS); Didier Uriot (CEA Saclay); Horst Klein, Holger Podlech, and Chuan Zhang (IAP Frankfurt); Paolo Pierini (INFN Milano); Frédéric Bouly, Christophe Joly, Alex C. Mueller, and Hervé Saugnac (IPN Orsay); and Dirk Vandeplassche (SCK•CEN). The research leading to these results has received funding from the European Atomic Energy Community (EURATOM) Fifth and Sixth Framework Programmes, and is being supported from the Seventh Framework Programme FP7/2007-2011 under grant agreement No. 269565 (MAX Project).

References

[1] Weibull.com, http://www.weibull.com/SystemRelWeb/blocksimtheory.htm

[2] M. Xie, K.-L. Poh, and Y.-S. Dai, *Computing System Reliability: Models and Analysis* (Springer, Berlin, 2004).

[3] P. Pierini and L. Burgazzi, *Reliab. Eng. Syst. Safety* **92** (2007) 449–463.

[4] European Technical Working Group, *The European Roadmap for Developing ADS for Nuclear Waste Incineration* (ENEA, Rome, 2001).

[5] SCK•CEN, http://myrrha.sckcen.be/

[6] SNETP, http://www.snetp.eu/

[7] J.-L. Biarrotte, A.C. Mueller, H. Klein, P. Pierini, and D. Vandeplassche, Accelerator reference design for the MYRRHA European ADS demonstrator, Proc. 25th LINAC Conf., Tsukuba, Japan, 2010.

[8] J. Galambos, T. Koseki, and M. Seidel, Commissioning strategies, operations and performance, beam loss management, activation, machine protection, Proc. 42nd ICFA Advanced Beam Dynamics Workshop, Nashville, TN, 2008.

[9] US Department of Energy, *Accelerator and Target Technology for Accelerator Driven Transmutation and Energy Production*, 2010, http://science.energy.gov/hep/news-and-resources/

[10] M. Seidel, Experience with the production of a 1.3MW proton beam in a cyclotron-based facility, Proc. 1st TC-ADS Workshop, Karlsruhe, Germany, 2010.

[11] L. Hardy *et al*., Operation and recent developments at the ESRF, Proc. 11th EPAC Conf., Genoa, Italy, 2008.

[12] J.-L. Biarrotte *et al*., *Nucl. Instrum. Meth. Phys. Res. A* **562** (2006) 565–661.

[13] MAX Project, http://ipnweb.in2p3.fr/MAX/

[14] C. Zhang *et al*., From Eurotrans to MAX: new strategies and approaches for the injector development, Proc. 2nd IPAC Conf., San Sebastian, Spain, 2011.

[15] F. Dziuba *et al*., *Phys. Rev. Spec. Top. Accel. Beams* **13** (2010) 041302.

[16] H. Saugnac *et al*., High energy beam line design of the 600MeV 4mA proton linac for the Myrrha facility, Proc. 2nd IPAC Conf., San Sebastian, Spain, 2011.

[17] J. Galambos *et al*., A fault recovery system for the SNS superconducting cavity linac, Proc. 23rd LINAC Conf., Knoxville, TN, 2006.

[18] J.-L. Biarrotte *et al*., Beam dynamics studies for the fault tolerance assessment of the PDS-XADS linac design, Proc. 9th EPAC Conf., Lucerne, Switzerland, 2004.

[19] CEA, http://irfu.cea.fr/Sacm/logiciels/index3.php

[20] J.-L. Biarrotte and D. Uriot, *Phys. Rev. Spec. Top. Accel. Beams* **11** (2008) 072803.

[21] F. Bouly *et al*., LLRF developments toward a fault tolerant Linac scheme for ADS, Proc. 25th LINAC Conf., Tsukuba, Japan, 2010.

[22] G. Dodson, The SNS reliability program, Proc. 3rd Accelerator Reliability Workshop, Cape Town, South Africa, 2011.

[23] M. Di Giacomo, Solid state RF amplifiers for accelerator applications, Proc. 23rd PAC Conf., Vancouver, Canada, 2009.

Participants

ALMALKI, M.	Johann Wolfgang Goethe Universitaet, Frankfurt am Main, DE
ALMOMANI, A.	Institute for Applied Physics, Frankfurt am Main, DE
ALONSO, J.	Tekniker, Eibar, ES
ARIZ, I.	Tekniker, Eibar, ES
BAER, T.	CERN, Geneva, CH
BARILLERE, R.	CERN, Geneva, CH
BARLOW, R.	Cockcroft Institute, Warrington, UK
BARTMANN, W.	CERN, Geneva, CH
BARTOSIK, H.	CERN, Geneva, CH
BAZIN, N.	CEA Saclay, Gif-sur-Yvette, FR
BHATTACHARYYA, A.	CERN, Geneva, CH
BRACCO, C.	CERN, Geneva, CH
BUSTINDUY, I.	ESS-Bilbao, Leioa, ES
CHEYMOL, B.	CERN, Geneva, CH
DALLOCCHIO , A.	CERN, Geneva, CH
DAMERAU, H.	CERN, Geneva, CH
DE COS, D.	ESS-Bilbao, Leioa, ES
DORDA, U.	CERN, Geneva, CH
DUPERREX, P.	PSI, Villigen, CH
EGBERTS, J.	CEA Saclay, Gif-sur-Yvette, FR
EGUIA, J.	Tekniker, Eibar, ES
ENPARANTZA, R.	Tekniker, Eibar, ES
ESHRAQI, M.	ESS, Lund, SE
FERNANDEZ CANOTO, D.	ESS-Bilbao, Zamudio, ES
FITTERER, M.	CERN, Geneva, CH
GARCIA TUDELA, M.	CERN, Geneva, CH
GONZALEZ, O.	ESS-Bilbao, Leioa, ES
GONZALEZ BERGES, M.	CERN, Geneva, CH
HOLM, A.	Institute For Storage Ring Facilities, Aarhus, DK
IZAOLA, Z.	ESS-Bilbao, Leioa, ES
JAROSZ, M.	ESS, Lund, SE
JEFF, A.	CERN, Geneva, CH
KAUFMANN, W.	GSI, Darmstadt, DE
LAFACE, E.	ESS, Lund, SE
LALLEMENT, J.B.	CERN, Geneva, CH
LARRANAGA, M.	Tekniker, Eibar, ES
LEONARDO, A.	ESS-Bilbao, Leioa, ES
LI, K.	CERN, Geneva, CH
MACPHERSON, A.	CERN, Geneva, CH
MAGLIONI, C.	CERN, Geneva, CH
MIRAPEIX, F.	UPV, Santander, ES
MOLENDIJK, J.	CERN, Geneva, CH
MONTESANO, S.	CERN, Geneva, CH
MUELLER, H.	PSI, Villigen, CH
PEREIRA, H.	CERN, Geneva, CH
PONTON, A.	ESS, Lund, SE
REGGIANI, D.	PSI, Villigen, CH
ROBIN, A.	Siemens AG, Erlangen, DE

RODRIGUEZ, I.	ESS-Bilbao, Leioa, ES
ROMERO SERRANO, A.J.	Centro Nacional De Aceleradores, Seville, ES
RONCAROLO, F.	CERN, Geneva, CH
SARGSYAN, E.	CERN, Geneva, CH
SHEEHY, S.	RAL, Didcot, UK
SMITH, H.	RAL, Didcot, UK
STADLBAUER, T.	CERN, Geneva, CH
STRASIK, I.	GSI, Darmstadt, DE
SWINIARSKI, J.	ESS, Lund, SE
TCHELIDZE, L.	ESS, Lund, SE
THOMSEN, H.	Institute for Storage Ring Facilities, Aarhus, DK
THORN, D.	Siemens AG, Erlangen, DE
UGENA TIRADO, P.	CERN, Geneva, CH
VALUCH, D.	CERN, Geneva, CH
VELEZ, A.	ESS-Bilbao, Leioa, ES
WILLIAMSON, R.	RAL, Didcot, UK
YAN, J.	University of Tokyo, Tokyo, JP
ZANGENBERG, N.	Danish Technological Institute, Aarhus, DK
ZENG, R.	ESS, Lund, SE
ZERLAUTH, M.	CERN, Geneva, CH
ZHANG, H.	PSI, Villigen, CH