SLAC-R-721

The Proceedings of the 29th SLAC Summer Institute On Particle Physics: Exploring Electroweak Symmetry Breaking (SSI 2001)

29th SLAC Summer Institute On Particle Physics: Exploring Electroweak Symmetry Breaking (SSI 2001),
13-24 Aug 2001, Stanford, California

Stanford Linear Accelerator Center, Stanford University, Stanford, CA 94309

Work supported by Department of Energy contract DE–AC03–76SF00515.

Proceedings of the

29th SLAC Summer Institute

Exploring Electroweak Symmetry Breaking

August 13-24, 2001

Stanford University, Stanford, California

Sponsored by Stanford University and Stanford Linear Accelerator Center under contract with the U.S. Department of Energy, Contract DE-AC03-76SF00515.

Printed in the United States of America. Available from National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Road, Springfield, VA 22161.

Table of Contents

CP VIOLATION AND THE ORIGINS OF MATTER

Andrew G. Cohen
Department of Physics, Boston University
Boston, MA 02215

ABSTRACT

I present a gentle introduction to baryogenesis, the dynamical production of a baryon asymmetry during the early universe. I review the evidence for a cosmic baryon asymmetry and describe some of the elementary ingredients necessary for models of baryon number production.

1 Introduction and Experiment

Even though the Universe has a size, age and complexity far beyond our everyday experience, the laws of physics determined in the laboratory can be extrapolated to the vast realms of the cosmos. This program, pursued since the earliest developments in the physical sciences, has seen enormous change over the last century. Especially important for particle physics has been the close interaction between the high energy frontier and the very early universe, and cosmological arguments are now routinely used to constrain the rampant imaginings of particle theorists. One area that is closely connected with the principle topic of this years school, CP violation, is baryogenesis, the dynamical production of a net baryon number during the early universe. This asymmetry, which is well established experimentally, is one of the most important features of the cosmos as a whole, and represents an enormous departure from the CP invariant state of equal matter and antimatter densities, with no net baryon number. The subject has been of concern to particle physicists since the discovery of microscopic CP violation, which encouraged the construction of concrete baryogenesis scenarios. The subject became a standard part of modern cosmology with the introduction of grand unified theories (GUTs), introduced in the 1970s, which establish a possible source for baryon number violation, an essential component of baryogenesis. More recent ideas have attempted to link the baron asymmetry with details of models of electroweak symmetry breaking, and offer the possibility of testing models of baryogenesis in future colliders such as the LHC.

There are many good reviews of baryogenesis at all levels*. Here we give only a brief overview of the subject and encourage further consultation of the references.

1.1 Initial Data

One of the fundamental questions concerning the large scale structure of our universe is surprisingly difficult to answer: What is the universe made of? In general terms this question reduces to the value of a single parameter, the total energy density of the universe, which is usually quoted in terms of a "critical" density related to the current Hubble expansion rate:

$$0.01 \lesssim \Omega_0 \equiv \frac{\rho}{\rho_{\rm crit}} \lesssim 3 \ , \tag{1}$$

*The book[1] by Kolb and Turner is a good (although somewhat dated) starting point. There are many more recent reviews,[2–4] as well as references therein.

where $\rho_{\text{crit}} \equiv 3H_0^2/(8\pi G_N)$, H_0 is the Hubble constant and G_N is Newton's gravitational constant. The lower value comes from the visible content of the universe, the mass-energy associated with stars, galaxies, *etc*. The larger value comes from various measurements of large scale structure, especially measurements of the potential associated with gravitating (but not necessarily visible) mass-energy. The discrepancy between these numbers suggests that the majority of the mass-energy of the universe is dark, possibly a completely new kind of material. But even for the visible mass, we have no direct experience of the stuff out of which distant stars are made, although we believe this stuff to be matter similar to that which makes up our own star. The detailed physics of distant stars, such as stellar evolution, spectral lines, *etc*, is convincing evidence that these objects are made of baryons and leptons much as ourselves, but there remains the possibility that they are constructed from *antimatter*, *i.e.* antiquarks and positrons, rather than quarks and electrons. The transformation CP acting on a state of ordinary matter (by which we mean baryons, objects made of quarks carrying a positive baryon number) produces a state of antimatter (with negative baryon number). Thus if all stars in the universe contain matter (in the form of baryons) rather than antimatter (in the form of antibaryons), then this matter antimatter (or baryon) asymmetry represents a departure from CP symmetry as well.

What evidence is there that distant objects are made of matter rather than antimatter? For that matter, how do we know that the earth itself is matter? Matter and antimatter couple electromagnetically with known strength. Contact between matter and antimatter leads naturally to annihilation into photons with characteristic energy of 100s of MeV. Casual observation easily demonstrates the absence of this radiation when matter (in the form of ourselves, say) comes in contact with another terrestrial object. Thus we easily deduce that the earth (and all its occupants) are made of matter. A similarly pedestrian argument indicates that the moon too is made of matter. Indeed our exploration of nearby space convincingly shows that the solar system is composed of matter.

In fact it is not necessary that a man-made item come into contact with distant objects to establish the nature of such objects. If anything known to be matter is in contact with an unknown object, the absence of gamma radiation from annihilations demonstrates the object is not antimatter. For example micro-meteorites are continuously bombarding the earth without such radiation, and are therefore not antimatter. But these objects also rain upon Mars, which is therefore also not antimatter. This argument can obviously be extended: as long as a sufficiently dense matter trail extends

from our solar system, absence of 100 MeV gamma rays demonstrates the absence of antimatter. This trail extends to distances comparable to the size of our local galactic cluster,[5] the Virgo cluster, a distance of 20 Mpc.

Unfortunately this region covers only a tiny fraction of the observable universe, which has a characteristic linear size several orders of magnitude larger than that of the Virgo cluster. Constraining the composition of objects beyond our local neighborhood requires a more complex analysis.

Experiments to search for cosmic antimatter from beyond this 20 Mpc distance have been proposed. The most ambitious of these, the Alpha Magnetic Spectrometer[6–11] (AMS) is scheduled to be deployed aboard the International Space Station sometime in the distant future†. This device, essentially a large mass spectrometer, will search for negatively charged nuclei in cosmic rays. The device should place a direct limit on antimatter in cosmic rays coming from a distance of nearly an order of magnitude beyond our local cluster. Although this distance scale remains small compared to the current visible universe, it is a significant step beyond our local cluster.

Lacking further direct experimental evidence against distant regions of antimatter, we must rely on alternative observational and theoretical analyses. Our original argument, the lack of gamma radiation emanating from points of contact between regions of matter and antimatter, fails when the density of both matter and antimatter becomes so small that the expected gamma ray flux falls below a detectable level. However this suggests an improvement on this argument: since the density of matter (and any putative antimatter) is decreasing with the cosmic expansion of the universe, we might expect that the flux of gamma radiation from such points of contact was larger in the early universe than it is today. Thus we might search for radiation from matter antimatter annihilation that occurred not today but sometime in the far past. A search for such radiation would differ from those which already place stringent limits on antimatter in our local neighborhood. Firstly, once produced as gamma rays, radiation would subsequently redshift as the universe expands. Consequently rather than searching for gamma rays with energies of 100s of MeV, we should search for lower energy radiation. Secondly, when we look out to large redshift (the distant past) on the night sky we are integrating over large portions of the universe. Consequently rather then seeking point sources we should search for a diffuse background of radiation coming from many points of intersection of domains of matter with those of antimatter.

† A prototype device has flown in the space shuttle. Although the exposure was insufficient to detect antimatter, this brief test has returned interesting cosmic ray physics.[12]

In order to use this technique to place limits on cosmic antimatter we must have some idea of how a diffuse photon spectral flux is related to the properties of domains of antimatter, in particular their size. We already know that such domains should be larger than the 20 Mpc limit we have in hand. The environment of this photon production, the interface between regions of matter and antimatter in the early universe, involves known principles of physics, and upper limits on the photon flux can be deduced. Although rather complicated in detail, the basic strategy is straightforward:

- The observed uniformity of the cosmic microwave background radiation implies that matter and antimatter must have been extremely uniform at the time when radiation and matter decoupled, a redshift of about 1100 or a time of about 10^{13} seconds. Thus at this time domains of matter and antimatter cannot be separated by voids, and must be in contact with each other. Prior to this time it is conceivable that matter and antimatter domains *are* separated by voids, and thus we do not include any annihilation photons prior to this epoch.
- Annihilation proceeds near matter antimatter boundaries through combustion, converting matter into radiation according to standard annihilation cross-sections. This change of phase in the annihilation region leads to a drop in pressure, and matter and antimatter then flow into this region. This leads to a calculable annihilation flux via the flow of matter and antimatter into this combustion zone. The annihilation process also gives rise to high energy leptons which deposit energy in the matter and antimatter fluids, significantly enhancing the annihilation rate.
- At a redshift of about 20 (approximately 10^{16} s after the big bang) inhomogeneities leading to structure formation begin to become significant. Although this likely does not affect the rate of annihilations significantly, rather than analyze this era in detail it is safer (more conservative) to ignore any further annihilation.
- The spectrum of photons produced prior to a redshift of 20 continues to evolve due to the expansion of the universe as well as subsequent scattering.

The results of this calculation[13] are shown in Figure 1. The upper curve represents the computed spectral flux of diffuse radiation from domains of antimatter with a characteristic size of 20 Mpc, the lower limit allowed by other analyses. The lower curve represents the spectral flux for a domain size of 1000 Mpc, a large fraction of the visible universe. In both cases this calculated flux is substantially larger then the observed diffuse gamma ray background (by balloon and satellite experiments). In particular such a flux would be in serious conflict with the results of the COMPTEL satellite

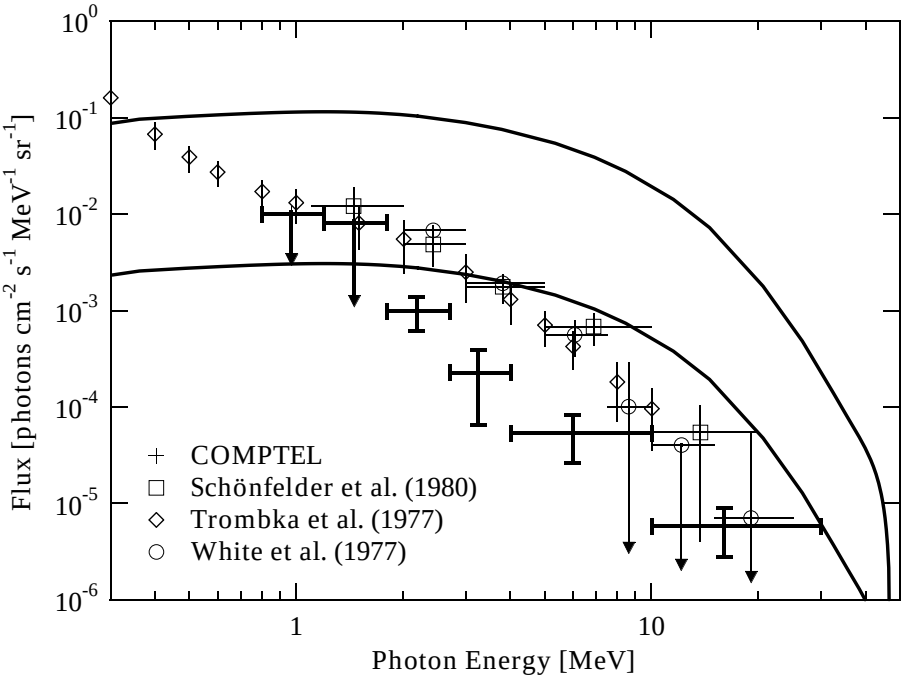

Fig. 1. Data[14] and expectations for the diffuse γ-ray spectrum.

experiments. We conclude that domains of antimatter of size less than 1000 Mpc are excluded.

1.2 A Baryon Asymmetry

The arguments of the preceding section indicate that the universe contains predominately matter and very little antimatter (or that matter and antimatter have been separated into several near universe-sized domains, a possibility[15,16] we will not consider here.) This asymmetry has been a focus of contemporary cosmology and particle physics principally because of its implied CP violation. To decide the significance of this asymmetry we need a quantitative measure of this departure from baryon antibaryon equality. Normally we will use the baryon density to photon number density ratio:

$$\eta \equiv \frac{n_B + n_{\bar{B}}}{n_\gamma} \ . \tag{2}$$

This choice is motivated partly by the dimensionless nature of the ratio, but more importantly, by the way in which this ratio scales with the expansion of the universe. Provided the expansion is isentropic (and ignoring baryon production or destruction) both the numerator and denominator densities dilute with the cosmic expansion in the

same way, inversely proportional to the change in volume, and thus the ratio η is time independent.

Since our previous arguments suggest that $n_{\bar{B}}$ is insignificant, we may use the observed (visible) baryon density and the microwave background radiation density to obtain an experimental lower limit on η

$$10^{-10} \lesssim \eta \ . \tag{3}$$

In fact a more constrained value may be obtained by using some additional theoretical information. The synthesis of the light elements in the early universe depends quite sensitively on the baryon density. Using the best observations on the primordial elements this constrains η[17]:

$$4 \cdot 10^{-10} \lesssim \eta \lesssim 7 \cdot 10^{-10} \ . \tag{4}$$

Is this value significant? To get a better idea of how large this number is, we might imagine its value in a baryo-symmetric universe. In this case, as the universe cools from temperatures above 1 GeV where baryons and antibaryons are in thermal equilibrium with a thermal number density proportional to T^3, baryon number is kept in thermal equilibrium by baryon antibaryon annihilation. Once the rate for this process becomes slower than the expansion rate, the probability of a subsequent annihilation becomes negligible. Using a typical hadronic cross-section, this equality of rates occurs at a temperature of about 20 MeV. At this time baryons, in the form of protons and neutrons, have an equilibrium number density proportional to:

$$n_B \propto (T m_N)^{3/2} e^{-m_N/T} \tag{5}$$

and give a value for η

$$\eta \sim 10^{-20} \ . \tag{6}$$

This value, in gross conflict with the experimental number, cannot be avoided with thermal equilibrium between equal number of baryons and antibaryons, reflecting the efficient and near total annihilation of all matter. However there is a simple path to obtain a much larger value. If the number of baryons exceeds that of antibaryons by even a small amount, than the inability of each baryon to "pair up" with an antibaryon prevents total annihilation. In fact this excess need be only a few parts per billion at high temperature (leading to one extra baryon for each several billion photons) to achieve an adequate value for η.

But where would such an excess come from? It might appear as an initial condition, set at the beginning of the universe in some way beyond our ken. Note that such an initial condition is irrelevant in the context of inflation; following the reheating phase at the end of inflation all memory of such an initial condition is erased. Without inflation this is a rather unpleasant possibility that we must acknowledge, but we will favor an explanation that does not rely on a *deus ex machina* of this type. What is preferable is a mechanism by which this peculiar excess arises dynamically during the evolution of the universe, a possibility known as *baryogenesis*.

As was first observed by Andrei Sakharov,[18] there are three conditions that must be met in order for baryogenesis to occur:

- Baryon violation. Obviously if the universe is going to evolve a non-zero baryon number from a time when the baryon number vanishes (at the end of inflation, say) than the laws of physics must allow the baryon number to change.
- C and CP violation. Whatever process changes the baryon number must do so in a way that favors baryon production, rather than antibaryon production. Since both C and CP transformations change the sign of the baryon number, the laws of physics must violate both C and CP in order to obtain a positive value. Fortunately nature has provided us with both of these elements. As an example:

$$\frac{\text{Rate}[K_L^0 \to e^+\pi^-\nu]}{\text{Rate}[K_L^0 \to e^-\pi^+\bar{\nu}]} \simeq 1.006 \tag{7}$$

- Departure from thermal equilibrium. Roughly speaking if we populate all levels according to a Boltzmann distribution, since CPT guarantees that each level with a positive baryon number has a corresponding level with a negative baryon number, the total baryon number must vanish. More formally, since $\hat{B}$ is CPT odd and the Hamiltonian CPT even, in thermal equilibrium

$$\langle \hat{B} \rangle = \text{Tr}\hat{B}e^{-\beta\hat{H}} = \text{Tr}\ \Omega_{CPT}\Omega_{CPT}^{-1}\hat{B}e^{-\beta\hat{H}} = -\langle \hat{B} \rangle = 0\ . \tag{8}$$

Discussions of baryogenesis are often, not surprisingly, focused on the origin of these three ingredients. Beginning in the late 1970s it was realized that all three arise in commonly considered extensions of the standard model:

- Baryon Violation. Grand Unified theories, in which quarks and leptons appear in the same representation of a gauge group, naturally give rise to baryon violation.
- C, CP violation. Kaon physics already implies a source of C and CP violation.

- Departure from thermal equilibrium. The universe is known to be expanding and cooling off. This change in the temperature with time *is* a departure from thermal equilibrium.

We will turn to an evaluation each of these items in somewhat more detail.

Baryon violation is severely constrained by its apparent absence in the laboratory: experiments searching for proton decay have already placed a limit on the proton lifetime greater than 10^{32} years. How can baryon violation be significant for baryogenesis yet avoid a disastrous instability of the proton? The key is the notion of an accidental symmetry: a symmetry of all possible local operators of dimension four or less constrained by the particle content and gauge invariance of a theory. The significance of accidental symmetries appears when we consider the effects of new physics at high energies. These effects may be incorporated at low energies by including all possible local operators that respect the symmetries of this new physics. By dimensional analysis all operators of dimension higher than four will be suppressed by powers of the ratio of the low energy scale to the high energy scale. Now imagine that new physics at high energies does not respect some symmetry, like baryon number. At low energies we must include all local operators, including those that violate baryon number, an apparently disastrous result. But if the theory has an accidental symmetry, the only such operators are of dimension greater than four (by the definition of accidental symmetry), and thus new physics at high energies which violates this symmetry is suppressed by the high energy scale. In the standard model baryon number is exactly such an accidental symmetry: no baryon violating operators of dimension four or less can be constructed out of the standard particles consistent with the $SU(3) \times SU(2) \times U(1)$ gauge invariance. In fact the leading baryon violating operator in this construction is dimension six. If we then contemplate new physics which violates baryon number at a high energy scale, such as in grand unified theories, baryon violating effects will be suppressed at low energies by two powers of this high energy scale. Thus if the scale is greater than 10^{16} GeV, proton decay (a low energy process taking place near 1 GeV) is hugely suppressed.

As already indicated, CP violation is present in the kaon system at a level which appears more than adequate to explain a baryon asymmetry of less than one part in one billion. However CP violation in the standard model arising from a phase in the CKM matrix (which may or may not account for the phenomena observed in the kaon system) is unlikely to be responsible for the baryon asymmetry of the universe. As we will see, the effects of this phase in the early universe are quite small.

If the CP violation in the standard model can not account for the observed baryon asymmetry of the universe, what can? In fact almost *any* new source of CP violation beyond that of the phase in the CKM matrix gives rise to significant effects in the early universe. From a particle physics perspective, this is the principal reason for interest in the cosmic baryon asymmetry: it is a strong indication of physics beyond the standard model.

Lastly, the expansion of the universe which characterizes a departure from thermal equilibrium is governed by the Hubble parameter:

$$\frac{\dot{T}}{T} = -H \tag{9}$$

(at least during periods of constant co-moving entropy.) Today the Hubble parameter is quite small; the characteristic time scale for expansion of the universe is 10 billion years. Since most microphysical processes lead to thermal equilibrium on much shorter time scales, baryogenesis must take place either at a time when H is much larger, or at a time when Eq. (9) doesn't hold.

2 Grand Unification

Together the items of the previous section suggest that baryogenesis occurs at relatively early times, when the universe was hot and baryon violation was important. In particular the ingredients on our list all fit quite naturally into many grand unified theories. In such theories, super-heavy gauge bosons associated with the grand unified gauge group, as well as super-heavy Higgs bosons associated with GUT symmetry breaking, can mediate baryon violating processes. Although suppressed at low energies, at the high temperatures prevalent in the early universe baryon violation rates can be large. In addition, the rapid expansion rate

$$H \sim \frac{T^2}{M_P} \tag{10}$$

allows for significant departure from thermal equilibrium. Finally the interactions associated with new scalar fields that all GUT models must have may include CP violating couplings.

To see how this works in more detail, consider a toy model consisting of bosons X (and $\bar{X}$) which couple to quarks and leptons in a baryon violating, and CP violating, way. For example imagine that the X ($\bar{X}$) boson decays into the two final states qq ($\bar{q}\bar{q}$)

and $\bar{q}\bar{l}$ (ql) with branching fractions r ($\bar{r}$) and $1-r$ ($1-\bar{r}$) respectively. The parameters of this toy are constrained by symmetry. For example, CPT insures that the masses of the bosons are equal $m_X = m_{\bar{X}}$, as are the total widths $\Gamma_X = \Gamma_{\bar{X}}$. The baryon number of each final state is conventional: $B(qq) = 2/3$, $B(\bar{q}\bar{l}) = -1/3$, *etc*. Finally C and CP symmetry would imply $r = \bar{r}$. However lacking these symmetries, generically r will differ from $\bar{r}$[‡].

If we now imagine starting with thermal number densities of X and $\bar{X}$ bosons, our CPT constraint insures that these densities are equal $n_{\bar{X}} = n_X$. Using the parameters of introduced in the preceding paragraph we can compute the net baryon number of the quarks and leptons which result from the X and $\bar{X}$ decays:

$$n_B + n_{\bar{B}} = n_X\Big[r\frac{2}{3} + (1-r)(-\frac{1}{3})\Big] + n_{\bar{X}}\Big[\bar{r}(-\frac{2}{3}) + (1-\bar{r})\frac{1}{3}\Big] = n_X(r-\bar{r})\ . \quad (11)$$

Although this formula is correct, it is the answer to the wrong question. If all interactions are in thermal equilibrium, the X and $\bar{X}$ bosons will be replenished at the same time that they decay. That is, the rate for the inverse process, production of X (and $\bar{X}$) bosons through qq or $\bar{q}\bar{l}$ fusion, will have a rate in equilibrium which is precisely the same as the decay rate, when the number densities of all the particles are equal to their thermal equilibrium values. For example, at temperatures small compared to the X boson mass, the production rate of quarks and leptons via X decay is small, since there are very few X bosons in equilibrium, $n_X \propto \exp(-m_X/T)$. Conversely the inverse process, creation of an X boson, is rare since the quarks and leptons are exponentially unlikely to have the energy necessary to produce a real X boson. So in equilibrium the baryon number does not change, and Eq. (11) is not relevant.

This suggests what turns out to be the key to baryon production—we need the number density of X and $\bar{X}$ bosons at $T << m_X$ to be much larger then the exponentially small equilibrium number density. Under these circumstances the X and $\bar{X}$ production processes will be much smaller than the decay processes. If the number density of X and $\bar{X}$ bosons is sufficiently large, we may even ignore the inverse process all together.

How do we arrange this miracle? Clearly we must depart from thermal equilibrium, something we already knew from our discussion of Sakharov's conditions. But as we have also discussed the universal expansion allows such a departure when the rate for an equilibrating process is slow compared to the expansion rate. In this case, we need the processes that keeps the number density of X and $\bar{X}$ bosons in equilibrium to be

[‡] Of course C and CP violation are not sufficient—interference with a scattering phase is also necessary.

slow compared to the expansion. There are two processes which decrease the number of bosons: the decay of the X and $\bar{X}$ bosons; and annihilation of the X and $\bar{X}$ bosons into other species. Both of these processes can be slow if the couplings of the X boson are weak. Of course "slow" means in comparison with the Hubble expansion rate, $H \sim T^2/M_P$. If this is indeed the case, the number density of X bosons will not track the equilibrium value proportional to $\exp(-m_X/T)$, but instead remain larger. Then once the age of the universe is larger than the lifetime of the X boson, decay will occur, leading to a baryon number according to Eq. (11).

There is one important constraint that we have overlooked. Even though the X and $\bar{X}$ bosons are not re-produced around the time that they decay, there are other processes we must not forget. In particular, there are processes which violate baryon number through the mediation of a (virtual) X boson. In our toy example these may be represented by the effective four-fermion operator $qqql$. This dimension six operator has a coefficient proportional to two inverse powers of the m_X mass, and thus at temperatures low compared to this mass the effects of this operator are small. Nevertheless processes of this type will change the baryon number, tending to equilibrate this number to zero. Therefore we must further require that baryon violating processes such as this one must also be out of equilibrium at the time the X and $\bar{X}$ bosons decay.

The procedure outlined above is usually called a "late decay", or "out-of-equilibrium decay" scenario. Developed extensively from late 1970s through the present, they have provided a framework in which to discuss baryogenesis, and have led to many concrete models that can explain the non-zero value of η. Although successful in principal, models of GUT baryogenesis often have difficulty obtaining the large baryon asymmetry we observe:

- *Rates:* We have seen that a number of rates must be slow compared to the expansion rate of the universe in order to depart sufficiently from equilibrium. These rates are typically governed by the GUT scale, while the expansion rate is proportional to T^2/M_P. The relevant temperature here is that just prior to the decay of the X bosons. Since we need these bosons to be long lived, this temperature is lower than the GUT scale, and the expansion rate is correspondingly slower. Thus the departure from equilibrium is far from automatic and detailed calculations in a specific GUT are necessary to determine whether these conditions can be satisfied.
- *Relics:* One problematic aspect of many GUTS is the presence of possible stable

relics. For example some GUTS have exactly stable magnetic monopoles which would be produced in the early universe at temperatures near the GUT scale. Unfortunately these objects are a cosmological disaster: the energy density in the form of monopoles would over-close the universe, in serious conflict with observation. One of the early great successes of inflation was a means for avoiding this catastrophe. At the end of inflation all matter in the universe has been "inflated away", leaving a cold empty space free from all particles (baryons as well as monopoles!). However following the end of inflation, the vacuum energy density in the inflaton field goes into reheating the universe, producing a thermal distribution of particles. If this reheating is fast, energy conservation tells us that the reheat temperature will be close to the original scale of inflation, near or above the GUT scale. Unfortunately this would reintroduce the monopoles. On the other hand if this reheating is slow (as would be the case if the inflaton is weakly coupled) then the energy density in the inflaton field decreases as the universe expands, leading to a much lower reheat temperature. Thus for inflation to solve the monopole crisis, the reheat temperature must be well below the GUT scale, in which case monopoles are not re-introduced during the reheating process. Unfortunately neither are the X and $\bar{X}$ bosons, and thus baryogenesis does not occur.

Neither of these objections are definitive—there are proposals for circumventing them both. For example much of our discussion has focused on small departures from thermal equilibrium. It may be possible to have huge departures, where particle distributions are not even remotely thermal. In this case the analysis of reaction rates is quite different. There may also be many more couplings which allow a greater range of reaction rates. Perhaps these are associated with Yukawa couplings of neutrinos or other sectors of the GUT. These objections do however make these scenarios less compelling. In addition there is another, more philosophical, problem. Often in these models the details of baryogenesis are pushed into very particular aspects of the GUT, physics at scales which are not accessible in the laboratory. Thus in many instances, whether or not GUT baryogenesis occurs is experimentally unanswerable. For these reasons it is advisable to investigate alternatives.

3 Electroweak Baryogenesis

In 1985 Kuzmin, Rubakov and Shaposhnikov[19] made the remarkable observation that all three of Sakharov's criteria may be met in the standard model. Firstly, and perhaps most surprisingly, the standard model of the weak interactions does not conserve baryon number!

The non-conservation of baron number in the standard model is a rather subtle effect. At the classical level, the conservation of baryon number is practically obvious—each term in the classical action respects a transformation of the baryon number. Nöther's theorem then applies, and we can construct a four-vector, the baryon number current, which satisfies the continuity equation, that is whose four-divergence vanishes. Nonetheless this naïve argument is wrong: this four vector does *not* have vanishing four-divergence in the full quantum theory.

This situation is not totally unfamiliar. In the simple case of quantum electrodynamics a corresponding phenomena occurs, known as the axial anomaly. QED has a symmetry of the classical action corresponding to an axial rotation of the electron field (that is, a rotation which is opposite on the left and right chirality electron fields). Aside from the electron mass term which we will ignore, this transformation leaves the action unchanged, and the Nöther procedure leads to a covariantly conserved four-vector, the axial current. However as is well known this current is *not* divergenceless:

$$\partial_\mu J^\mu_a = \frac{e^2}{32\pi^2} F_{\mu\nu}\tilde{F}^{\mu\nu} \propto \vec{E}\cdot\vec{B}\ , \tag{12}$$

where $\tilde{F}^{\mu\nu} \equiv \epsilon^{\mu\nu\alpha\beta}F_{\alpha\beta}/2$. This remarkable equation, which can be derived in a number of different ways, embodies the violation of axial charge due to quantum effects in the theory[§]. Note that ignoring spatial variations this equation implies that the time derivative of the baryon density will be non-zero in the presence of a non-zero $\vec{E}\cdot\vec{B}$. Note that the chiral nature of the current couplings are important for obtaining this result; the current with non-chiral couplings, the electromagnetic current, is strictly conserved.

The situation in the standard model is similar. The baryon current derived via the Nöther procedure is vectorial, and thus would seem an unlikely candidate for an

[§]The axial anomaly in QED has a long and well-known history. Eq. (12) may be obtained for example by evaluating the triangle diagram, by computing the change in the functional integral measure under an axial rotation, or by an exact calculation of the electron propagator in a constant background electric and magnetic field.

anomaly. However the weak interaction couplings *are* chiral, which leads to an equation for the divergence of the baryon current corresponding to Eq. (12):

$$\partial_\mu J^\mu_B = 3\Big[\frac{g^2}{32\pi^2} W^a_{\mu\nu}\tilde{W}^{a\mu\nu} + \frac{g'^2}{32\pi^2} F_{Y\mu\nu}\tilde{F}^{\mu\nu}_Y\Big] = \partial_\mu J^\mu_L \ . \tag{13}$$

In this equation W and F_Y are the gauge field strengths for the $SU(2)$ and $U(1)$ hypercharge gauge potentials, g and g' are the corresponding gauge couplings and the 3 arises from a sum over families. We have also noted that the lepton number current has the same divergence as the baryon number current. Consequently the current $J_{B-L} \equiv J_B - J_L$ is divergenceless, and the quantum number $B - L$ is absolutely conserved.

What does Eq. (13) really mean? To gain some understanding of this equation, imagine constructing an electroweak solenoid surrounding an electroweak capacitor, so that we have a region in which the quantity $\vec{E}^a \cdot \vec{B}^a$ is non-zero. In practice this is rather difficult, primarily because we live in the superconducting phase of the weak interactions, and therefore the weak Meissner effect prevents the development of a weak magnetic field. But lets ignore this for the moment. Now perform the following gedanken experiment: start with no weak electromagnetic fields, and the region between the capacitor plates empty. If we solve the Dirac equation for the quarks and leptons, we obtain the usual free particle energy levels. In this language, we fill up the Dirac sea, and leave all positive energy levels unoccupied. Now imagine turning on the weak $\vec{E}^a$ and $\vec{B}^a$ fields adiabatically. In the presence of these slowly varying fields, the energy level solutions to the Dirac equation will flow, while the occupation of any given level does not change. But according to Eq. (13) the baryon number will change with time. This corresponds to the energy of some of the occupied levels in the Dirac sea flowing to positive energy, becoming real particles carrying baryon number. Although surprising at first, this is not very different from ordinary pair production in a background field. What is peculiar is the creation of quarks in a way different from antiquarks, so that a net baryon number is produced.

By itself this effect is intriguing but not sufficient. After all what we are really after is a transition which changes baryon number without changing the state of the gauge field, much as the four-fermion operator in our grand unified example did. That is, what we would like to do is begin our gedanken experiment as above, but at the end of the day turn off the electric and magnetic fields. Naïvely this would leave us with zero baryon number: if we turn the fields off as the time reverse of how we turned them on, we produce baryon number at first, and then remove it later on. Indeed this

is what happens with axial charge in the quantum electrodynamics example. But the non-abelian example contains another wrinkle: it is possible to turn the electric and magnetic fields on and then off in a way which leaves a non-zero baryon number!

The trick as realized by 't Hooft[20,21] follows from noticing that, unlike the abelian case, there are a large number of non-trivial gauge potentials which have vanishing electric and magnetic fields. It is possible in our gedanken experiment to begin with one of these potentials, and finish with another, thus tying a "knot" in the gauge field[¶]. The result is a transition from a state with no weak electric and magnetic fields and no baryon number (a "vacuum"), and ending with no weak electric and magnetic fields but non-zero baryon number. Making such a transition requires a "large" gauge field, one in which the field strength is of order $1/g$. In addition, the total change in baryon number is quantized in units of the number of families, presumably 3.

If we accept this fancy formalism, we have an obvious question: why is the proton stable? If the weak interactions violate baryon number, shouldn't the proton lifetime be a characteristic weak time scale? In fact, the proton is absolutely stable even in the presence of this baryon violation, because each process changes the baryon number by 3. Since the proton is the lightest particle carrying baryon number, its decay would require changing the baryon number by 1, which cannot occur if all baryon violating process change the baryon number by multiples of 3. Thus there is a selection which accounts for the stability of the proton.

What about other baryon violating processes? In fact these too are unimportant. In our gedanken experiment above we ignored the fact that the weak interactions are broken, that we live in a superconducting phase of the weak interactions. But this means that there is a large potential energy cost in creating a weak $\vec{E}^a$ and $\vec{B}^a$ field which interpolates between our states with different baryon number. That is, there is a potential barrier that we must overcome in order to change the baryon number by a weak interaction. Since the gauge field must change by order $1/g$, the height of this barrier (the cost of overcoming the Meissner effect) is

$$E_s \sim \frac{M_Z}{\alpha_{wk}} \sim \text{a few TeV} \tag{14}$$

where α_{wk} is the weak analog of the electromagnetic fine structure constant. The gauge field configuration at the peak of this barrier is called the "sphaleron", and hence this

[¶]This argument is a bit tricky. In order to discuss the physics of gauge potentials it is necessary to gauge fix. Even after gauge fixing there are gauge potentials which begin in the far past with one "vacuum" potential, and end with a different one.

energy is known as the sphaleron energy.

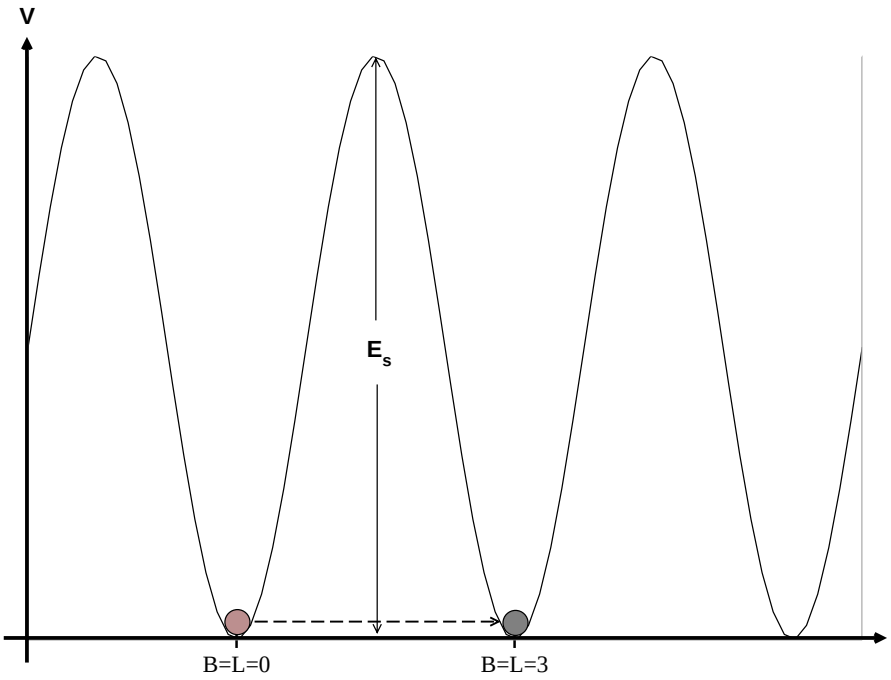

Fig. 2. The potential energy in one direction in gauge field space. This direction has been chosen to go from one zero energy gauge field configuration to another through the pass of lowest energy.

The presence of this barrier means that processes with energies below the barrier height are highly suppressed; they are strictly forbidden classically, but can occur through quantum tunneling. Like all tunneling processes, the probability of such a transition will be proportional to a semi-classical barrier penetration factor:

$$\text{Prob} \propto e^{-4\pi/\alpha_{wk}} \sim 10^{-40} \,, \tag{15}$$

an utterly negligible effect. In contrast to the grand unified case where baryon violation was suppressed at low energies by powers of the ratio of the energy to the grand unification scale, here the baryon violation is exponentially suppressed by the presence of a barrier.

If our interest were only sensitive tests of baryon number conservation in the laboratory, we would safely move on to another area of research. But since our interest is in baryogenesis in the early universe, we must take this picture of baryon violation in the weak interaction by transiting this barrier more seriously. At temperatures comparable

or larger than the barrier height we would expect a significant population of states with energies above the barrier. These states could make a transition without the quantum tunneling suppression by simply evolving classically over the top of the barrier. The rate for such a baryon violating process will be controlled by the probability of finding a state with energy at least as large as the sphaleron energy:

$$\Gamma \propto e^{-E_s/T} \ . \tag{16}$$

When the temperature is larger than E_s this exponential is no longer a suppression at all. Hence we expect that at temperatures above a few TeV baryon violation in the weak interactions will occur at a characteristic weak interaction rate. Note that at temperatures of a few TeV weak interactions are extremely rapid compared to the Hubble expansion rate, and thus baryon violating interactions would be in thermal equilibrium.

We come to the first important consequence of baryon violation in the weak interactions: grand unified baryogenesis does not necessarily produce a baryon asymmetry! Even if a late decaying X boson would produce a baryon asymmetry at temperatures near the GUT scale, this asymmetry will be equilibrated away by baryon violating weak interactions. Our discussion of grand unified baryogenesis concluded that baryon violation from virtual X boson exchange must be slow for baryogenesis to succeed, but the real requirement is that *all* baryon violation must be slow; we must take into account *all* sources of baryon violation, including that of the weak interactions.

There is a simple way of avoiding this effect. As indicated in Eq. (13) the baryon and lepton number currents have exactly the same divergence. Hence their difference, the $B-L$ current, is strictly conserved. Therefore if the X boson decay produces a net $B-L$, weak interactions cannot equilibrate this quantum number to zero. The result will be both a net baryon number and a net lepton number. However baryon and lepton number violating weak interactions must be taken into account when calculating the baryon asymmetry produced.

Rather surprisingly we have concluded that baryon violation is present in the standard model, at least at temperatures above a few TeV. In principle this opens the possibility of baryogenesis taking place at temperatures well below the GUT scale. Unfortunately we face another obstacle: departure from thermal equilibrium. As discussed earlier, the expansion rate of the universe at temperatures near a TeV is quite slow: $H \sim T^2/M_P \sim 10^{-16}$ TeV. All standard model interactions lead to reaction rates much larger than this expansion rate, typically of order $\Gamma \sim \alpha_{wk} T \sim 10^{-3}$ TeV. Thus departure from thermal equilibrium is impossible with such a leisurely expansion. For-

tunately there are occasions during the early universe in which the smooth variation of the temperature with the expansion, Eq. (9), is invalid. This typically occurs when the equation of state for the content of the universe undergoes an abrupt change, such as during a change in phase structure. For example when the temperature falls below the mass of the electron, electrons and positrons annihilate into photons, converting their energy from a non-relativistic form (the mass-energy of the leptons) into a relativistic form (radiation). But there may be other phase changes in the early universe. With a phase transition there exists the possibility of significant departure from thermal equilibrium, at least if the transition is discontinuous, or first order.

Is there any reason to expect a phase transition in the early universe? At temperatures much higher than a few TeV we have very little idea of the state of the universe; until we probe physics at these high energies in the laboratory we cannot say whether or not phase transitions occur. Of course we are permitted to speculate, and indeed there are many proposals for new physics beyond the standard model which lead to interesting dynamics in the early universe. But beyond speculation, we already expect that there is at least one phase transition in the context of the standard model: the electroweak phase transition.

As we have already mentioned we currently live in a superconducting phase of the electroweak interactions. The W and Z boson masses arise from the interaction of the gauge fields with a non-zero order parameter, an object that carries electroweak quantum numbers and has a non-zero expectation value in the vacuum. The short range nature of the weak force is a consequence of this interaction, just as the electromagnetic interaction is short range in ordinary superconductors. In fact it is this property of the weak interactions which leads us to deduce the existence of a non-zero order parameter. We know the value of the order parameter, the weak vev, is approximately 250 GeV; we also know that the order parameter is a weak doublet, from the relation between the W and Z masses and the weak mixing angle. However unlike electromagnetic superconductivity where the order parameter is known to be a composite of two electrons, a so-called "Cooper pair", the weak order parameter remains mysterious. One possibility is that the order parameter is simply some new field with its own physical excitations, the Higgs field. Another is that it is a composite of two fermions, like the Cooper pair. But until we have probed the details of electroweak symmetry breaking in detail, as we hope to do in future collider experiments, we can not say with any confidence what form the detailed physics of this order parameter takes.

One thing we do expect, in analogy with ordinary superconductivity, is the change

in phase of the weak interactions at high temperatures. Just as an electromagnetic superconductor becomes non-superconducting as the temperature is increased, so too the weak interactions should revert to an unbroken phase at high temperature. When the temperature is on the order of 100 GeV, the order parameter should vanish, the weak gauge symmetry will be unbroken and the W and Z (and the quarks and leptons) will become massless. In our discussion of baryon violation in the weak interactions we suggested that at temperatures larger than the sphaleron energy baryon violation would be unsuppressed, as transitions could take place above the barrier. But the barrier itself was a consequence of the Meissner effect, a sign of superconductivity. Indeed Eq. (14) clearly shows the relationship with symmetry breaking: the sphaleron energy is proportional to M_Z which in turn is proportional to the order parameter. At temperatures of a few hundred GeV, well below the sphaleron energy, when the weak symmetry is restored and the order parameter goes to zero, the barrier disappears. Consequently baryon violation will occur rapidly just on the unbroken side of the phase transition.

In order for any of this to play a role in baryogenesis, we require significant non-equilibrium effects at the phase transition. According to the usual classification of phase transitions, such non-equilibrium effects will arise if the phase transition is first order. Under these circumstances the transition itself may proceed in a classic first order form, through the nucleation of bubbles of broken phase[‖]. Indeed as the universe cools from high temperature, we begin with a homogeneous medium in the unbroken phase of the weak interactions. Quarks and leptons are massless, weak interactions are long range (aside from thermal screening effects) and, most importantly, baryon violation is rapid. Calculating the rate for baryon violation requires understanding the details of the classical thermodynamics of the gauge fields, a difficult subject. The result however is relatively simple:

$$\Gamma_{\Delta B} \sim \alpha_{wk}^5 T \tag{17}$$

This is a rather crude approximation; for example there are logarithmic corrections to this relation that may be significant, as well as a potentially large dimensionless coefficient. Nevertheless the exact formula may in principle be obtained numerically in terms of α_{wk} and the temperature.

As the universe cools we eventually reach a moment in which the free energy of the unbroken phase is equal to that of the broken phase, as indicated by the free energy

[‖]This is not the only possibility; for example it may proceed through spinodal decomposition, or some more complicated mechanism. In all these circumstance non-equilibrium phenomena are likely.

curve labeled by T_c in Fig. (3). However if the transition is first order, these two phases are separated by a free energy barrier and the universe, unable to reach the broken phase, remains in the unbroken phase. As the universe continues to expand, the system *supercools*, remaining in the unbroken phase even though the broken phase has a lower free energy. Finally we reach a point where bubbles of the preferred, broken, phase nucleate and begin to grow. Eventually these bubbles percolate, completing the transition.

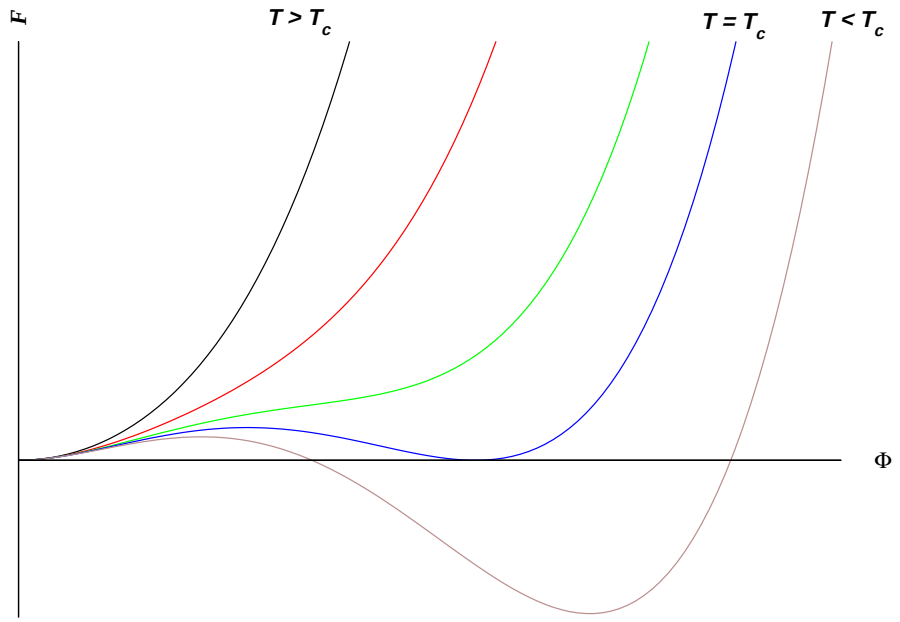

Fig. 3. The free energy versus the order parameter for a classic first order phase transition.

Clearly these expanding bubbles represent a departure from thermal equilibrium. From the point of view of Sakharov's condition the most relevant fact is the discontinuity in the order parameter, the weak vev. In the region outside the bubbles the universe remains in the unbroken phase where the weak order parameter is zero. As discussed previously there is no barrier between the states of different baryon number, and baryon violation is rampant. In the bubble interior the weak vev is non-zero, the W and Z bosons are massive, and *the barrier between states of different baryon number is in place.* In this case the rate of baryon violation is exponentially suppressed according to a Boltzmann factor $\exp(-E_b/T)$ where E_b is the barrier height. Naïvely we might expect E_b to be the sphaleron energy. However the sphaleron energy represented

the barrier height at zero temperature; at finite temperature the barrier is generically different, evolving to the zero temperature shape as the universe cools. But E_b is still controlled by the order parameter, the weak vev. If this vev is large, near its vacuum value of 250 GeV, baryon violation will be essentially shut off in the bubble interior. On the other hand if the vev is too small, baryon violation will proceed rapidly inside the bubble as well as out.

The difference in the weak vev in the bubble interior and the bubble exterior, the discontinuity in the weak order parameter, is a measure of the strength of the transition. If these two values are nearly equal, the phase transition is nearly continuous, a second order transition. If on the other hand the discontinuity is large, the phase transition is said to be strongly first order. For electroweak baryogenesis to occur, baryon violation must be out of thermal equilibrium in the bubble interior, a situation that will transpire only if the vev is sufficiently larger. Thus we need a strongly first order electroweak phase transition.

What do we know about the electroweak phase transition? Unfortunately almost nothing. This is due in small measure to our inability to understand the complex thermal environment in a relativistic quantum field theory. Over the past decade there has been a great deal of progress in simulating field theories at finite temperature, deducing details of phase transitions and reaction rates. However these advances are of little use if we don't know what theory to simulate. The main reason we can't say definitely whether the electroweak phase transition is first or second order, whether it is strongly or weakly first order, or practically anything else about it is simple: we have no idea what physics is responsible for electroweak symmetry breaking.

We do have some theories of electroweak symmetry breaking, and huge effort has been invested in determining the details of the phase transition in these cases. The original theory of electroweak symmetry breaking relied on the introduction of a fundamental weak doublet scalar field, the Higgs field. In this rather simple case, the electroweak phase transition is first order only if the physical Higgs scalar is very light, with a mass well below the current experimental bound. But this theory is not the most popular alternative for electroweak symmetry breaking due to its theoretical shortcomings. Of somewhat greater appeal is the minimal supersymmetric standard model, the MSSM. In this case there are a host of new particles: supersymmetric partners of the quarks, leptons and gauge bosons, as well as two Higgs multiplets. In fact this theory also requires some of these new states to be relatively light in order to obtain a sufficiently strongly first order phase transition. As the LEP bound on the MSSM Higgs mass im-

proves, the region of parameter space for which the phase transition is appropriate is rapidly disappearing.

Should we take this to mean the weak phase transition is probably inappropriate for electroweak baryogenesis to take place? That depends a bit on our philosophy. Given that these are but 2 ideas out of a nearly infinite variety we should not necessarily become disheartened. More importantly there have been analyses of modest alternatives of the above theories: non-supersymmetric theories with multiple Higgs fields, extensions of the MSSM including singlets, and even strongly interacting theories of electroweak symmetry breaking. In most of these cases a sufficiently strong first order phase transition is easy to arrange, if not generic. In fact this is perhaps one of the more positive aspects of electroweak baryogenesis. The physics responsible for electroweak symmetry breaking is intimately related with the possibility of electroweak baryogenesis: some models of electroweak symmetry breaking do not produce a baryon asymmetry (or not one of sufficient size) while others do. This is one of the few places that the forefront of electroweak physics, electroweak symmetry breaking, may have a profound effect on cosmology (or vice versa).

3.1 Baryon Production

We now have all of Sakharov's ingredients in place, all in the weak interactions: baryon violation, C and CP violation and a departure from thermal equilibrium. But we still have not explored how these ingredients combine to produce a baryon asymmetry.

Clearly we require all three ingredients to work together—the absence of any one implies the absence of baryogenesis. The non-equilibrium requirement, satisfied by the nucleation and subsequent expansion of bubbles of broken phase, is most importantly realized as a spatial separation of baryon violation: baryon violation is rapid outside the bubble, and non-existent in the bubble interior. C and CP violation, at least in the standard model, take place through the Yukawa couplings in the Lagrangian. That is, C and CP violation appear in the form of non-trivial phases in the couplings of quarks (and possibly leptons in extensions of the standard model) to the Higgs field, the order parameter for electroweak symmetry breaking. But it is precisely this field which represents the electroweak bubbles which appear at the phase transition.

The details of how the baryon asymmetry may be calculated in the context of these expanding bubbles is complicated, and we will not discuss it at any length. The ingredients are clear: the CP violating interaction of quarks and leptons with the expanding

bubbles can in principle bias the production of various quantum numbers (including but not limited to baryon and lepton number); all that is required is an interaction that allows the creation or destruction of a net value for such a quantum number. For example, the interaction with the expanding bubble may bias the production of left-chirality top quarks over right-chirality top quarks (to pick a random example). Provided CP violation (either directly or in the form of one of these quantum number asymmetries) biases baryon number in a region outside the bubble where baryon violation is rapid, a net baryon number will be produced. Following our example, an excess of left-chirality top quarks (which have a weak interaction) over right-chirality top quarks (which do not) biases the weak interactions in the direction of increasing baryon number. An important element which complicates the discussion is the transportation of quark and lepton charges from one region of space to another. The transport properties of the plasma are crucial in understanding how the baryon violating interactions, which take place outside the bubble, are biased by CP violation, which is dominant where the Higgs field is changing inside the bubble. Depending on the details of the bubble profile the analysis looks a bit different, although the results are qualitatively similar.

3.2 CP Violation

We finally must come to grips with CP violation; now that we understand how it is relevant to electroweak baryogenesis, we can ask what the characteristic size of CP violating effects of the sort described in the last paragraph will be. In fact this question is not as difficult as might be supposed. CP violation in the standard model arises from a non-trivial phase in the Yukawa couplings of the quarks. The only tricky issue is that this phase has no unique location: we may move it from one coupling to another by making field redefinitions. More physically this means that an interaction will only violate CP when the interaction involves enough couplings such that we cannot remove this phase from all these couplings simultaneously. For example, if a process involves only two families of quarks, the CP violating phase may be put in the third family, and this process will be CP conserving.

Since the Yukawa couplings are relatively small (even the top quark coupling), perturbation theory should be an adequate guide to the size of CP violating effects. To estimate this size we must construct an object perturbatively out of the various coupling constants of the standard model in a way which involves an (irremovable) CP violating phase. Clearly there must be a large number (8) of Yukawa couplings from all

three families as well as a large number (4) of weak interactions in order to get an irremovable phase. This product of small dimensionless coupling constants is an invariant measure of CP violation in any perturbative process. One such example, involving the largest Yukawa couplings, is

$$\delta_{CP} \sim \alpha_{wk}^2 \lambda_t^4 \lambda_b^2 \lambda_s \lambda_d \sin^2\theta_1 \sin\theta_2 \sin\theta_3 \sin\delta \sim 10^{-16} \ . \tag{18}$$

This remarkably small number, many orders of magnitude smaller than the observed baryon asymmetry, is a consequence of the detailed symmetries of the standard model, where CP violation is intimately connected with flavor violation. As long as the flavor physics of baryogenesis is perturbative, the standard model has no hope of producing a baryon asymmetry large enough. Although we have consistently maintained that the standard model has CP violation, and that this is one of the most interesting reasons to investigate baryogenesis, it now seems that we have been misled, that this CP violation is far too small to be relevant for baryon production in the early universe.

Why did we argue earlier that CP violation in the kaon system, Eq. (7), was so much larger than this perturbative estimate? In fact we have been careful to argue that the estimate of CP violation, Eq. (18), only applies when the standard model Yukawa interactions can be used perturbatively. This is not the case for CP violation in the kaon system. If we wish to compute CP violating effects at kaon energies, $E << 250$ GeV, we must first construct the effective theory appropriate to these energy scales by integrating out modes with energies larger than E. This includes for example the W and Z, the top and bottom quarks, *etc*. As usual this process introduces inverse powers of these heavy masses, such as $1/M_W^2$ and $1/m_t^2$. Since these masses are proportional to the weak couplings g and λ_t appearing above, this effective theory has interactions which can *not* be represented as a power series in couplings (although it is easy enough to construct this effective theory and keep track of the Yukawa couplings), and the estimate Eq. (18) does not apply[**].

But we have now come to the crux of the matter, and if it were not for the interesting physics associated with baryon violation, cosmic expansion, *etc*. that we wished to discuss we could have started (and ended) our discussion of baryogenesis here. The most important message from this analysis is that it is highly unlikely that CP violation from the phase in the CKM matrix has anything at all to do with the cosmic baryon asymmetry. Although we have chosen to mention this in the context of electroweak

[**] A more old-fashioned language for the same phenomena would note the enhancement of perturbative matrix elements by small energy denominators in perturbation theory.

baryogenesis, there is nothing special about this scenario in our analysis of the size of CP violating effects. Everything we have said applies to standard model CP violation in any theory of baryogenesis that takes place at high energies where our perturbative argument applies. This is certainly the case in grand unified baryogenesis as well as electroweak baryogenesis.

Once more, with feeling: standard model CP violation in the form of a phase in the CKM matrix is not likely to produce a significant baryon asymmetry of the universe. Why is this so important? As we have argued there *is* a cosmic baryon asymmetry, and if it didn't come from CP violation in the standard model, where did it come from? The obvious conclusion is that there is CP violation (and hence new physics) beyond the standard model. This is one of the strongest pieces of evidence we have that the standard model is incomplete.

One comment is in order. We have now repeatedly said that standard model CP violation is inadequate for baryogenesis. This is sometimes confused with the (incorrect) statement that the CP violation observed in the kaon system is too small to produce the observed baryon asymmetry. At the moment our knowledge of CP violation is not extensive enough to say definitively that the observed CP violation is associated with a phase in the CKM matrix. It is perfectly possible that CP violation in the kaon system is dominated by physics beyond the standard model. This would likely show up as a discrepancy between CP violation measured in the B system relative to the expectations from the K system.

If the standard model must be augmented with new CP violation to create the baryon asymmetry, what form is this new CP violation likely to take? We don't know. However it is worth noting that CP violation in the standard model, with its intimate connection to flavor symmetries, is rather special. In almost any extension of the standard model, new interactions and new particles allow for new sources of CP violation. Under these circumstances this new CP violation is not constrained by the standard model flavor symmetries and will typically give large effects. Indeed the apparent smallness of CP violation at low energies is a strong constraint on physics beyond the standard model, since most extensions of the standard model lead to large, even unacceptable, CP violating effects.

Most investigations of baryogenesis have focused on models proposed for reasons other than CP violation and the baryon asymmetry. For example, a natural extension of the original fundamental Higgs standard model includes multiple Higgs fields. With one or more new Higgs fields there are new CP violating couplings, the flavor structure

of the model is different, and baryogenesis is certainly possible. A particularly popular extension of the standard model, the MSSM, has a number of new CP violating phases, and can easily have large CP violation at the electroweak scale. As we have discussed, the phase transition in this model may be too weak (depending on the latest bounds on the parameters of the Higgs potential) to allow electroweak baryogenesis, but most non-minimal extensions of this model (for example the inclusion of a new singlet superfield), allow a strongly first order phase transition consistent with current supersymmetry bounds. In grand unified models new CP violation may be associated with the scalar fields necessary to break the grand unified symmetry. Many examples of this type have been proposed.

This is in fact the best news from baryogenesis, especially electroweak baryogenesis. By bringing the physics of baryon production down to energies that we are currently probing in the laboratory, we have an opportunity to verify or falsify these ideas in detail. For example CP violation in the extensions of the standard model mentioned above, particularly supersymmetry, lead to observable effects at low energies, both CP conserving and CP violating. If the next round of collider experiments determine the nature of electroweak symmetry breaking, then the nature of the phase transition and its suitability for electroweak baryogenesis may be determined. If new CP violation is observed in experiments like the B factory, or in electric dipole moment experiments, it will be especially interesting to determine the flavor structure of this CP violation and its possible connection with the baryon asymmetry of the universe.

Although we have only touched on two broad areas of baryogenesis, electroweak and grand unified, there are a variety of other interesting ideas, including spontaneous baryogenesis, topological defects, *etc*. One of the more interesting variants, leptogenesis, involves the production of an asymmetry in lepton rather than baryon number. Subsequent production of baryon number then relies upon further processing of the lepton number asymmetry by interactions, like the electroweak interaction we have already discussed. These models are especially timely since the lepton asymmetry may be connected with the physics of neutrinos, an area where we are now beginning to obtain a great deal of experimental information.

The only bad news here, is the rather vague connection between baryogenesis and *specific* laboratory experiments. There is no single smoking gun; new CP violation large enough to produce the observed baryon asymmetry will almost certainly have low energy effects, but not decisively so. And where these effects show up, be it in EDMs, B or D mixing, or top quark physics, is highly model dependent. Without

more experimental information constraining our current theoretical ideas, baryogenesis does not suggest that any one experiment is more likely than another to see new CP violation. But these are minor quibbles. Baryogenesis is already a strong indication of new physics to come, and even tells us that this new physics should emerge in one of the most fascinating areas of current research, CP violation.

Baryogenesis has been a fruitful cross-roads between particle physics and cosmology. Uniting ideas of early universe phase transitions, electroweak symmetry breaking and CP violation, it is an area that touches on many of the most exciting experiments that we look forward to in the coming decade. The B factory, the LHC, the Tevatron and even tabletop atomic physics experiments, may provide provide the clues that help explain the presence of matter in the universe. Unraveling the mystery of the cosmic baryon asymmetry remains one of the most exciting tasks for particle physicists and cosmologists alike.

References

[1] E. W. Kolb and M. S. Turner. *The Early universe*. Addison-Wesley, 1990.

[2] A. G. Cohen, D. B. Kaplan, and A. E. Nelson. Progress in electroweak baryogenesis. *Ann. Rev. Nucl. Part. Sci.*, 43:27–70, 1993.

[3] V. A. Rubakov and M. E. Shaposhnikov. Electroweak baryon number non-conservation in the early universe and in high-energy collisions. *Usp. Fiz. Nauk*, 166:493–537, 1996.

[4] Antonio Riotto and Mark Trodden. Recent progress in baryogenesis. 1999. hep-ph/9901362 submitted to Ann. Rev. Nucl. Part. Sci.

[5] G. Steigman. Observational tests of antimatter cosmologies. *Ann. Rev. Astron. Astrophys.*, 14:339–372, 1976.

[6] S. P. Ahlen. The ams experiment to search for antimatter and dark matter. In *Proceedings of the 5th Annual LeCroy Conference on Electronics for Particle Physics, Chestnut Ridge, NY*, 1995.

[7] S. P. Ahlen. Ams: A magnetic spectrometer for the international space station. In S. J. Ball and Y. A. Kamyshkov, editors, *Proceedings of the International Workshop on Future Prospects of Baryon Instability Search in p decay and n—anti-n Oscillation Experiments, Oak Ridge, TN*, 1996.

[8] R. Battiston. The alpha magnetic spectrometer (ams): Search for antimatter and dark matter on the international space station. *Nucl. Phys. Proc. Suppl.*, 65:19, 1998.

[9] R. Battiston. The alpha magnetic spectrometer (ams). *Nucl. Instrum. Meth.*, A409:458, 1998.

[10] V. Plyaskin. Antimatter and dark matter search with the alpha magnetic spectrometer (ams). *Surveys High Energ. Phys.*, 13:177, 1998.

[11] J. Casaus. The alpha magnetic spectrometer (ams). *Acta Phys. Polon.*, B30:2445, 1999.

[12] U. Bekcer. Alpha magnetic spectrometer ams report on the first flight in space june 2-12. In *Proceedings of the 29th International Conference on High-Energy Physics (ICHEP 98), Vancouver, British Columbia, Canada*, 1998.

[13] A. G. Cohen, A. De Rujula, and S. L. Glashow. A matter-antimatter universe. *Astrophys. J.*, 495:539, 1998.

[14] S.C. Kappadath et al. In *Proc. 12th Int. Cosmic Ray Conf.*, page 25, 1995.

[15] William H. Kinney, Edward W. Kolb, and Michael S. Turner. Ribbons on the cbr sky: A powerful test of a baryon symmetric universe. *Phys. Rev. Lett.*, 79:2620–2623, 1997.

[16] A. G. Cohen and A. De Rujula. Scars on the cbr? *Astrophys. J.*, 496L:63, 1998.

[17] Keith A. Olive, Gary Steigman, and Terry P. Walker. Primordial nucleosynthesis: Theory and observations. 1999. astro-ph/9905320 submitted to Phys. Rep.

[18] A. D. Sakharov. Violation of cp invariance, c asymmetry, and baryon asymmetry of the universe. *JETP Letters*, 5:24, 1967.

[19] V. A. Kuzmin, V. A. Rubakov, and M. E. Shaposhnikov. On the anomalous electroweak baryon number nonconservation in the early universe. *Phys. Lett.*, B155:36, 1985.

[20] G. 't Hooft. Symmetry breaking through bell-jackiw anomalies. *Phys. Rev. Lett.*, 37:8–11, 1976.

[21] G. 't Hooft. Computation of the quantum effects due to a four- dimensional pseudoparticle. *Phys. Rev.*, D14:3432–3450, 1976.

ACCELERATOR PHYSICS ISSUES AT THE LHC AND BEYOND

Frank Zimmermann
CERN, SL Division
1211 Geneva 23, Switzerland

ABSTRACT

I review the past performance of hadron colliders and their limitations, discuss the accelerator physics challenges faced by the Large Hadron Collider (LHC) now under construction, and, finally, present an outlook into the future, covering upgrades of the LHC as well as a Very Large Hadron Collider.

1 Introduction

This lecture is structured as follows. First, past and future hadron colliders and the effects limiting their performance are reviewed. Then, I discuss the accelerator physics challenges being confronted by the Large Hadron Collider (LHC). Lastly, an outlook onto the future is given, which includes scenarios for an LHC upgrade and the proposed two stages of a Very Large Hadron Collider (VLHC).

1.1 Collider Performance

The two primary parameters characterizing the performance of a collider are its energy and its luminosity. The maximum beam energy of a hadron collider grows linearly with the strength of the magnets and with the ring circumference. The second parameter, the luminosity L, characterizes the reaction rate R. One can write

$$R = L\sigma \tag{1}$$

where σ is the cross section for a particular reaction. The luminosity L is conventionally quoted in units of cm^{-2} s^{-1}. The particle physicists desire a large value of L and, thus, one task of the accelerator physicist is to increase L as much as possible. If one approximates the transverse beam profile by a Gaussian distribution, the luminosity can be expressed in terms of beam parameters as

$$L \approx \frac{N_b^2 n_b f_{\rm rev} \gamma}{4\pi \epsilon_{x,N} \beta_x^* \kappa} \tag{2}$$

where N_b denotes the number of particles per bunch, n_b the number of bunches per ring, $f_{\rm rev}$ the revolution frequency, γ the particle energy divided by the rest mass, $\epsilon_{x,N}$ the normalized (subindex 'N') horizontal emittance, and $\kappa = \sigma_y/\sigma_x$ the beam-size aspect ratio at the collision point.

The emittance specifies the area in phase space occupied by the beam. A vertical, horizontal, and longitudinal emittance are defined for the three degrees of motion. These are denoted by $\epsilon_{x,N}$, $\epsilon_{y,N}$, and $\epsilon_{z,N}$ (or $\epsilon_{L,N}$). Without diffusion due to scattering processes or synchrotron radiation, the normalized emittances are conserved quantities under acceleration.

More precisely, the emittances are equal to the area of the ellipse in a 2-dimensional phase space which is encircled by a particle launched at an amplitude equal to the rms

beam size, divided by π and by the particle rest energy, *e.g.*, in the horizontal plane,

$$\epsilon_{x,N} = \frac{\oint p_x dx}{\pi m_0 c^2} \ , \tag{3}$$

where x and p_x are the horizontal position and momentum of the particle, as viewed at one location in the ring on successive turns, and m_0 is the particle mass.

The *unnormalized* or *geometric* horizontal emittance is defined as $\epsilon_x = \epsilon_{x,N}/(\gamma\beta)$, or, equivalently, as

$$\epsilon_x = \oint x' dx/\pi \ , \tag{4}$$

where $x' \equiv p_x/p_z$ is the slope of the particle trajectory, p_z the longitudinal momentum, and $\beta = v/c$ the velocity in units of the speed of light.

At any given location around in the ring, the emittance is proportional to the square of the rms beam size, *e.g.*, for the horizontal plane at location s we have

$$\sigma_x^2(s) = \frac{\beta_x(s)\epsilon_{x,N}}{\gamma} = \beta_x(s)\epsilon_x \tag{5}$$

where $\beta_x(s)$ is the horizontal beta function. Equation (2) shows that a small beam size corresponds to a higher luminosity, and in view of Eq. (5), this implies a small beta function at the collision point, a small emittance, and a high energy. In particular, Eq. (2) indicates that for a constant normalized emittance, $\epsilon_{x,N}$, and for a constant beta function the luminosity increases linearly with the beam energy.

We mention in passing that Eq. (2) is an approximation because it ignores variations in the beam-beam overlap which may arise from (1) a crossing angle between the two beams, (2) the change of the transverse beam size over the length of the two colliding bunches, also known as the 'hour-glass effect', and (3) the change in the optics due to the beam-beam collision. The approximation of Eq. (2) is good, if the crossing angle θ_c is small compared with the bunch diagonal angle σ_x/σ_z, if the bunch length is small compared with the beta functions $\beta^*_{x,y}$ at the collision point, and if the additional tune shift induced by the collision is small.

The primary luminosity limitations of present and future hadron colliders are imposed by a number of effects, each of which constrains one or several of the parameters on the right-hand side of Eq. (2), or even the value of the luminosity, on the left, itself. The most prominent of these effects include:

1. the beam-beam interaction which refers to either the nonlinear or the coherent interaction of the two colliding particle beams, and which is important for all hadron colliders;

2. the number of available particles which is a concern for $\mathrm{p\bar{p}}$ and ion colliders;
3. the emittance growth due to intrabeam scattering, *i.e.*, scattering of the particles inside a bunch off each other;
4. the luminosity lifetime;
5. the heat load inside the cold superconducting magnets due to synchrotron radiation and electron cloud (we will discuss the electron cloud in a later section);
6. the number of events per crossing, which is limited by the capacitiy of the detector; and
7. quenches (transitions into the normal state) of superconducting magnets due to localized particle losses near the interaction region.

In the course of this lecture, we will describe or give examples for all of these effects.

There could be other parameters relevant to the collider performance, for example the beam polarization. However, this option is presently not foreseen for the next and next-to-next generations of energy-frontier machines, *i.e.*, LHC and VLHC, the only exception being the Relativistic Heavy Ion Collider (RHIC) on Long Island, and we will not discuss it here.

1.2 Past and Future

So far 4 hadron colliders have been in operation, namely the ISR, SPS, Tevatron, and RHIC. A 5th is under construction, the LHC.

The CERN ISR started operation in 1970. A double ring ppcollider, it reached a peak luminosity of 2.2×10^{32} cm^{-2}s^{-1} and a maximum beam energy of 31 GeV with coasting beams of 38–50 A current each. The ISR luminosity was limited by space-charge tune shift and spread (due to the defocusing force of the beam field), coherent beam-beam effects, proton-electron two-stream instabilities, pressure bumps, detector background, and accumulation efficiency.[1] The ISR also provided the first $\mathrm{p\bar{p}}$ collisions, and, when operated with bunched beams, it reached a beam-beam tune shift of $\xi = 0.0035$ per interaction point (IP) with 8 crossings.[2] The beam-beam tune shift is a parameter which characterizes the strength of the beam-beam collision, which we will define further below. The ISR first produced the J/ψ particle and the b quark, though these particles were identified among the ISR collision products only after their discovery elsewhere.

The second hadron collider was the CERN S$\mathrm{p\bar{p}}$S operating since 1981 at ten times higher energy than the ISR. The S$\mathrm{p\bar{p}}$S discovered the W and Z bosons. Its luminosity was limited by beam-beam interaction, loss of longitudinal Landau damping (the term 'Landau damping' refers to the stabilizing effect of a frequency spread within the beam), number of available antiprotons, hourglass effect, and intrabeam scattering.[3] A typical beam-beam tune shift was $\xi = 0.005$ at each of three interaction points.

The FNAL Tevatron is the first collider constructed from superconducting magnets. Colliding-beam operation here started in 1987.[4] Tevatron luminosity is limited by antiproton intensity, beam-beam interaction including long-range effects, luminosity lifetime, number of events per crossing, and intrabeam scattering. The Tevatron reached an antiproton beam-beam tune shift above $\xi = 0.009$. It discovered the b and t quarks.

RHIC at BNL, the first heavy-ion collider, delivers luminosity since 2000. The main limiting factor is intrabeam scattering. Other factors again are beam-beam interaction, luminosity lifetime, and the number of events per crossing.

The Large Hadron Collider (LHC) is scheduled to start operation in 2006. As for the Tevatron, limits will be the beam-beam interaction, luminosity lifetime, and the number of events per crossing. Possibly, in addition, the electron cloud produced by photoemission or beam-induced multipacting,[5] and local magnet quenches induced by the collision products[6] may prove important. The LHC centre-of-mass energy is 14 TeV and its design luminosity 10^{34} cm^{-2}s^{-1}. The LHC will be the first machine where radiation damping is stronger than intrabeam scattering. The scarcity of antiprotons is no longer a problem, as LHC and all future machines will collide protons on protons.

If stronger magnets become available in the future, the LHC energy could be raised, *e.g.*, by a factor of 2. In the following, we call this energy increase, combined with a luminosity upgrade to 10^{35} cm^{-2}s^{-1}, the 'LHC-II'. Finally, there exist design concepts for two stages of a Very Large Hadron Collider (VLHC),[7] reaching an energy of up to 175 TeV centre of mass, and the Eloisatron Project.[8]

Tables 1, 2, and 3 list parameters for all these colliders, except for the ISR and the Eloisatron. The ISR was a rather special machine, whose parameters are not easily compared with the others. The properties of the Eloisatron are similar to those considered for the VLHC.

Table 1. Example parameters for heavy-ion ion colliders: gold collisions at RHIC and lead ions in LHC.

accelerator	RHIC	LHC
ion species	gold	lead
energy per charge E/Z [TeV]	0.25	7
energy per nucleon E/A [TeV]	0.1	2.76
total centre of mass E_{CM} [TeV]	39	1148
dipole field B [T]	3.46	8.4
circumference C [km]	3.83	26.66
no. of bunches n_b	57	608
number of ions per bunch N_b [10^7]	100	6.8
rms beam size at IP $\sigma^*_{x,y}$ [μm]	110	15
IP beta function $\beta^*_{x,y}$ [m]	2	0.5
tune shift per IP $\xi_{x,y}$	0.0023	0.00015
rms bunch length σ_z [cm]	18	7.5
bunch spacing $L_{\rm sep}$ [m]	63.9	124.8
rms transv. emittance $\gamma\epsilon_{x,y}$ [μm]	1.7	1.5
rms longit. emittance ϵ_L/Z [eVs]	0.12	0.2
IBS emittance growth $\tau_{\rm IBS}$ [hr]	0.4	9.8
initial luminosity L [10^{27} cm^{-2}s^{-1}]	0.2	1.0
luminosity lifetime τ [hr]	$\sim$10	9.3

Table 2. Example parameters for pp or $\mathrm{p\bar{p}}$ colliders: $\mathrm{Sp\bar{p}S}$, Tevatron run IIa ('TeV2a'),[9] and LHC. † The bunches are split in 3 trains, separated by 2.62 μs; ‡ The total LHC dipole heat load is about 0.8 W/m including the electron cloud. ⋆Equilibrium determined by radiation damping and intrabeam scattering. Arrows refer to dynamic changes during the store.

accelerator	$\mathrm{Sp\bar{p}S}$	TeV2a	LHC
beam energy E [TeV]	0.32	0.98	7
dipole field B [T]	1.4	4.34	8.39
total energy/beam [MJ]	0.05	1	334
circumference C [km]	6.9	6.28	26.7
number of bunches n_b	6	36	2800
bunch population N_b [10^{11}]	1.7 (p)	2.7 (p)	1.05
	0.8 ($\bar{\mathrm{p}}$)	$\sim$1.0 ($\bar{\mathrm{p}}$)	
no. of IPs	3	2	2 (4)
rms IP beam size $\sigma^*_{x,y}$ [μm]	80, 40	32	15.9
rms IP div. $\sigma^*_{x',y'}$ [μrad]	136, 272	91	31.7
IP beta $\beta^*_{x,y}$ [m]	0.6, 0.15	0.35	0.5
beam-beam tune shift / IP $\xi_{x,y}$	0.005	0.01	0.0034
crossing angle θ_c [μrad]	0	0	300
rms bunch length σ_z [cm]	30	37	7.7
bunch spacing L_{sep} [m]	1150	119 †	7.48
SR power P_{SR} [kW]		$< 10^{-3}$	3.6
dipole heat load dP/ds [W/m]		$\ll 10^{-3}$	0.2 ‡
betatron tune Q_β	26	$\sim$20	63
rms transv. emittance $\gamma\epsilon_{x,y}$ [μm]	3.75	~ 3	3.75
eq. horiz. emittance $\gamma\epsilon_x^{eq}$ [μm]		$\sim 10^\star$	2.03⋆
longit. emittance ϵ_L (σ) [eVs]	0.11	0.11	0.2
damp. time $\tau_{x,\mathrm{SR}}$ [hr]		1200	52
IBS growth time $\tau_{x,\mathrm{IBS}}$ [hr]	10	50(?)	142
damping decrement per IP [10^{-10}]		0.025	2.5
events per crossing		$\sim$6	18
peak luminosity L [10^{34} cm^{-2}s^{-1}]	0.0006	$\sim$ 0.02	1.00
lum. lifetime τ [hr]	9	9	10

1.3 Empirical Scaling

The empirical parameter scaling of past, present and future colliders may give an indication of the design optimization and possibly provide a guidance for the future development.

Figures 1 and 2 illustrate that both the circumference and the dipole field have increased roughly with the square root of the beam energy. This implies that, at least in the past, half of the energy gain has been realized by advances in magnet technology and the other half by expanding the real estate. We note that LHC-II is consistent with the historical trend, whereas for the VLHC a different scaling is assumed.

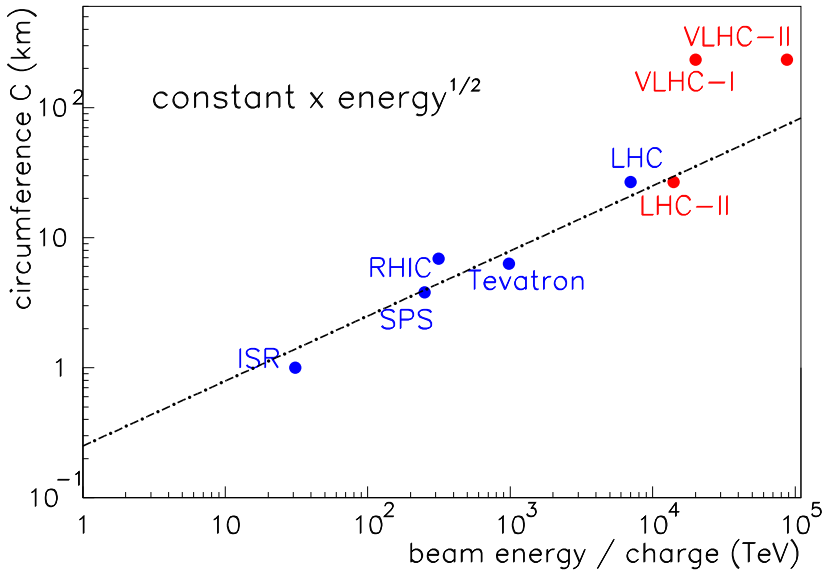

Fig. 1. Ring circumference as a function of beam energy. The solid line indicates the scaling $C \propto \sqrt{E}$.

At the same time, the luminosity has roughly followed the ideal scaling, $L \propto E^2$, as is demonstrated in Figure 3. This would ensure a constant rate of reactions, $R = L\sigma$, in case the cross section decreases inversely with the square of the energy, *i.e.*, $\sigma \propto 1/E^2$.

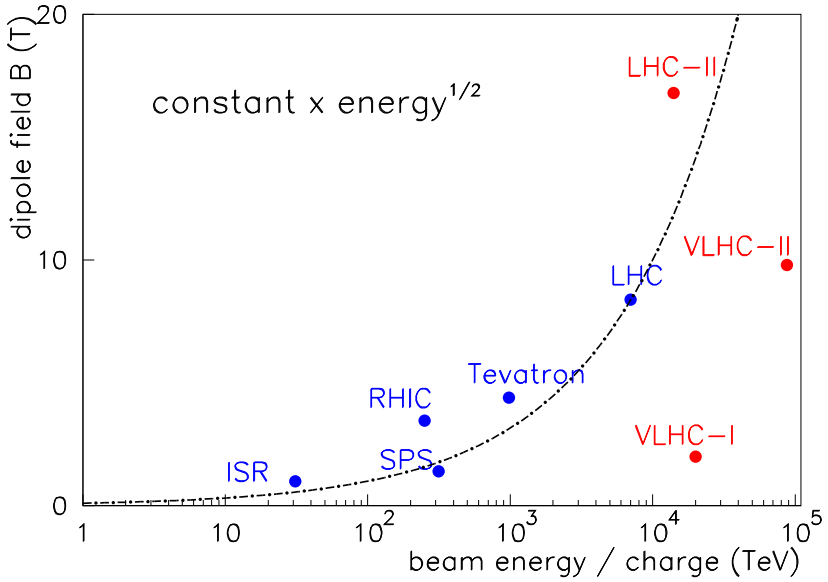

Fig. 2. Bending field as a function of beam energy. The solid line indicates the scaling $B \propto \sqrt{E}$.

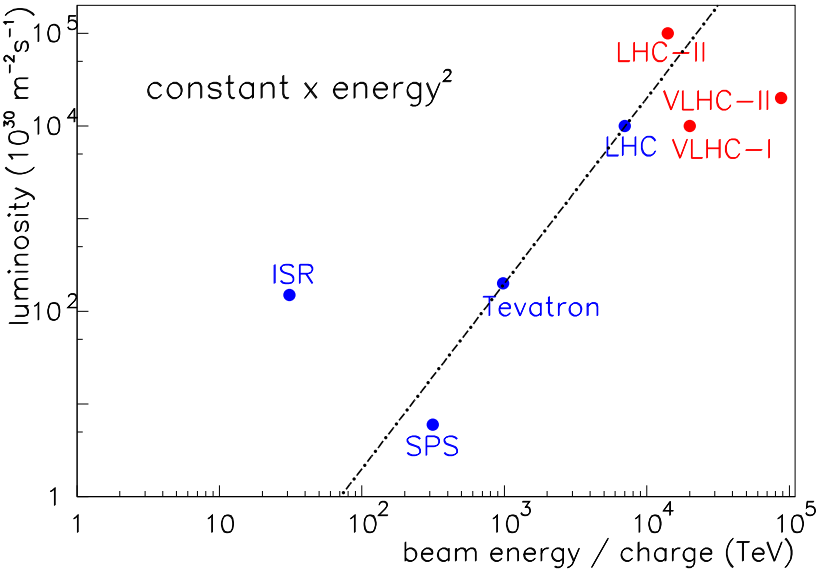

Fig. 3. Luminosity as a function of beam energy. The solid line indicates the scaling $L \propto E^2$.

Table 3. Example parameters for pp colliders: LHC-II, VLHC-I, and VLHC-II. *Assuming a dipole packing factor 0.8 for HF-VLHC, and 0.65 for LHC-II, and ignoring possible contributions from electron cloud. ⋆Equilibrium determined by radiation damping and intrabeam scattering. Arrows refer to dynamic changes during the store. The suffix 'in' indicates initial values.

accelerator	LHC-II	VLHC-I	VLHC-II
beam energy E [TeV]	14	20	87.5
dipole field B [T]	16.8	2	9.8
total energy/beam [MJ]	1320	3328	4200
circumference C [km]	26.7	233	233
number of bunches n_b	5600	40000	40000
bunch population N_b [10^{11}]	1.05	0.26	0.075
no. of IPs	2 (4)	2	2
rms IP beam size $\sigma^*_{x,y}$ [μm]	7.4⋆	4.6	3.4 →0.79
rms IP div. $\sigma^*_{x',y'}$ [μrad]	34⋆	15	5→1
IP beta $\beta^*_{x,y}$ [m]	0.22	0.3	0.71
beam-beam tune shift / IP $\xi_{x,y}$	0.005	0.002	→0.008
crossing angle θ_c [μrad]	300	153	10
rms bunch length σ_z [cm]	4.0⋆	3	→1.5
bunch spacing L_{sep} [m]	3.74	5.645	5.645
SR power P_{SR} [kW]	114	7	1095
dipole heat load dP/ds [W/m]	6.6*	0.03	4.7
betatron tune Q_β	63	220	220
rms transv. emittance $\gamma\epsilon_{x,y}$ [μm]	3.75 →1.0	1.5	1.6 →0.04
eq. horiz. emittance $\gamma\epsilon_x^{eq}$ [μm]	1.07⋆	1.0	0.06
longit. emittance ϵ_L (σ) [eVs]	0.15⋆	0.4	0.4 →0.1
damp. time $\tau_{x,\mathrm{SR}}$ [hr]	6.5	200	2
IBS growth time $\tau_{x,\mathrm{IBS}}$ [hr]	345 (in.)	400	4000→10
damping decrement per IP [10^{-10}]	20	5	400
events per crossing	90	21	54
peak luminosity L [10^{34} cm^{-2}s^{-1}]	10.	1.0	2.0
lum. lifetime τ [hr]	3.2	24	8

1.4 Accelerator Fundamentals

In a storage ring the beam particles execute transverse betatron oscillations as they circulate around the circumference. This is illustrated schematically in Fig. 4. The betatron oscillation with respect to an ideal reference particle on the 'closed orbit' is described by a quasi-harmonic oscillator equation,

$$\frac{d^2x}{ds^2} = -k(s)\ x \tag{6}$$

with the quadrupole focusing force k [m^{-2}]:

$$k = \frac{eB_T}{pa}, \tag{7}$$

where B_T denotes the pole-tip field, a the pole-tip radius of the quadrupole magnet, and p the particle momentum.

The *betatron tune* is defined as the number of betatron oscillations executed per turn. If the betatron tune is near an integer, a particle trajectory will sample a local perturbation on every turn at the same phase of oscillation, and its amplitude may grow until the particle is lost to the chamber wall. Therefore the tune should not be exactly equal to an integer. Similarly, deflections experienced by higher-order fields, *e.g.*, fields with transverse sextupole or octupole symmetry, will accumulate over many turns whenever the horizontal and vertical tunes fulfill the resonance condition

$$kQ_x + mQ_y = p \tag{8}$$

where k, m, and p are integers. In a collider, the largest perturbations of the particle motion usually are the fields of the opposite beam, which 'excite' resonances. The lower the order of a resonance the stronger is its effect. In the CERN S$\mathrm{p\bar{p}}$S collider all resonances of order $(|k| + |m|) \leq 12$ had to be avoided, in order to obtain a good lifetime.

A further complication arises, since the different particles in the beam oscillate at slightly different tunes. The tunes of all particles have to be kept away from the low-order resonances. The beam-beam collision itself, for example, generates such *tune spread*.

Figure 5 shows that the betatron tune Q_β grows with the square root of the circumference, implying a similar scaling for the cell length and the arc beta function.[10] For a constant normalized emittance, the transverse beam sizes in the arc then decrease weakly with beam energy as $\sigma_{x,y\ \mathrm{arc}} \propto 1/E^{1/4}$.

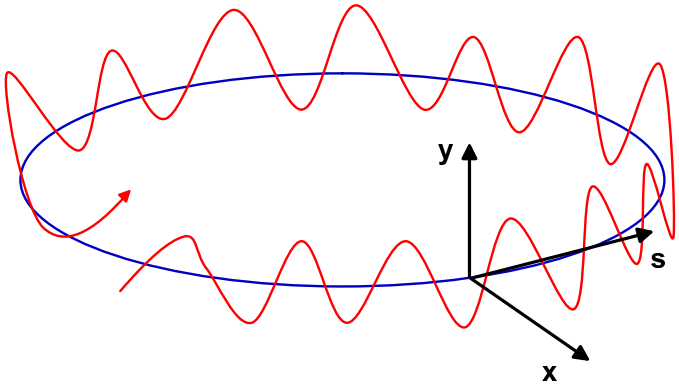

Fig. 4. Schematic of a betatron oscillation in a storage ring. The betatron tune $Q_{x,y}$ is equal to the number of transverse oscillation periods per revolution.

At this occasion, we may recall that the *geometric emittance* refers to the phase space area of the beam distribution, namely

$$\epsilon = \oint x' \, dx \, / \pi \tag{9}$$

where $x'(x)$ is the phase space trajectory of a particle at a transverse amplitude of 1 σ and $x' \equiv dx/ds$ is the slope of the physical trajectory, which here serves as the canonical momentum, and that the *beta function* $\beta_x(s)$ determines the local rms beam size via

$$\sigma_{x,y}(s) = \sqrt{\beta_{x,y}(s)\epsilon_{x,y}}. \tag{10}$$

2 The Large Hadron Collider (LHC)

With 14 TeV centre-of-mass energy, the Large Hadron Collider (LHC) now under construction at CERN will be the highest-energy collider ever built.

In the following, I describe the accelerator physics challenges which are faced by the LHC project. Starting with the choice of machine parameters, and then addressing the issues of superconducting magnets, commissioning schedule, accelerator layout and optics, I proceed to the effects of head-on and long-range beam-beam collisions, and their impact on luminosity and potential loss of Landau damping. Next, I discuss the dynamic aperture, *i.e.*, the particle-orbit stability, at injection, and give a few examples for the ongoing experimental tests of novel beam diagnostics and analysis. I then briefly mention several technical developments, such as power converters, vacuum system,

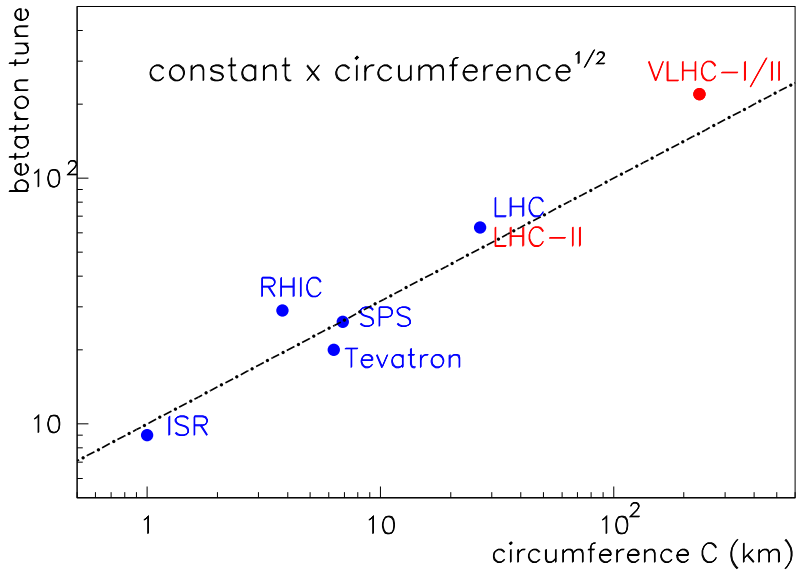

Fig. 5. Betatron tune as a function of the ring circumference.

machine protection and beam dump, including the heat load inside the cold magnets and the requirements for the LHC beam collimation. This is followed by an overview of the LHC injectors and pre-injectors, and the beams they can provide, as well as a brief discussion of luminosity limitations for heavy-ion collisions. Finally, I will describe a new phenomenon that may determine the LHC commissioning strategy and also constrain the ultimate beam parameters, that is the *electron cloud*. This refers to a rapid accumulation of electrons inside the beam pipe during the passage of a bunch train and its consequences.

For more detailed informations on accelerator physics at the LHC, the reader may consult the LHC project web page,[11] the proceedings of the workshops Chamonix X and Chamonix XI,[12] and the web page of the accelerator physics group in the CERN SL Division.[13]

2.1 LHC Parameter Choice

The circumference of the existing LEP tunnel (26.7 km) and the highest possible magnetic field confine the maximum beam energy according to

$$E\ [\mathrm{TeV}] \approx 0.84\ B\ [\mathrm{T}]. \tag{11}$$

For a nominal field B of 8.4 T this yields a beam energy of 7 TeV.

The beam-beam collision induces a betatron-tune spread, whose size is characterized by the beam-beam tune shift parameter ξ. The latter is proportional to the ratio of bunch population N_b and emittance ϵ_x, *i.e.*, $\xi \propto N_b/\epsilon_x$. The maximum tolerable value for the emittance is imposed by the aperture of the magnets, especially at injection.[14] Hence the number of N_b is limited, to about $N_b \approx 1.1 \times 10^{11}$, in the nominal LHC parameter table.

The desired LHC luminosity is $L \approx 10^{34}$ cm^{-2}s^{-1}. Since β_x^*, the beta function at the collision point, cannot be reduced arbitrarily (in particular is should remain larger than the bunch length), and since $\kappa = 1$, the only free parameter in Eq. (2) is the number of bunches n_b. This is chosen as 2808 to match the LHC luminosity target value. The high number of bunches implies a large average beam current, and a high synchrotron radiation power, which has to be absorbed inside the cold magnets.

2.2 Superconducting Magnets

Table 4 shows that the LHC dipoles represent a significant step forward in magnet technology. This is in line with the scaling of Fig. 2. In order to arrive at a compact and cost-efficient design, the LHC magnets are of a new 2-in-1 type where both beam pipes are placed inside the same support structure and cryostat.

Table 4. Dipole magnetic fields in varius hadron colliders. For the Superconducting Super Collider (SSC) only magnet prototypes were built.

accelerator	dipole field
SPS	1.8 T
Tevatron	4 T
HERA	5 T
SSC	6 T
LHC	8.4 T

The heartpiece of the magnets is a superconducting cable, called Rutherford cable, which can support a high current density of 400 A/mm^2, in case of the LHC, to be compared with current densities of order 1 A/mm^2 for normal conductors.[15] The cable itself is made from about 20 strands, each of which consists of hundreds of NbTi

filament islands embedded in a copper matrix. The cable is arranged around the beam pipe in a geometry which produces the desired field shape without introducing large errors and nonlinearities. For example, a $\cos\theta$ arrangement yields a pure dipole field. The cable is surrounded by an iron yoke placed inside a non-magnetic collar. Several layers of superinsulation and a vacuum vessel form the outer shell. The first pre-series magnets were delivered to CERN by industry, and have exceeded the nominal field.

2.3 Commissioning Schedule

According to the commissioning schedule as of summer 2001 a complete octant of the LHC will be cooled down and tested in 2004. The last dipole magnet is due to be delivered in March 2005. First beam is foreseen in February 2006, and a 1-month pilot run in April 2006. The first full physics run should start in the fall of the same year. Already for 2007 a few weeks of lead ion collisions are planned.

2.4 Layout and Optics

Figure 6 illustrates the overall layout of the LHC. There are 8 long straight sections. The two largest experiments, CMS and ATLAS, are located in the North and South straight sections, called interaction point 5 (IP5) and 1 (IP1), respectively. The two straight sections adjacent to ATLAS accommodate the experiments LHC-B (IP8) and the ion experiment ALICE (IP2). They simultaneously serve for beam injection. Two of the remaining straight sections are devoted to beam cleaning, another houses the rf, and the last one is needed for beam extraction and dump. The two beams pass alternately thought the inner and outer beam pipe, interchanging their locations in the 4 experimental IPs. Each beam travels for half of the circumference on the outer and the other half on the inner side, such that the revolution times are identical and the beams remain synchronized.

Development of the LHC optics has been a challenging task, as the length of the straight sections was pre-defined by the geometry of the LEP tunnel. In addition, due to a large number of magnets common to both rings, new optics tools had to be developed which allow for a simultaneous ‘matching’ of both rings.

As an illustration of the final achievement, Fig. 7 shows the beta functions $\beta_{x,y}$ and the horizontal dispersion D_x as a function of longitudinal position for beam no. 1 in IP5. The optics in IP1 is basically identical. The optics for beam no. 2 is always the mirror image of that for beam no. 1. The minimum beta functions of $\beta_{x,y} = 0.5$ m

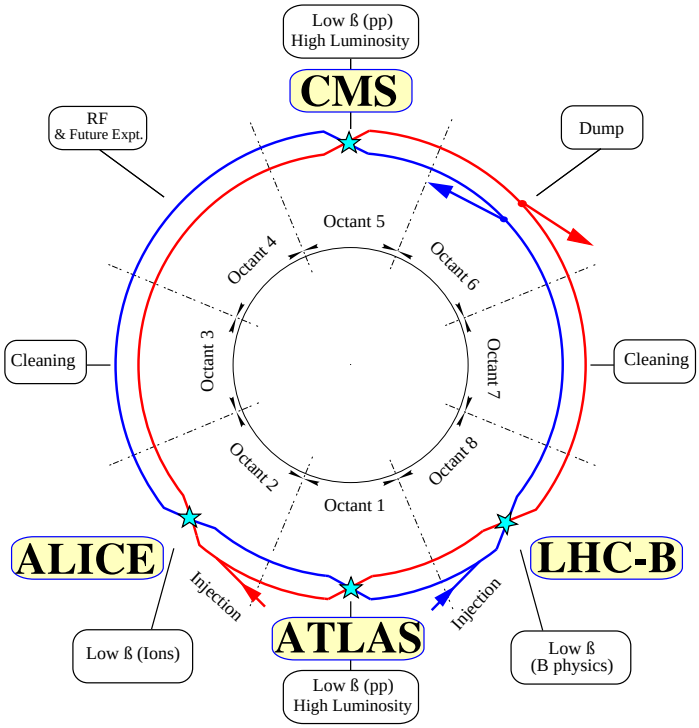

Fig. 6. LHC layout.

are assumed at the collision point (the center of the picture). The dispersion function D_x, which describes the horizontal orbit offset x for a relative momentum error $\Delta p/p$ via the relation $x = D_x(\Delta p/p)$, is almost zero around the collision point. It takes on noticable values only at the entrance to the arcs, on either side of the picture.

Figure 8 displays the orbit in the interaction region. The orbit is not flat, because the bunches collide with an angle, in order to separate them as quickly as possible before and after the main collision point. Otherwise, unwanted collisions with earlier or later bunches of the opposing beam would equally contribute to the beam-beam tune shift and tune spread, and possibly to the background, but not to the luminosity. The nominal full crossing angle is 300 μrad. The orbit of each beam must provide half this angle, as indicated.

Figure 9 shows a top view of magnets around the ATLAS detector (IP1). The collision point is at the center. The beams are focused by superconducting quadrupole triplets, consisting of the three quadrupoles Q1, Q2 and Q3. The free distance between the exit face of the last quadrupole and the collision point is about 23 m. Outside the triplet, a dipole magnet D1 separates the two beams, so that they are guided into the two

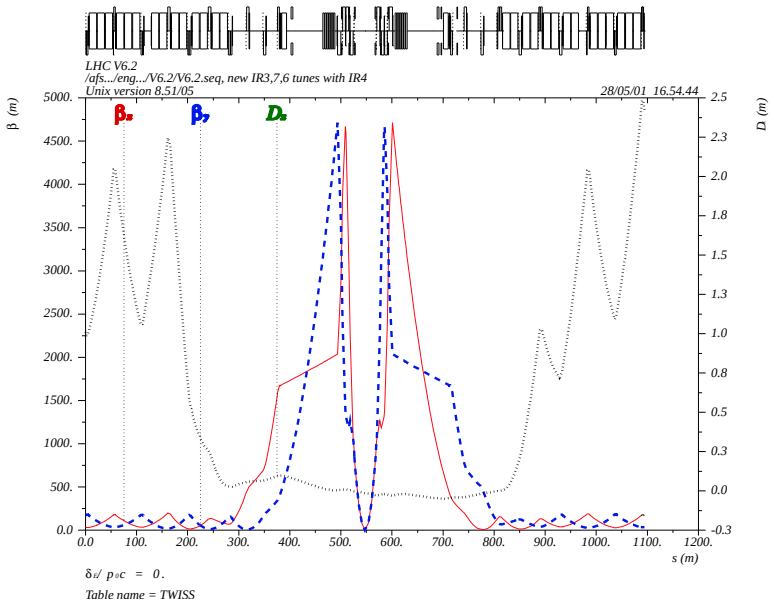

Fig. 7. Collision lattice for beam 1 at IP5. Both beta functions $\beta_{x,y}$ and horizontal dispersion are shown. (Courtesy A. Faus-Golfe, 2001)

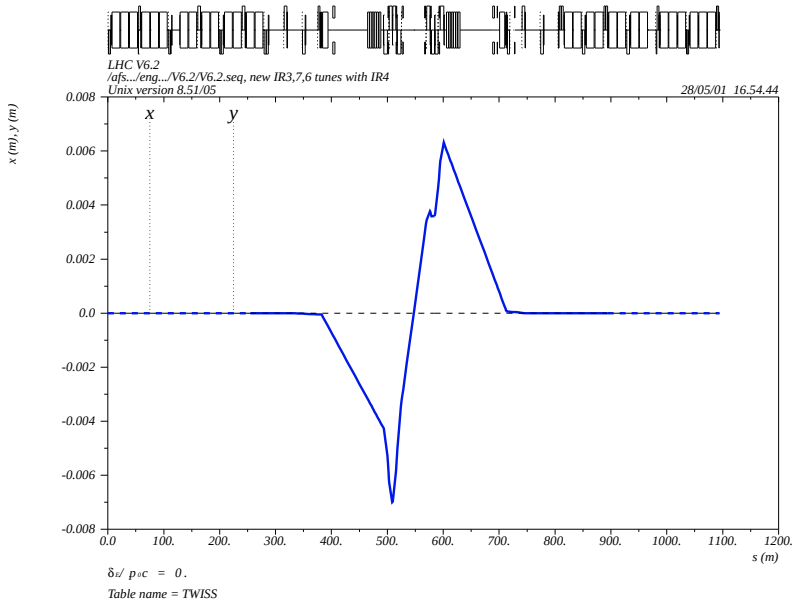

Fig. 8. LHC design orbit for beam 1 near IP5 (CMS) in collision. (Courtesy A. Faus-Golfe, 2001)

beam-pipe channels of the arc magnets. A second dipole D2 further outwards, reverses the deflection imparted by D1, such that the beams are again perfectly aligned in the direction of the arc magnets.

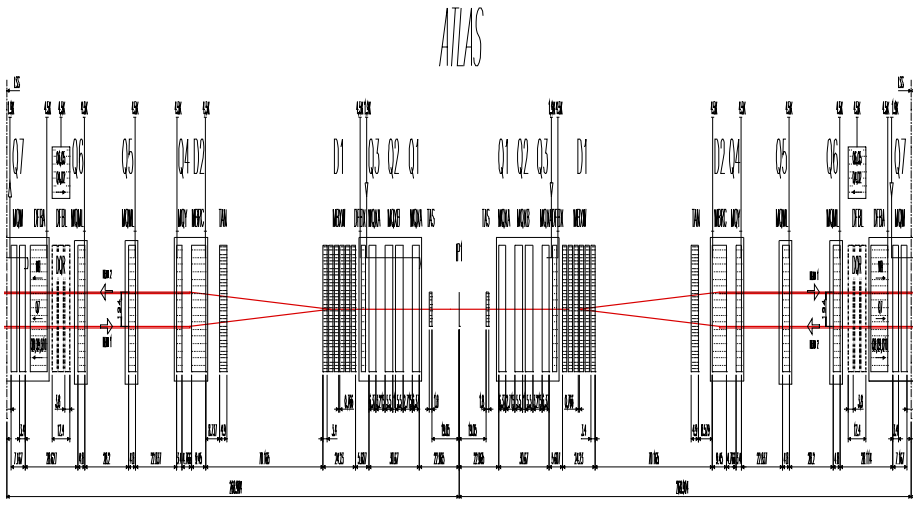

Fig. 9. Magnet layout (top view) around IP 1 (ATLAS). (Courtesy A. Faus-Golfe, 2001)

2.5 Head-On Beam-Beam Collision

In the main collision point, the repelling force of the opposing beam acts like a defocusing lens, as is illustrated in Fig. 10. The defocusing force decreases the betatron tune of all particles. However, for large amplitudes the beam fields decrease inversely with the transverse distance, so that particles at large amplitudes experience a smaller effect than particles near the center of the other beam. The nonlinearity of the beam lens thus induces a tune spread. The maximum acceptable tune spread gives rise to the so-called *beam-beam limit*.

The tune shift and maximum tune spread $\Delta Q_{x,y}$ induced by the collision with the opposing beam is characterized by the beam-beam tune shift parameter:

$$\xi_{x,y} \equiv \Delta Q_{x,y} = -\frac{r_p \beta^*_{x,y} N_b}{2\pi\gamma\sigma^*_{x,y}(\sigma^*_x + \sigma^*_y)}. \tag{12}$$

Note that the horizontal effect of the other beam is similar to that of a single defocusing quadrupole of integrated strength $kl_{\rm quad}$, and that the latter would produce a (horizontal) tune shift

$$\Delta Q_{x,\rm quad} \approx \frac{1}{4\pi}\beta_{x,q}\; kl_{\rm quad}, \tag{13}$$

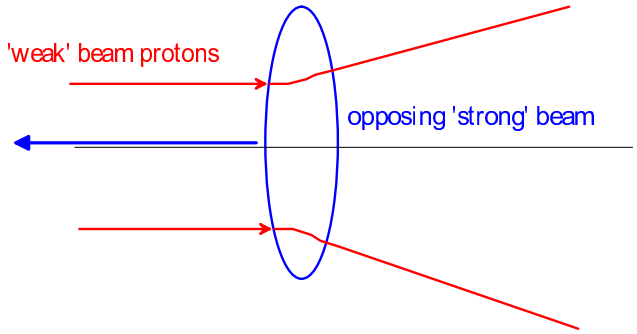

Fig. 10. Schematic of head-on beam-beam collision.

Table 5. Comparison of single-IP and total beam-beam tune shifts for selected hadron colliders.

	SPS	TeV-IIa	LHC
ξ/IP	0.005	0.01	0.0034
$\xi_{\rm tot}$	0.015	0.02	0.009

where k was defined in Eq. (7), $l_{\rm quad}$ is the quadrupole length, and $\beta_{x,q}$ the beta function at the quadrupole. Indeed, Eq. (12) for ξ_x can be derived from Eq. (13), if one replaces $kl_{\rm quad}$ by $-\Delta x'/x$ where $\Delta x'$ denotes the kick imparted by the opposing beam to a particle with a small horizontal offset x.

Table 5 compares the beam-beam tune shift for the LHC with the tune shifts achieved at the SPS and the Tevatron. Both the tune shift per collision point and the total tune shift (adding contributions from all interaction points) are listed. The table demonstrates that either number is smaller for the LHC than what has already been reached elsewhere. In this regard, the LHC parameters appear rather conservative.

The head-on beam-beam tune shift for one IP, given in Eq. (12), can be rewritten as

$$\xi_{x,y} = \frac{r_p N_b}{2\pi\epsilon_{x,N}(1+\kappa)}, \tag{14}$$

where $\kappa = \sigma_y/\sigma_x$ denotes the aspect ratio. Assuming that $\beta_y^*/\beta_x^* = \epsilon_y/\epsilon_x = \kappa$, so that the beam-beam tune shift is of the same value in both planes, $\xi \equiv \xi_x = \xi_y$, we can reexpress the luminosity of Eq. (2) as

$$L = (f_{\rm rev} n_b N_b)\,\frac{1+\kappa}{\beta_y^*}\,\gamma\,\frac{\xi}{2r_p} \tag{15}$$

This demonstrates that there are only four factors which can be optimized for high luminosity: (1) the emittance ratio κ, (2) the IP beta function $\beta_y^* = \kappa\beta_x^*$, (3) the maximum beam-beam tune shift ξ, and (4) the total beam current $(f_{\rm rev} n_b N_b)$.

For flat beams $\kappa \ll 1$ and one finds that the luminosity is half that of the round-beam case, $L_{\rm flat} \approx L_{\rm round}/2$, unless β_y^* can be reduced, which seems more difficult for $\rm pp$ than for $\rm p\bar{p}$ colliders.[16]

2.6 Long-Range Beam-Beam Collisions

Both on the incoming and outgoing side of the IP, each bunch encounters several bunches of the opposing beam, which are transversely displaced due to the crossing angle. The perturbation from these long-range encounters further increases the tune spread and can destabilize particles oscillating at amplitudes of a few σ, *i.e.*, particles which come closer to the other beam during their betatron motion.

Each bunch experiences up to 15 long-range collisions on either side of each head-on interaction point. Bunches with a smaller number of long-range encounters at the head and tail of a bunch train will likely have a poor lifetime. These bunches are therefore called PACMAN bunches,[17] alluding to the computer game of the same name.

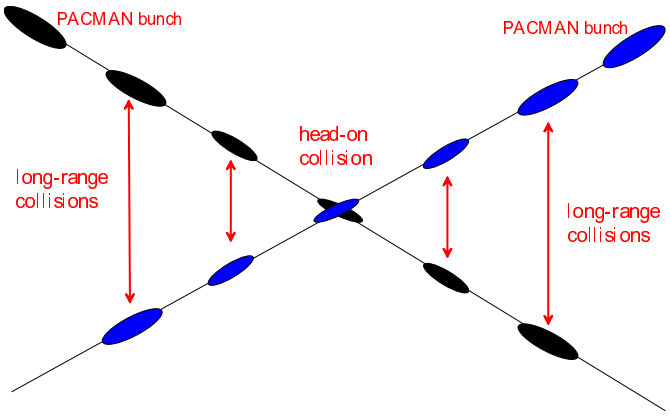

Fig. 11. Schematic of long-range collisions on either side of the main interaction point.

The linear tune shift introduced by the long-range collisions exactly cancels if half of the beam-beam crossings are in the vertical and the other half in the horizontal plane. For this reason the LHC beams will be crossed horizontally at two IPs and vertically at the other two.

However, the higher-order effects of the long-range collisions do not cancel, but

instead can cause a strong diffusion at larger betatron amplitudes. Indeed, the LHC will enter a new regime of the beam-beam interaction, where the long-range encounters on either side of the interaction point may be the dominant perturbation, rather than the head-on collisions as in the past colliders.

These long-range collisions give rise to a well defined *diffusive aperture*.[18,19] This diffusive aperture, $x_{\rm da}$, is smaller than the beam-centroid separation at the long-range collision points, $x_{\rm sep}$, by an amount Δ. In other words, we can write

$$x_{\rm da} = x_{\rm sep} - \Delta, \tag{16}$$

where[19,20]

$$\frac{\Delta}{\sigma} \propto \sqrt{\frac{N_b}{\epsilon_N}}. \tag{17}$$

In particular, if quoted in units of the rms beam size σ, the diffusive aperture is independent of the IP beta function and the beam energy.[20] For the nominal LHC parameters, the beams are separated by $x_{\rm sep} \approx 9.5\sigma$ and the diffusive aperture may be as low as $x_{\rm da} \approx 6\sigma$.[19,20]

Figures 12 illustrates the head-on *tune footprint*, as well as the additional tune spreads due to the long-range effects at LHC IP 1 and 5, respectively. These tune footprints show the tunes for particles with transverse amplitudes extending between 0 and 6 times the rms beam size (6 σ). The figure confirms that the alternating crossings in IP1 and IP5 results in a partial cancellation of the long-range tune shifts. Figure 13 compares the total LHC tune spread, due to all 4 collision points, for a nominal bunch and for a PACMAN bunch, *i.e.*, for a bunch which only encounters half of the nominal number of long-range collisions. The total tune spread of the entire LHC beam, including the PACMAN bunches, must fit between harmful resonances in the tune plane. This requirement will limit the maximum achievable tune shift parameter ξ and thus the bunch intensity N_b.

Figure 14 displays further tune footprints, this time extending up to 10 σ, and calculated with and without long-range collisions, head-on collisions, or field errors in the final quadrupoles. The figure demonstrates that for amplitudes larger than a few σ the effect of the long-range collisions is dominant.

Of more immediate concern than the tune spread is the diffusion rate of particles. In unstable (chaotic) regions of phase space, the particle amplitude increases randomly until the particle is lost. Approximately one can describe this behavior as a diffusion in the action variables I_x and I_y, the latter being defined as the square of the horizontal or

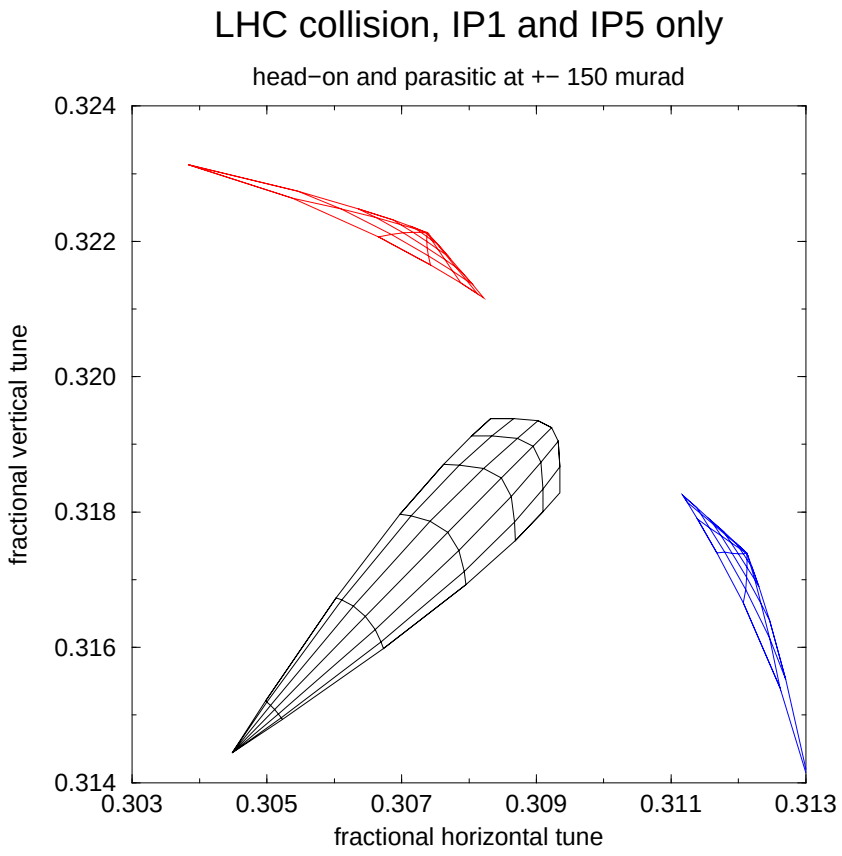

Fig. 12. Tune footprints due to head-on and long-range beam-beam effects in LHC IPs 1 and 5. Vertical axis refers to the vertical tune, horizontal axis to the horizontal. (Courtesy H. Grote)

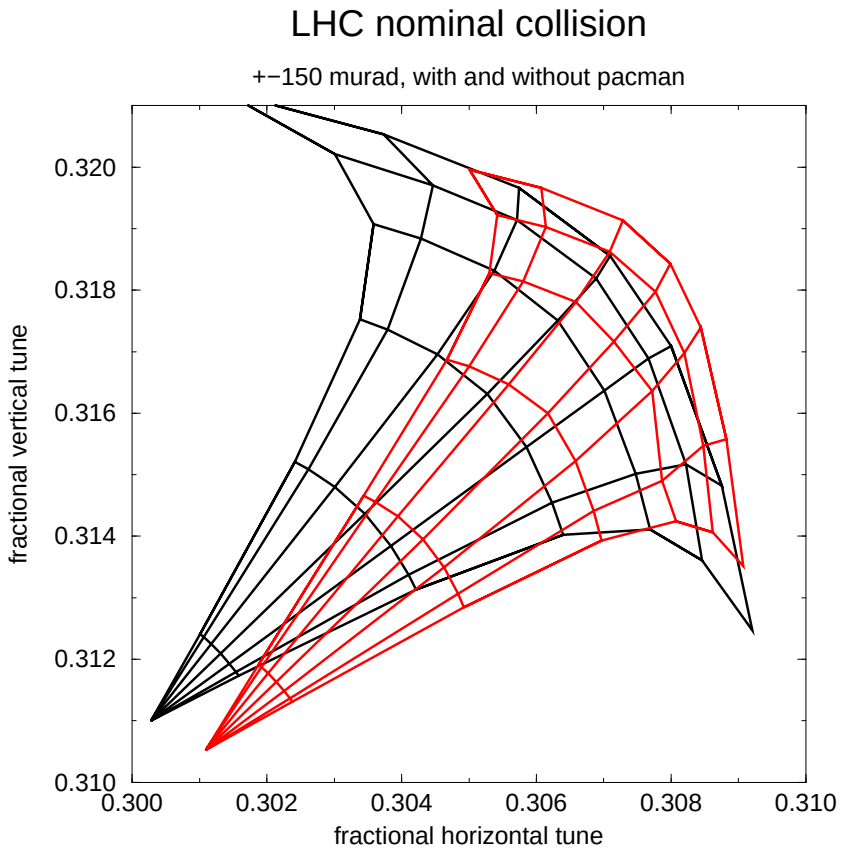

Fig. 13. Total tune footprints in the LHC for a regular bunch and for a PACMAN bunch. (Courtesy H. Grote)

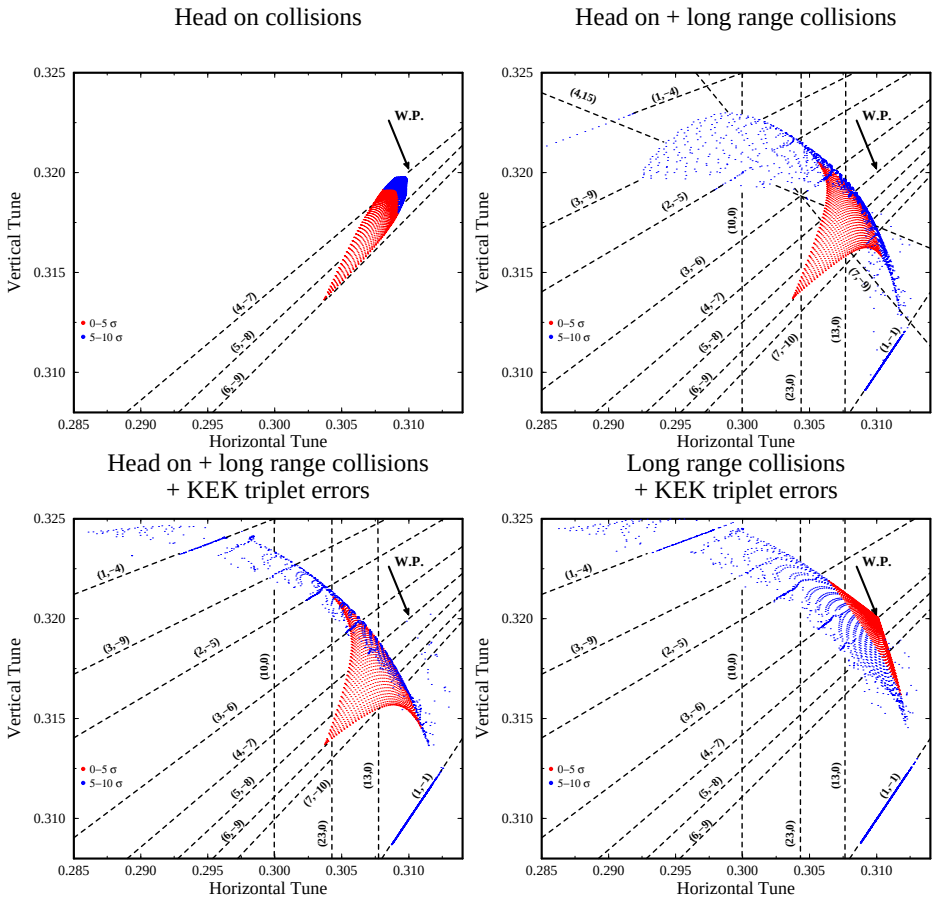

Fig. 14. LHC tune footprints with head-on & long-range collisions and triplet errors.[19] Red dots: x, y_{in} up to $5\sigma_{x,y}$; blue dots: x, y_{in} up to 10 $\sigma_{x,y}$. Top left: head-on collisions only; top right: head-on and long-range collisions; bottom left: head-on plus long-range collisions and triplet (magnet) errors; bottom right: long-range collisions and triplet errors, but no head-on collision.

vertical oscillation amplitude divided by $(2\beta_{x,y})$. In the simulation, the diffusion can be computed by calculating the change in the action variance of a group of particles per unit time.[19] An example is displayed in Fig. 15 for various conditions. Note that the vertical axis has a logarithmic scale. Whenever the long-range collisions are included, the diffusion increases by many orders of magnitude at amplitudes larger than about 6σ. We call this threshold aperture the diffusive aperture. It is due to the long-range conditions. Outside of the diffusive aperture particles will be lost within a few seconds.

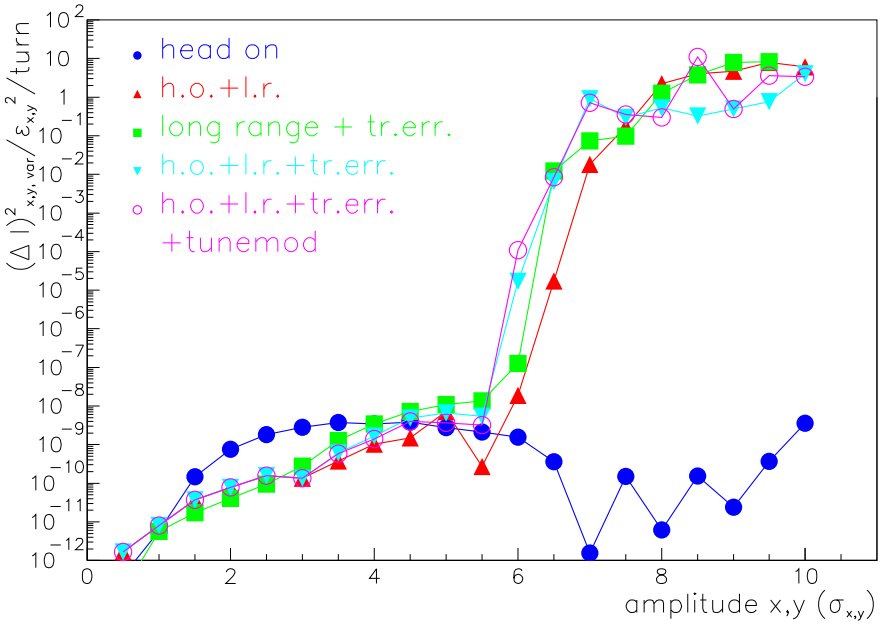

Fig. 15. Change of action variance per turn as a function of starting amplitude in units of the rms beam size, for the LHC.[19] Compared are different combinations of head-on collisions, long-range collisions, triplet-field errors, tune modulation, and even a hypothetical 'Moebius twist', where the horizontal and vertical particle coordinates are exchanged on each turn.

Figure 15 presents further simulation results, illustrating the variation of the diffusive aperture with the bunch charge. The right picture summarizes the data on the left-hand side. The simulation confirms that Δ varies with the square root of the bunch population, consistent with Eq. (17). This scaling behavior was first noted by J. Irwin.[18]

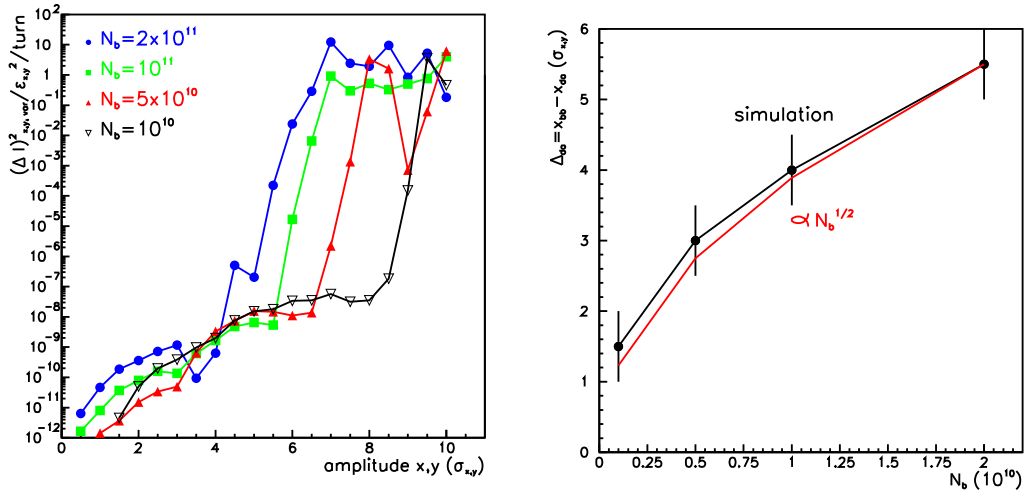

Fig. 16. Dependence of diffusion due to long-range collisions on the beam current.[19] Left: change of action variance per turn vs. bunch population; right: approximate diffusive aperture vs. bunch population; vertical axis describes the distance to the other beam at the parasitic collision points in units of the rms beam size; a square root dependence is also indicated for comparison (dashed line).

2.7 Minimum β^*

A first limit on the IP beta function arises from the hourglass effect. In order to avoid luminosity loss, the IP beta function should be larger than the rms bunch length

$$\beta^*_{x,y} \geq \sigma_z, \tag{18}$$

since on either side of the IP the beta function increases as

$$\beta_{x,y}(s) = \beta^*_{x,y} + \frac{s^2}{\beta^*_{x,y}}, \tag{19}$$

where s denotes the distance to the IP.

A second limit is set by the long-range collisions. As we have just seen, the dynamic aperture caused by parasitic collisions is $x_{\rm da} \approx (n_{\rm sep}\sigma - \Delta)$ where $n_{\rm sep}$ is the separation in units of the beam size. For the LHC the separation is chosen as

$$n_{\rm sep} \geq 9.5, \tag{20}$$

and the simulations indicate that $\Delta \approx 3\sigma$.

If we want to limit the luminosity loss due to the crossing angle, we must demand

$$\theta_c \equiv n_{\rm sep}\sqrt{\frac{\epsilon_{x,y}}{\beta^*_{x,y}}} < 2\,\frac{\sigma_x}{\sigma_z}. \tag{21}$$

Combining Eqs. (20) and (21), we find that

$$\beta^*_{x,y} \geq \frac{n_{\rm sep}\sigma_z}{2} \approx 5\sigma_z, \tag{22}$$

which for the LHC yields $\beta^*_{x,y} > 0.38$ m to be compared with a design value 0.5 m. However, this may not be the full story. Ongoing studies suggest that, if one also includes the constraints from the head-on beam-beam tune shift, it might actually prove advantageous to operate with a crossing angle and an rms bunch length exceeding the limits of Eq. (21) and accept a loss in geometric luminosity, in exchange for a decreased beam-beam tune shift ξ.[21]

Two schemes are presently being explored for compensating the effects of the beam-beam collision. The field of a pulsed electric wire is similar to the beam field experienced at a long-range collision point, and such wire can, therefore, be used to exactly compensate the effect of the long-range encounters. This scheme was proposed by J.-P. Koutchouk.[22] Simulations confirm that a compensating wire is highly effective. An example result is shown in Fig. 17, where the field of the wire increases the diffusive aperture by about 2 σ, even if the betatron phase at the wire location differs by a few degrees from that at the long-range collision points.

A complementary approach is the electon lens built and tested at Fermilab.[23] This lens consists of a low-energy electron beam, which is collided with the antiproton beam. If beta functions and electron current are correctly adjusted, the focusing field of the electrons compensates the focusing force experienced by the antiprotons in the two proton-antiproton collision points. In order to obtain a controllable compensation in both transverse planes, two lenses at locations with different beta functions are needed. If the electron current is modulated, the central tune shift of each bunch can be controlled independently, thereby avoiding PACMAN bunches. Transverse shaping of the electron beam profile should even allow reducing the beam-beam tune spread inside the bunch. The interaction takes place in a strong longitudinal solenoid, in order to suppress transverse two-stream instabilities, which otherwise might develop. During a first beam test in the spring of 2001, the electron lens successfully changed the tune of the Tevatron proton beam by about 0.005, in accordance with the prediction.

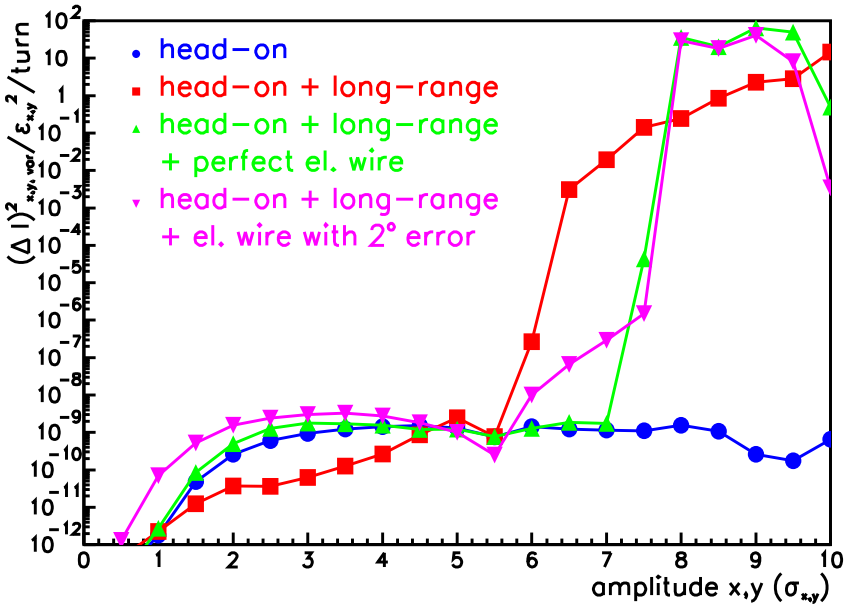

Fig. 17. The diffusion in action variance per turn as a function of the start amplitude, illustrating the effect of an electric wire which mimics long-range encounters of opposite charge.

2.8 Strong-Strong Beam-Beam Effects at LHC

In addition to the effect of a strong opposing beam on a single particle in the other beam, which we have considered above, there also exist strong-strong beam-beam effects, *i.e.*, effects where a collective motion develops due to the coherent interaction of the two beams.

In the case of two colliding bunches two coherent modes are observed: the σ or 0 mode, for which the oscillations of the two bunches are in phase, and the π mode, where the bunches oscillate in counter-phase. These two modes are illustrated for a coupled pendulum in Fig. 18. The oscillation frequency of the σ mode is equal to the unperturbed betatron tune, whereas the frequency of the π mode is shifted downwards (in LHC) by an amount $\Delta Q = Y\xi$. The amount of the downward tune shift is proportional to the tune-shift parameter ξ. The coefficient Y, of the order 1.2–1.3, is sometimes called the Yokoya factor or the Meller-Siemann-Yokoya factor.[24,25]

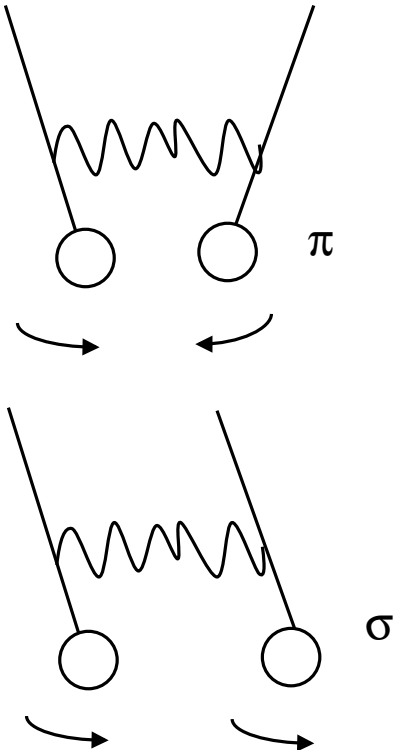

Fig. 18. Simple Model of π and σ modes for a system of two coupled oscillators.

For the following we need to introduce the notion of *Landau damping*. This refers to the phenomenon that a spread of oscillation frequencies of individual particles tends

to stabilize the coherent beam motion of the particle ensemble against excitation frequencies within the frequency spread. An illustration employing three swings with either equal or different frequencies on the same support is shown in Fig. 19.

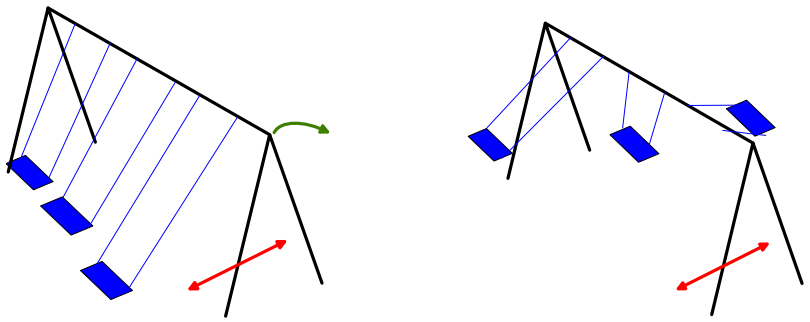

Fig. 19. Schematic of Landau damping, from A. Hofmann.[26]

Mathematically, the driven particle motion is described by

$$\ddot{x} + \omega^2 x = Ae^{-i\Omega t}. \tag{23}$$

If the eigenfrequencies ω of many particles are distributed according to a density $\rho(\omega)$, the centroid response of the particle ensemble to the external perturbation $A\exp(-i\Omega t)$ is[27]

$$< x >= \frac{A}{2\bar{\omega}} e^{-i\Omega t} \int d\omega \, \frac{\rho(\omega)}{\omega - \Omega - i\epsilon} \tag{24}$$

where $\epsilon \rightarrow 0^+$.

For LHC worriesome is a prediction by Y. Alexahin,[28] according to which the coherent π mode in the LHC will not be Landau damped. His argument is that, for bunch intensities of the two beams which are equal to within 40%, the frequency shift of the coherent π mode is larger than the incoherent beam-beam tune spread ξ. A possible reason why this loss of Landau damping was not observed in the SPS or Tevatron is that the antiproton intensities in these machines were always much smaller than the proton intensities, as can be seen in Table 6.

Table 6. Comparison of bunch intensity ratios in SPS, TeV-II and LHC.

	SPS	TeV-II	LHC
intensity ratio N_1/N_2	2	$9 \rightarrow 2$	1

The original argument[28] applied to the head-on collision only. It was speculated that the long-range collisions may act either stabilizing or de-stabilizing. Extensive simulation studies by M. Zorzano[29,30] support Y. Alexahin's predictions, and do show the loss of Landau damping. An example simulation result is shown in Fig. 20. These simulations also indicate that the long-range collisions will not stabilize the π mode. Further analytical work by Y. Alexahin has since confirmed this conclusion.

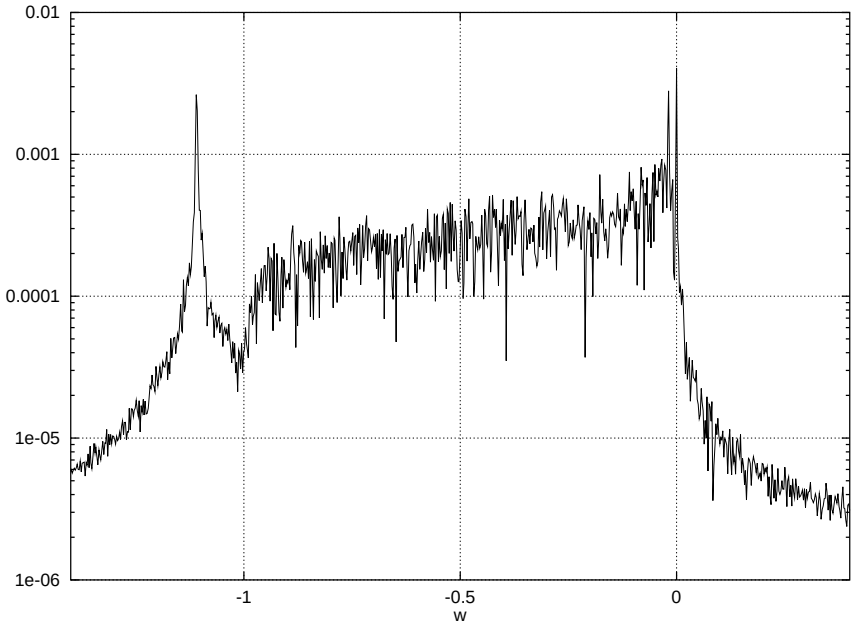

Fig. 20. Simulation of coherent modes (M. Zorzano): frequency spectrum of the bunch centroid motion; vertical axis is on a logarithmic scale with arbitrary units, plotted along the horizontal axis is the normalized frequency $w = (\nu - Q)/\xi$. The π- and σ-oscillation modes are clearly visible.[29,30]

A possible cure suggested by A. Hofmann is to separate the tunes in the two rings. Simulations for separated tunes suggest that Landau damping may be restored, provided that the betatron tune split between the two rings is larger than the beam-beam tune shift. However, at most or all of the alternative asymmetric working points higher-order coherent resonances may be encountered.[30]

Both theory and simulations rely on various approximations and assumptions. Experimental studies of the π mode stability have been performed in LEP, where the π mode was clearly observed, and are planned at RHIC.

2.9 Single-Bunch Collective Effects

There are a number of single-bunch collective effects. They all are driven by the impedance of the vacuum chamber, *i.e.*, by electro-magnetic fields excited by the beam and acting back on it. Here we do not discuss these effects in detail, but merely mention the most important ones.

The coherent synchrotron tune shift with intensity (the synchrotron tune describes the longitudinal oscillation frequency and is defined in analogy to the betatron tune for the transverse plane) may cause a loss of Landau damping at high bunch intensity.[31,32] A possible countermeasure is the controlled blow up of the longitudinal emittance. For a longer bunch the synchrotron frequency spread increases due to the nonlinearity of the sinusoidal rf wave in the rf cavities.

The predicted threshold of the longitudinal microwave instability is far above the nominal LHC parameters. Similarly, the calculated threshold for transverse mode coupling at injection of $N_b \approx 5.9 \times 10^{11}$ is safely above the design current.

The transverse resistive wall instability is important, however. For the nominal LHC parameters the growth time of the lowest multi-bunch mode is $\tau \approx 30$ ms, which corresponds to 300 turns; for twice the number of bunches and the ultimate bunch population (1.7×10^{11}) it decreases to $\tau \approx 10$ ms or 100 turns.

The tune shift variation for a partially filled ring due to the ac magnetic field leakage and a finite resistive wall is a small effect, as shown by J. Gareyte.[33]

Incoherent tune shift due to collective fields was recognized as a potential problem for the VLHC.[34,35] It might also be noticable at the LHC. For the nominal LHC parameters at injection the incoherent tune shift is $\Delta Q_y \approx 0.02$; for higher intensity it may approach $\Delta Q_y \approx 0.07$. This could cause potential problems such as (1) a reduction of dynamic aperture, or (2) resonance crossings of the coherent multi-bunch modes.

2.10 Dynamic Aperture at Injection

Nonlinear field errors can destabilize particle motion after 1000s of turns. Error sources include persistent currents (eddy currents in the superconductor), the geometry of the superconducting coil, and the current redistribution during acceleration.

The maximum stable area in phase space is called the *dynamic aperture*. The approach that was taken to guarantee a sufficiently large dynamic aperture for the LHC consisted of three parts[36]: (1) computer simulations of the particle motion under the influence of nonlinear field errors were performed over 10^6 turns, (2) the computer

simulation were calibrated against measurements at the SPS and HERA, which showed that the simulation and measurements deviate at most by a factor of two, and (3) a 12σ dynamic aperture was required in the simulation, so as to assure that the actual aperture will be larger than 6σ.

2.11 Persistent Currents

The persistent currents decay during injection. This will cause a change in chromaticity Q' by some 300 units, due to a change in the sextupole fields generated by the persistent currents. Here, the chromaticity is defined as the change in betatron tune ΔQ per relative momentum error $\Delta p/p$. At the start of acceleration the eddy currents are rapidly reinduced, within 100 s, and the chromaticity accordingly changes back to its initial value. This is called the '*snap-back*'. A chromaticity of several hundred units would imply a tune spread of the order 1, clearly unacceptable. In order to maintain a good beam lifetime and large dynamic aperture, the chromaticity must be controlled to within about 5 units.

The strategy to cope with the decay and snap-back is twofold. First, it is important that the acceleration starts slowly and reproducibly. Precise digital controllers for the LHC main power converters have been designed and built to accomplish this goal,[37] and an optimized excitation curve has been computed.

Second, new diagnostics enabling a fast measurement of chromaticity for immediate correction was developed and has already been tested at the CERN SPS. This is discussed next.

2.12 Novel Diagnostics

The conventional way of measuring the chromaticity is to detect the tune variation with rf frequency. This technique is rather time consuming.

A new method invented for the LHC measures the change in the phase of the betatron oscillation at the head and tail of a bunch following a kick excitation,[38] as illustrated in Fig. 21. If the chromaticity is zero, the head and tail always oscillate in phase. If the chromaticity is nonzero, on the other hand, a phase shift builds up between head and tail due to the integrated energy difference between particles passing these two locations during their slow oscillations in the longitudinal phase space. The longitudinal oscillations are called synchrotron oscillations, and the associated tune is called the

synchrotron tune Q_s. The value of Q_s is much smaller than the betatron tunes. In the LHC at injection, it is 0.006.

In the new chromaticity measurement, the phase difference which emerges between head and tail is proportional to the chromaticity. It is maximum after half a synchrotron period, and decreases again to zero after a full period.

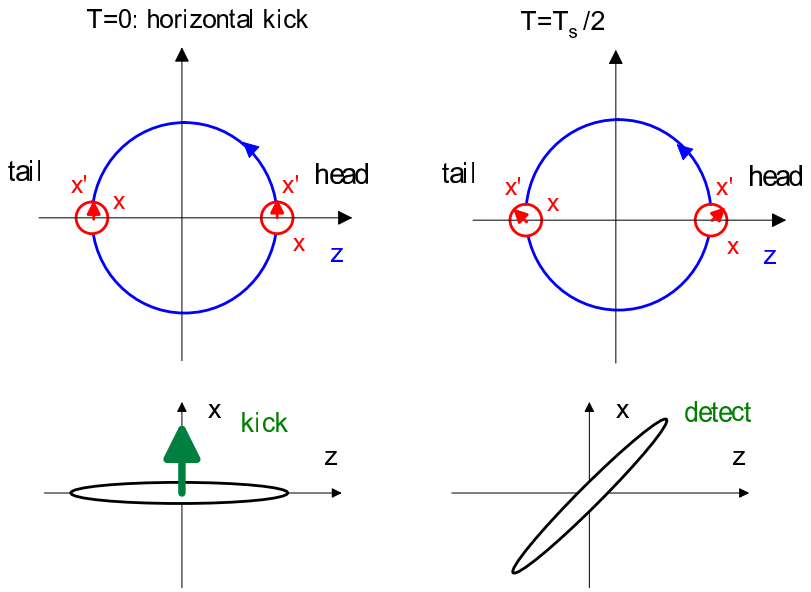

Fig. 21. Principle of chromaticity measurement via head-tail phase shift.[38]

The chromaticity inferred at turn n after the kick is

$$Q'_{x,y} = \frac{\eta\ \Delta\phi(n)}{\omega_0 \Delta\tau(\cos(2\pi n Q_s) - 1)}, \tag{25}$$

where $\Delta\phi(n)$ is the head-tail phase difference measured at the nth turn, $\Delta\tau$ the difference in arrival time between head and tail, and η the slippage factor, an optical parameter that can be calculated analytically ($\eta \equiv \alpha_c - 1/\gamma^2$ is defined as the relative change in revolution time per relative momentum change, and α_c is the so-called momentum compaction factor).

In principle, this technique might measure the chromaticity in about 10 ms, which is much shorter than the time scale of the snap back. A test measurement using a wideband pick up at the SPS is shown in Fig. 22.

Another diagnostics which has been developed in view of LHC is the processing of data from multi-turn beam-position monitors (BPMs) taken after deflecting a bunch to

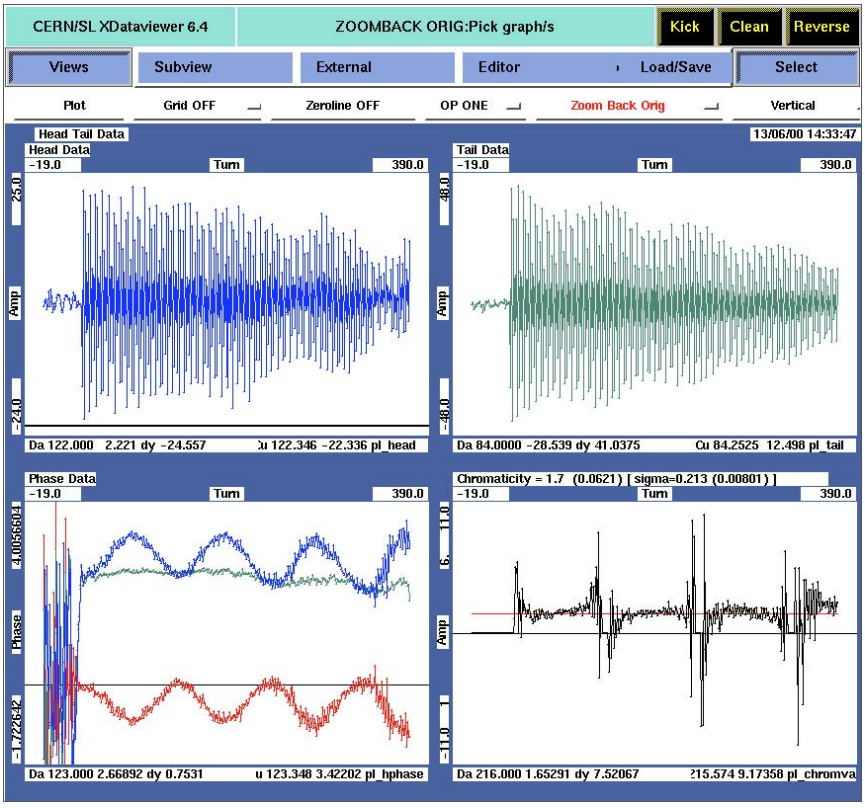

Fig. 22. Chromaticity measurement via head-tail phase shift in the SPS.[38] Top: raw oscillation data of bunch head and bunch center, bottom left: individual phases and phase difference $\Delta\phi$ (red), bottom right: inferred chromaticity. (Courtesy R. Jones, 2000)

a large amplitude, so as to extract informations about the nonlinear resonances and to localize nonlinear field errors all around the machine.[39]

The basic idea is to identify for each line in the position Fourier spectrum the corresponding higher-order resonance. From the relative height of such lines, their variation with the kick amplitude, and their change from one BPM to the next, informations can be obtained which may allow identifying problematic regions in the ring and minimization of the residual nonlinearities, thereby maximizing the dynamic aperture.

2.13 Power Converters

The LHC power converters were newly developed to meet the stringent demands on resolution, stability, and accuracy. The power converters for the main bending magnets and quadrupoles have demonstrated a resolution of 1 ppm.[37] Stability over a day is of the order of 5 ppm.[37]

2.14 Heat Load inside the Cold Magnets

Four primary sources of heat load have been identified, and require special remedies.

The first are lost beam particles. These can be particles which are scattered either off the other beam at the collision point or off residual gas nuclei. Another possibility are particles on unstable trajectories diffusing outwards.

In order to limit the rate of particle losses inside the cold magnets, halo collimation is performed in two straight sections which accommodate warm magnets. Complementarily, the cold magnets are cooled by superfluid helium at 1.9 K, which at this temperature has a remarkable heat capacity.

The second source of heat load is synchrotron radiation. For a bunch intensity of $N_b \approx 1.6 \times 10^{11}$, the synchrotron radiation amounts to about 0.27 W/m. This radiation does not directly shine onto the 1.9-K cold bore, but it is intercepted by a beam screen at a higher temperature varying between 4 and 20 K.

A third source are beam image currents in the resistive chamber wall. For the ultimate bunch intensity $N_b \approx 1.6 \times 10^{11}$, this contributes about 0.46 W/m. Also here the beam screen absorbs most of the heat. The screen is coated with a thin Cu layer to improve the surface conductivity.

A fourth source is the electron cloud, *i.e.*, electrons, generated by photoemission from synchrotron radiation or by secondary emission, which are accelerated in the field of the beam. The heat deposited on the walls by these electrons should not exceed

the residual cooling capacity, after accounting for the image-charge effects and direct synchrotron radiation. For $N_b \approx 1.6 \times 10^{11}$, the heat load due to the electron cloud must stay below 0.56 W/m.

2.15 Thermodynamic Considerations

Heat capacities C of the various magnet components have a strong influence on the quench limit. The heat capacity relates the temperature rise ΔT to the energy deposition ΔU per mass Δm via

$$\Delta T = \frac{1}{C}\frac{\Delta U}{\Delta m}. \tag{26}$$

For copper at 1.9 K, the heat capacity is only $C_{\mathrm{Cu}} \approx 0.03$ J/kg/K, which could easily give rise to premature quenches. To raise the quench limit, in the LHC magnets the s.c. cable is permeated by superfluid helium at 1.9 K, whose heat capacity is much higher than that of copper, *i.e.*, $C_{\mathrm{He}} \approx 4000$ J/kg/K.[36] With a measured helium content in the s.c. cable of $\sim 4.5\%$, the average heat capacity is significantly increased.[40] The helium absorbs deposited energy and transports it away from the magnet coils.

Another important point to recall is the refrigeration (Carnot) efficiency, which is given by

$$\eta = \frac{T_{\mathrm{cold}}}{T_{\mathrm{warm}}}\ , \tag{27}$$

and relates the optimum (minimum) power P_{warm} required at room temperature for absorbing a heat influx P_{cold} at a lower temperature:

$$P_{\mathrm{warm}} = \frac{P_{\mathrm{cold}}}{\eta} = \frac{T_{\mathrm{warm}}}{T_{\mathrm{cold}}} P_{\mathrm{cold}}. \tag{28}$$

It is evident that the absorption of heat at $T_{\mathrm{cold}} = 1.9$ K is not efficient. For this reason, a beam screen at higher temperature (4–20 K) is installed inside the magnets, which absorbs the proton synchrotron radiation power as well as the energy from the electron cloud. Two rows of pumping slots on either side — *i.e.*, horizontally outwards or inwards — of the beam-screen center connect the beam vacuum with the cold bore of the magnets, which is held at 1.9 K. This arrangement also enables a highly efficient cryopumping, where desorbed gas molecules diffuse through the pumping slots and then stick to the cold part of the magnet.

2.16 Quench Limits and Collimation

If too many protons are lost in a superconducting magnet, it will *quench*, which means it will become normal conducting. Then the machine protection system acts, and the beam will be dumped. Recovery from a quench is time consuming, and the number of quenches should therefore be minimized, ideally avoided. Taking into account the contributions to the heat capacity from the superfluid helium, the quench limit of an LHC magnet corresponds to a local temperature increase of 7 K at injection and 1 K at top energy.

A quench can be generated by local proton losses. Proton loss mechanisms include[41] (1) injection errors, where the losses occur within a few turns, (2) protons outside of the rf bucket which are lost at the start of the ramp in a 'flash', and (3) continuous losses in collision.

Table 7 compares the expected losses with the quench limit. In view of these numbers, a dedicated beam cleaning system is considered as indispensible for the LHC.

Table 7. Expected total losses and quench limit.[41]

process	exp. total losses	quench limit
injection	$\Delta N = 1.25 \times 10^{12}$	$\Delta N_q = 10^9$ m^{-1}
ramping	$\Delta N = 9 \times 10^{12}$	$\Delta N_q = 2.5 \times 10^{10}$ m^{-1}
collision	$\dot{N} = 3 \times 10^9$ s^{-1}	$\dot{N}_q = 6 \times 10^6$ m^{-1}s^{-1}

The chosen design is a 2-stage system, consisting of primary and secondary collimators.[41]

The primary collimation comprises 3 betatron collimators at an amplitude of 6σ and 1 energy collimator. Each of these is followed by a set of three secondary collimators at an amplitude of 7σ. The collimation inefficiency sensitively depends on the ring aperture $A_{\rm ring}$;

If $A_{\rm ring} = 8\sigma$, the efficiency is about $\eta_{\rm coll} \approx 10^{-4}$, which means that from 10^4 protons in the beam halo, all but one are intercepted by a collimator, before hitting the beam pipe.

At the LHC the collimation must be in the working position already at injection, and all through the acceleration. The tolerance on the dynamic closed orbit stability is rather stringent, namely < 30 μm ($1/10\sigma$), and must be met at all times. This condition assures that the secondary collimators are in the shadow of the primary collimators.

2.17 Machine Protection

The total energy stored in the LHC magnets is about 11 GJ, and the LHC beam energy is 0.7 GJ.[42] These amounts of energy, if liberated in an uncontrolled way, could cause a considerable damage to the machine components. Therefore, a reliable machine protection system is crucial.[42]

There are many aspects to the protection system. We mention only two.

In case of a magnet quench, the ensuing resistive heating further increases the temperature and the rapid heating could destroy the magnet. In order to avoid this, quench heaters will be fired, which induce additional quenches in the adjacent magnets and distribute the energy dissipation over a larger region. At the same time, switches are activated, so that the main current bypasses the region of the quench.

However, the heartpiece of the machine protection is the beam dump. Since the rise time of the extraction kickers is finite and long, an adequately long gap in the stored LHC beam is needed. The kickers can only be fired during this gap, since otherwise several bunches would be deflected by the rising edge of the kicker pulse to intermedate amplitudes without being extracted, and these bunches would damage the collimators or some magnets.

The protection philosophy is that whenever an error is detected, *e.g.*, the beam deviates too much from its nominal orbit, the beam is extracted from the ring and sent onto the dump, before it can destroy any machine components.

The design of the beam extraction system is itself not simple, since the beam density is so high that it can also destroy the beam dump. To prevent this, the extraction system comprises several dilution kickers which deflect the beam in both transverse planes and are activated at the same time as the extraction kickers.[43] Different bunches are deflectd by different amounts, such that the bunch impact point on the entrance face of the beam dump traces a nearly circular path over the length of the bunch train. The diameter of the sweep profile is about 15 cm, which provides for sufficient dilution of the beam density.[43]

The extraction kickers consist of many units. The most serious conceivable failure mode in the LHC is the accidental spontaneous firing of one of these kicker units. The protection system will then also fire all other kicker modules, in order to sent the beam to the dump. However, in this case the kick is not synchronized with the position of the beam gap, and component damage due to the impact of several bunches on collimators or septum cannot be excluded in the present design.

Table 8 compares the melting temperature, the maximum temperature rise $T_{\max}$ expected in case of a single bunch impact, as well as the front temperature rise T_{front} of the dump, if hit by the full LHC beam without dilution, for different candidate materials. The only material for which both $T_{\max}$ and T_{front} are smaller than the melting temperature is carbon. Thus, carbon has been selected as the LHC dump material.[43]

Table 8. Candidate materials for the LHC beam dump.[43]

material	T_{melt} [°C]	$T_{\max}$ [°C/bunch]	T_{front} [°C/beam]
Be	1280	75	3520
C	4500	320	3520
Al	660	360	3390
Ti	1670	1800	3250
Fe	1540	2300	3120
Cu	1080	4000	2980

2.18 LHC Filling Pattern

The filling pattern of bunches around the machine determines the time structure of events seen by the experiments. The nominal LHC bunch spacing is 25 ns, and the total revolution time is 88.924 μs. The 25-ns spacing is interrupted by various gaps, which are needed for injection and extraction between the different injector storage rings and the LHC itself. A gap of 111 missing bunches is required for extraction from the LHC, gaps of 30 or 31 missing bunches correspond to the rise time of the LHC injection kickers, and various gaps of 8 missing bunches are related to the injection into the SPS.

The final nominal bunch pattern is complicated due to all these gaps. It can be expressed in mathematical notation as[44]

$$\begin{aligned}
&(((72 \times b + 8 \times e) \times 3) &+ \quad &30 \times e) \times 2) \\
&+((72 \times b + 8 \times e) \times 4) &+ \quad &31 \times e) \times 3) \\
&+((72 \times b + 8 \times e) \times 3) &+ \quad &30 \times e) \times 3) \\
& & &+81 \times e),
\end{aligned}$$

where e refers to an empty place and b to a bunch.

Because of the many gaps different bunches in LHC experience different numbers of long-range collisions around the primary IPs. Even the number of head-on collisions in IP 2 and IP 8 is not the same for all the bunches. Indeed, less than half of the bunches are nominal ones, and all the others belong to one or another type of PACMAN bunch. This means that all these bunches will have different betatron tunes and different orbits.

According to the number and types of opposing bunches encountered, bunch equivalence classes can be defined.[44] Their number is almost comparable to the total number of bunches.[44]

Fortunately, careful analysis and simulations suggest that although different, the bunch orbits and tunes are still sufficiently similar that the lifetime and luminosity should not be much degraded.[45]

2.19 LHC Injectors

Before the beam is injected into the LHC it must be produced and accelerated in the injectors and pre-injectors.

These comprise, in order of decreasing energy, the Super Proton Synchrotron (SPS) the Proton Synchrotron (PS), and 4 PS Booster rings. In order to provide the high-quality beam demanded by the LHC a number of upgrades were necessary and new operational procedures and techniques of beam manipulation were introduced.

Historically, multiple bunches were generated in the PS by debunching (switching off the rf) and recapturing in a higher-harmonic rf system. A schematic of phase space evolution during slow debunching is shown in Fig. 23.

The problem with this scheme is that during the debunching process the microwave instability threshold is reached. Namely, while a bunch debunches, the density dN/ds and the local energy spread $\delta_{\rm rms}$ decrease by the same factor. The local instability threshold scales as

$$\left(\frac{dN}{ds}\right)_{\rm thr} \propto \delta_{\rm rms}^2, \tag{29}$$

which follows from the so-called Boussard criterion. Then, the beam becomes unstable as soon as its energy spread $\delta_{\rm rms}$ is small enough that this threshold condition is reached. The unwanted results are an unequal filling pattern and the non-reproducibility of the bunch intensities.

The new method developed for the LHC is a controlled bunch splitting without ever turning off the rf.[46] Instead, the relative amplitudes of various rf systems operating at different frequencies are varied as a function of time in such way that each bunch is

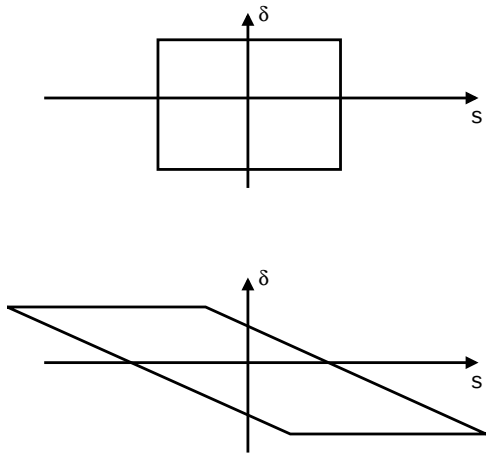

Fig. 23. Schematic of phase space evolution during slow debunching.

Table 9. Status of the PS for the LHC nominal beam.[47]

	achieved	nominal
protons per bunch	1.1×10^{11}	1.1×10^{10}
hor. emittance $\gamma\epsilon_x^{1\sigma}$ [μm]	2.5	3
vert. emittance $\gamma\epsilon_y^{1\sigma}$ [μm]	2.5	3
long. emittance $\epsilon_l^{2\sigma}$ [eVs]	0.35	0.35
total bunch length l_b [ns]	≤ 4	4
momentum spread $2\sigma_p/p$ [10^{-3}]	2.2	2.2

smoothly divided into 2 or 3 bunches. The nominal scheme for producing the LHC beam now starts with six high-intensity bunches injected into the PS. Each of these bunches is split into three, which is later followed by two further double splittings. The entire process thus transforms the original 6 into 72 bunches.

As an illustration, Fig. 24 shows a simulation of triple bunch splitting in the PS. The entire procedure has been successfully demonstrated experimentally, and since 2000 is routinely used to produce the LHC beam for machine studies in the SPS.

Table 9 demonstrates that the PS already delivers an LHC beam which meets all the design parameters.[47]

Work is also progressing in the SPS. In the winter shut down 2000/2001 a few thousand pumping ports were shielded in an attempt to reduce the longitudinal impedance.

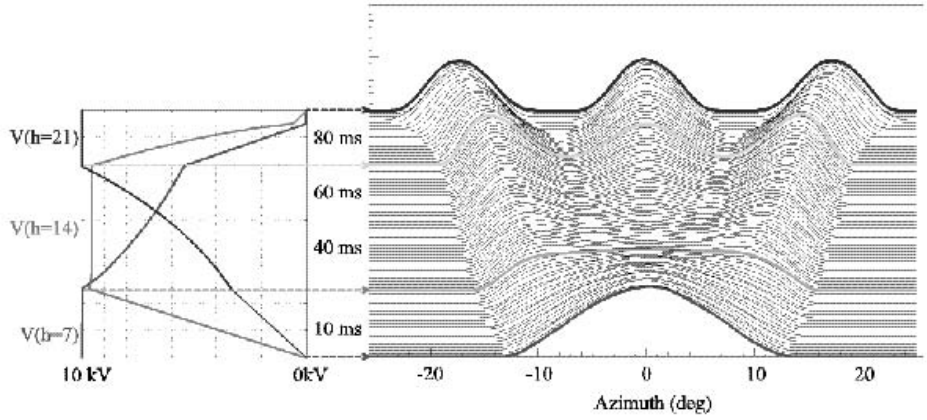

Fig. 24. Simulation of bunch splitting in the CERN PS in preparation for injection into the LHC.[46] The vertical axis is the time. The left picture shows the amplitudes of three rf systems operating at different frequencies (harmonic number h), which are used for this process. (Courtesy R. Garoby, 1999)

As a result of this effort, in 2001, bunch lengthening and strong impedance signals at 400 MHz are no longer observed.[48]

The evolution of the transverse impedance is also monitored by measurements of the coherent betatron tune shift with current and of the head-tail growth rates as a function of chromaticity.[49]

2.20 LHC as Heavy Ion Collider

Parameters for the LHC ion operation have been compared with the only existing ion collider, RHIC, in Table 1. The main limitation for ion operation varies with the ion mass.[6]

Heavy ion operation is limited by the electromagnetic processes occurring in the collision, namely by e^+e^- pair production and subsequent e^- capture.

The cross section of this process is about $\sigma_c \approx 100$ barn for Pb^{81+}-Pb^{81+} collisions, which corresponds to a rate of $\dot{N}_c \approx 10^5$ ions s^{-1} per side of IP at a luminosity of $L \approx 10^{27}$ $cm^{-2}s^{-1}$.

The cross section increases strongly with the atomic number $\sigma_c \propto Z^7$, whereas the energy deposition in a material only increases linearly with Z.

From the beam-optics point of view, for Pb ions a change in the ion charge by 1 unit is equivalent to a change in the relative momentum error of $\Delta\delta = 1.2\%$. Ions

with a momentum error of this magnitude are lost in a region of about 1 m length at the entrance to the LHC arcs, where s.c. dipole magnets are located. The predicted loss rate for the nominal LHC ion parameters is close to the quench limit, thus setting a limit for the maximum luminosity.

Potential remedies might be a dynamic squeeze of the IP β function during the store, so as to optimize the integrated luminosity, or the installation of local collimators, which could reduce the loss rate in the magnets of the dispersion suppressor.

For light ions, the cross section for the above electromagnetic process is negligibe, and, for these ions, the main limitation for luminosity operation is the growth of the longitudinal emittance due to intrabeam scattering (IBS). For nominal parameters the IBS growth time is 10 hours. It scales as

$$\frac{1}{\tau_{\rm IBS}} \propto \frac{N_b Z^3}{A} \ , \tag{30}$$

where A is the ion mass in units of the proton mass. Taking into account the two limiting factors from above, the projected initial luminosities are 1.0×10^{27} cm^{-2}s^{-1} for Pb$^{82}_{208}$ ions, 6.6×10^{28} cm^{-2}s^{-1} for Kr$^{36}_{84}$ ions, and 3.1×10^{31} cm^{-2}s^{-1} for O$^{8}_{16}$ ions.

2.21 Electron Cloud

In 1999, the build up of an electron cloud was observed with the LHC beam in the SPS, and in 2000 also in the PS and in the PS-SPS transfer line.

Observations in the SPS are illustrated in Fig. 25, which shows beam loss in the last 4 bunches of a 72-bunch LHC batch, occuring about 5 ms after injection. The beam loss only occurs above the threshold current of multipacting, which manifests itself by a large vacuum pressure rise and by electron signals seen on dedicated electron-cloud monitors.

The electron cloud in the SPS is generated as follows. A small number of primary electrons is generated, *e.g.*, by gas ionization or beam loss. For the narrow LHC bunch spacing of 25 ns and typical vacuum-chamber half apertures of 2–3 cm, the number of electrons exponentially amplifies during the single passage of a 72-bunch train, by a process called beam-induced multipacting, already observed in the CERN ISR almost 30 years ago.[51]

In this section, we discuss build up, saturation, and decay of the electron cloud, then the wake fields and instabilities induced by the electrons, finally the heat load

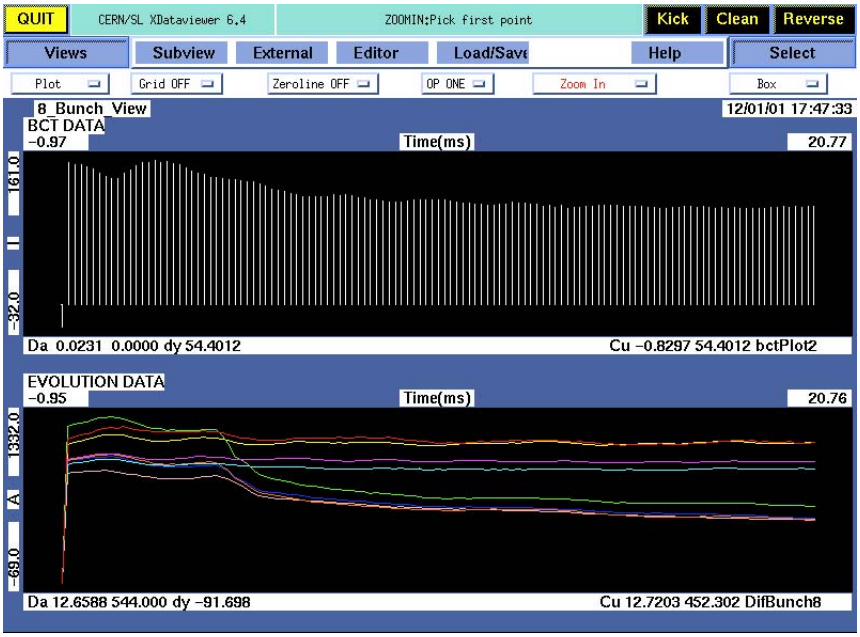

Fig. 25. Intensity of 72-bunch LHC beam in SPS vs. time. Batch intensity (top) and bunch intensity for the first 4 bunches and the last 4 bunches (where losses are visible after about 5 ms) of the batch (bottom).[50] (Courtesy G. Arduini, 2001)

from the electrons, which is the largest concern for the LHC, and the presently foreseen countermeasures.

We use the variable λ to denote the electron line density; t is the time and s the position along the beam line. For a beam current of 0.5–1 A, the processes contributing to the generation of electrons are (1) residual gas ionization, with a typical rate of $d^2\lambda_e/(ds\ dt) \approx 5 \times 10^{11}$ e$^-$ m^{-1}s^{-1}; (2) synchrotron radiation and photo-emission, with a typical rate $d^2\lambda_e/(ds\ dt) \approx 5 \times 10^{18}$ e$^-$ m^{-1}s^{-1}; and (3) secondary emission, consisting of true secondaries and also of elastically reflected or rediffused electrons. If the average secondary emission yield is larger than 1, the secondary emission leads to an exponential growth.

Indeed, the key process of the electron-cloud formation in the LHC is the secondary emission, and the most important parameter the secondary emission yield. The latter depends on the energy of the primary electron. A parametrization for the LHC vacuum chamber[52] is shown in Fig. 26. The secondary electrons consist of two components. The true secondaries are emitted at low energies, of the order of a few eV. Their yield reaches a maximum value $\delta_{\max}$ at a certain impact energy $\epsilon_{\max}$, and the yield curve is well approximated by a universal function with only these two free parameters. A certain fraction of the incident electrons is elastically reflected. The lower the energy of the incident electron, the larger is the proportion of the reflected electrons. These reflected electrons are responsible for the fact that the total secondary emission yield does not approach zero if the energy of the primaries approaches zero, but remains at a finite value. The contribution from elastically reflected electrons to the total yield is also illustrated in Fig. 26. The nonzero yield value for low energies implies that a certain number of low-energetic electrons will survive for a long time inside the vacuum chamber, even if there is a large gap in the bunch train.

The build up of the electron cloud due to beam-induced multipacting does not continue indefinitely, but it saturates, roughly at the moment when the average number of electrons per unit length is equal to the average line density of beam protons or positrons. In other words, the order of magnitude of the saturated electron line density can be estimated as

$$\lambda_{\rm sat} \approx \frac{N_b}{L_{\rm sep}}, \tag{31}$$

where N_b is the bunch population, and $L_{\rm sep}$ the bunch spacing. The corresponding volume density is obtained by dividing with the beam-pipe cross section.

In the SPS the small number of primary electrons produced via gas ionization is

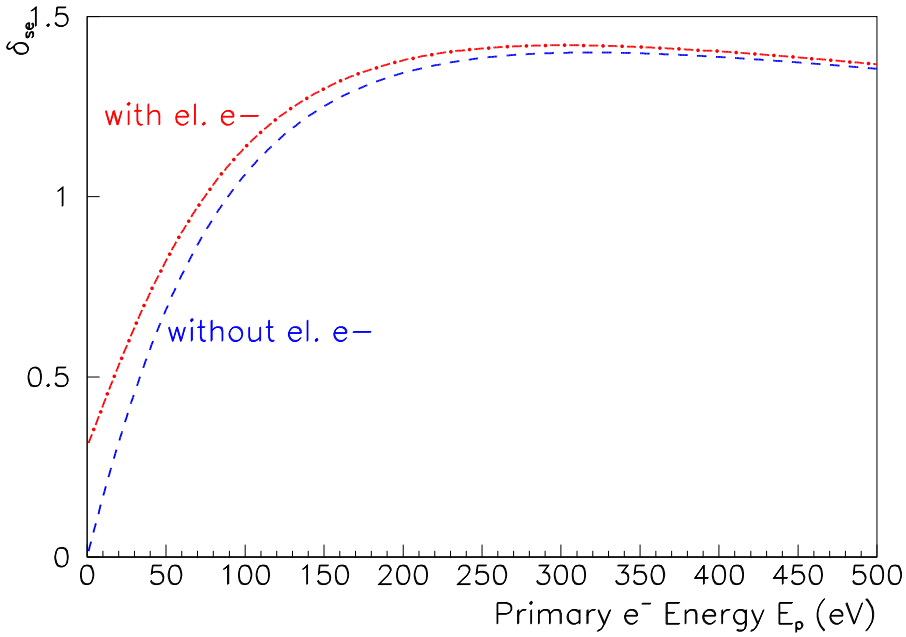

Fig. 26. Secondary emission yield for perpendicular incidence vs. the primary electron energy with and without elastically scattered electrons. The parametrization is based on measurements for copper surfaces.[52]

amplified by multipacting so strongly that saturation is reached already after about 30 bunches. In the LHC the number of primary photoelectrons will be much larger than that of ionization electrons in the SPS. Figure 27 shows a schematic of the electron-cloud build up in the LHC beam pipe.

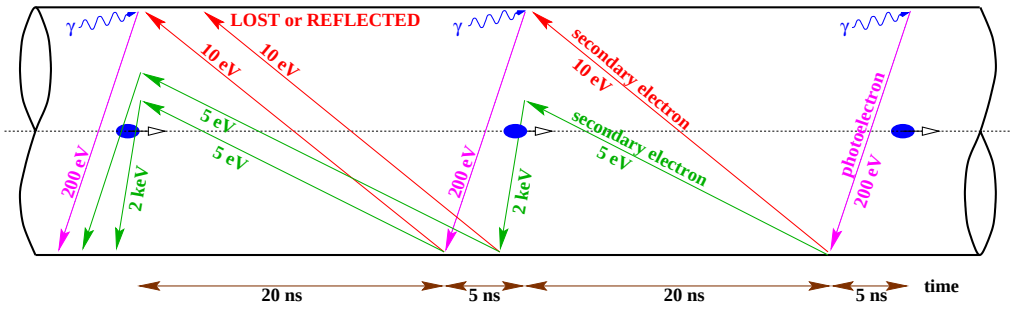

Fig. 27. Schematic of electron-cloud build up in the LHC beam pipe. (Courtesy F. Ruggiero)

The condition for proper multipacting is[53]

$$n_{\min} \equiv \frac{h_y^2}{N_b r_e L_{\rm sep}} = 1, \tag{32}$$

which describes the situation that the travel time of electrons across the chamber exactly equals the time between two bunches, and includes the assumption that the electrons are close to the chamber wall when the bunch arrives. However, it should be noted that the condition (32) is neither necessary nor sufficient to observe electron amplification. If the parameter $n_{\min}$ is smaller than one, low-energetic secondary electrons are produced before the next bunch arrives. They will move slowly through the chamber and are accelerated only when a bunch passes by. On the other hand, if $n_{\min}$ is larger than 1, an electron will interact with more than 1 bunch. In either situation electron amplification can still occur. This is illustrated in Table 10 which lists parameters for several accelerators where electron clouds have been observed or are predicted to occur. The values of $n_{\min}$ are also listed for each ring. They extend over several orders of magnitudes.

Thus $n_{\min}$ is not a reliable parameter to assess the possibility of multipacting. In order to predict the occurrence and magnitude of multipacting, detailed computer simulations are required. Figure 28 illustrates the ingredients of such simulations.[55] Both bunches and interbunch gaps are split into slices. For each slice, the motion of elec-

Table 10. Comparison of parameters related to the electron-cloud build up for the LHC beam in the CERN PS, SPS, and the LHC with those of several other proton and positron storage rings, in which an electron cloud is observed or expected.[54]

accelerator	PEP-II	KEKB	PS	SPS	LHC	PSR	SNS
species	e^+	e^+	p	p	p	p	p
population N_b [10^{10}]	10	3.3	10	10	10	5000	10000
spacing $L_{\rm sep}$ [m]	2.5	2.4	7.5	7.5	7.5	(108)	(248)
bunch length σ_z [m]	0.013	0.004	0.3	0.3	0.077	25	30
h. beam size σ_x [mm]	1.4	0.42	2.4	3	0.3	25	0.6
v. beam size σ_y [mm]	0.2	0.06	1.3	2.3	0.3	7.5	0.6
ch. $\frac{1}{2}$ size h_x [mm]	25	47	70	70	22	50	100
ch. $\frac{1}{2}$ size h_y [mm]	25	47	35	22.5	18	50	100
synchrotron tune Q_s	0.03	0.015	0.004	0.006	0.002	0.0004	0.0007
circumf. C [km]	2.2	3.0	0.63	6.9	27	0.09	0.22
beta function β	18	15	15	40	80	5	6
parameter $n_{\rm min}$	1	10	0.58	0.24	0.15	0.0002	0.0001

trons is computed under the influence of the beam field, external magnetic fields, electron space-charge field, and the image forces induced by both beam and electrons. For each passing bunch slice, a certain number of primary electrons is created. Whenever an electron is lost to the wall, its charge state is changed according to the secondary emission yield computed for its energy and impact angle, and the electron is re-emitted representing either a true secondary or an elastically scattered electron.

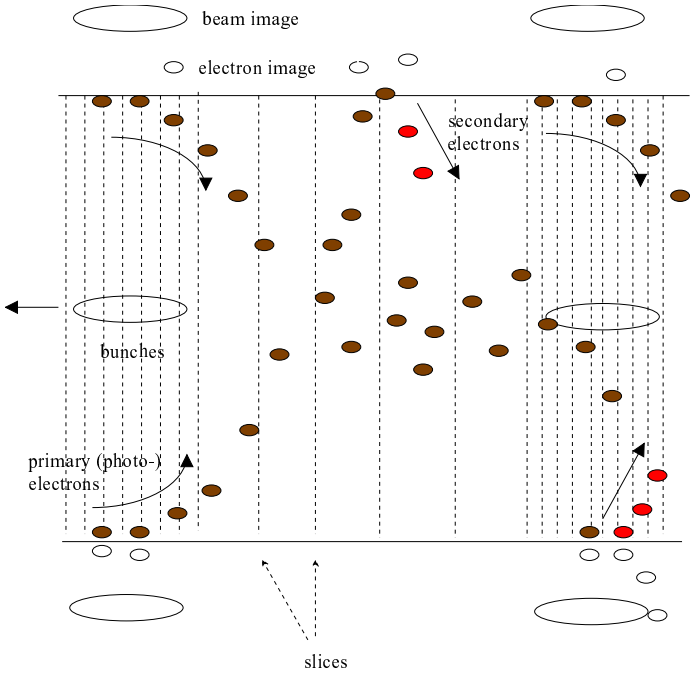

Fig. 28. Schematic illustrating various ingredients of the electron-cloud simulations.

In the actual accelerator, various indicators can signal the electron-cloud build up, such as (1) a nonlinear pressure rise with beam current, (2) current data from electrostatic pick ups or dedicated electron monitors, (3) the measured tune shift along the train, (4) the beam-size blow up along the train, and where applicable, (5) a drop in luminosity. All these items entail direct informations about the electron density.

As an example, we mention two estimates of the electron-cloud density in the SPS. The first is based on the pressure rise, and is due to O. Gröbner.[56] The equation for the pressure balance reads

$$S_{\text{eff}} P/(k_B T) = Q \tag{33}$$

where S_{eff} denotes the pumping speed in volume per meter per second, $Q = \alpha d\dot{\lambda}_{t,e}/ds$ the total flux of molecules per unit length (α is the desorption yield per electron, and

$\lambda_{t,e}$ the number of electrons hittng the chamber wall per unit length and per bunch train passage) and $P = k_B T N/V$, where N/V is the number of gas molecules per unit volume, and P the pressure. We insert the expression for Q into Eq. (33), and solve for $d\lambda_{t,e}/ds \approx T_{\rm rev}(d\dot{\lambda}_{t,e}/ds)$

$$\frac{d\lambda_e}{ds} \approx \frac{T_{rev}}{\alpha k_B T}\, S_{\rm eff} P, \tag{34}$$

where $T_{\rm rev}$ is the revolution period.

With an enhanced pressure of $P = 100$ nTorr, $\alpha \approx 0.1$ and $S_{\rm eff} \approx 20\,{\rm l\,s^{-1}\ m^{-1}}$ one estimates

$$\frac{d\lambda_e}{ds} \approx 10^{10} \frac{\rm electrons}{\rm bunch-train\ meter}\ . \tag{35}$$

The second estimate is directly related to the signal seen on the transverse damper pick up in the SPS, which indicates that a few 10^8 electrons per bunch passage are deposited on the pick-up.[57] This number amounts to $10^9 - 10^{10}$ per train, or, for an effective pick-up length of about 10 cm, to[57]

$$\frac{d\lambda_e}{ds} \approx 10^{10} \frac{\rm electrons}{\rm bunch-train\ meter}\ . \tag{36}$$

The two estimates, (35) and (36), are consistent.

Figure 29 shows two difference signals measured between the plates of two identical electro-static pick ups in the SPS. Without perturbation from the electron cloud, the difference signal should be proportional to the beam offset in the chamber. One of the two signals in Fig. 29 is processed at low frequencies, the other in a higher frequency band around 120 MHz. The shift in the baseline of the low-frequency signal, which is seen near the center of the 1.8-μs bunch train and persists in the 20-μs gap without beam, indicates a net charge transfer between the pick-up plates, due to the multipacting electrons. The same distortion is not visible in the high-frequency signal, which may suggest that the frequency spectrum of the electron cloud current between the plates of the pick up does not extend up to 120 MHz.

Figure 30 displays a simulation of the electron-cloud build up for the SPS parameters. The simulation can reproduce the observed saturation of the electron-cloud build up at the center of the bunch train, provided that the elastically reflected electrons are included in addition to the true secondaries.

In the SPS also a positive tune shift is observed which starts between the 10th and 20th bunch of the train and is of order $\Delta Q \approx 0.01$. The tune shift permits an independent estimate of the electron density, which is consistent with the other two estimates from above.

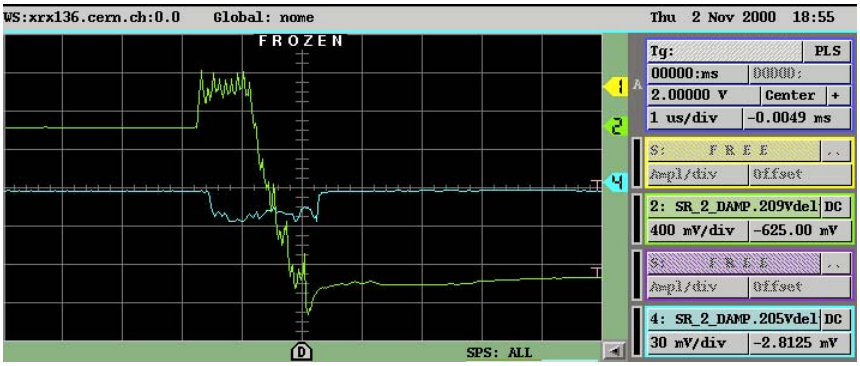

Fig. 29. Difference signals on damper pick-up during the passage of an LHC batch in the SPS (1μs/div); the signal observed at low frequency (green line, shifted due to electron cloud) and the downmixed signal sampled at 120 MHz (blue line, without final offset). (Courtesy W. Hofle, 2000)

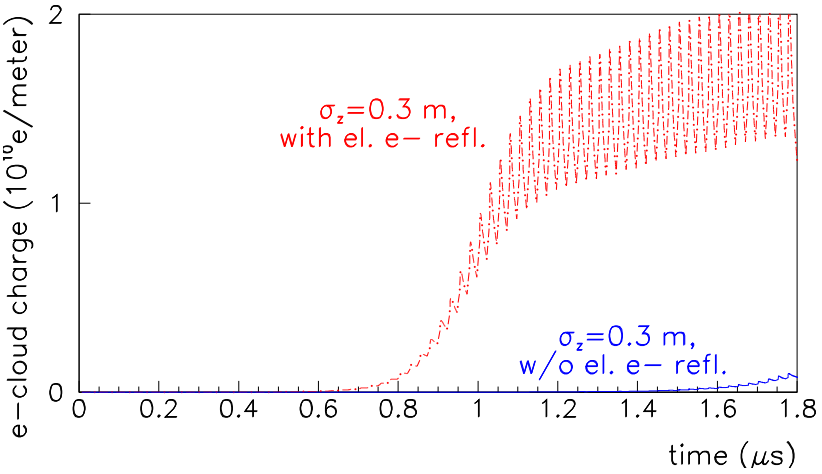

Fig. 30. Simulated evolution of the electron line density in units of 10^{10} m^{-1} as a function of time in s, for an SPS dipole chamber, with and without elastic electron reflection.[55]

The electron cloud, once generated, can gives rise to wake fields and instability. Namely, the electrons couple the motion of subsequent bunches and can thereby couple the motion of successive bunches. They also introduce coherent and incoherent tune shifts. However, in the SPS the most harmful effect is a single-bunch instability, which presumably is of a similar nature as those observed in the positron rings of the PEP-II and KEKB B factories. This instability appears to be the analogue of the strong head-tail instability caused by a conventional wake field. (The strong head-tail instability is also called the transverse-mode coupling instability, abbreviated as TMCI.) In addition, the electron cloud might excite the regular head-tail instability, and also induce longitudinal electric fields, albeit recent analysis suggests that the latter two effects are small.

We limit the following discussion to the TMCI-like instability. Several dedicated computer programmes were written, at KEK and CERN, which model this instability. In the simulation both the electrons and a bunch of the beam are represented by macroparticles. On each turn the bunch interacts with a fresh uniform cloud of electrons, assumed to be generated by the preceding bunches. The electron cloud can act like a wake field and enhance an initial head-tail perturbation in the beam.

During the bunch passage, the electrons oscillate in the beam potential. Figure 31 shows a snapshot of the simulated electron phase space at the end of a bunch passage.

If the electrons perform several transverse oscillations over the length of the bunch, they may be adiabatically trapped in the beam potential and remain at the center of the chamber for a long time.[59]

Using the WKB approximation the adiabaticity condition for this trapping process can be written as

$$A \equiv \sigma_z \omega_{e,y} \sqrt{8e}/c \gg 1 \ , \tag{37}$$

where $\omega_{e,y}$ is the vertical electron angular oscillation freqency, and $e = 2.718\ldots$ Inserting the accelerator parameters into Eq. (37), we obtain $A \approx 10$ for KEKB, PEP-II, PS, SPS, and LHC. Hence, in all these accelerators, electrons may be trapped.

If simulations are performed with the electron cloud as the only perturbation, the beam size increases smoothly but signficantly with time. If the position-dependent tune shift due to the proton space-charge force in the SPS at 26 GeV/c is also included, the simulated instability becomes more violent.[60] This is illustrated in Fig. 32.

Figure 33 compares the simulated emitance growth in both transverse planes, computed using two different models of the beam field, namely a soft-Gaussian approximation and a particle-in-cell (PIC) code.[61] The two results are comparable. Either

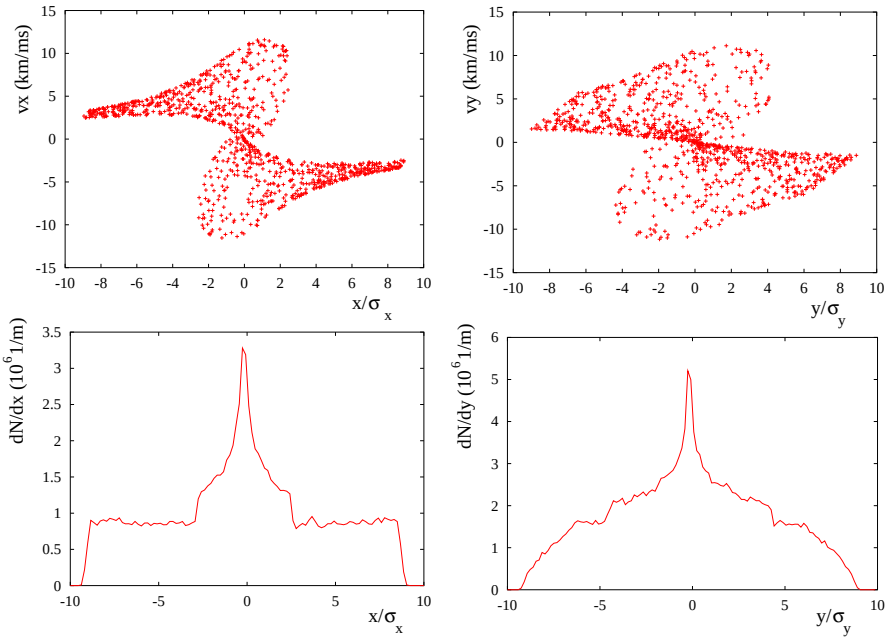

Fig. 31. Snapshots of the horizontal and vertical electron phase space (top) and their projections onto the position axes (bottom).[58] (Courtesy G. Rumolo, 2001)

simulation predicts a rapid emittance growth within a few ms, which is consistent with the time scale of the observed beam loss.

The effective transverse wake field of the electron cloud can be obtained from the simulation, by displacing a slice of the bunch transversely, and computing the resulting force on the subsequent bunch slices. A typical result is shown in Fig. 34. Because of the electron accumulation inside the bunch during its passage and due to the nonlinearities of the forces acting between beam and electrons, the computed wake fields depend on the position of the displaced slice, as illustrated in this example.

Either using a two-particle model,[62] or approximating the simulated wake field by a broadband resonator,[63] one can estimate the TMCI threshold. Table 11 compares the estimated threshold cloud density with the expected saturation density for various accelerators. The table demonstrates that almost all the accelerators listed may operate above the electron-cloud instability threshold.

At the SPS direct evidence for the head-tail instability comes from a wideband pickup which measures the transverse position every 0.5 ns, compared with a total bunch length of 4 ns. In the vertical plane, significant motion is detected inside the bunch. The oscillations of subsequent bunches are uncorrelated. The wave length of

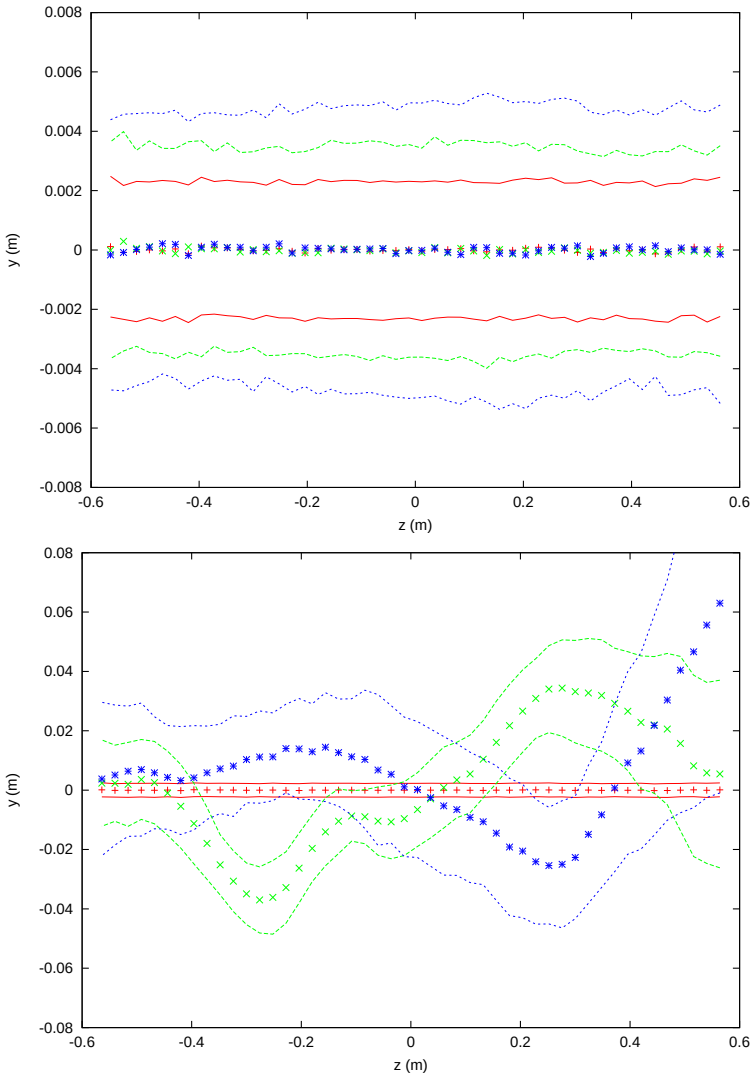

Fig. 32. Simulated bunch shape after 0, 250 and 500 turns (centroid and rms beam size shown) in the CERN SPS with an e^- cloud density of $\rho_e = 10^{12}$ m^{-3}, without (top) and with (bottom) proton space charge.[60] (Courtesy G. Rumolo, 2001)

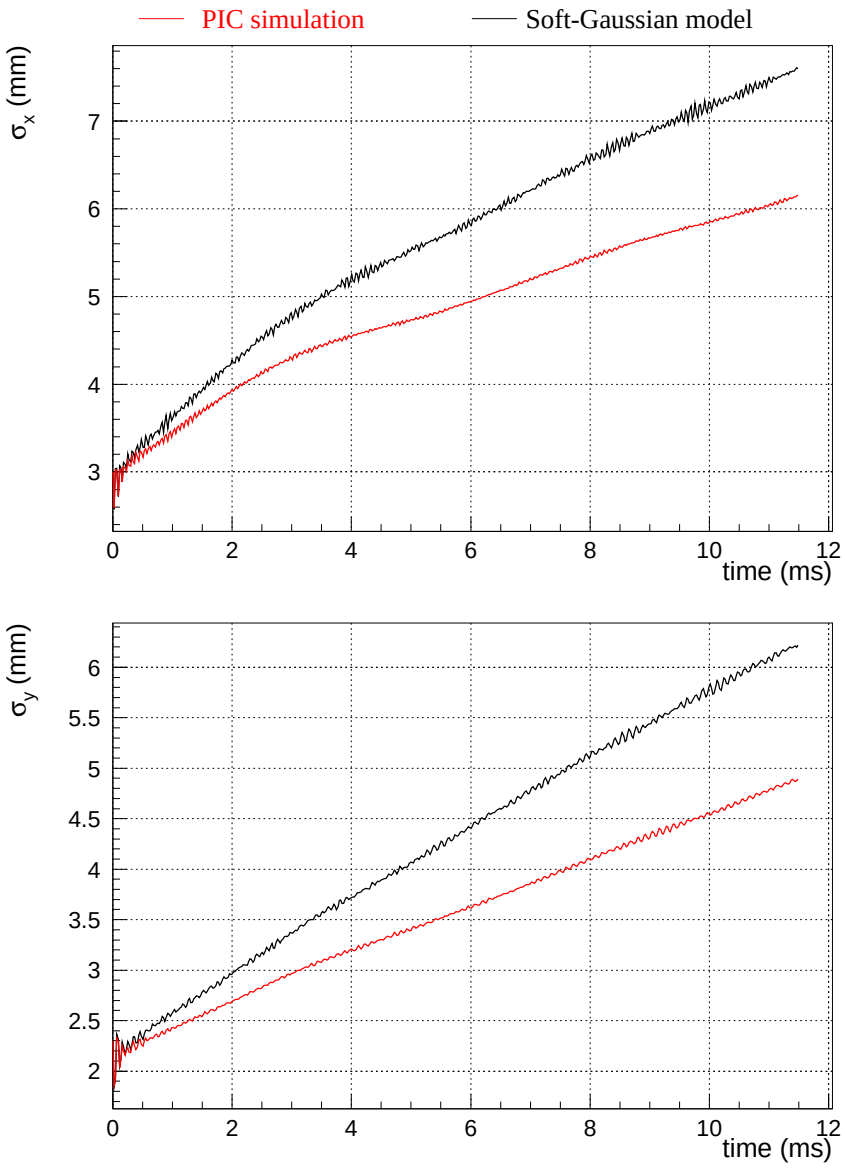

Fig. 33. Beam size evolution for an SPS bunch interacting with an electron cloud as predicted by different simulation approaches,[61] for a cloud density of $\rho_e = 10^{12}$ m^{-3}. (Courtesy G. Rumolo, 2001)

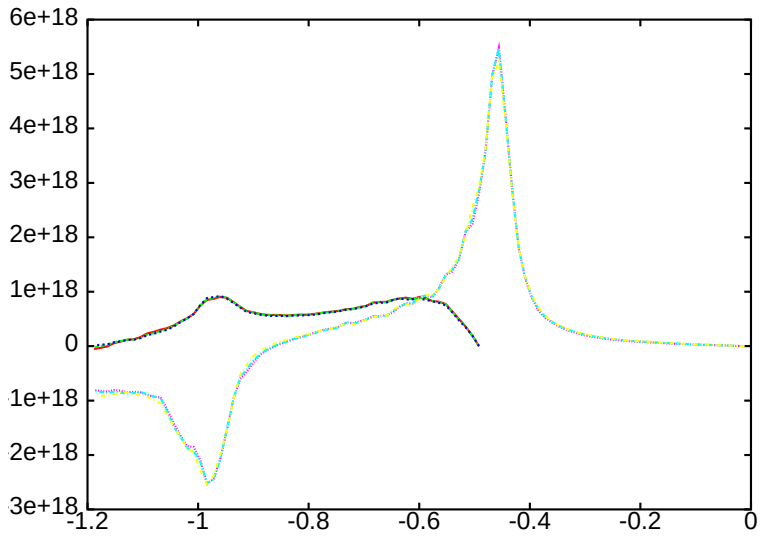

Fig. 34. Simulated wake force in V/m/C computed by displacing slice 1 and 40 (out of 100) of a Gaussian bunch with rms length 0.3 m, as a function of longitudinal position along the bunch in units of meter. The bunch center is at -0.6 m and the bunch head on the right.[58] (Courtesy G. Rumolo, 2001)

Table 11. The number of electron oscillations during a bunch passage, the estimated electron-cloud TMCI threshold, and the ratio of electron equilibrium density to threshold density, for various accelerators.[54]

accelerator	PEP-II	KEKB	PS	SPS	LHC	PSR	SNS
e^- osc./bunch $n_{\rm osc} \equiv \omega_e \sigma_z/(\pi c)$	0.8	1.0	1	0.75	3	34	970
TMCI threshold ρ_e [10^{12} m^{-3}]	1	0.5	5	0.25	3	(0.6)	(0.5)
density ratio $\rho_{e,\rm sat}/\rho_{e,\rm thresh}$	19	4	0.35	11	4	(92)	(27)

the wake field was fitted from the data by K. Cornelis, and it agrees with the calculated wavelength of electron oscillations.[64]

Simulations including both the electron cloud and, in addition, a regular broadband impedance show that the instability can be suppressed by a large positive chromaticity,[65] in accordance with observations.

The heat deposited by electrons on the beam screen is a major concern for the LHC. Simulated electron impact energies extend up to several 100s of eV. This is much larger than the typical emission energy of secondaries of only a few eV. In other words, the electron cloud extracts a significant energy from the beam, and transfers it to the chamber wall.

The importance of this issue for the LHC is illustrated in Fig. 35, which shows the simulated arc heat load, averaged over dipoles, field-free regions and quadrupoles, as a function of bunch population for various values of the maximum secondary emission yield $\delta_{\max}$. Also indicated is the maximum cooling capacity available for the electron cloud. The figure demonstrates that in order to reach the design bunch intensity of $N_b = 1.1 \times 10^{11}$ the secondary emission yield must not be much larger than 1.1.

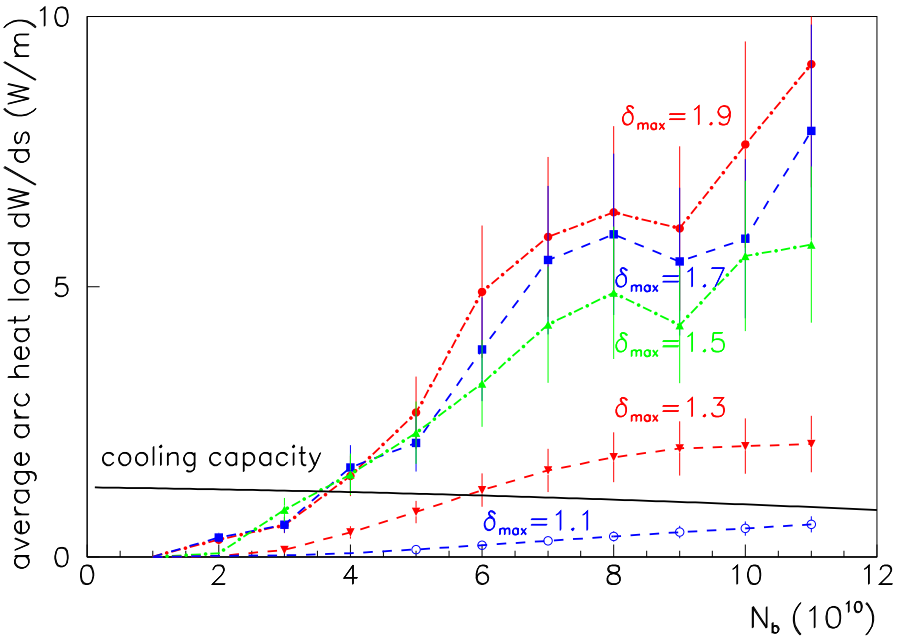

Fig. 35. Average arc heat load and cooling capacity as a function of bunch population N_b, for various values of the maximum secondary emission yield $\delta_{\max}$.

Figure 36 shows the electron distribution simulated for an LHC dipole. The vertical stripes with enhanced electron density correspond to the regions with maximum multipacting. If such electron stripes would lie on top of the beam-screen pumping slots, electrons could pass directly to the 1.9-K cold bore, instead of being absorbed by the beam screen. The cooling capacity for the cold bore is much smaller, and a quench would be a likely consequence.

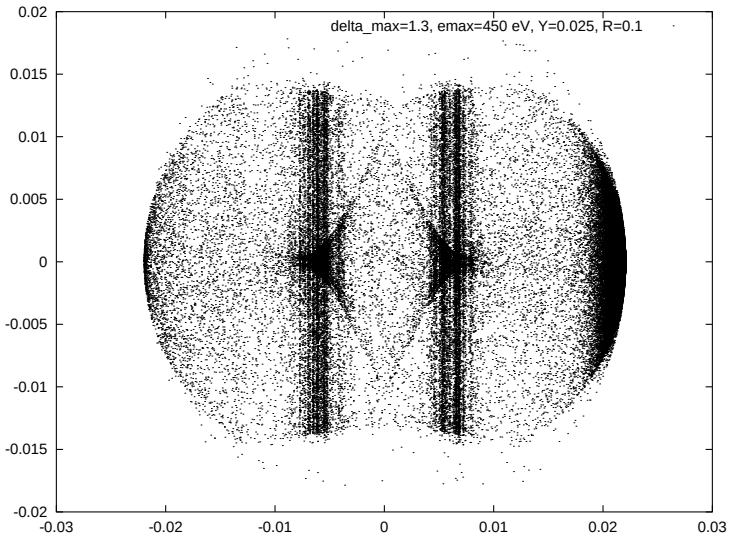

Fig. 36. Snapshot of the transverse electron distribution in an LHC dipole chamber, for a maximum secondary emission yield of $\delta_{\max} = 1.3$.[66]

Three main countermeasures against the electron cloud are foreseen, *i.e.*,

- to install a 'sawtooth chamber' (with a height of about 35 μm and a period of 200 μm), which reduces the photon reflection in the arc dipoles;
- to coat all warm sections with a getter material TiZr, that exhibits a low secondary emission yield;
- and to rely on surface conditioning during the commissioning, which should reduce the maximum secondary emission yield to a value of 1.1; an electron dose of about 10 C/mm^2 is needed to reach this target value.[67]

In 2001 several novel electron cloud detectors were installed in the SPS by G. Arduini, J.M. Jimenez, et al., whose purpose is to serve as a benchmark for the simulation and to directly provide measurements under conditions very similar to those in the LHC.

The newly installed SPS electron-cloud detectors include[68]: (1) pick-ups which measure the electron characteristics, in particular, the e^- cloud build up and the e^- energy distribution; they also allow for triggering on the batch; (2) monitors which characterize the behavior of the electrons in a dipole magnetic field; 2 different designs were developed for this purpose, the first is a 'strip detector', the second a so-called 'triangle detector'; (3) an in-situ measurement of secondary emission yield, which can verify the effect of surface processing; (4) ion detectors to exclude ion-stimulated desorption as a source of the pressure rise; and (5) a so-called WAM_PAC Cu calorimeter which directly measures the heat load from the electron cloud.

First observations with these detectors are promising. The strip monitor clearly reveals the horizontal position and width of the multipacting electrons. Above a bunch intensity of $N_b \approx 5 \times 10^{10}$ protons, the single strip splits into two, which for further increasing bunch current move towards the outside of the chamber. This behavior agrees well with the simulations. Preliminary measurements using a biasing grid and the triangular detector suggest average electron energies of the order of 75 eV. The calorimeter measures a power deposition, which, scaled to the LHC, might correspond to a heat load of the order of 1 W/m, comparable to typical predictions.

The in-situ change in the secondary emission yield was also measured. After about 24-hours of effective conditioning time with an LHC beam, the maximum secondary emission yield $\delta_{\max}$ had decreased from 2.3 to 1.8, which demonstrates that surface scrubbing is acting as foreseen.

3 Beyond LHC: LHC-II and VLHC

Once the LHC is operating, the particle physicists will push for higher luminosity and higher energy. A feasibility study for an 'LHC-II' has been launched at CERN.[69]

The LHC luminosity can be raised by increasing the number of bunches, which might imply a larger crossing angle. As a next step, one might contemplate more exotic schemes, where, *e.g.*, 'crab' cavities on either side of the collision point deflect the head and tail of the bunches transversely in opposite directions such that the bunches collide effectively head on.

The availability of stronger or cheaper magnets will facilitate the path towards higher energy and indicate the direction to follow.

Synchrotron radiation and emittance control will become an important issue, as the higher-energy machines will operate in a new regime, where the effects of synchrotron

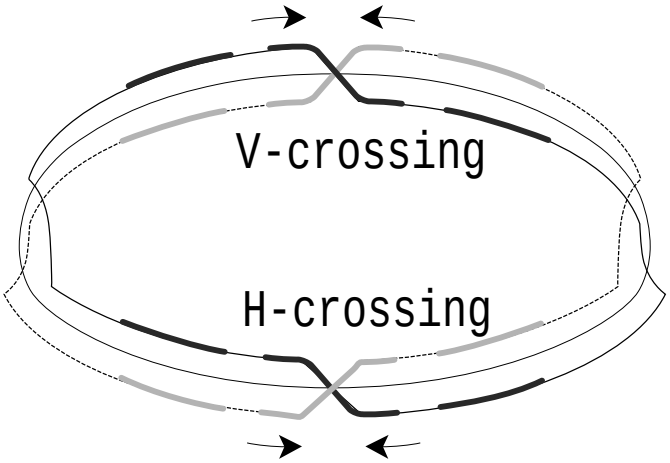

Fig. 37. Schematic of Super Bunches in a High-Luminosity Collider.[71]

radiation become more and more noticable, and where the radiation equilibrium emittance is much smaller than the injected emittance. The shrinking of the emittance during the store could be a nuisance, since the beam-beam tune shift increases with time and the beam-beam limit may be reached soon after injection, potentially leading to an unstable situation. On the other hand, intrabeam scattering is still significant for these energies and emittances, and may balance the radiation damping. Thus, a careful study of emittance control is clearly an important topic for the future hadron machines.[70]

Some collective effects can also prove more severe. If the circumference is large, the coupled-bunch resistive wall instability may require several local feedbacks. More worriesome still is the electron cloud, which might introduce an ultimate limitation.

Also the debris from the IP, the quench limits and the question of a safe beam abort will be major challenges.

Lastly, there is an option to collide continuous beams as in the ISR. In reality these would be rather ‘quasi-continuous beams’ or ‘superbunches’, occupying only a small fraction of the total circumference and being confined by barrier rf buckets,[71] as illustrated in Fig. 37. The barrier buckets may be generated by induction acceleration modules.[71]

Such continuous beams or superbunches are based on the successful experience of the ISR and hold various promises:

- in conjunction with alternating crossing at two IPs superbunches can provide a higher luminosity with acceptable beam-beam tune shift;
- PACMAN bunches are absent and the beam particles almost identical to one an-

other, each sampling all longitudinal positions with respect to the opposing beam;

- the electron-cloud build up should be strongly suppressed, since electrons cannot gain any energy in the constant beam potential;
- superbunches might allow for stochastic cooling, which is not an option for bunched beams.[72]

A primary reason why coasting beams were abandoned, despite of the ISR success, was the scarcity of antiprotons. This is no longer a problem for proton-proton colliders.

A large number of questions need to be answered, however, before a superbunch scheme could be envisioned for a future LHC upgrade. Ongoing work at CERN includes the optimization of the beam parameters, such as length, line density, and total number of superbunches, which would maximize the luminosity, while maintaining a tolerable beam-beam tune shift and an acceptable heat load, and obeying the timing constraints imposed by the (induction) rf system, the capacity of the injectors, the filling time, and the beam abort system.

3.1 Prospects for Luminosity and Beam-Beam Tune Shifts

Considering round beams, *i.e.*, $\beta_x^* = \beta_y^*$ and $\epsilon_x \approx \epsilon_y$, and including the hourglass effect and a horizontal crossing angle θ_c, the luminosity for both normal and superbunches can be expressed as

$$L = \frac{f_{\rm coll}\lambda_1\lambda_2\beta^*}{2\pi\sigma_0^2}\int_{-l_{\rm det}/(2\beta^*)}^{l_{\rm det}/(2\beta^*)}\frac{1}{1+u^2}\,\exp\left[-\frac{\beta^{*2}\theta_c^2/\sigma_0^2}{4}\frac{u^2}{1+u^2}\right]\,f(u)\,du \qquad (38)$$

where

$$f(u) = l_{\rm bunch} \qquad (39)$$

for a superbunch whose length $l_{\rm bunch}$ is much larger than the effective length of the detector $l_{\rm det}$, and

$$f(u) = \sqrt{\pi}\sigma_z\exp\left[-\frac{\beta^{*2}u^2}{\sigma_z^2}\right]. \qquad (40)$$

for a regular Gaussian bunch of rms length σ_z. The coefficients λ_1 and λ_2 denote, for a normal bunch, the maximum line density $\lambda = N_b/(\sqrt{2\pi}\sigma_z)$, and, for a superbunch, the constant line density $\lambda = N_b/l_{\rm bunch}$.

Considering a single collision point with horizontal crossing, the maximum beam-beam tune shifts, experienced by a particle at the center of the bunch, are

$$\Delta Q_x \;=\; -\frac{\lambda r_p}{\pi\gamma}\int_{-l/2}^{l/2}\left(\beta^*+\frac{s^2}{\beta^*}\right)\left[\left(\frac{1}{(\beta^*+s^2/\beta^*)\,\epsilon}+\frac{1}{\theta_c^2 s^2}\right)\right.$$

$$\exp\left(-\frac{\theta_c^2 s^2}{2\left(\beta^* + s^2/\beta^*\right)\epsilon}\right) - \frac{1}{\theta_c^2 s^2}\right]\, g(s)\, ds$$

$$\Delta Q_y = -\frac{\lambda r_p}{\pi\gamma}\int_{-l/2}^{l/2}\left(\beta^* + \frac{s^2}{\beta^*}\right)\left[\frac{1}{\theta_c^2 s^2}\,(1 - \exp\left(-\frac{\theta_c^2 s^2}{2\left(\beta^* + s^2/\beta^*\right)\epsilon}\right)\right)\right]\, g(s)\, ds\ , \tag{41}$$

where

$$g(s) = \exp\left(-2\frac{s^2}{\sigma_z^2}\right) \tag{42}$$

for a regular bunch, and $g(s) = 1$ for a superbunch, and λ is the (maximum) line density of the opposing, equal to either λ_1 or λ_2 in the luminosity formula. Here, the electrostatic interaction between the two bunches is assumed to occur between $-l/2$ and $l/2$. Outside of this range the beams are either separated by a bending magnet, or shielded from each other. The distance l can be much larger than the effective detector length $l_{\rm det}$.

Figure 38 shows the luminosity and beam-beam the shifts as a function of crossing angle as computed from Eqs. (38) and (41) for the so-called ultimate LHC bunch intensity of $N_b \approx 1.7 \times 10^{11}$ with regular Gaussian bunches of 7.7-cm rms length, and considering collision-point beta functions which are reduced from the nominal value of 0.5 m to 0.25 m. The number of bunches is unchanged compared with the nominal scenario. Assuming two interaction points with alternating crossing, the maximum total beam-beam tune shift is given by the sum $\Delta Q_{\rm tot} = (\Delta Q_x + \Delta Q_y)$. This total tune shift is also shown in the figure. For crossing angles of 300–400 μrad, it is quite moderate, and much below the highest values achieved elsewhere (compare Table 5).

Figure 39 shows the corresponding curves for a coasting beam or for a superbunch scheme. If the entire ring is filled, with 40 A dc current, the luminosity is of the order of 5×10^{36} cm^{-2}s^{-1}. If only a 1/40th of the ring is occupied the luminosity could still 10^{35} cm^{-2}s^{-1} with an average current of 1 A. These parameters have not yet been fully optimized.

3.2 Crab Cavities

As shown in Fig. 38 the luminosity decreases for larger crossing angles. This luminosity loss can be avoided by means of 'crab crossing', a scheme which was first proposed for linear colliders,[73] and will be tested at the KEK B factory. The basic idea of crab

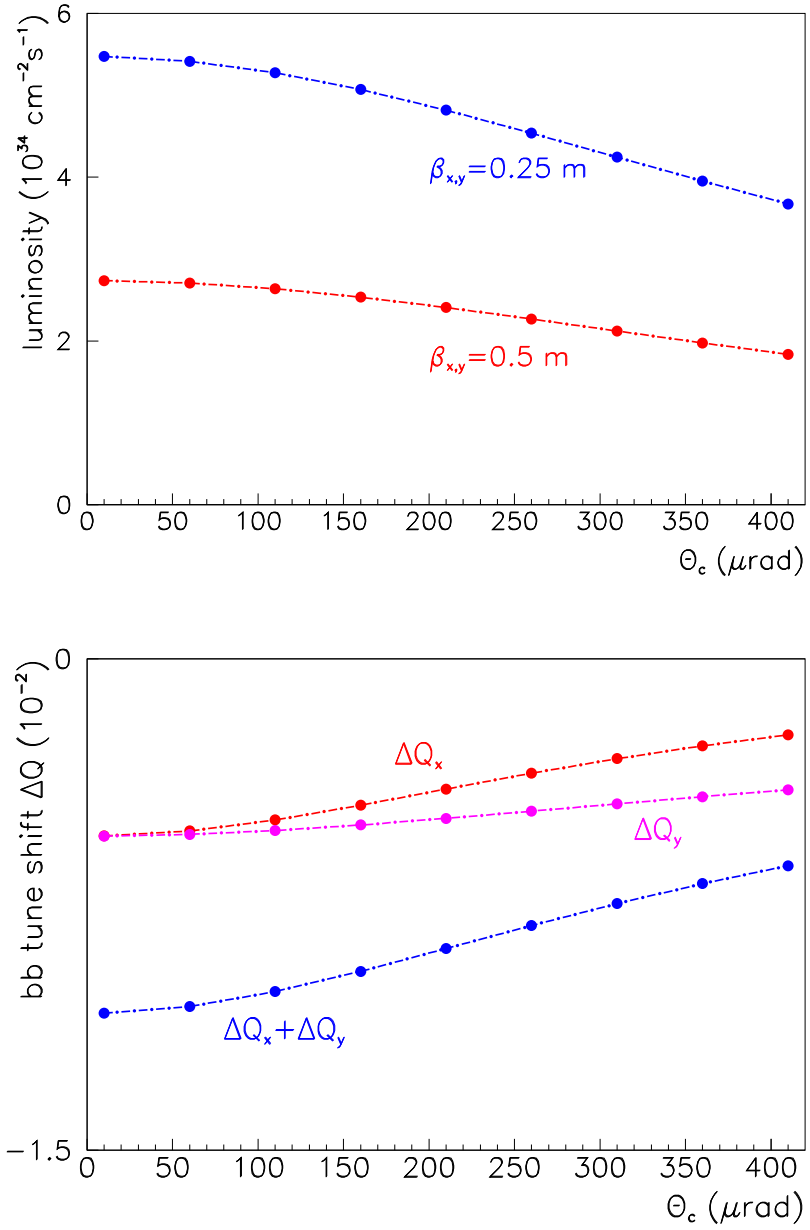

Fig. 38. Luminosity (top) and total beam-beam tune shift (bottom) vs. crossing angle; parameters: $N_b = 1.7 \times 10^{11}$, $\beta^* = 0.25$ m, $\sigma_z = 7.7$ cm, $n_b = 2800$, $\gamma\epsilon_\perp = 3.75$ μm.

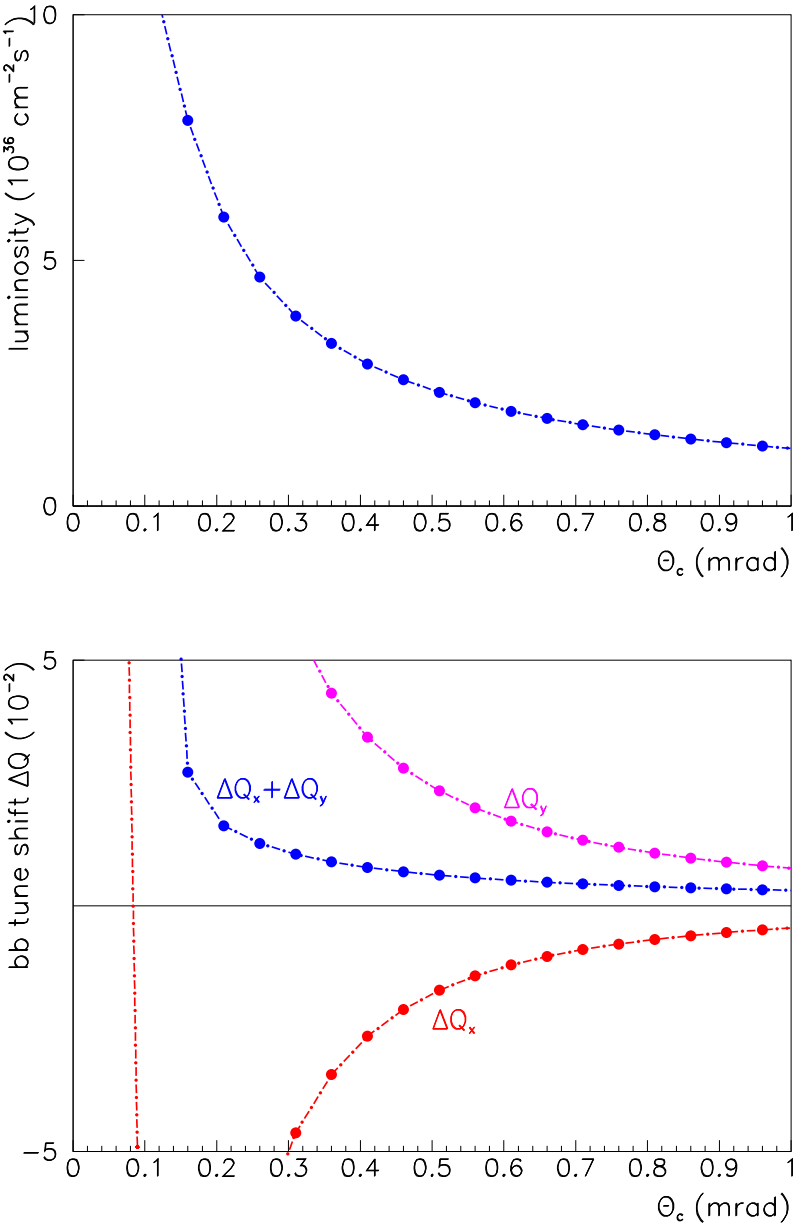

Fig. 39. Luminosity (top) and total beam-beam tune shifts (bottom) vs. crossing angle, for a continuous beam with a line density $\lambda = 8.8 \times 10^{11}$ m^{-1} (40 A current), $\beta^* = 0.25$ m, $l_{\rm det} = 1$ m, $l = 20$ m, and $\gamma\epsilon_\perp = 3.75$ μm.

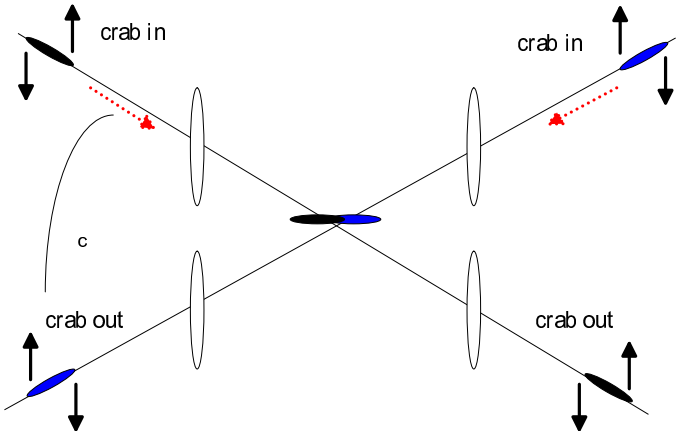

Fig. 40. Applying a deflection of opposite sign to the head and tail of each bunch avoids luminosity loss due to the crossing angle.

crossing is illustrated in Fig. 40. The differential deflection received in the dipole cavities aligns the bunches at the collision point, so that the luminosity is the same as for head-on collisions.

The crab cavities would be most useful if they would allow separating the beams after the collision into two disjoint quadrupoles. Assuming that the distance between the last quadrupole and the IP remains about 20 m, and considering an outer quadrupole radius of 25 cm, a scheme with two separate final quadrupoles requires a minimum crossing angle of $\theta_c \geq 25$ mrad. The corresponding transverse crab deflecting voltage is

$$V_\perp = \frac{cE \tan\theta_c/2}{e\omega_{\rm rf}\sqrt{\beta_x^* \beta_x^{\rm crab}}} \,, \tag{43}$$

where E is the beam energy, $\omega_{\rm rf}$ the angular rf frequency, and $\beta_x^{\rm crab}$ the beta function at the cavity.

Table 12 compares the crab-cavity parameters required for such type LHC upgrade with those designed for the KEK B factory. The LHC requires about 100 times more deflecting voltage, primarily due to the increased beam energy. Note also that the rf frequency of 1.3 GHz chosen would be too high for the present nominal LHC bunch length.

Table 12. Comparison of crab cavities parameters for KEKB with those for an advanced LHC upgrade.

variable	symbol	KEKB HER	LHC
beam energy	E	8.0 GeV	7 TeV
RF frequency	$f_{\rm rf}$	508.9 MHz	1.3 GHz
half crossing angle	$\theta_c/2$	11 mrad	12.5 mrad
IP beta function	β_x^*	0.33 m	0.25 m
cavity beta function	$\beta_x^{\rm crab}$	100 m	2000 m
required kick voltage	$V_\perp$	1.44 MV	144 MV

3.3 Stronger Magnets

In order to reach a higher energy in the LEP/LHC tunnel, stronger magnets are absolutely needed. These stronger magnets would also be in line with the historial trend, evidenced in Fig. 2. There exist s.c. magnet materials which can sustain much higher fields and current densitities than NbTi, the material used so far for all accelerator magnets. A candidate material which could approximately double the maximum field of the magnets is Nb_3Sn. Table 13 summarizes the historical evolution of the field strengths achieved in Nb_3Sn magnets. Nb_3Sn is more brittle than NbTi, which complicates the cable fabrication and the processing procedures, but recent progress bodes well for the future.

3.4 Emittance Evolution

The synchrotron radiation amplitude damping time is[75]

$$\tau_z J_z = \left(\frac{3(m_p c^2)^3}{e^2 c^3 r_p Z^2}\right) \frac{1}{B^2 E} \left(\frac{C}{2\pi\rho}\right) \approx \frac{16644\text{hr}}{E[\text{TeV}]B[\text{T}]^2} \left(\frac{C}{2\pi\rho}\right) \frac{A^4}{Z^4}. \tag{44}$$

The damping decrement is defined as

$$\delta = \frac{T_0}{n_{\rm IP}\tau_{x,y}} \approx 5.7 \times 10^{-13}\, E[\text{TeV}]^2 B[\text{T}] \frac{Z^3}{A^4} \tag{45}$$

where we have assumed $n_{\rm IP} = 2$ interaction points.

Radiation damping could improve the beam-beam limit, a hypothesis which is supported by the much higher tune shifts achieved in electron-positron colliders as com-

Table 13. Evolution of Nb_3Sn Magnets.[74]

year	group	type	field/gradient
1982	CERN	quad	71 T/m
1983	CERN/Saclay	dipole	5.3 T
1985	LBL	dipole D10	8 T
1986	KEK	dipole	4.5 T
1988	BNL	dipole	7.6 T
1991	CERN-ELIN	dipole	9.5 T
1995	LBNL	hybrid dipole D19H	8.5 T
1995	UT-CERN	dipole MSUT	11.2 T
1996	LBNL	dipole D20	13.3 T
2001	LBNL	common coil dipole	14.4 T

pared with hadron colliders. Measurements and simulations have been fitted by[76]

$$\xi_{\max} \propto 0.009 + 0.021\ (\delta/10^{-4})^{0.5}. \tag{46}$$

Figure 41 illustrates this dependence. Superimposed on the curve representing Eq. (46) are data from hadron colliders and from LEP. The points for the future hadron colliders were chosen on top of the predicted curve. The figure demonstrates that even for the next and the next-to-next generations of hadron colliders, the damping decrement is still too small to noticably enhance the maximum beam-beam tune shift.

A more important consequence of synchrotron radiation is the shrinkage of the emittance during the store. As mentioned earlier the situation is still different from electron storage rings, as the damping time is of the order of hours and not milliseconds.

The equilibrium emittance due to synchrotron radiation is[75]

$$\epsilon_{x,N}^{\mathrm{SR}} \approx \frac{55}{32\sqrt{3}} \frac{\lambda\!\!\!^{-}_A}{J_x} \left(\frac{\gamma^3}{Q_\beta^3}\right) \left(\frac{C}{2\pi\rho}\right)^3. \tag{47}$$

For both LHC-II and VLHC, this 2–3 orders smaller than the design emittance, implying the possibility of excessive beam-beam forces, and the generation of beam halo and background.

However, an equilibrium emittance of much larger value will be reached much earlier, namely at the time when the radiation damping is balanced by intrabeam scattering.

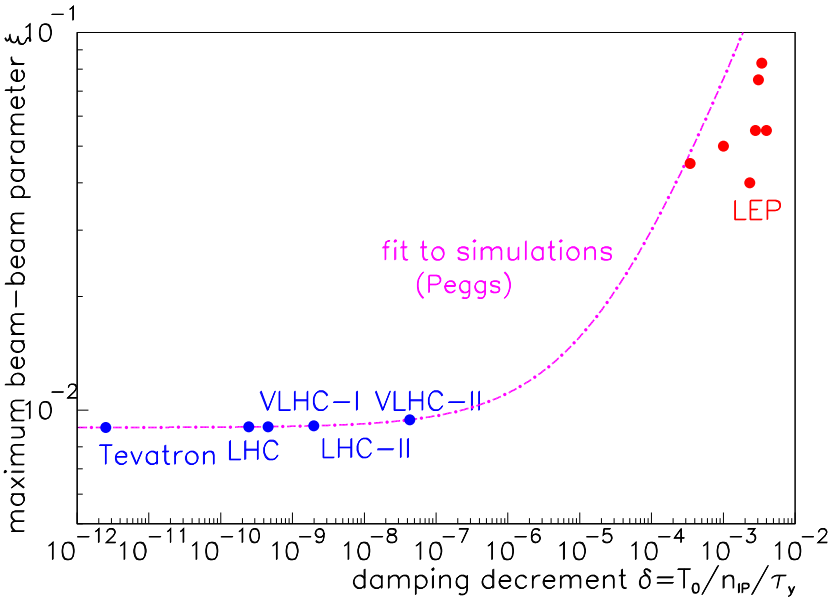

Fig. 41. Tune shift parameter vs. damping decrement. (LEP data courtesy of R. Assmann; LEP was not beam-beam limited)

The beam size growth rate from intrabeam scattering is[77]

$$\frac{1}{\tau_{x,\mathrm{IBS}}} \approx \frac{c r_p^2 N_b L_c}{16 Q \epsilon_{x,N}^2 \sqrt{\kappa}\sqrt{\kappa+1}\gamma\sigma_z\sigma_\delta} \tag{48}$$

where $L_c \approx 20$. Asymptotically, for $\gamma \gg Q_\beta$, one expects that $1/\tau_{\delta,\mathrm{IBS}} \approx 1/\tau_{x,\mathrm{IBS}}$ and $\sigma_\delta \approx Q_\beta^{3/2}\sqrt{\epsilon_x/\rho}$.[77] Equating the intrabeam-scattering growth rate and the radiation damping yields the following expression for the equilibrium emittance[78]:

$$\epsilon_{x,N}^{\mathrm{IBS}} = \frac{\rho^{5/6} N_b^{1/3}}{Q_\beta \gamma^{7/6}} \left(\frac{Z f_{\mathrm{rf}} e V_{\mathrm{rf}}}{cE\kappa(\kappa+1)}\right)^{1/6} \left(\frac{C}{2\pi\rho}\right)^{1/6} \left(\frac{3 r_p L_c}{16}\right)^{1/3}, \tag{49}$$

where f_{rf} denotes the rf frequency, V_{rf} the total rf voltage, $\kappa = \epsilon_y/\epsilon_x$ the asymptotic emittance ratio as determined by linear coupling and spurious vertical dispersion.

To give a concrete example, we take the LHC-II parameters of Table 3. Note that these assume $\kappa = 1$, which can be achieved by skew qudrupoles and/or a proper choice of betatron tunes. Figure 42 shows the predicted emittance variation as a function of time, and Figs. 43, 44, and 45 the bunch population, beam-beam tune shift, and luminosity, respectively. The result is encouraging: the luminosity initially stays high and almost constant, while the beam-beam tune shift only slowly and moderately increases.

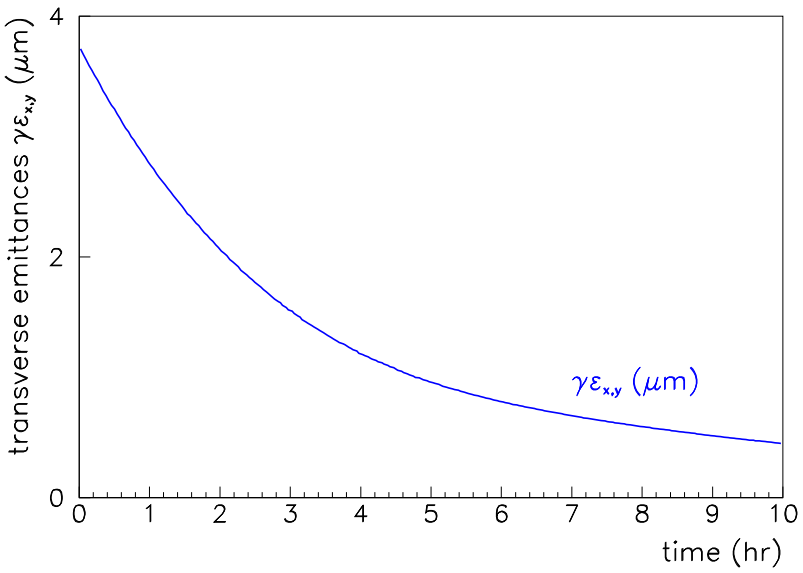

Fig. 42. Evolution of transverse emittance vs. time in LHC-II at 28 TeV centre-of-mass energy, for the parameters of Table 3. The simulation includes synchrotron radiation damping, intrabeam scattering, and particle consumption in the collision.

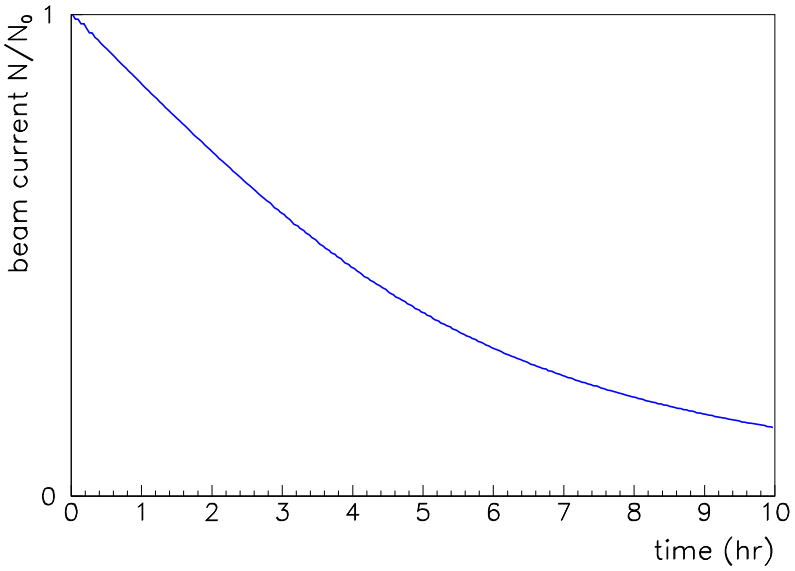

Fig. 43. Evolution of beam current vs. time in LHC-II at 28 TeV centre-of-mass energy, for the parameters of Table 3.

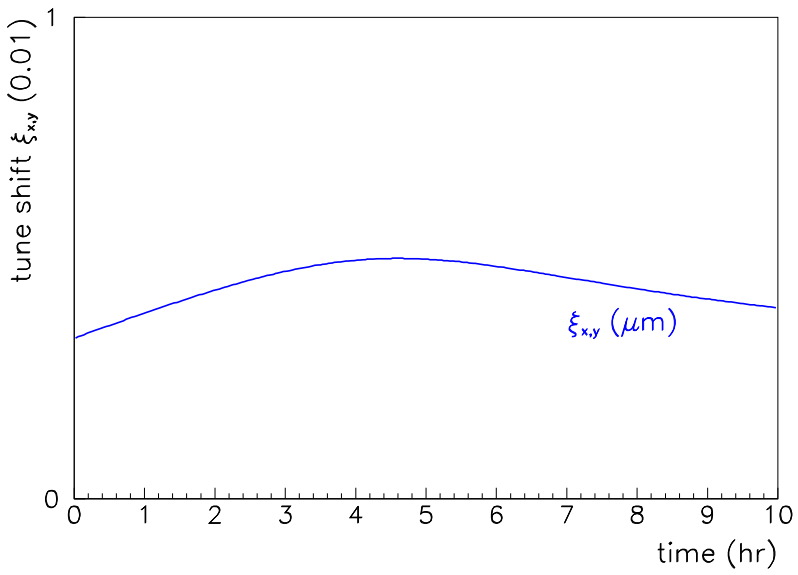

Fig. 44. Evolution of beam-beam tune shift vs. time in LHC-II at 28 TeV centre-of-mass energy, for the parameters of Table 3.

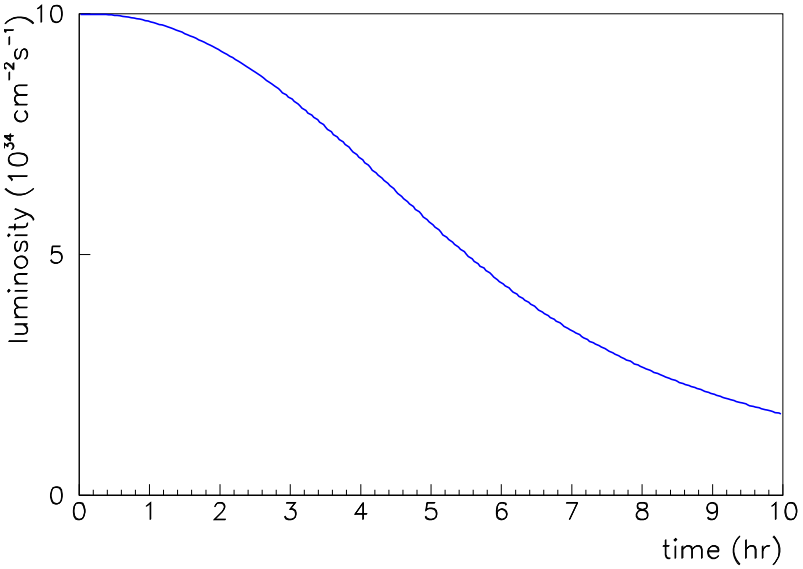

Fig. 45. Evolution of luminosity vs. time in LHC-II at 28 TeV centre-of-mass energy, for the parameters of Table 3.

3.5 Collective Effects

One of the most harmful collective effects is the loss of Landau damping for higher-order longitudinal modes. If the bunch becomes too short the frequency spread due to the nonlinearity of the rf decreases, and Landau damping may be lost. The condition for stability is

$$\sigma_s \geq \frac{C}{2\pi}\left[\frac{\pi^3}{6}\frac{N_b f_{\rm rev} e}{h_{\rm rf}^3 V_{\rm rf}}\,{\rm Im}\left(\frac{Z_L}{n}\right)_{\rm eff}\right]^{1/5} . \qquad (50)$$

Figure 46 shows that in the contemplated scenario for LHC-II Landau damping would be lost after about 3 hours. Longitudinal noise excitation[79] could maintain a minimum bunch length and thereby stabilize the beam, as is illustrated in Fig. 47.

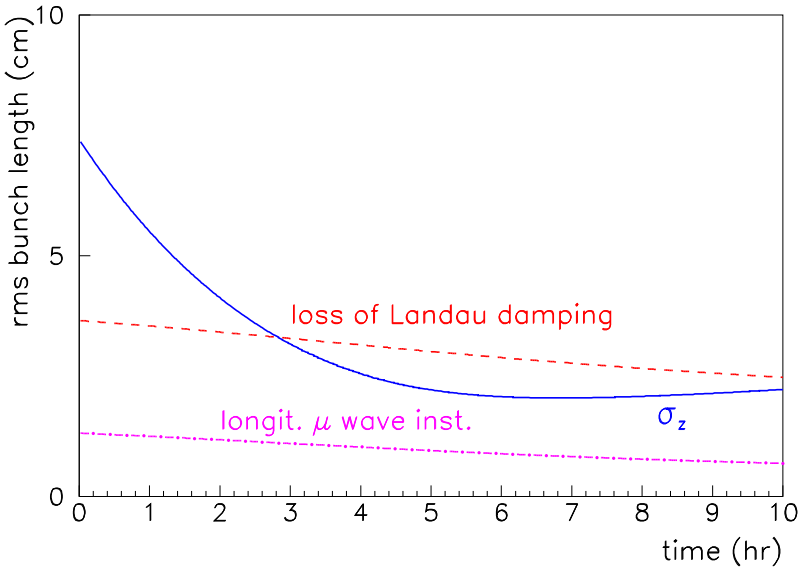

Fig. 46. Evolution of the rms bunch length during a store in LHC-II, and the instability thresholds for ${\rm Im}(Z_L/n)_{\rm eff} \approx 0.1\,\Omega$ (as estimated for LHC), for 28 TeV centre-of-mass energy, and the parameters of Table 3.

As for the LHC, other collective effects that may occur are the longitudinal microwave instability, the transverse coupled-bunch resistive-wall instability, and the electron cloud.

Figure 48 displays the simulated arc heat load in the LHC due to the electron cloud as a function of bunch spacing. For bunch spacings shorter than the nominal, the heat load can easily increase by an order of magnitude. Only when the spacing becomes

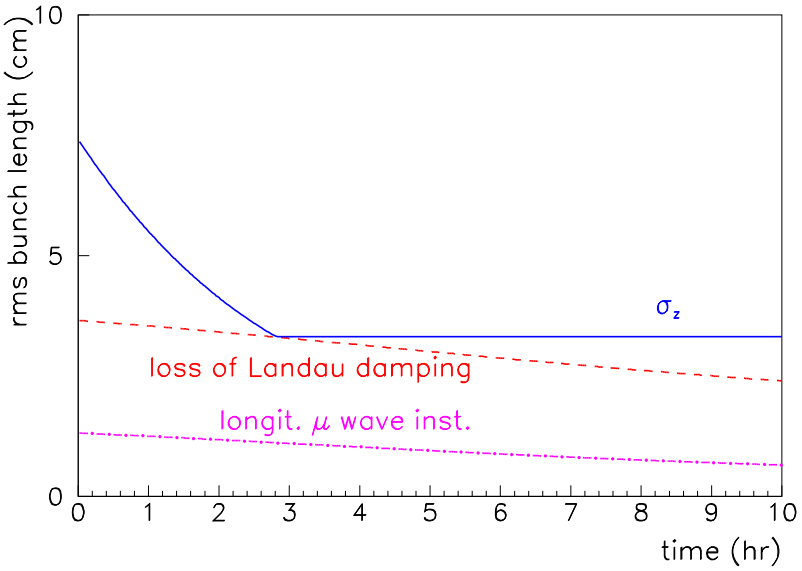

Fig. 47. Evolution of the rms bunch length during a store in LHC-II, and the instability thresholds for $\mathrm{Im}(Z_L/n)_{\mathrm{eff}} \approx 0.1\ \Omega$ (as estimated for LHC) when after 3 hours rf noise is added to maintain a constant longitudinal emittance of $\epsilon_L \geq 0.104$ eVs.

comparable to the bunch length, and we approach the limit of a coasting beam, does the heat load again decrease.

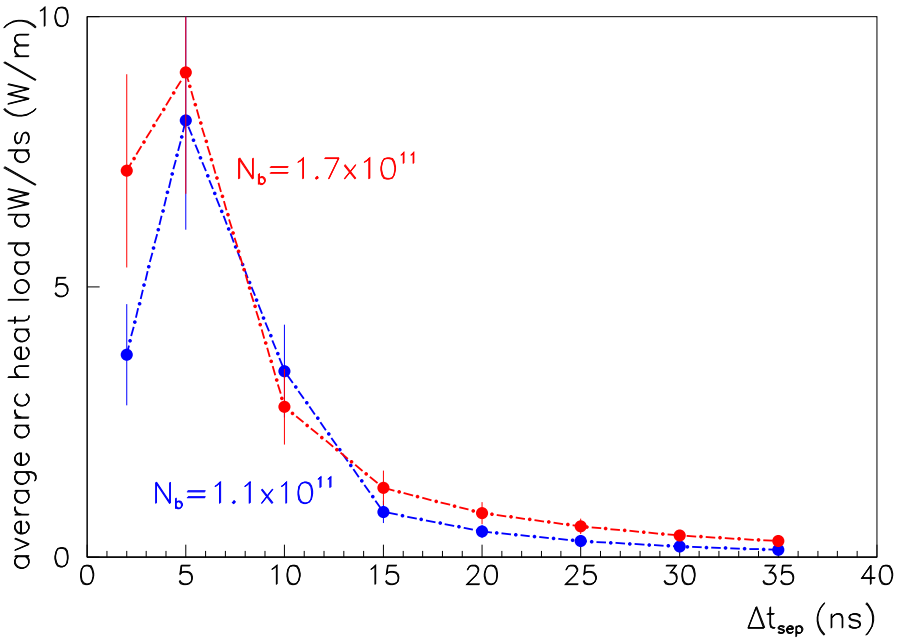

Fig. 48. Average LHC arc heat load as a function of bunch spacing, for a maximum secondary emission yield of $\delta_{\max} = 1.1$, a beam energy of 7 TeV, and two different bunch populations N_b.

3.6 Total Current and Synchrotron Radiation

The total beam current could be limited either by magnet quenches due to gas scattering, or by the maximum tolerable synchrotron radiation power,

$$P_{\mathrm{SR}} = \frac{C_\gamma E^4 N_b n_b c}{C\rho} = U_0 f_{\mathrm{rev}} n_b N_b, \tag{51}$$

where U_0 denotes the energy loss per turn, and $C_\gamma = 4\pi/3\ r_p/(m_p c^2)^3 \approx 0.778 \times 10^{-17}\ \mathrm{m/GeV^3}$ (m_p is the proton mass, and r_p the classical proton radius). Using the expressions for the luminosity and the beam-beam tune shift, Eq. (51) can be rewritten as[75]

$$P_{\mathrm{SR}} = \left(\frac{8\pi r_p^{3/2}}{\sqrt{3cE_A}}\right) \frac{\kappa}{1+\kappa^2} \frac{E^{3/2} L \beta_x^*}{\xi\sqrt{J_z \tau_z}} \sqrt{\frac{C}{2\pi\rho}}\ . \tag{52}$$

This implies the following scaling.[80] If the magnetic field is held constant, then $J_z\tau_z \propto 1/E$ and the radiation power increases as $P_{\rm SR} \propto E^2 L$. On the other hand, if the magnetic field follows the historical evolution, $B \propto E^{1/2}$, we obtain $J_z\tau_z \propto 1/E^2$ and the power grows as $P_{\rm SR} \propto E^{5/2} L$. In the next generation of hadron colliders, the power per unit length deposited by synchrotron radiation is already of the order of 1 W/m, and this scaling indicates much higher power levels for the machines which follow.

It is not easy and rather inefficient to absorb this energy inside the magnets at a temperature of a few Kelvin, even using a beam screen. For the VLHC, P. Bauer et al. have proposed a more efficient scheme, which is based on discrete warm photon-stops inserted into the beam pipe.[81] Such photon stops would considerably improve the efficiency and could reduce the wall plug power required for cooling by an order of magnitude. However, the stops have to be retracted at injection and they contribute to the beam-pipe impedance.

3.7 The VLHC

The VLHC design study has made great progress recently,[82] and a complete report has been published before Snowmass 2001.[7] The VLHC circumference is almost 10 times that of LHC, and the costs are kept low, by staging the project, and by economizing the magnets. The first stage uses 2-T magnets, whose design comprises a small 100-kA superconducting transmission line surrounded by the beam pipe and by warm iron yokes, which determine the shape of the field. The operating margin of several such designs has been verified in a 100-kA test loop at Fermilab.

A single tunnel can house the stage-1 magnets, and at a later time the higher-field stage-2 magnets, which will increase the centre-of-mass energy to values close to 200-TeV.

A complete site layout has been proposed, adjacent to the Tevatron, the latter serving as an injector. The layout foresees two collision points, both close to the Fermilab site, and it includes a bypass line for the lower-energy stage-1 ring around the detectors, which is needed, once the stage-2 is operational.

4 Conclusions

Hadron colliders have performed exceedingly well in the past. The LHC will break new territory. With 14 TeV centre-of-mass energy, and a luminosity of 10^{34} cm^{-2}s^{-1}, it

will surpass all previous colliders. The LHC design is based on the experience gained at the ISR, SPS, the Tevatron, HERA, RHIC, and other machines. The underlying assumptions are rather conservative.

Yet, the accelerator physicists face various exciting challenges, *e.g.*, related to magnet design, cryogenics, long-range beam-beam collisions, strong-strong beam-beam collisions, radiation damping — which for the first time is stronger than intrabeam scattering —, and the electron cloud.

Beyond the nominal LHC, studies have started on LHC luminosity and energy upgrades. A more ambitious 2-stage Very Large Hadron Collider has been proposed in the US. The second stage of the VLHC could reach an energy of 175 TeV in the centre of mass. The problems confronted by these future projects include the development of new magnets with either higher field or much reduced production costs, the possibly large circumference, the increased synchrotron radiation, and again the electron cloud. A further new development are 'quasi-continuous' beams or superbunches. These might provide a path towards significantly higher luminosity. They may also allow for a reduced beam-beam tune shift, suppress the electron-cloud build up and avoid PACMAN bunches, *i.e.*, bunches with unfavorable long-range collisions.

In conclusion, profiting from enhanced synchrotron radiation, the LHC upgrades and the VLHC hold the promise of further substantial advancements in energy and luminosity at sustainable power levels and costs.

Acknowledgements

I would like to thank the organizers for inviting to me to lecture at this school and the participants for enlightening and lively discussions. I also thank G. Arduini, R. Assmann, P. Bagley, P. Bauer, F. Bordry, L. Bottura, D. Brandt, O. Brüning, I. Collins, K. Cornelis, A. Faus-Golfe W. Fischer, J. Gareyte, O. Gröbner, H. Grote, G. Guignard, W. Herr, J.B. Jeanneret, J.M. Jimenez, C. Johnstone, J. Jowett, E. Keil, J.-P. Koutchouk, K.-H. Mess, K. Ohmi, S. Peggs, F. Pilat, B. Richter, L. Rossi, F. Ruggiero, G. Rumolo, F. Schmidt, R. Schmidt, E. Shaposhnikova, V. Shiltsev, M. Syphers, T. Taylor, R. Thomas, A. Verdier, L. Vos, J. Wei, and many others for material, inspiration, and helpful comments.

References

[1] E. Keil, CERN 72-14, 1972.

[2] A. Hofmann et al., CERN ISR-OP-TH/80-19, 1980.

[3] L. Evans, 1987 IEEE PAC, Washington, 1987; M. Harrison, R. Schmidt, EPAC 88 Nice, 1988; V. Hatton, 1991 IEEE PAC, San Francisco, 1991.

[4] G. Dugan, 14th HEACC, Tsukuba, 1989.

[5] K. Ohmi, Phys. Rev. Lett. 75, 1526, 1995; F. Zimmermann, CERN LHC PR 95, 1997; O. Gröbner, IEEE PAC97 Vancouver; K. Ohmi, F. Zimmermann, Phys. Rev. Lett. 85, 3821, 2000; G. Rumolo et al., PRST-AB 012801, 2001; CERN SL workshops Chamonix X & XI, 2000 and 2001.

[6] S. Klein, LBL-PUB-45566, 2000; J.B. Jeanneret, SL/AP Beam Physics Note 41; D. Brandt, LHC Project Report 450, 2000.

[7] See web site *http://vlhc.org*

[8] A.G. Ruggiero, *Hadron Colliders at Highest Energies and Luminosities*, World Scientific, Singapore, 1998.

[9] P. Bagley, private communication (2001).

[10] A. Chao, S. Peggs, R. Talman, presentations at Mini-Workshop: "The effect of synchrotron radiation in the VLHC," BNL, Sept. 18–20, 2000.

[11] The web page of the LHC project is *http://wwwlhc.cern.ch*

[12] Proceedings of the Chamonix workshops X and XI are accessible at *http://cern.web.cern.ch/CERN/Divisions/SL/news/news.html*

[13] Accelerator Physics Group of the CERN SL Division at *http://wwwslap.cern.ch*

[14] O. Brüning, private communication (2001).

[15] L. Bottura, Proc. Chamonix X, p. 238, CERN-SL-2000-007 DI (2000); Proc. Chamonix XI, p. 238, CERN-SL-2001-003-DI (2001).

[16] E. Keil, private communication (2001).

[17] W. Herr, "Effects of PACMAN bunches in the LHC," LHC Project Report 39 (1996).

[18] J. Irwin, SSC-233 (1989).

[19] Y. Papaphilippou and F. Zimmermann, PRST-AB **2**, 104001 (1999).

[20] F. Zimmermann, CERN LHC Project Note 250 (2001).

[21] F. Ruggiero and F. Zimmermann, “Luminosity Optimization by Increasing Bunch Length or Crossing Angle,” report in preparation.

[22] J.-P. Koutchouk, Proc. IEEE PAC 2001 Chicago (2001).

[23] V.D. Shiltsev et al., PRST-AB, 2:071001,1999

[24] R.E. Meller, R.H. Siemann, IEEE Trans. Nucl. Sci. **28**, p. 2431 (1981)

[25] K. Yokoya, Y. Funakoshi, E. Kikutani, H. Koiso, J. Urakawa, KEK-PREPRINT-89-14 (1989).

[26] O. Brüning, private communication (2001).

[27] A.W. Chao and B. Zotter, ‘Landau Damping,’ in ‘Handbook of accelerator physics and engineering’, by A.W. Chao (ed.), M. Tigner (ed.), Singapore (1999).

[28] Y. Alexahin, CERN-SL-96-064-AP (1996).

[29] M.P. Zorzano and F. Zimmermann, PRST-AB **3**, 044401 (2000).

[30] Y. Alexahin, H. Grote, W. Herr, M.-P. Zorzano, Proc. HEACC’01 Tsukuba (2001).

[31] F. Ruggiero, CERN SL/95-09, 1995.

[32] J.T. Rogers, CBN 96-14, 1996.

[33] J. Gareyte, CERN SL-Note-2000-056 AP (2000).

[34] V. Lebedev, VLHC miniworkshop, SLAC March 2001.

[35] M. Blaskiewicz et al., VLHC miniworkshop, SLAC March 2001. SLAC-PUB-8800 (2001).

[36] L.R. Evans, IEEE PAC New York (1999).

[37] F. Bordry, Proc. Chamonix XI, p. 250, CERN-SL-2001-003-DI (2001).

[38] D. Cocq, O.R. Jones, H. Schmickler, 8th Beam Instrumentation Workshop, SLAC, USA, CERN SL-98-062 BI (1998).

[39] F. Schmidt, R. Tomas, A. Faus-Golfe, Proc. IEEE PAC 2001 Chicago (2001).

[40] F. Sonnemann, Ph.D. Thesis, CERN-THESIS-2001-004 (2001).

[41] J.B Jeanneret, Chamonix X, CERN-SL-2000-007 DI (2000).

[42] K.H. Mess, R. Schmidt, Proc. Chamonix XI, p. 284, CERN-SL-2001-003-DI (2001).

[43] J.M. Zazula and S. Péraire, LHC Project Report 112 (1997).

[44] J. Jowett, Proc. LHC99 beam-beam workshop, CERN/SL/99-039 AP, p. 63 (1999).

[45] W. Herr, H. Grote, in LHC Project Report 502 (2001).

[46] R. Garoby, Proc. Chamonix XI, p. 32, CERN-SL-2001-003-DI (2001).

[47] R. Cappi et al., Proc. Chamonix XI, p. 29, CERN-SL-2001-003-DI (2001).

[48] T. Bohl et al., CERN SL-Note-2001-040 MD (2001).

[49] G. Arduini et al., PAC 2001, Chicago (2001).

[50] G. Arduini, Chamonix XI, CERN-SL-2001-003 DI (2001).

[51] O. Gröbner, HEACC'77, Protvino (1977).

[52] V. Baglin et al, Electron Emission of Copper," LHC-Project-Report-472 (2001).

[53] O. Gröbner, "Beam induced multipacting," IEEE PAC 97, Vancouver (1997).

[54] F. Zimmermann, Proc. PAC'2001 Chicago, USA, CERN-SL-2001-035 (AP) (2001).

[55] F. Zimmermann, in Proc. Chamonix X, CERN-SL-2000-007 (2000); Chamonix XI, CERN-SL-2001-003-DI (2001).

[56] O. Gröbner, private communication (2000).

[57] W. Höfle, Chamonix X, CERN-SL-2000-007-DI (2000).

[58] G. Rumolo and F. Zimmermann, Proc. Int. workshop in Two-Stream Instabilities," KEK Tsukuba (2001).

[59] B. Richter, SLAC memo, unpublished (2000).

[60] G. Rumolo and F. Zimmermann, Proc. Int. Workshop on Two-Stream Instabilities," KEK Tsukuba (2001).

[61] G. Rumolo, F. Zimmermann, H. Fukuma, K. Ohmi, Proc. PAC'2001 Chicago, USA, and CERN-SL-2001-040 (AP) (2001).

[62] K. Ohmi, F. Zimmermann, Phys. Rev. Lett. 85, 3821, 2000;

[63] K. Ohmi, F. Zimmermann, E. Perevedentsev, HEACC2001, Tsukuba, Japan, and CERN-SL-2001-011 AP (2001).

[64] K. Cornelis, Chamonix XI, CERN-SL-2001-003-DI.

[65] G. Rumolo, F. Zimmermann, Proc. PAC'2001 Chicago, USA, and CERN-SL-2001-041 (AP) (2001).

[66] F. Zimmermann, LHC Project-Report 95, and SLAC-PUB-7425 (1997).

[67] V. Baglin et al., LHC-Project-Report-472 (2001).

[68] J.M. Jimenez, 'The SPS as Vacuum Test Bench for the Electron Cloud Studies with LHC Type Beams,' August 2001.

[69] F. Ruggiero (ed.) et al., 'Feasibility Study of LHC Luminosity and Energy Upgrades,' report in preparation (2001).

[70] J. Gareyte, private communication (2001).

[71] K. Takayama, J. Kichiro, M. Sakuda, M. Wake, 'Superbunch Hadron Colliders,' submitted to Physical Review Letters (2001).

[72] F. Ruggiero, private communication (2001).

[73] R. Palmer, 'Energy Scaling, Crab Crossing, and the Pair Problem,' in the DPF Summer Study Snowmass '88, 'High Energy Physics in the 1990s,' SLAC-Pub-4707 (1988).

[74] A. den Ouden, U. of Twente, CERN LHC Seminar (2001).

[75] E. Keil, Proc. PAC 97, p. 104, 1996, and Proc. 34th Eloisatron Workshop, Erice 1996.

[76] E. Keil and R. Talman, Particle Accelerators vol. 14, 1993; S. Peggs, LHC99 beam-beam workshop, Geneva, 1999; R. Assmann and K. Cornelis, EPAC 2000 Vienna, 2000.

[77] J. Wei, "Intrabeam Scattering Scaling for Very Large Hadron Colliders," unpublished draft (2001); J. Bjorken and S. Mtingwa, Part. Acc. 13, 115 (1983).

[78] F. Zimmermann, Proc. HEACC'2001 , Tsukuba, Japan, CERN-SL-2001-009 AP (2001).

[79] T. Toyama et al., Proc. EPAC 2000, Vienna; similar methods were used at the S$p\bar{p}$S by T. Linnecar.

[80] F. Zimmermann, HEACC2001, Tsukuba, Japan, CERN-SL-2001-009 AP (2001).

[81] P. Bauer, et al., IEEE PAC 2001 Chicago (2001).

[82] M. Syphers, S. Peggs, M4 Group Summary Talk at Snowmass 2001.

RECENT RESULTS FROM FOCUS

Brian O'Reilly

University of Colorado
Boulder, CO 80309

Representing the FOCUS Collaboration

ABSTRACT

Some recent results from the Fermi National Accelerator Laboratory (Fermilab) fixed target experiment FOCUS are presented. In particular we discuss a study of the decay $D^0 \to K^+\pi^-$ and its implications for mixing, a search for direct CP violation and some new measurements of charm particle lifetimes.

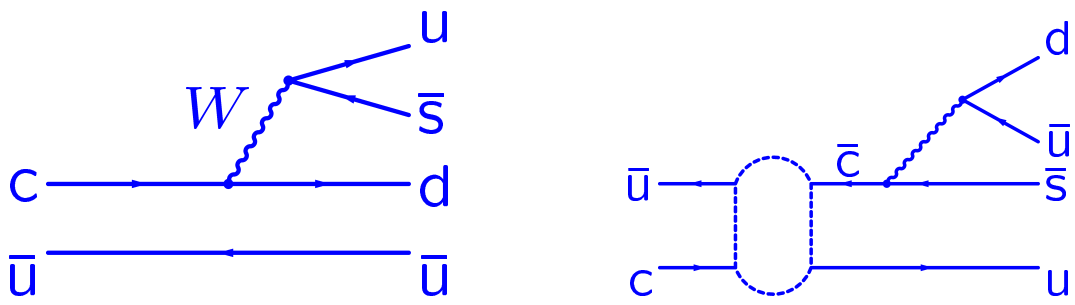

Fig. 1. Feynman diagrams of the DCS and mixing processes for $D^0 \to K^+\pi^-$

1 Introduction

Precise measurements of charmed particle decays challenge existing theoretical methods of calculating the dynamics of heavy quark decays. Additionally mixing and CP violation are expected to be small in this sector making it an ideal place to search for non-Standard Model physics. FOCUS is a photoproduction experiment which took data during the 1996-1997 fixed target run at Fermilab. Bremsstrahlung of electrons and positrons with an endpoint energy of approximately 300 GeV produces a photon beam. These beam photons interact in a segmented beryllium-oxide target and produce charmed particles. The average photon energy for events which satisfy our trigger is $\simeq$ 180 GeV. FOCUS uses an upgraded version of the E687 spectrometer which is described in detail elsewhere.[1] Charged decay products are momentum analyzed by two oppositely polarized dipole magnets. Tracking is performed by a system of silicon vertex detectors in the target region and by multiwire proportional chambers downstream of the interaction. Particle identification is performed by three threshold Čerenkov counters, two electromagnetic calorimeters, an hadronic calorimeter, and by a system of muon detectors.

2 The decay $D^0 \to K^+\pi^-$

The decay $D^0 \to K^+\pi^-$ (throughout this article the charge conjugate mode is implied unless otherwise indicated) may occur either as a doubly Cabibbo suppressed (DCS) decay or through mixing of the D^0 into a $\bar{D^0}$ followed by the Cabibbo Favored (CF) decay $\bar{D^0} \to K^+\pi^-$. Therefore the wrong-sign (WS) decay rate R_{WS} can have contributions from both DCS and from mixing. The time-dependent rate for WS decays relative to the CF process is:

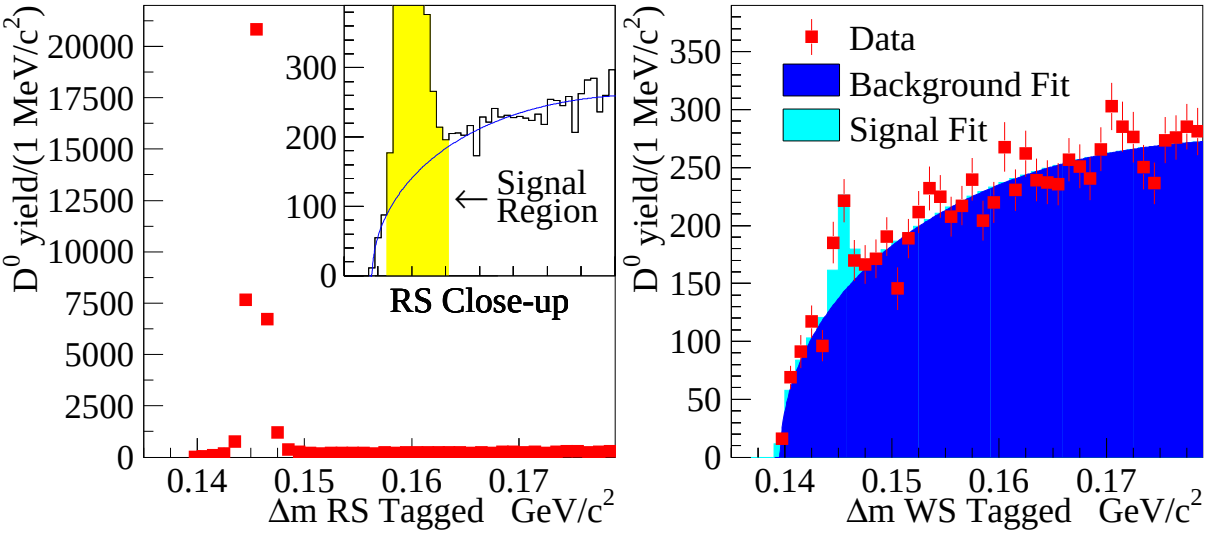

Fig. 2. RS and WS signals for the decay $\bar{D}^0 \to K^+\pi^-$

$$R(t) = \left[R_{DCS} + \sqrt{R_{DCS}}y\prime t + \frac{(x\prime^2 + y\prime^2)}{4}t^2\right]e^{-t} \tag{1}$$

where t is in units of the D^0 lifetime and we have used the strong phase (δ) rotated convention of CLEO[2] where $y\prime = y\cos\delta - x\sin\delta$ and $x\prime = x\cos\delta + y\sin\delta$. $x = \Delta m/\Gamma$ and $y = \Delta\Gamma/2\Gamma$ are the usual mixing parameters. Using Monte Carlo (MC) generated sample of $\bar{D}^0 \to K^+\pi^-$ decays, (with an input lifetime of 413 fs for the D^0[3]), we can calculate the expected number of WS events by re-weighting each accepted MC event with a weight given by:

$$W_i = \frac{N_{data}}{N_{MC}}\left(R_{DCS} + \sqrt{R_{DCS}}y\prime t_i + \frac{(x\prime^2 + y\prime^2)}{4}t_i^2\right), \tag{2}$$

where t_i is the generated proper time for event i, and $N_{data}(N_{MC})$ is the number of accepted RS events in the data(MC). Summing Equation 2 over all accepted MC events and dividing by N_{data} we obtain:

$$R_{WS} = R_{DCS} + \sqrt{R_{DCS}}y\langle t\rangle + \frac{(x\prime^2 + y\prime^2)}{4}\langle t^2\rangle. \tag{3}$$

The averages $\langle t\rangle$ and $\langle t^2\rangle$ are obtained from the generated lifetime of the accepted MC events. We find $\langle t\rangle = 1.578 \pm 0.008$ and $\langle t^2\rangle = 3.61 \pm 0.03$ where the error is a systematic obtained by comparing the reconstructed MC averages to those obtained in the data. We now have an expression for R_{WS}, which is the quantity we measure experimentally, in terms of R_{DCS} and the mixing parameters $x\prime$ and $y\prime$.

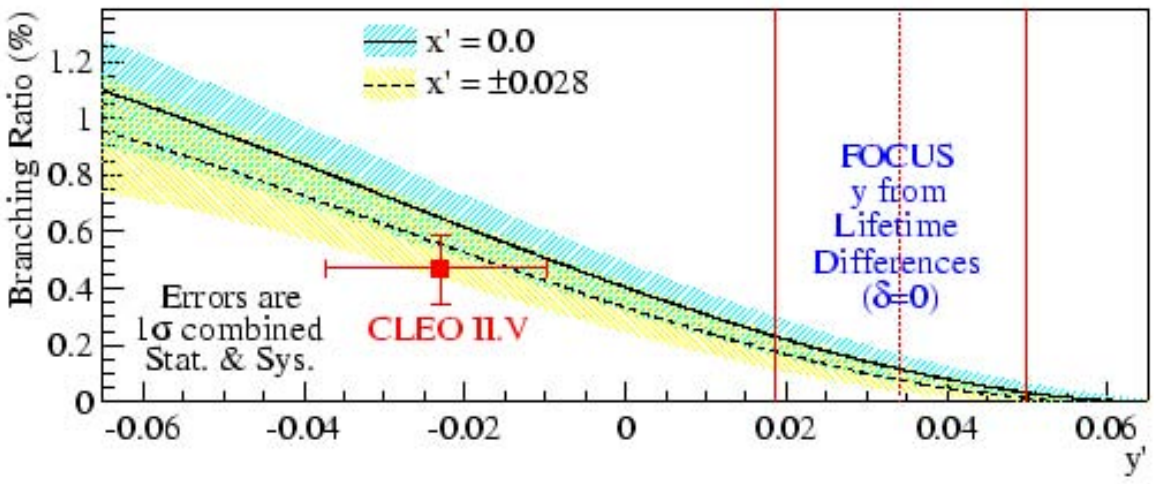

Fig. 3. R_{DCS} vs. $y\prime$. Contours are plotted for two values of $x\prime$ which cover the 95% CL of the CLEO.II.V result.

We identify right sign (RS) and WS decays by "tagging" the soft pion in the decay $D^{*+} \to D^0(\to K^-\pi^+)\pi^+$. In Figure 2 we show the signals obtained. The WS signal is obtained by fitting the D^0 yield in bins of the $D^{*+} - D^0$ mass difference and the fit is a sum of a background contribution and a scaled signal shape from the RS. We measure $R_{WS} = (0.404 \pm 0.085 \pm 0.025)\,\%$ with a WS yield of 149 ± 31 events.

In Figure 3 we use our measured value for R_{WS} to plot R_{DCS} as a function of $y\prime$. The CLEO.II.V and FOCUS[4] results are also included for comparison purposes. The FOCUS result comes from a measurement of y using the lifetime difference between CP even and CP mixed final states. The CLEO.II.V result comes from a direct measurement of R_{DCS}. One can only compare the FOCUS y value to the others by assuming that the strong phase $\delta = 0$.

If charm mixing is sufficiently small then Equation 3 tells us that $R_{WS} \approx R_{DCS}$. In Table 1 we list the existing measurements of this branching ratio under the assumption of no mixing or CP violation. Our analysis of the decay $D^0 \to K^+\pi^-$ has been published in Reference 5.

Table 1. Measurements of R_{DCS} assuming no charm mixing or CP violation.

Experiment	R_{DCS}(%)	Events
CLEO[6]	$0.77 \pm 0.25 \pm 0.25$	19.1
E791[7]	$0.68^{+0.34}_{-0.33} \pm 0.07$	34
Aleph[8]	$1.77^{+0.60}_{-0.56} \pm 0.31$	21.3
CLEO.II.V[2]	$0.332^{+0.063}_{-0.065} \pm 0.040$	44.8
This Study[5]	$0.404 \pm 0.085 \pm 0.025$	149

3 Search for Direct CP violation in the decays $D^+ \to K_S\pi^+$ and $D^+ \to K_S K^+$

CP is violated when the decay rate of a particle differs from that of its CP conjugate.[9] In the Kobayashi-Maskawa ansatz this arises due to the non-vanishing phase in the Cabibbo-Kobayashi-Maskawa matrix when the decay amplitude has contributions from at least two quark diagrams with differing weak phases. In addition final state interactions (FSI) must provide a strong phase shift. In the Standard Model direct CP violation in the charm meson system is predicted to occur at the level of 10^{-3} or below.[10] The mechanism usually considered is the interference of the tree and penguin amplitudes in singly-Cabibbo suppressed (SCS) decays. In the decay $D^+ \to K_S\pi^+$, (The charge conjugate state is implied unless stated otherwise), the Cabibbo favored (CF) and doubly-Cabibbo suppressed (DCS) amplitudes contribute coherently with, perhaps, a different weak phase. In addition the isospin content of the DCS amplitude differs from that of the CF case so we can expect a non-trivial strong phase shift. Several authors have commented on the effect of K^0 mixing on the CP asymmetry for this decay mode and the possibility of using it to search for new physics.[11,12]

Differences in the weak two-body non-leptonic decay amplitudes of charmed mesons are almost certainly due to FSI. These effects tend to be large in the charmed system making it an ideal laboratory for their study.[13] The isospin amplitudes and phase shifts in $D \to KK$, $D \to K\pi$ and $D \to \pi\pi$ decays can be extracted from measurements of the branching fractions.[14] For example the magnitude of the I=3/2 amplitude can be obtained directly from the $D^+ \to \bar{K}^0\pi^+$ partial width.[15]

Previous studies of $D^+ \to K_S\pi^+$ and $D^+ \to K_S K^+$ have concentrated on measuring relative branching ratios.[16,17] FOCUS has made the first measurement of the CP asymmetry for these decays.

Table 2. Yields and relative efficiencies for $D^+ \to K_S\pi^+$, $D^+ \to K_SK^+$ and $D^+ \to K^-\pi^+\pi^+$. Efficiency numbers are quoted relative to the average of the $D^+ \to K^-\pi^+\pi^+$ and $D^- \to K^+\pi^-\pi^-$ efficiencies. We generated a very large Monte Carlo sample to render the statistical error on the efficiencies negligible.

Decay Mode	$D^+ \to K_S\pi^+$ cuts		$D^+ \to K_SK^+$ cuts	
	Yield	Eff.	Yield	Eff.
$D^+ \to K_S\pi^+$	5080 $\pm$110	0.58	4487 $\pm$ 96	0.51
$D^- \to K_S\pi^-$	5518 $\pm$110	0.56	4770 $\pm$ 96	0.50
$D^+ \to K_SK^+$	-	-	495 $\pm$ 38	0.26
$D^- \to K_SK^-$	-	-	454 $\pm$ 42	0.25
$D^+ \to K^-\pi^+\pi^+$	84750 $\pm$512	1.01	84750 $\pm$ 512	1.01
$D^- \to K^+\pi^-\pi^-$	91520 $\pm$508	0.99	91520 $\pm$ 508	0.99

To correct for production induced asymmetries we make a double ratio using a CF decay where no CP violation is expected to occur. We measure

$$A_{CP} = \frac{\eta\left(D^+\right) - \eta\left(D^-\right)}{\eta\left(D^+\right) - \eta\left(D^-\right)}; \tag{4}$$

where (for example)

$$\eta\left(D^+\right) = \frac{N(D^+ \to K_S\pi^+)}{N(D^+ \to K^-\pi^+\pi^+)} \tag{5}$$

i.e. the ratio of the yields in each decay mode corrected for efficiency and acceptance. This last quantity is equivelant to the relative branching ratio for the decay in question.

The invariant mass signals for the decays $D^+ \to K_S\pi^+$ and $D^+ \to K_SK^+$ can be seen in Figures 4 and 5. The reconstruction efficiencies, relative to that of the $D^+ \to K^-\pi^+\pi^+$ normalizing mode are listed, together with the yields, in Table 2.

In Table 3 we present our relative branching ratio measurements and compare them to the current world average. Finally in Table 4 we show our A_{CP} measurements for the $D^+ \to K_S\pi^+$ and $D^+ \to K_SK^+$ decay modes. This work has now been published in reference [18].

Fig. 4. $D^+ \to K_S\pi^+$ and $D^- \to K_S\pi^-$ signals.

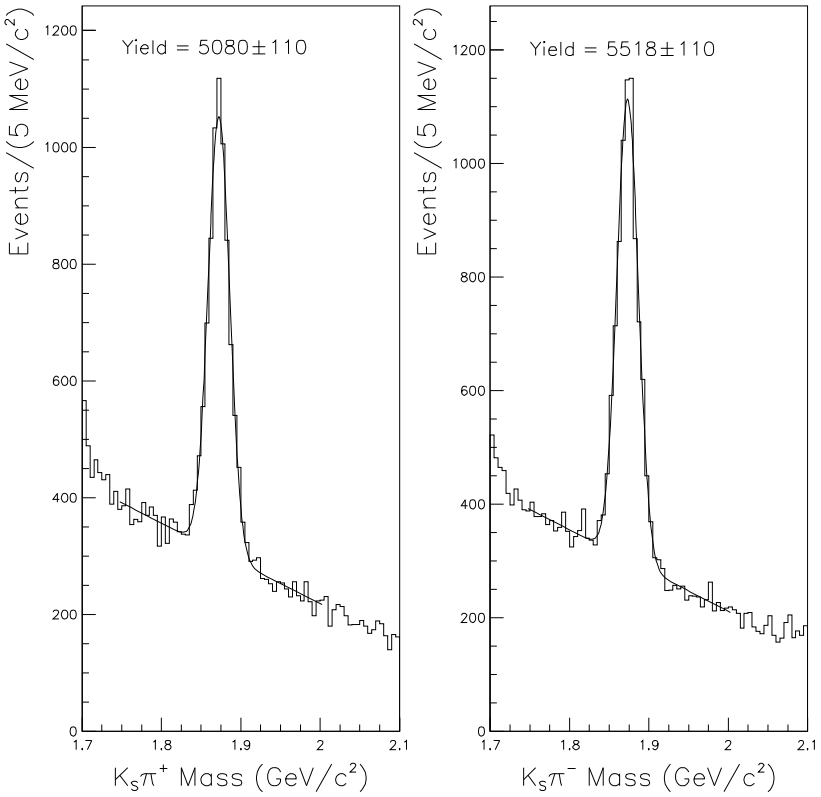

Table 3. Relative branching ratio results. The first error is statistical and the second is systematic. We account for the decay chain $\bar{K}^0 \to K_S \to \pi^+\pi^-$ by multiplying our K_S numbers by a factor of 2.91 assuming that $\Gamma(D^+ \to \bar{K}^0\pi^+) = 2 \times \Gamma(D^+ \to K_S\pi^+)$; we then quote these results in terms of $\bar{K}^0$.

Measurement	Result	PD G Average [3]
$\frac{\Gamma(D^+\to\bar{K}^0\pi^+)}{\Gamma(D^+\to K^-\pi^+\pi^+)}$	$(30.60 \pm 0.46 \pm 0.32)\%$	$(32.0 \pm 4.0)\%$
$\frac{\Gamma(D^+\to\bar{K}^0K^+)}{\Gamma(D^+\to K^-\pi^+\pi^+)}$	$(6.04 \pm 0.35 \pm 0.30)\%$	$(7.7 \pm 2.2)\%$
$\frac{\Gamma(D^+\to\bar{K}^0K^+)}{\Gamma(D^+\to\bar{K}^0\pi^+)}$	$(19.96 \pm 1.19 \pm 0.96)\%$	$(26.3 \pm 3.5)\%$

Fig. 5. $D^+ \to K_S K^+$ and $D^- \to K_S K^-$ signals.

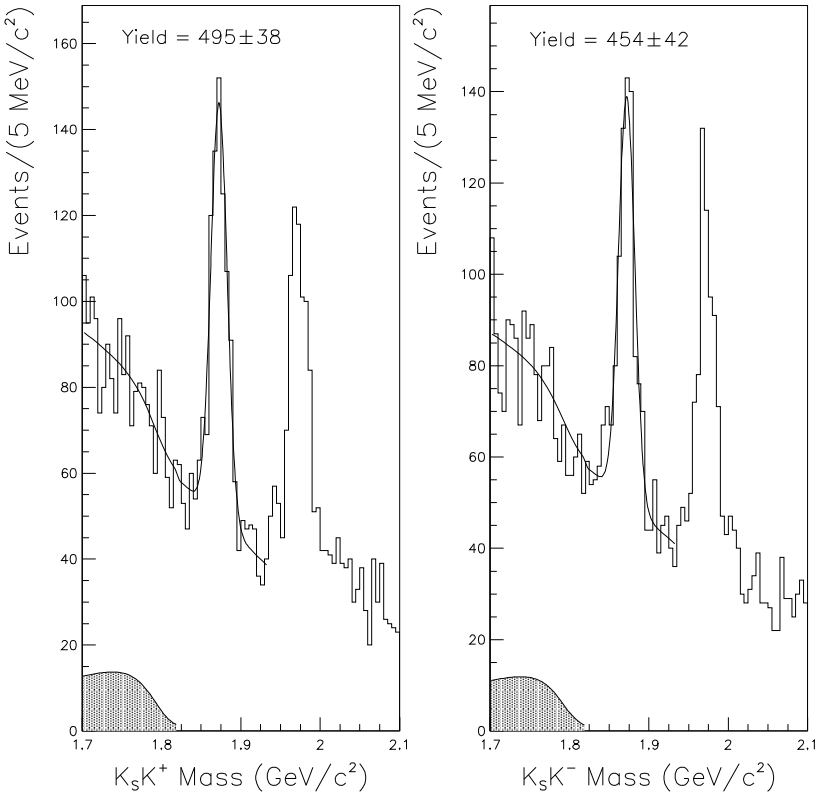

Table 4. CP asymmetry measurements. The first error is statistical and the second is systematic.

Measurement	Result
$A_{CP}(K_S\pi^+)$ w.r.t. $D^+ \to K^-\pi^+\pi^+$	$(-1.6 \pm 1.5 \pm 0.9)\%$
$A_{CP}(K_S K^+)$ w.r.t. $D^+ \to K^-\pi^+\pi^+$	$(+6.9 \pm 6.0 \pm 1.5)\%$
$A_{CP}(K_S K^+)$ w.r.t. $D^+ \to K_S\pi^+$	$(+7.1 \pm 6.1 \pm 1.2)\%$

4 Charm Lifetimes

Precise measurements of the lifetimes of charmed mesons and baryons provide an important test of our theoretical understanding of the dynamics of heavy quarks. Heavy Quark Effective theory relies on expansions in the heavy quark mass, extensions to the charm sector may be complicated by the lower mass of the charm quark. Lifetime differences between mesons and baryons in the beauty sector tend to be significantly scaled down relative to those of charm. Thus it has been said that "the decays of charm hadrons act as nature's microscope into the decays of beauty hadrons".[19]

Historically, FOCUS is the only collaboration to have measured all of the weakly decaying charm particle lifetimes. Our excellent lifetime resolution (on the order of 30fs for some decays), and high statistics ensure that our new measurements will once again dominate the world average. Only with the advent of high statistics charm analyses from the e^+e^- factories will more precise measurements be forthcoming. In that event our precision measurements with tightly controlled systematics should serve as a benchmark by which to evaluate and control systematic effects unique to the collider regime.

Currently we have published results for the Ξ_c^+ and are in the process of finalizing the Λ_c^+,D^+,D^0,D_s^+,Ξ_c^0 and Ω_c^0 lifetime analyses.

4.1 Ξ_c^+ Lifetime

We have measured the Ξ_c^+ lifetime using five different decay modes which occur in eight distinct topologies. In Figure 6 we show the signal distributions and the lifetime fit is shown in Figure 7. Our analysis was based on a yield of 532.4 ± 30.4 events. We measured a lifetime of $439 \pm 22 \pm 9$ fs where the first error is statistical and the second is systematic. In Figure 8 we compare this result to previous experimental measurements. The improvement over previous results is obvious as is the fact that the world average for the Ξ_c^+ lifetime will increase. Several authors[20–23] predict that $\tau(\Xi_c^+) > \tau(\Lambda_c^+)$ where the inequality represents a factor of about 1.3. Using the Λ_c^+ lifetime average of PDG, CLEO and SELEX,[3,24,25] (0.1916 ± 0.0054 ps) and the Ξ_c^+ lifetime reported in this paper, one obtains a ratio $\tau(\Xi_c^+)/\tau(\Lambda_c^+) = 2.29 \pm 0.14$, which differs significantly from the prediction. This work is now published in reference [26].

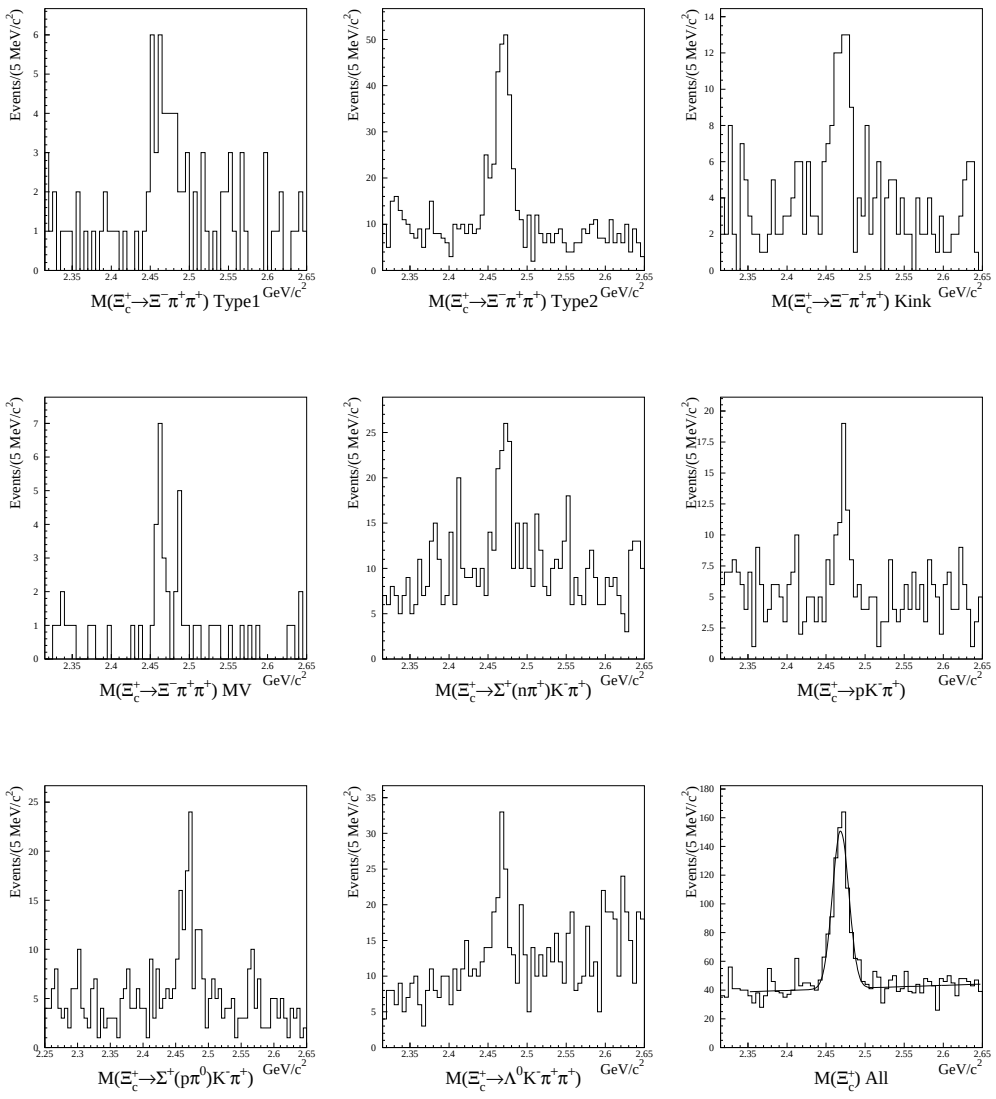

Fig. 6. Signals for the five different decay modes used in our determination of the Ξ_c^+ lifetime. The bottom right plot is the sum of all the modes.

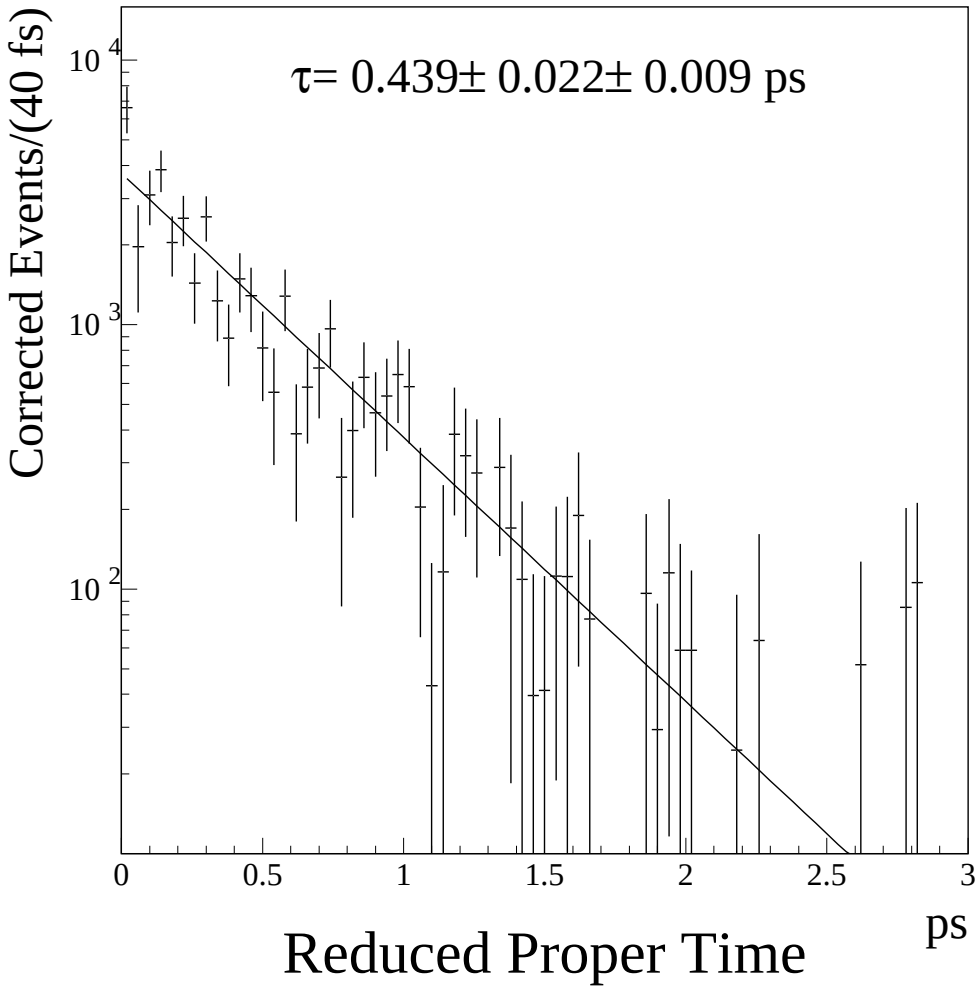

Fig. 7. The combined lifetime fit to the background subtracted, Monte Carlo corrected, reduced proper time distribution obtained from all the studied decay modes.

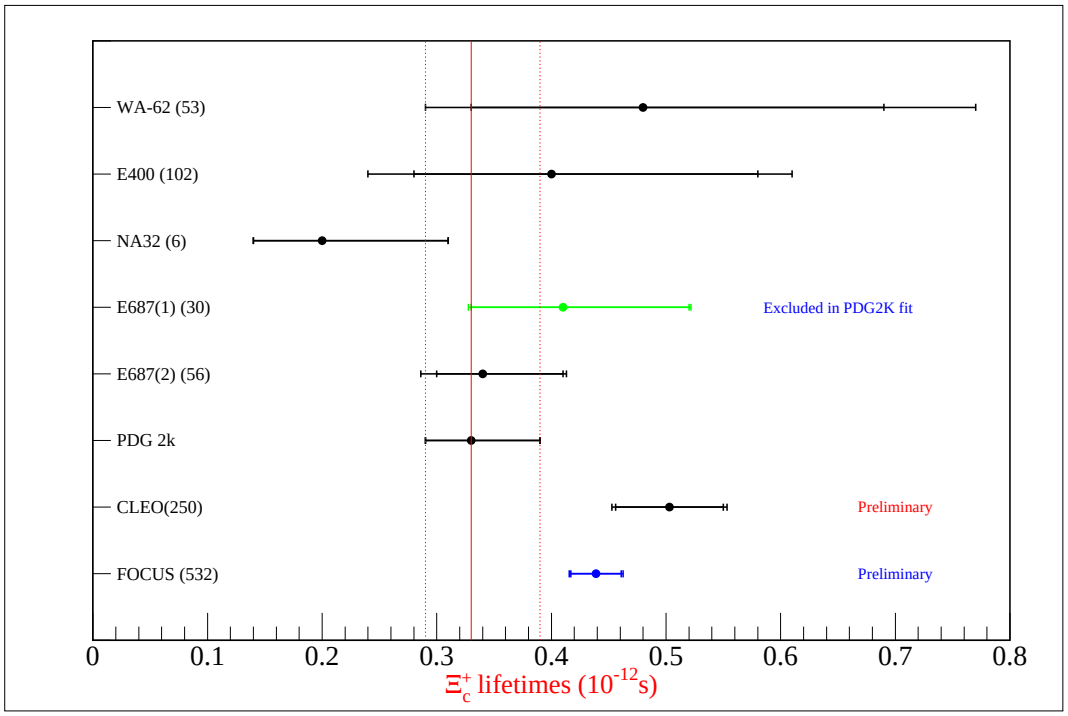

Fig. 8. Comparison of experimental measurements of the Ξ_c^+ lifetime. Note that the CLEO and FOCUS numbers are no longer preliminary. The number in parentheses after each experiment's name represents the number of events they used to make their measurement.

4.2 Λ_c^+ Lifetime

We have also measured the lifetime of the Λ_c^+ from the decay mode $\Lambda_c^+ \to \mathrm{p}K^-\pi^+$. We reconstructed 8034 ± 122 events and determined the lifetime to be $204.6 \pm 3.4 \pm 2.5$ fs. This result has been submitted for publication.[27]

We are analysing two decay modes for the Ξ_c^0 which occur in five separate topologies. In Figure 11 the signals used in our preliminary determination of this lifetime are plotted. Using 137 ± 18.8 events we measure the lifetime to be 109^{+10}_{-9} fs.

In addition to these analyses we are also working on the lifetime measurements for the D^0, D^+ and Ω_c^0.

5 Summary

We have presented some recent results from FOCUS on mixing, direct CP violation limits and charm lifetimes. Many of these analyses are soon to be, or have already been

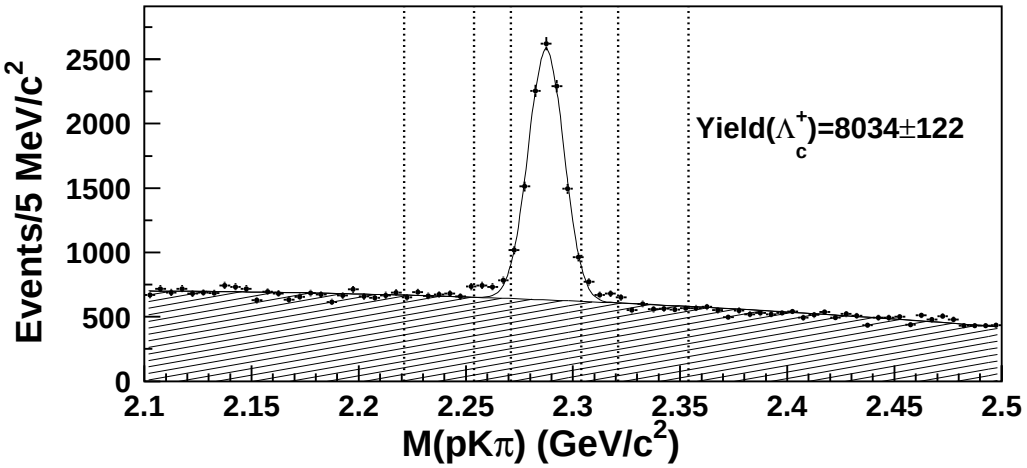

Fig. 9.

published. In addition we are working on a wide variety of other topics such as Dalitz analyses, $D\bar{D}$ production, semileptonic branching ratios and form factors, five-body hadronic decays and the spectroscopy of excited charm mesons.

References

[1] P. L. Frabetti *et al.*, Nucl. Instrum. Meth. **A320**, 519 (1992).

[2] R. Godang *et al.*, Phys. Rev. Lett. **84**, 5038 (2000).

[3] D. E. Groom *et al.*, Eur. Phys. J. **C15**, 1 (2000).

[4] J. M. Link *et al.*, Phys. Lett. **B485**, 62 (2000).

[5] J. M. Link *et al.*, Phys. Rev. Lett. **86**, 2955 (2001).

[6] D. Cinabro *et al.*, Phys. Rev. Lett. **72**, 1406 (1994).

[7] E. M. Aitala *et al.*, Phys. Rev. **D57**, 13 (1998).

[8] R. Barate *et al.*, Phys. Lett. **B436**, 211 (1998).

[9] I. Bigi and A. Sanda, *CP Violation* (Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK, 2000).

[10] F. Buccella, M. Lusignoli, G. Miele, A. Pugliese, and P. Santorelli, Phys. Rev. **D51**, 3478 (1995).

[11] I. I. Bigi and H. Yamamoto, Phys. Lett. **B349**, 363 (1995).

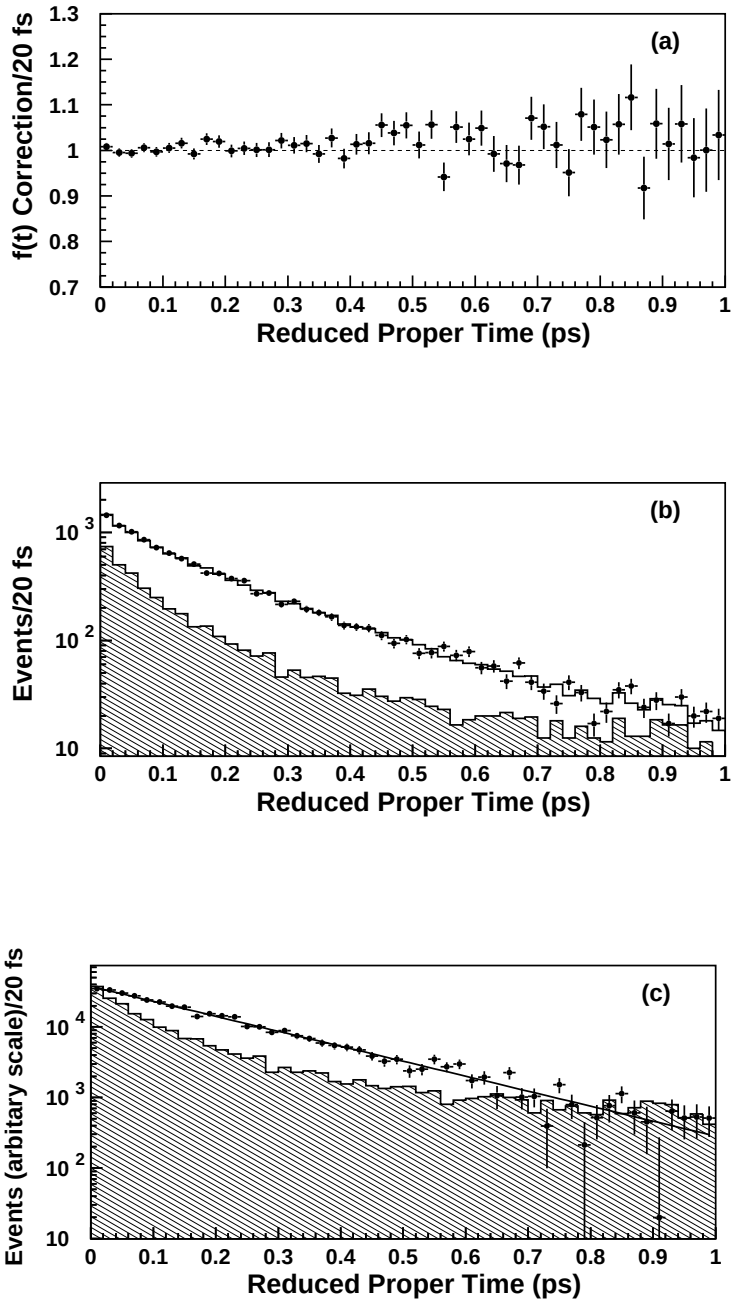

Fig. 10. (a) The f(t′) correction function. Deviation from a flat line indicates the correction from a pure exponential; (b) the lifetime distribution for all decays in the data signal region (points) and the fit (histogram). The shaded distribution shows the lifetime distribution of the background component in the signal region; (c) The lifetime distribution for Λ_c^+ decays (points), $i.e.$ the sideband subtracted and f(t′) corrected yield. The line is a pure exponential with the fitted lifetime and the shaded region gives the background. An arbitrary yield scale is used because of the particular normalization of f(t′).

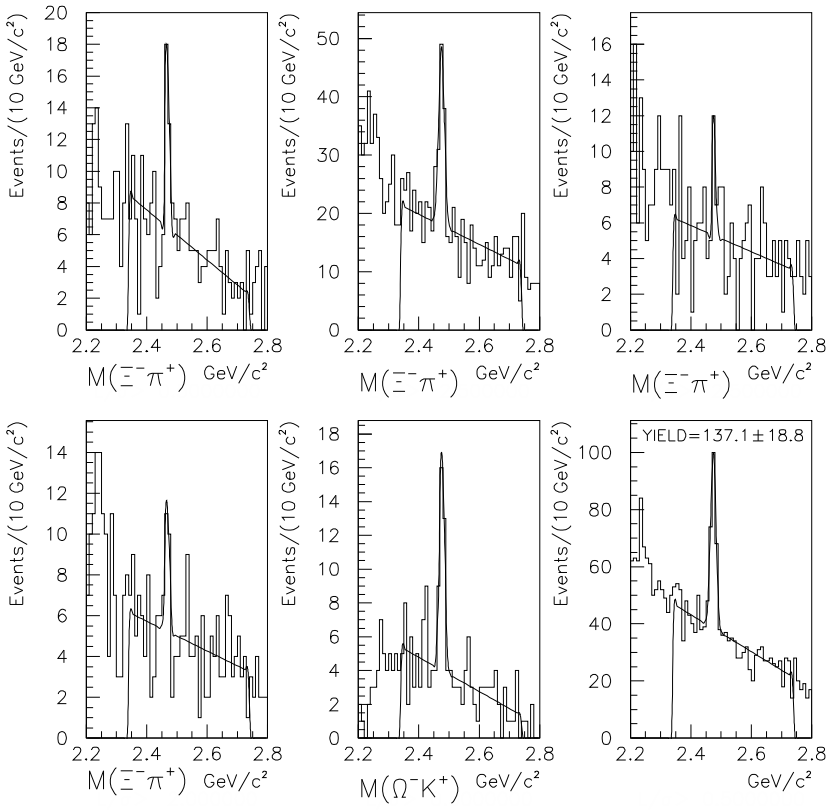

Fig. 11. Signals for the two decay modes used in our determination of the Ξ_c^0 lifetime. The bottom right plot is the sum of all the modes.

[12] H. J. Lipkin and Z.-Z. Xing, Phys. Lett. **B450**, 405 (1999).

[13] J. L. Rosner, Phys. Rev. **D60**, 114026 (1999).

[14] M. Bishai *et al.*, Phys. Rev. Lett. **78**, 3261 (1997).

[15] M. Bauer, B. Stech, and M. Wirbel, Z. Phys. **C34**, 103 (1987).

[16] J. C. Anjos *et al.*, Phys. Rev. **D41**, 2705 (1990).

[17] P. L. Frabetti *et al.*, Phys. Lett. **B346**, 199 (1995).

[18] J. M. Link *et al.*, Phys. Rev. Lett. **88**, 041602 (2002).

[19] G. Bellini, I. I. Y. Bigi, and P. J. Dornan, Phys. Rept. **289**, 1 (1997).

[20] B. Blok and M. A. Shifman, (1991).

[21] B. Guberina and B. Melic, Eur. Phys. J. **C2**, 697 (1998).

[22] I. I. Y. Bigi, (1996), talk given at Workshop on Heavy Quarks at Fixed Target (HQ 96) , St. Goar, Germany, 3-6 Oct 1996 (UND-HEP-96-BIG06).

[23] H.-Y. Cheng, Phys. Rev. **D56**, 2783 (1997).

[24] A. H. Mahmood *et al.*, Phys. Rev. Lett. **86**, 2232 (2001).

[25] A. Kushnirenko *et al.*, Phys. Rev. Lett. **86**, 5243 (2001).

[26] J. M. Link *et al.*, Phys. Lett. **B523**, 53 (2001).

[27] J. M. Link *et al.*, Submitted to Phys. Rev. Lett., hep-ex/0202001 (2002).