DEVELOPMENTS IN SOLID STATE VERTEX DETECTORS

C. J. S. Damerell

Rutherford Appleton Laboratory

Chilton, Didcot, Oxon, England

LECTURES PRESENTED AT THE SLAC SUMMER INSTITUTE, 1984

CONTENTS

1. PHYSICS MOTIVATION

Physicists long ago concluded that the variety of observable particles is so great that one could reasonably expect nature to have provided a simplifying sub-structure. All theoretical progress regarding this sub-structure, of quarks and gluons and beyond, is based on clues provided by the composite particles which we can observe in experiments. One property of the observable particles is their lifetimes, and high precision vertex detectors enable the lifetime range to be pushed below the previously measurable limits. Such measurements may at first sight seem to be unrelated to the interesting problem of hadron structure, but in fact they can be of great importance in precisely this area, as we shall see. But let us start by taking a global look at the physically observable particles and their lifetimes.

The term stable particles is usually taken to include those having lifetimes in excess of about 10^{-8} s. If produced in high energy collisions, such particles have decay lengths of the order 1 metre or more, which means that the electrically charged ones ($\mu^{\pm}$, $\pi^{\pm}$, $K^{\pm}$, $\overset{(-)}{p}$, ...) can be tracked in conventional detectors and often identified by means of Čerenkov counters. The neutral ones (γ, K_L°, n, ...) are in general observable by calorimetry; however, neutrinos can usually be inferred only by means of missing energy.

We shall use the term long-lived particles to describe those having lifetimes of the order of 10^{-10} s (K_s°, Λ°, $\Sigma^{\pm}$, ...) which (depending on the experiment) may be visible directly or at least will be recognised by having decay products whose tracks clearly do not point back to the production vertex. Such decay products have a projected distance of closest approach to the primary vertex (impact parameter) of typically 1 cm.

Until about 10 years ago, all known particles had lifetimes in one or other of these categories, or else were subject to what we shall call prompt decays, ie. lifetimes too short to allow for direct experimental observation by any known technique. In this class we have the π^0 ($\tau \sim 10^{-16}$ s) and all the resonances (η°, Σ°, ω, ρ, Δ, ...) with lifetimes 10^{-18} to 10^{-23} s or (more relevant to experiments) mass widths of 1 keV to 100 MeV. Such particles are observable only via their decay products as peaks in effective mass distributions, or in formation experiments via the energy dependence of a measured cross-section, such as $\pi^+p \to \Delta^{++}$, or $e^+e^- \to J/\psi$. The observation of resonances in high multiplicity inelastic processes is notoriously difficult due to the problem of combinatorial background; one may for example have so many possible $\pi^+\pi^-$ combinations that the recognition of which pairs (if any) result from the decay $\rho^{\circ} \to \pi^+\pi^-$ becomes impossible. Quite apart from the non-observation of the quark substructure, one may often be unable to disentangle these first generation hadronic states, and be left only with the measured stable particles which are one stage further removed from the fundamental physical processes of interest. This is an unavoidable fact associated with high energy experiments.

During the past decade, a sequence of particles (the τ lepton, and hadrons such as D, F, Λ_c, B, ...) have been discovered which have lifetimes in the region 10^{-13} to 10^{-12} s. Such particles were first observed as effective mass peaks in favourable situations (low combinatorial background) but can be seen much more extensively in experiments where special high precision vertex detectors are used to recognise the finite lifetimes of the parent particles, which we shall hereafter refer to as <u>short-lived</u>. By recognising which of the charged particle tracks emerge from the decay vertex, the parent particle can be reconstructed without the combinatorial background which otherwise could completely obscure the signal. The measurement of the lifetime is a by-product, possibly a very important one.

Thus vertex detectors of the type we shall be discussing are useful in recognising short-lived particles, but are of no help in sorting out the large class of promptly decaying ones. Given that these detectors are not entirely straightforward, one might reasonably ask why we bother to do this at all. The reason is simply that the short-lived charm and bottom particles achieve their 10 orders of magnitude lifetime extensions by being ground states of matter containing quarks of higher flavours. As such they are particularly interesting. Not only that, but the predominance of <u>sequential decays</u>,

$c \to s$

$b \to c$

$t \to b$ probably

$x \to t$ possibly (where x is a quark from a hypothetical fourth generation)

.
.
.

ensures that a c or b tag will also enrich the signals for t, x, ... whose lifetimes may well be too short for direct measurement.

We shall use the nomenclature c, b, t, ... to signify the heavy quarks and C, B, T, ... to signify hadrons containing these quarks.

As already mentioned, some of the short-lived states may be observed in clean conditions without vertex detection. This becomes more difficult as the energy is raised, and some presently marginal signals (such as $t \to b$ in UA1) could be transformed into definitive experimental results with the aid of vertex detectors able to see the decays of the short-lived particles.

As a specific physics area where vertex detectors may have an important role, we consider the case of $e^+e^- \to Z^\circ$ in SLC or LEP. The Z° will decay via all kinematically allowed $q\bar{q}$ final states including the higher flavours. The Z° decay involves a large release of energy, the charged particle multiplicities will be high, and decays such as $D \to K\pi$ will be swamped by combinatorial background unless the K and π tracks can be recognised as not coming from the primary vertex. Let us now look at some areas of physics which can be studied provided we are able to distinguish the heavy flavour decays.

1.1 <u>Neutral Current Weak Coupling of Quarks.</u> The aim is to measure separately the couplings for the processes $Z^\circ \to q\bar{q}$ where $q = u, d, s, c, t, b, \ldots$ over a good angular range (say $\theta(q) \gtrsim 20^\circ$, where θ is the polar angle of the produced quark) including the distinction between quark and anti-quark. Being unable to observe the decays at the quark level, we cannot really hope to distinguish the processes involving the light quarks

$$Z^\circ \rightarrow u\bar{u}$$
$$d\bar{d}$$

or $s\bar{s}$ due to the ease with which such $q\bar{q}$ pairs are generated out of the vacuum. The situation is much more promising for tagging the Z° decays to massive quark pairs (which are very *unlikely to be generated from the vacuum) viz:*

$$Z^\circ \rightarrow c\bar{c}$$
$$b\bar{b}$$
$$t\bar{t}$$
$$\cdot$$
$$\cdot$$

Indeed, we have some techniques which can be extrapolated from our experience at lower energy. Muons may be used as a signature for D or B decay, the process $D^* \rightarrow D\pi$ may be used to enrich the $c\bar{c}$ sample, and kinematic tests based on the mass of the B or T states may be useful. But the increasing complexity of events with energy, and the presence of sequential decays, make these procedures at best problematical. What is needed for a clean signature is (for example) kaon identification in conjunction with the vertex topology. As shown in Figure 1, the emission of a positively charged kaon from the final vertex can be used to cleanly distinguish the $\bar{c}$ or $\bar{b}$ jet and *also (in conjunction with kinematic tests on particles from the* primary vertex) the $\bar{t}$ jet. A good vertex detector at SLC can give flavour tagging efficiencies of 25-50% for all of these quark states, in contrast with approximately 1% for the branching ratio times efficiency for c tagging with a D* tag.

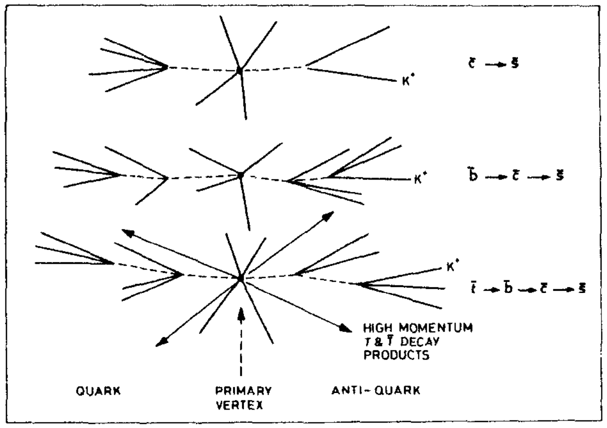
$\bar{c} \rightarrow \bar{s}$
K+
$\bar{b} \rightarrow \bar{c} \rightarrow \bar{s}$
K+
K+
$\bar{t} \rightarrow \bar{b} \rightarrow \bar{c} \rightarrow \bar{s}$
HIGH MOMENTUM
t & t̄ DECAY
PRODUCTS
QUARK
PRIMARY
VERTEX
ANTI-QUARK

Figure 1 In events where the vertex topology is observed and a charged kaon emerges from the final decay vertex, the primordial quark content of the jet may be inferred.

Once one has a good sample of $q\bar{q}$ events, one can deduce the weak couplings as follows[1]:

$g_A^2 + g_V^2$ from the partial width $Z^\circ \rightarrow q\bar{q}$,

g_V/g_A from the forward-backward asymmetry defined to be

$$A_{FB} = \left\{ \frac{\frac{d\sigma}{d\Omega}(\theta) - \frac{d\sigma}{d\Omega}(\pi - \theta)}{\frac{d\sigma}{d\Omega}(\theta) + \frac{d\sigma}{d\Omega}(\pi - \theta)} \right\}_q .$$

This is best measured with a longitudinally polarized electron beam, which will be available at SLC. Then at the Z° peak,

$$A_{FB} = P_q \left\{ \frac{P(Z^\circ) + P_e}{1 + P_e \, P(Z^\circ)} \right\} \frac{2\cos\theta}{1 + \cos^2\theta} \text{, where}$$

P_e is the polarization of the incoming electron beam (positron polarization assumed to be zero),

P_q is the natural quark polarization from unpolarized Z°s,

viz
$$\frac{2g_V/g_A}{1 + (g_V/g_A)^2} ,$$

$P(Z^\circ)$ is the Z° polarization from unpolarized beams, viz

$$\frac{2v_e/a_e}{1 + (v_e/a_e)^2} \text{, and}$$

a_e and v_e are the axial vector and vector weak couplings of the electron. Since v_e is small (≈ -0.02) the forward-backward asymmetry is small unless P_e is non-zero. In fact, 50% electron polarization is worth a factor 10 in luminosity. Given control of the electron polarization, it is preferable[2] to measure the longitudinal asymmetry in the quark production process, defined by

$$A_L = \frac{\frac{d\sigma}{d\Omega}(P_e = +) - \frac{d\sigma}{d\Omega}(P_e = -)}{\frac{d\sigma}{d\Omega}(P_e = +) + \frac{d\sigma}{d\Omega}(P_e = -)}$$

since this is more sensitive to model parameters, and less sensitive to backgrounds, detector asymmetries, radiative corrections and energy variations than a measurement of the forward-backward asymmetry.

Polarized electron beams and clean quark jet tagging will provide an important testing ground for the standard model, which makes specific predictions for both g_A and g_V for the up-type (u, c, t) and down-type (d, s, b) quarks.

1.2 <u>Flavour-Mixing Matrix</u>. The measurement of "the B lifetime" (in reality there may be different lifetimes) and the tightened bounds on $\Gamma(b \to u)/\Gamma(b \to c)$ have considerably improved the knowledge of the Kobayashi-Maskawa mixing matrix. Theoretical papers abound, and generally explain such features as the increasing mass separation and decoupling between the higher flavours. In some theories (see for example Stech[3]), the mass of the top quark and the amplitude ratio $b \to u/b \to c$ are predicted. A detailed measurement of lifetimes and decay modes of the many undiscovered physical particles (mesons and baryons) containing charm, bottom and top quarks (decay modes only in the case of top) will be of great importance. The signals from these particles in general depend on vertex detection in order that they should be pulled out of the background, and of course the vertex detector is needed for any lifetime measurements.

1.3 <u>Particle-Antiparticle Mixing</u>. The measurement of the off-diagonal elements of the K° mass matrix gave one of the first clues for the existence of a fourth (charmed) quark. In the same way, measurements of $D^\circ \bar{D}^\circ$ mixing, $B_d^\circ \bar{B}_d^\circ$ mixing or $B_s^\circ \bar{B}_s^\circ$ mixing (the first expected to be small and the second to be large)[4] could disprove the 6-quark model even before the direct observation of a higher flavour.

The distinction between B° and $\bar{B}^\circ$ or between D° and $\bar{D}^\circ$ is given unambiguously by the identification of the strange particle which emerges in Cabibbo favoured decays from the charm vertex, except in the case where this is a neutral kaon. In all cases, very useful numbers of events should be accessible at SLC, particularly since the vertex system should result in their being observed on negligible background.

According to Monte Carlo calculations (and there are many uncertainties) a sample of 10^5 events of the type $Z^\circ \to b\bar{b}$ would yield about 500 $B^\circ\bar{B}^\circ$ events with found secondary vertices and full reconstruction of the B° and $\bar{B}^\circ$. Such a sample (and the numbers are about a factor 2 higher for $D^\circ\ \bar{D}^\circ$) would not allow a precision measurement of the mixing (which theoretically cannot anyway be calculated precisely) but the observation of a several percent mixing in $D^\circ\ \bar{D}^\circ$ or $B_d^\circ\ \bar{B}_d^\circ$ would invalidate current models. Just a few events (on zero background) would be sufficient.

Another very powerful approach to the search for anomalous particle-antiparticle or flavour mixing is based on the study of semileptonic decays. Figure 2 from Reference 5 indicates the richness of information available, and demonstrates that the simple rules applicable in hadronic collisions where only $c\bar{c}$ production is significant no longer apply. The presence of like sign dileptons (for example) signifies nothing unless one knows from which vertex each lepton emerged. One should further note that decays including τ leptons are particularly characteristic topologically, and will stand out very clearly.

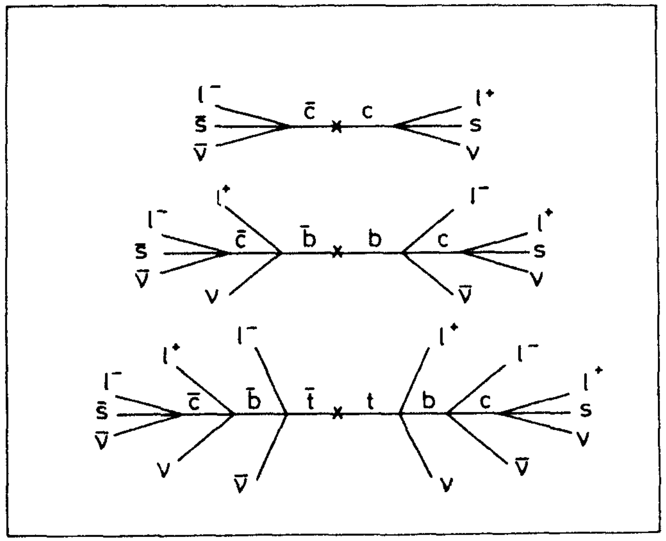

Figure 2 Semileptonic decay chains. Any lepton pair can be replaced by $q\bar{q}$.

1.4 Higgs Boson Production. This is an interesting example where a vertex detector would be used in "veto mode," to exclude events where the particles of interest emerge from a background of short-lived particle decays.

A promising production mode in searching for Higgs bosons is

$$Z^\circ \to H^\circ \ell^+ \ell^-$$

which would be seen as 2 jets (the Higgs decaying to $b\bar{b}$ or higher flavours if kinematically allowed). Given the tiny sample of H° events expected (less than 100) most estimates suggest major backgrounds from leptons from semileptonic decays of B and C. Using the vertex detector to establish the prompt origin of each lepton is a very powerful element in background rejection. The vertex detector will in addition provide evidence for the expected secondary vertices in the H° decay, where the precise vertex structure of course depends on the H° mass.

The production process $Z^\circ \to H^\circ \nu\bar{\nu}$ has a factor 3 rate advantage over the first process, but is generally considered to have very large background. Estimates for SLD[6] taking advantage of the almost complete coverage with compensated calorimetry, and using the vertex detector to single out $b\bar{b}$ final states, suggest that it may be possible to reduce backgrounds to the point where this becomes a viable process for discovering the Higgs boson.

1.5 Other New States. Particle lifetimes are very difficult to calculate theoretically. As late as mid-1982, the B lifetime was estimated[7] to be certainly less than 10^{-13} s. We have learned that increasing mass by no means implies decreasing lifetimes (as was commonly believed before the discovery of the J/ψ). It has been pointed out[8] that if there is a fourth generation with its charge $-\frac{1}{3}$ member below the top quark mass, this could well have a long lifetime. Long-lived heavy neutrinos are possible[9] and some supersymmetric particles may have measurable lifetimes.[10] Independent of theory, it is obviously an experimental necessity to take advantage of the SLC environment and set up the most sensitive vertex detection system which can be built, in order to look for the unexpected (and therefore most significant) short-lived particles which may be produced by Z° decays. The fact that all heavy flavours and leptons so far discovered have observable lifetimes in a favourable experimental environment lends confidence that there may be more to be found.

In summary, the first centimetre away from the Z° production point is a rich source of information which will have a vital bearing on many areas of physics, some of which may be completely unexpected on current theory. A vertex detector capable of defining the event topology would provide orthogonal information to that given by all other detectors in the spectrometer. Apart from measuring lifetimes, such information can provide orders of magnitude suppression of combinatorial background which would otherwise obliterate signals from heavy quark states.

In the case of e^+e^- production of the Z°, the experimental situation is particularly favourable due to the democratic coupling of the Z° to many $q\bar{q}$ states. However, in hadronic production experiments there is an additional problem, that of triggering, since the production of heavy quarks is perhaps 10^{-3} or less of the total cross-section. Vertex detectors may have a role in this area also, by (for example) recognising non-pointing tracks or changes of multiplicity. This possibility will not be pursued in these lectures, largely because electronic vertex detectors are still in their infancy and their use in triggering systems has only begun to be developed. Nevertheless, there is a substantial rate for production of charm and (probably) bottom particles in high energy hadronic collisions, and it is extremely probable that vertex detectors with some form of lifetime trigger capability will be developed in the coming years, in order to exploit the very interesting physics possibilities.

2. IMPLICATIONS FOR DETECTORS

2.1 General Remarks. We would like to specify the required precision, maximum amount of material, etc, for the ideal vertex detector. Unfortunately there are no simple answers to these questions, for two main reasons.

Firstly, one should specify the experimental application. Thus in a typical high energy fixed target experiment, multiple scattering is generally no problem. Due to the lower momenta of final state particles in a collider such as SLC, one has to be more careful about the amount of material in the vertex detectors. In LEP (same momenta as SLC, but 10 times larger beam pipe) the effects of multiple scattering are very large, imposing serious limits on the precision of vertex reconstruction.

The second important question is the one of aims. Most modestly, one could design a vertex detector which allowed only a short-lifetime tag by measuring the impact parameter of the track for which this happens to be maximal. Such a system can in fact do useful physics. It provides an enriched sample of charm and bottom events, which may be refined by other cuts. But we may wish to do better. Most ambitiously, we may design a vertex detector which can efficiently assign the tracks to their respective vertices (primary vertex, B decay vertex, C decay vertex, ...). The efficiency for doing this can never be 100% due to decay tracks which happen to be colinear with the parent track, and very short-lived decay products; remember that the most probable lifetime is always zero. But we shall take this most ambitious aim as our guide to the quality of a given vertex detector, since it corresponds in general to the total physics information which is potentially available from such a detector. Setting a more modest aim can in fact be quite misleading. Thus one may demonstrate that some detector has an 80% efficiency for heavy flavour tagging. This gives the impression that a more precise detector could only pick up an additional 20%, whereas in fact it is also able to sweep aside the ambiguities associated with all the tracks not used for the flavour tagging. A fairer representation of the two situations is provided by quoting for each detector the percentage of tracks which can be uniquely assigned to their correct parent vertices, for various classes of event ($e^+e^- \to c\bar{c}$, $b\bar{b}$, etc). Such criteria will be used in the discussions which follow.

2.2 Historical Background. The period 1976-1980 provided some essential experience regarding the importance of vertex detectors in seeing charm particles in high energy collisions. Charm had been discovered in e^+e^- annihilation, and there were indications from the ISR that the hadronic production cross-sections might be large. Experiments at Fermilab and the CERN SPS using multiparticle spectrometers were able to accumulate tens of millions of events. From these experiments, likely effective mass distributions $M(K\pi)$, $M(K\pi\pi)$, $M(pK\pi)$ etc.were plotted. Unfortunately, they showed absolutely no interesting structures at all. We were learning the hard facts that (even given a good charm production rate) the combinatorial background is overwhelming except in the very favourable environment of e^+e^- interactions just above $D\bar{D}$ threshold.

But there were some high energy experiments being run with vertex detectors, and these were making real progress. Using nuclear emulsions or special purpose bubble chambers, a small number of clean charm events began to emerge, in which the decays were seen on virtually no background. For example, Figure 3 shows a beautiful event of the type

$$\begin{tikzcd}[row sep=small, column sep=small]
\gamma p \to & \bar{D}^{\circ} \arrow[ddd, dash] & \Lambda_c^+ \arrow[d, dash] & X & & \\
& & \arrow[r] & \pi^+ & \Lambda \arrow[d, dash] & \\
& & & & \arrow[r] & p\pi^- \\
& \arrow[r] & K^+\pi^-\pi^-\pi^+ & & &
\end{tikzcd}$$

induced in emulsion by a photon of energy 25 GeV. The event, from an experiment[11] at the SPS is fully reconstructed with the aid of the Omega spectrometer, where the Λ decay products are seen.

Not only did the experiments with interaction triggers and no vertex detectors fail to see charm, but even experiments with charm-enriched triggers had problems. For example, the ACCMOR

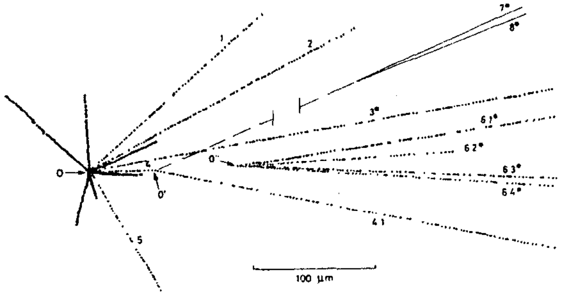

Figure 3 Associated charm production seen in nuclear emulsion.

Collaboration, in SPS experiment NA11 had a sophisticated single-electron trigger which enriched the charm content by about a factor of 20. The idea was to trigger on the semi-leptonic decay of one of the pair-produced charm particles, and to observe the hadronic decay of the other in a multi-particle spectrometer;

$$\begin{tikzcd} \pi p \arrow[r] & D \arrow[dd] & \bar{D} \arrow[d] & X \\ & & eX' & \\ & K\pi & & \end{tikzcd}$$

Inclusive $D^{\pm}$ and $\overset{(-)}{D}{}^{\circ}$ signals were visible only as 1 or 2 σ effects on backgrounds which were still large. Only in the special case of D^{*} production were the much tighter kinematic constraints adequate to yield effectively background-free events.

The ACCMOR Collaboration proceeded to upgrade the spectrometer by adding 6 planes of silicon microstrip detectors after the beryllium target. With the help of these detectors, it was possible to select the particles emerging from the secondary vertex for making mass fits. The precision on the impact parameter was typically 15 μm, and the effect on the backgrounds was dramatic, due to the rejection of most events (for which all tracks came from the primary vertex) and the rejection of most tracks from the remaining events. Given mean multiplicities for tracked charged particles in these 200 GeV/c πBe interactions of about 10, the background rejection factors from the secondary vertex cuts were as follows:

Decay Mode	Background Rejection Factor
$K\pi$	300
$K\pi\pi$	1500
$K3\pi$	7100

The result was twofold: clean charm signals on almost no background, and measurement of lifetimes.[12] This is the first example where high precision electronic tracking detectors have been successful in the observation of charm events through their finite lifetimes, and it is to be expected that there will be many more.

2.3 Implications for SLC. We now turn to the technical requirements for a vertex detector which could be used in the most interesting area of Z° decay physics at the SLC. This has the disadvantage (compared with hadronic production) that the momenta are lower and so one has to be more careful about multiple scattering, but the great advantage of the high rate for heavy flavour production and the special physics interest which we discussed in Section 1.

The detector requirements are evaluated on the basis of generated events (using the Lund Monte Carlo) of the type

$$e^+e^- \rightarrow Z^\circ \rightarrow q\bar{q} \text{ (i.e. 2-jet events).}$$

For $c\bar{c}$ events we put in the experimental lifetimes for $D^{\pm}$, D°, Λ_c and F.

For $b\bar{b}$ events, we take $\tau_B = 9 \times 10^{-13}$ s for all B states (since we know no better).

For $t\bar{t}$ we assume $M_T = 30$ GeV and prompt T decay to B states.

For decays including τ leptons, we put in the experimental τ lifetime.

Decay modes and branching ratios are left at the default values which emerge from the Lund Model (JETSET 5.21 of February 1984 in the CERN Program Library). For charm decays, these branching ratios typically agree within a factor 2 with measured values, for those decay modes for which measurements exist.

In deciding what detector characteristics are needed, we shall use the criterion of topological efficiency already mentioned. To get a feeling for the problem, let us begin by looking at some individual events from the generation program. These events are generally displayed in a "beam's eye view" with the primordial q jet directed vertically upwards in azimuth, and the $\bar{q}$ jet downwards, for clarity. The line lengths represent the track momenta, with the full-scale momentum (primary vertex to edge of figure) being approximately 5 GeV/c. Only the long-lived tracks are shown as lines; the bottom and charm particles (whether charged or neutral) are not shown explicitly since their tracks are not directly observable in the detectors. Figure 4(a) shows what is in fact a quite typical hadronization according to the Lund model. The tracks emerging from the primary vertex include several soft ones; most of the energy is carried off by the charm quarks and appears in the decay products of the D and $\bar{D}$. The charm particles travel on the order of a few mm before decay. Figure 4(b) shows a $b\bar{b}$ event in which the B happens to decay very close to the primary vertex, a common occurrence. Figure 4(c) shows a $t\bar{t}$ event; the track multiplicities are higher, the track angles are greater, and the track momenta are reduced. In this XY view the event topology is unclear. This again is frequently observed. Figure 4(d) shows the same event rotated by 90°. The topology is now clear, illustrating the desirability of building a vertex detector capable of providing more than one viewing direction for the event. For these topological

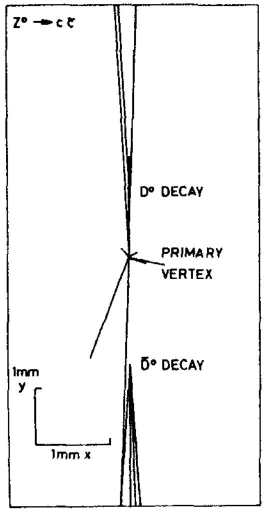

Figure 4(a) Hadronization from the Lund Monte Carlo; $Z^\circ \to c\bar{c}$. Beam's eye view of the event.

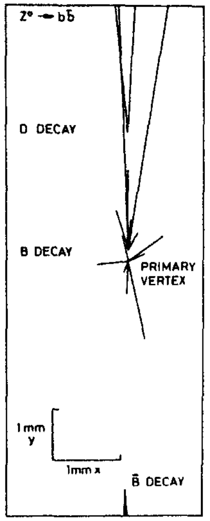

Figure 4(b) A $Z^\circ \to b\bar{b}$ decay.

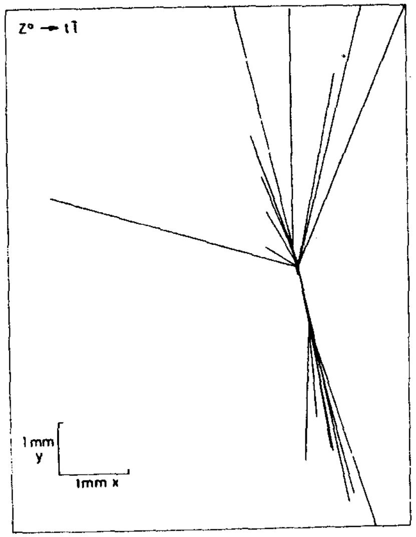

Figure 4(c)

A Z° → $t\bar{t}$ decay. The event topology is rather confused in this XY view.

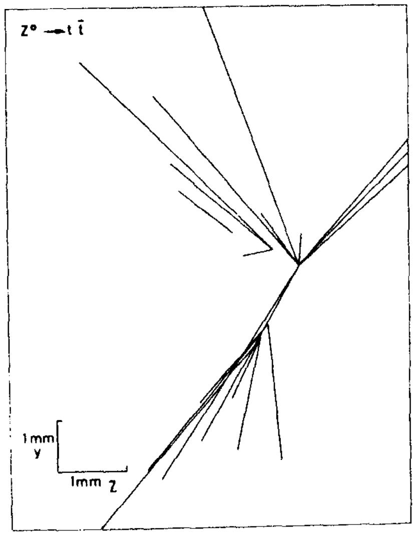

Figure 4(d)

A Z° → $t\bar{t}$ YZ projection (the Z axis is the beam direction). The event topology is clearly seen in this view.

reconstructions it is helpful if one can rotate the event in space in order to make a clean assignment of tracks to vertices.

After getting a general idea of the events we shall be trying to reconstruct, let us now look at some statistical information based on large numbers of generated events. Figure 5 shows the distribution of impact parameters (distances of closest approach) between the decay tracks for $c\bar{c}$ events and the primary vertex, as seen in the XY projection. The mean value is 44 μm, which is the sort of number which has induced in some areas the idea that a precision of (say) 10 μm represents "overkill." If one is aiming for efficient assignment of tracks to vertices, however, this is a very misleading statement; 26% of the tracks have impact parameters below 10 μm, and 16% below 5 μm. The point is that many charm particles decay with lifetimes well below the mean, and that kinematics can lead to situations in which tracks really do pass very close to vertices other than their own (e.g. D* → Dπ, where the π tends to follow the D direction). If one looks at the impact parameter distribution for tracks from the primary vertex in $c\bar{c}$ events, then this is somewhat broader than Figure 5. This becomes obvious from Figure 4(a) where one sees that the prompt tracks are of low momentum and appear reasonably isotropic, whereas the decay tracks from the high momentum charm decays point back in the general direction of the primary vertex. Figure 5 relates to $c\bar{c}$ events, i.e. a mixture of Ds, Fs, Λ_cs, etc. If we look at $b\bar{b}$ events we might expect a broader distribution due to the longer lifetime. But if we maintain our aim of assigning tracks to their correct vertices, we have to consider the impact

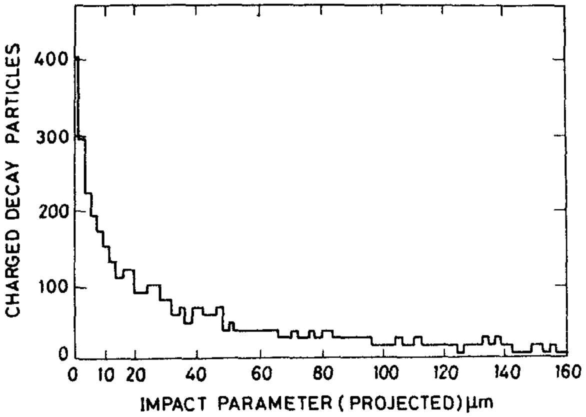

Figure 5 Impact parameter distributions (in XY projection) of decay products in Z° → $c\bar{c}$ events, with respect to the primary vertex.

parameters of the B decay tracks to the primary vertex and also to the same side charm decay vertex. If we plot the impact parameters of decay tracks with respect to all potentially confusing vertices, then we find an impact parameter distribution which is closely similar to Figure 5. The advantage of the longer B lifetime is offset by the more complicated topology (see Figure 4(b)). This remains true also for $t\bar{t}$ events which are similar to the $b\bar{b}$ ones apart from lower momenta in the decay products and higher momenta in the *primary* tracks. As is well known, the *impact parameter* (other things being equal) is momentum independent; higher momentum gives longer decay lengths but smaller decay angles.

Turning now to the momentum distributions of the decay tracks, these are shown in Figures 6(a) to (c). We have the following

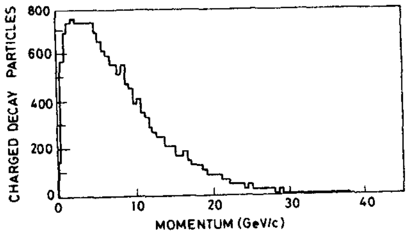

Figure 6(a) Momenta of charged decay particles from events of the type Z° → $c\bar{c}$.

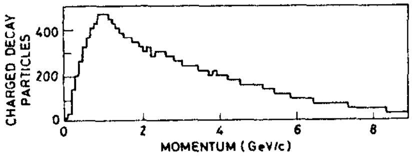

Figure 6(b) Momenta of charged decay particles from events of the type Z° → $b\bar{b}$. (Note the change of momentum scale.)

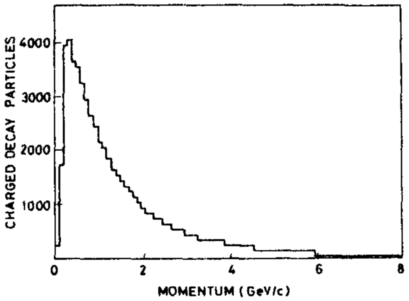

Figure 6(c) Momenta of charged decay particles from events of the type Z° → $t\bar{t}$.

general trend:

	Mean Decay Track Momentum (GeV/c)
$Z^\circ \to c\bar{c}$	8.3
$Z^\circ \to b\bar{b}$	3.3
$Z^\circ \to t\bar{t}$	1.6

The bottom and charm decay products in the $t\bar{t}$ events will be of even lower momentum if the top mass is greater than 30 GeV, as suggested by UA1. Note also from Figure 6 that the distributions peak well below the mean momentum. Multiple scattering is a very serious consideration for these events, especially for $b\bar{b}$ and $t\bar{t}$ production.

2.4 Conclusions. We may summarise the following implications for vertex detectors which aim to have good topological efficiency for reconstructing Z° decay to heavy flavours:

- A σ (b) $\lesssim$ 5 μm.
- If possible, 2-view reconstruction so that kinks in space are always visible.
- Minimal material, and the first measuring plane close to the primary vertex, so that the measurement precision is useful over as wide a momentum range as possible. (This may place tight requirements on the 2-track resolution.)

If one considers vertex detectors for high energy hadronic collisions, the first 2 conditions continue to apply. But the momenta are normally much higher, and multiple scattering is much less serious.

A quantitative evaluation of how well we might be able to do in practice will be given in Section 7, once we have an idea of the range of detectors which are available, and their various capabilities.

3. GENERAL CHARACTERISTICS OF SILICON DETECTORS

3.1 Why Silicon? (And Othef General Remarks). Nuclear emulsions and high precision bubble chambers have an excellent record as vertex detectors for short-lived particles. Yet they have in total accumulated only a handful of events and are not obviously amenable to operation in colliders. There is, for the obvious reasons of rate capability and triggerability, a strong interest in electronic rather than visual detectors, provided they can be developed to have the necessary spatial precision.

Gaseous detectors (specially those operated at high pressure to reduce diffusion) can satisfy some of the requirements for vertex detection (eg. B tagging) but do not really have the potential for efficient topological reconstruction of the events; they tend to work on the high tail of impact parameter distributions. Current detectors have σ (b) ≈ 100 μm, with 25 μm precision having been obtained in a beam test under rather artificial conditions (all tracks normal to the detector, tracks selected to be not too close to sense wires). In addition, gaseous detectors are normally 1-dimensional (or give a much poorer measurement in the orthogonal direction) and have very limited 2-track resolution which means that the measurements can only begin many centimetres from the interaction point.

There are condensed-matter detectors which have considerable potential as electronic vertex detectors. Among these are liquid multiwire proportional chambers, scintillating fibres and germanium detectors. Some of these are under active development and may have a very important role in the future.

What is unique about silicon detectors (and what accounts for their current lead in the field of high precision electronic detectors) is the fact that they are based on the highly developed planar technology. This technology encompasses all aspects of the precise processing (< 1 μm feature sizes) which can be applied to one face of a single crystal of silicon. The planar technology has revolutionised electronics (leading to integrated circuits through MSI, LSI, VLSI and now WSI-wafer scale integration), and is particularly suited to the fabrication of detectors for visible light, X-rays, min-I particles, stopping particles (alphas, etc). Such detectors can have very high precision in spatial position and in the measurement of energy deposited.

Semiconductor detectors have had a long and important history in the field of nuclear physics. The first signals were seen in 1951, from α particles impinging on a reverse-biased point contact germanium diode.[13] This principle - the detection of charge generated within the depletion region of a reverse-biased junction - has been retained in every semiconductor detector since then.

A long-standing advantage of silicon detectors is their intrinsic energy resolution. An ionizing particle in plastic scintillator has to expend 300 eV for every photoelectron generated at the photocathode. A gaseous detector (argon) requires 30 eV per electron liberated. In contrast, the lightly bound valence electrons in silicon are very easily excited into the conduction band; on average, an ionizing particle expends only 3.6 eV for every electron-hole pair liberated. For many nuclear physics applications, the stopping power of silicon is a major advantage. A 10 MeV proton stops in 1 mm of a silicon detector, but has a range of 1 metre in argon gas. This feature will not be directly useful to us, but the high density has the related advantage of yielding a large signal from a very thin detector, and of greatly reducing the range of δ-electrons.

Over 25 years, semiconductor detectors evolved in several forms (intrinsic silicon and germanium, and lithium-drifted varieties) generally in the direction of increasing sensitive volume (up to many cubic centimetres) and improved energy resolution. In some cases detectors were provided with subdivided surface electrodes to achieve modest spatial resolution in 1 dimension ($\gtrsim$ 1 mm). With the November Revolution (the discovery of the J/ψ in November 1974) an enormous interest focused on charm production experiments. By about 1980, it was apparent that high precision vertex detectors would be an enormous asset in such experiments, and some groups started to work on the problem of building tracking detectors with the necessary precision, based on the planar technology which had become increasingly refined since its inception in 1960. These efforts have gone in 3 related directions:

- by incorporating the planar technology into conventional nuclear physics detectors (leading to microstrip detectors);
- by adapting existing photosensitive detectors (leading to particle-detecting CCDs) ; and
- by developing new detector types (leading to the silicon drift chamber).

Before looking at the performance of these detectors in turn, we consider two general topics:

a. What are the fundamental limits to precision in tracking a min-I particle through silicon?

b. What are the principles of operation of silicon detectors? Here I shall spend some time on the basic operating features of MOS devices, of which all our detectors are particular examples.

3.2 Limits to Spatial Precision in Silicon Detectors

3.2.1 Mean Energy Loss. Consider a min-I particle traversing a thin sheet of silicon. It generates electron-hole pairs directly by ionization and indirectly by excitation of atoms and groups of atoms (phonons). These de-excite, emitting X-rays, light, etc. We shall ignore long-range coherent effects such as channelling and Čerenkov radiation since these are at a relatively very low level. We shall discuss these processes with models starting from the very simple (but least physical) and progressing to the more complex (and adequately physical).

We start with the free electron model, in which the electrons of the medium are supposed to be suspended in the material; all atomic binding is ignored. Consider the case of the non-relativistic passage of a charged particle through the material (charge equal in magnitude to the electron charge e, and velocity $v = \beta c$), so that the interaction can be handled by classical electrostatics. Take also the case of the particle mass m being much larger than the electron mass m_e, so that the main effect on the incident particle is energy loss rather than scattering through large angles. The maximum force experienced by an electron will be inversely proportional to the square of the impact parameter (see Figure 7)

$$F_{MAX} = \frac{e^2}{b^2} .$$

The components of force in the direction of the particle trajectory average to zero after the passage of the particle, and the overall effect is of a transverse force ~ F_{MAX} for a time duration $t = \frac{2b}{v}$, giving it a momentum

$$p = \frac{2e^2}{bv} \quad \text{and energy } T = \frac{p^2}{2m_e} = \frac{2e^4}{m_e c^2 b^2 \beta^2} \qquad (3.1)$$

$$\therefore \; 2b\,db = \frac{2e^4}{m_e c^2 \beta^2} \frac{dT}{T^2} .$$

Now the probability F(b) of a collision with impact parameter b in thickness dx of material is given by the number of electrons in a cylinder of radius b

$$F(b)\,db\,dx = 2\pi b\,db\,NZ\frac{\rho}{A}\,dx$$

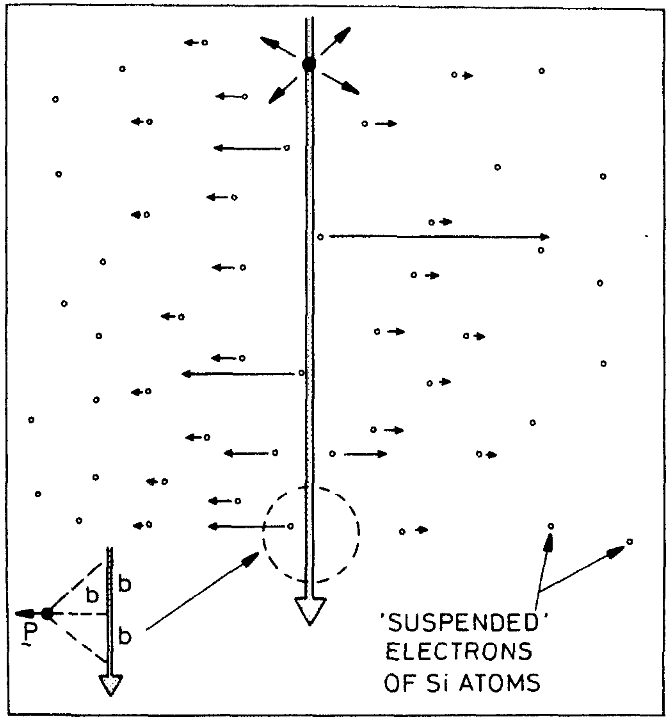

Figure 7   Passage of a heavy charged particle through matter. Close electrons receive a powerful brief impulse; distant electrons receive a weak impulse which is also much more extended in time.

where N is Avogadro's number

A, Z are the atomic weight and atomic number of the material

and ρ is the density.

This is equal to the probability $\phi(T)$ of a collision which imparts energy T to an electron where T and b are related by (3.1).

Thus $$\phi(T)dT\,dx = \frac{2\pi e^4}{m_e c^2\beta^2}\,\frac{NZ}{A}\,\rho\,\frac{dT}{T^2}\,dx\,. \qquad (3.2)$$

The term $\frac{1}{m_e}$ expresses the fact that work is most effectively done on the electrons due to their lightness. The energy transfer to nuclei is reduced by an enormous factor $\sim \frac{m_p + m_n}{m_e} \sim 4000$, and can be entirely neglected.

The term $\frac{1}{\beta^2}$ expresses the fact that slow particles have a higher rate of energy loss.

The term $\frac{1}{T^2}$ expresses the fact that collisions with large impact parameter (low T) take place with a much higher rate than close collisions. But even if we put $T = 1$ eV (and atomic binding will certainly dominate below this) we find from (3.1) that for $\beta = 1$,

$$b = 3 \times 10^{-10}\ \text{cm}.$$

Thus even the most remote ionizing collisions of the most weakly bound electrons will pinpoint the particle trajectory at the level of atomic dimensions. Thus the limit to precision will certainly not be set by the primary ionization process.

Note from (3.2) that in the free electron model, defining $\frac{dE}{dx}$ as the specific energy loss due to all collisions, we have

$$\frac{dE}{dx} = \int T\phi(T)\,dt \propto -\ln(T_{min}) = \infty\,.$$

Due to the long range of the coulomb interaction, this free-electron material would have infinite stopping power.

In moving to a more physically reasonable model, two developments are needed. Firstly, the change to relativistic quantum mechanics for the particle-electron collisions, and secondly a treatment of the atomic binding of the electrons. The calculation proceeds by choosing an energy η (typically 10 to 100 KeV) such that for

$T > \eta$ we have a 'close collision' in which the atomic electrons are effectively free; and

for $T < \eta$ we have a 'distant collision' in which the primary particle is effectively a point, and we handle all transitions leading to excitation or ionization of the atom.

For electron energies $T > \eta$, Formula (3.2) holds with little modification. We have[14]

$$\phi(T) = \frac{d^2N}{dTdx} = \frac{2\pi e^4}{m_e c^2\beta^2}\,\frac{NZ}{A}\,\rho\,\frac{1}{T^2}\,F \qquad (3.3)$$

where $F = 1 - \beta^2\,\frac{T}{T_{MAX}}$ for spin-0 projectiles.

The maximum kinematically allowed energy transfer is T_{MAX}, and is given by

$$T_{MAX} = \frac{2\,m_e\beta^2\gamma^2c^2}{1 + 2\gamma\,m_e/m}\,.$$

At low energies, the $\frac{1}{T^2}$ form holds down to energies of the order of the mean ionization potential of the atomic electrons, as first defined by Bethe

$$I(Z) \approx 16(Z)^{0.9}\ \text{eV} .$$

For still lower energies, the electron energy distribution must flatten and turn over in a way which depends on the details of the atomic structure of the medium. In the case of silicon, the crystalline structure is important. The importance of the low energy electrons can be gauged from the fact that integrating the δ-electron energy distribution (3.3) down to $T = I(Z)$ yields only half of the energy loss given by the Bethe-Block formula[15]

$$\frac{dE}{dx} = \frac{4\pi e^4}{m_e c^2 \beta^2} \frac{NZ}{A} \rho \left[\ln\left(\frac{2\ m_e \gamma^2 \beta^2 c^2}{I(Z)}\right) - \beta^2 \right] . \quad (3.4)$$

In low density gaseous media this formula holds, including the 2 ln γ term which implies a relativistic rise in the specific ionization, due partly to the increasing effective range of the electromagnetic interaction with γ, and partly to the increasing kinematic limit on the δ-electron energy. For a condensed material such as silicon, however, the density effect gives rise to an additional term $-D$ in the parentheses, where D rises to ln γ for highly relativistic projectiles; the rise in $\frac{dE}{dx}$ is thus limited to the effect of the increasing kinematic limit for δ-electron production.

3.2.2 Fluctuations in Energy Loss. So far we have not dealt specifically with the statistical fluctuations in the energy loss, which give rise to the limits to precision in silicon detectors.

For thin detectors, the $\frac{1}{T^2}$ form for the δ-electron energy distribution leads to the situation that collisions over much of the kinematically allowed energy transfer range occur with probability $<< 1$. Most frequently, traversal of the detector will be characterised by a large number of low energy collisions, with an occasional high energy transfer being seen. This gives rise to the familiar asymmetric Landau distribution[16] in energy loss. According to the Landau theory, energy loss fluctuations result only from the collisions described by the Rutherford spectrum (3.3), i.e. by energy transfers much greater than $I(Z)$. As a result, it underestimates the widths of the energy loss distribution for thin samples, for which even energy transfers around $I(Z)$ occur sufficiently infrequently that statistical fluctuations cannot be ignored.

Blunck and Leisegang[17] introduced corrections to the Landau theory to take account of the atomic binding of the electrons. However, this theory considerably overestimates the widths of the energy loss distribution for thin samples. Chechin and Ermilova[18] showed that both theories were inapplicable in such cases, and Ermilova et al.[19] using experimental photoabsorption coefficients, were able to obtain excellent agreement with experiment from samples as thin as 1.5 cm of argon at atmosphere pressure.

The essential point is that the 'distant collisions' previously referred to ($T < \eta$) involve the exchange of soft, nearly on-shell, virtual photons, whose interaction with the material is well described by the measured photoabsorption coefficients. The model is described in detail by Allison and Cobb[20] (applied to gaseous detectors) and has been successfully applied to silicon detectors by G.Hall.[21] When used in a Monte Carlo approach (as in Reference 19) it is possible to determine the precise shapes of the energy loss distributions. When used with numerical integration (as in Reference 21) the broadening of the Landau distribution is described by a convoluting normal distribution of which the width is explicitly calculated.

The thinnest samples in which min-I signals have been detected are the depletion layers of CCDs which amount to only about 20 μm of silicon. Figure 8 shows the experimental $\frac{dE}{dx}$ distribution[22] and indicates the good fit achieved by the calculation of Hall. Several earlier calculations predicted much broader distributions, to the point that it would have been impossible to achieve high efficiency with such thin silicon detectors. Fortunately, both experiment and the most recent calculations are in agreement, and achieving high efficiency is not a problem.

Before leaving this topic, we should look briefly at the experimental photoabsorption data on which the calculations are based (Figure 9). Apart from getting a feeling for the range of virtual photon energies which are important, the figure is instructive in showing the usefulness of silicon for real photon detection.

Almost any conceivable silicon detector has a suspect or dead region due to surface layers etc.of depth ~ 1 μm, and most have a total thickness of << 1 mm. If one converts the absorption cross-sections to absorption/μm or absorption/mm, one sees that for much of the photon energy range silicon is either too opaque or too transparent to be useful. The solid curves (a) on Figure 9 show the efficiency (linear scale on the right) for a typical CCD detector (10 μm depletion) and the broken curves (b) show the effect of

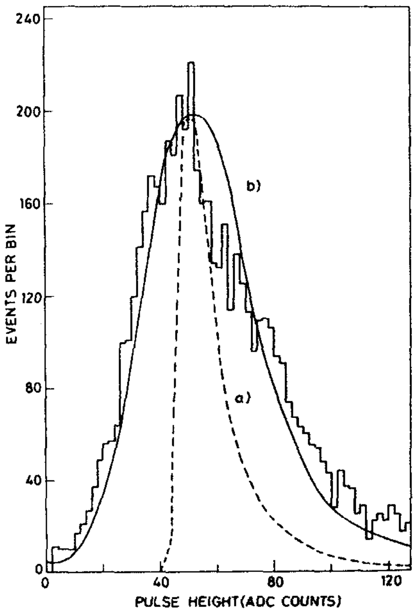

Figure 8 Experimental energy loss distribution from min-I pions traversing a CCD detector (20 μm of silicon equivalent) together with
(a) the Landau distribution and
(b) the distribution calculated by Hall.

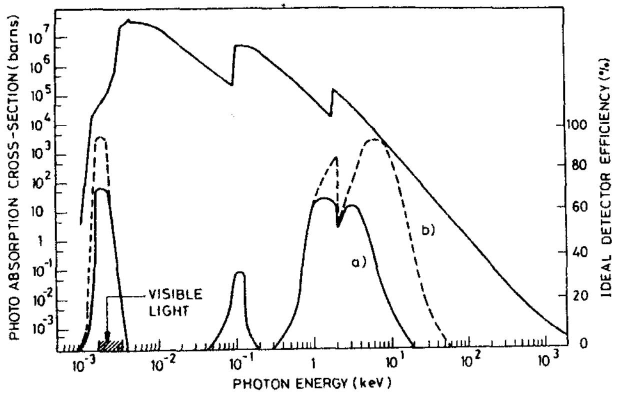

Figure 9 Experimental photoabsorption cross-section for silicon, indicating the energy ranges of use for the detection of real photons.

increasing the depletion depth to 200 μm (achievable with difficulty). From these one can see that silicon detectors are of no use in the infra-red, where the material is entirely transparent, but are useful for visible light and over a region in the X-ray spectrum from 1 to 10 or 20 keV. While of no direct relevance to the detection of min-I particles, this information can be very important when thinking about test procedures and backgrounds (eg. synchrotron radiation) as well as providing industry with its main impetus for the development of silicon detectors.

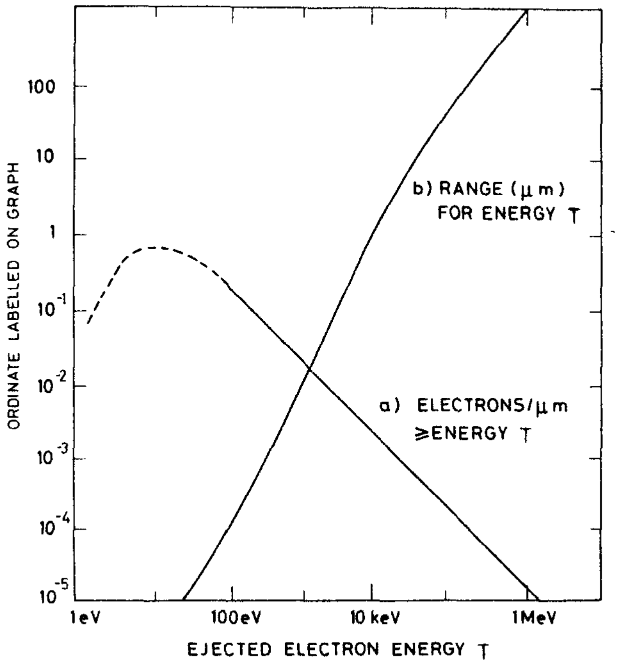

Figure 10 (*a*) *The number of electrons ejected by a min-I particle per μm of path with energy > T.*
(b) The range vs. T.

3.2.3 Limits to Precision. The basic situation, which follows from the previous section, is summarised in Figure 10. This shows firstly the probability per micron of ejecting an electron of energy greater than T, and contains the information needed to evaluate the high energy fluctuations in energy loss. As will be seen, these fluctuations in themselves cause problems for position measurement, but these problems are exacerbated once the ranges of the ejected electrons become significant. The electron ranges are included on the figure. For each decade in electron energy up to the kinematic limit, the flux of ejected electrons falls as $\frac{1}{T^2}$, but each ejected electron releases an increasing number of secondaries (one electron-hole pair per 3.6 eV of primary electron energy) with the result that the mean charge Q released in the silicon (averaged over many samples) increases linearly with ln (T), eventually tailing off in the last decade before the kinematic limit T_{MAX}.

Due to the asymmetric nature of the Landau distribution, the mean energy loss in a sample (which is weighted by the rare cases with very large energy loss) is significantly different from the most probable energy loss. This difference diminishes for thick samples; the mean energy loss scales with thickness but the most probable loss grows faster. For 5 GeV/c pions incident on 100 μm of silicon, the mean energy loss is 400 eV/μm, while the most probable is 240 eV/μm (110 and 67 electron-hole pairs per micron, respectively). From Figure 10(b) we see that one such measurement has a probability of 10% of including a δ-electron of energy > 20 keV, i.e. of range > 5 μm. The 5500 secondary electrons from this δ-electron will pull the centroid of the charge distribution off its true position by typically

$$\frac{5500 \times 2.5}{6700 + 5500} \approx 1 \text{ μm} .$$

Figure 11 shows the probability that the centroid be displaced by some specified values (1 μm and 5 μm are taken as examples) as a function of detector thickness. A detector with a thin active thickness of silicon is in principle to be preferred to a thicker one, since in the thin detector there is a much lower probability of generating a δ-electron of dangerously long range. Although the amount of charge in the main column of ionization increases with thickness, the effect of δ-electrons becomes more serious since their range increases approximately as the square of their energy.

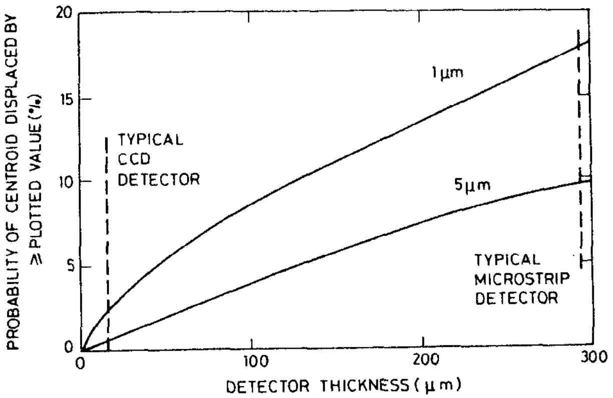

Figure 11 Detector precision limitations from δ-electrons for tracks at normal incidence as a function of detector thickness.

Finally, it should be noted that the Landau fluctuations in energy loss are particularly serious in determining the positions of angled tracks. Here, one can do no better than assign the centroid of the charge distribution to the mid-plane of the detector. Even ignoring the effect of δ-electron range, fluctuations in energy loss along the track immediately cause errors in the assigned position, as illustrated in Figure 12 for tracks at 45°. In a thin detector there is a 10% probability of producing a δ-electron which, if it occurs near one end of the track, pulls the co-ordinate across by 4 μm. In the thick detector, there is the same probability of producing a δ-electron which can pull the co-ordinate by 87 μm.

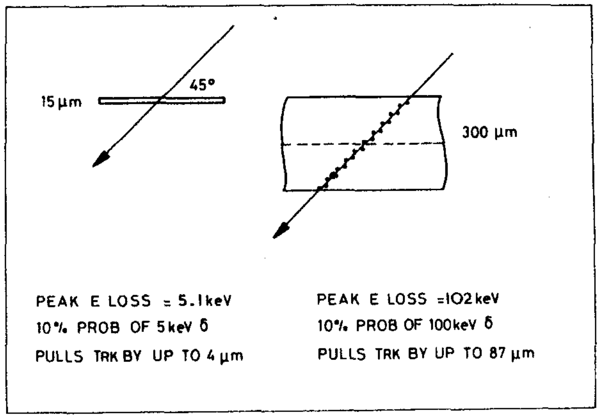

Figure 12 Effect on detector precision for angled tracks due to energy loss fluctuations.

Of course, the advantages of thin detectors must be weighed against the relative difficulty of extracting the information from them due to such effects as a poorer signal to noise ratio. These points will be discussed in the context of specific types of detectors.

In summary, the fine trail of electron-hole pairs in a thin silicon detector allows in principle an unprecedented precision of tracking (<< 1 µm) compared even with nuclear emulsion, in which a δ-electron of ~ 400 eV is needed to blacken a grain. Our task is to find the best method of taking advantage of this information; how can the positions of these minute quantities of charge be measured? There is in fact no completely satisfactory answer at present; the field is wide open for new ideas. But in order to explore the problem further, we shall consider those principles of construction and operation of MOS devices which are relevant to the silicon detectors which have been built so far.

3.3 Principles of Operation of Silicon Detectors. Gaseous silicon has a typical structure of atomic energy levels (see Figure 13). It has an ionization potential of 8.1 eV, i.e. it requires this much energy to release a valence electron, compared with 15.7 eV for argon. As silicon condenses to the crystalline form, the discrete energy levels of the individual atoms merge into a series of energy bands in which the individual states are so closely spaced as to be essentially continuous. The levels previously occupied by the valence electrons develop into the valence band, and those previously unoccupied become the conduction band. Due to the original energy

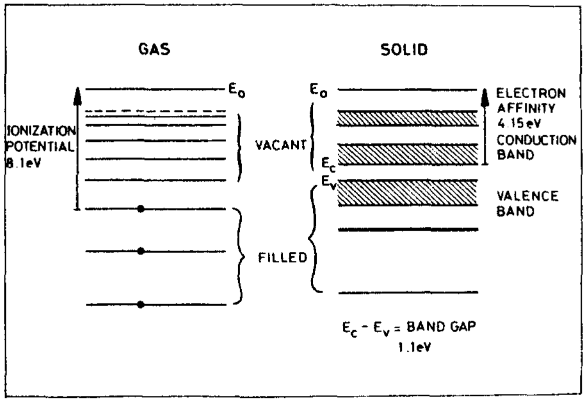

Figure 13 Diagrammatic sketch of allowed energy levels in gaseous silicon which become energy bands in the solid material.

level structure in gaseous silicon, it turns out that there is a gap between these two bands. In conductors, there is no such gap, in semiconductors there is a gap (1.1 eV in silicon, 0.7 eV in germanium) and in insulators there is a large band gap. In particular, the band gap in silicon dioxide is 9 eV. This makes it an excellent insulator and, coupled with the ease with which the surface of silicon can be oxidised in a controlled manner, accounts partly for the pre-eminence of silicon in producing electronic devices.

We shall denote as E_v and E_c the energy levels of the top of the valence band and the bottom of the conduction band (relative to whatever zero we like to define). The energy needed to raise an electron from E_c to the vacuum E_o is called the electron affinity. For crystalline silicon this is 4.15 eV.

3.3.1 Conduction in Pure and Doped Silicon. To understand the conduction properties of pure silicon, the liquid analogy is helpful. This is illustrated in Figure 14 where (a) shows the energy levels in silicon under no applied voltage with the material at absolute zero temperature. All electrons are in the valence band and under an applied voltage (b) there is no change in the population of occupied states, and so no flow of current; the material acts like an insulator. At a high temperature (c) a small fraction of the electrons are excited into the conduction band, leaving the same number of vacant states in the valence band. Under an applied voltage (d) the electrons in the conduction band can flow to the right and there is a re-population of states in the valence band which can be visualized as the left-ward movement of a bubble (hole) in response to the applied voltage.

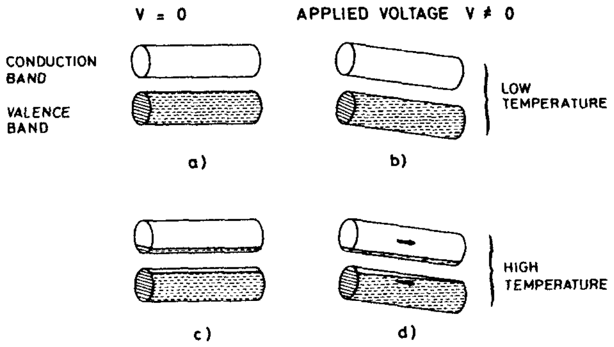

Figure 14 Liquid analogy for a semiconductor.

Now kT at room temperature is approximately 0.03 eV. This is small compared with the band gap of 1.1 eV, so the conductivity of pure silicon at room temperature is very low. To make a quantitative evaluation, we need to introduce the Fermi-Dirac distribution function $f_D(E)$ which expresses the probability that a state of energy E is filled by an electron. Figure 15(a) shows the form of this function

$$f_D(E) = \frac{1}{1 + e^{(E - E_f)/kT}} ; \qquad (3.5)$$

E_f, the Fermi level, is the energy level for which the occupation probability is 50%. Figure 15(b) shows the density of states g(E) in

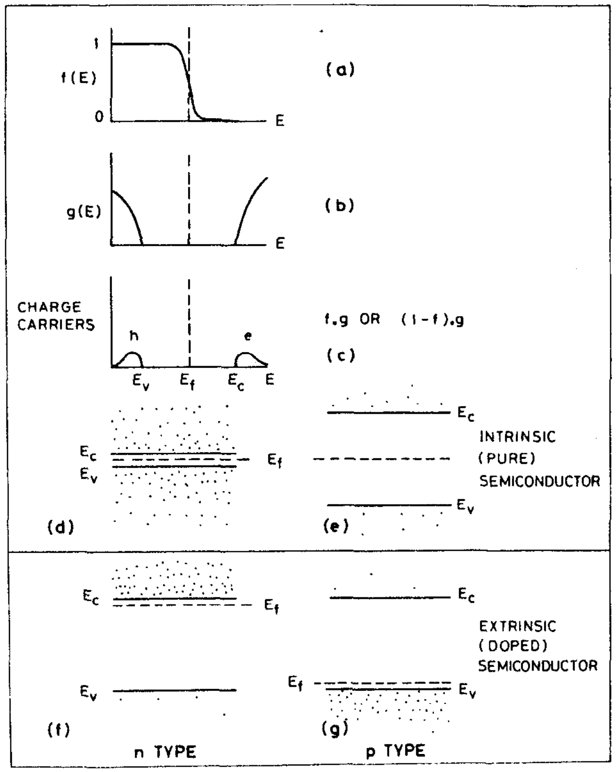

Figure 15 (a) Fermi-Dirac distribution function. The slope increases as the temperature is reduced.
(b) The density of states below and above the forbidden band-gap.
(c) Concentration of electrons and holes (charge carriers) available for conduction.
(d) and (e) Charge carrier distributions in narrow and wide band-gap semiconductors.
(f) and (g) Charge carrier distributions in n- and p-type semiconductors.

silicon. The concentration of electrons in the conduction band is given by the product f.g, and the density of holes in the valence band by (1 - f).g, as shown in Figure 15(c). In pure silicon, the Fermi level is approximately at the mid-band gap, and the concentrations of electrons and holes are of course equal. These concentrations, due to the form of f_D, are much higher in a narrow band-gap semiconductor (d) than in a wide gap material (e).

So far we have been discussing pure (so-called intrinsic) semiconductors. Next we have to consider the doped or extrinsic semiconductors. These allow us to achieve high concentrations of electrons (n-type, Figure 15(f)), or of holes (p-type, Figure 15(g)), by moving the Fermi level very close to the conduction or valence band edge. The procedure for doing this is to replace a tiny proportion of the silicon atoms in the crystal lattice by dopant atoms with a different number of valence electrons.

For example, phosphorus has 5 valence electrons in contrast with 4 for silicon. At absolute zero it holds all 5 and phosphorus-doped silicon is still an insulator. But at a very low temperature the extra electron is shaken free, and at room temperature most of the extra phosphorus electrons are available for conduction.

Conversely, boron has 3 valence electrons, leaving one vacant bond, easily filled by the movement of electrons, and so an available path for conduction. This is best visualised as the contrary motion of the vacant bond (hole).

Figure 16 shows the levels associated with various donor atoms (measured relative to E_c) and acceptor atoms (measured relative to E_v). Figure 17 shows the concentration of electrons in n-type silicon (1.15 x 10^{16} dopant atoms, arsenic, per cm^3) as a function of temperature. Below about 100° K one sees the phenomenon of carrier freeze-out, loss of conductivity due to the binding of the donor electrons. This is followed by a wide temperature range over which the electron concentration is constant, followed above 600° K by a further rise as the thermal energy becomes sufficient to add a substantial number of intrinsic electrons to those already provided by the arsenic atoms. This general behaviour is typical of all doped semiconductors.

The resistivity ρ of the material depends on the concentration of free holes and electrons and on their mobilities. These are a function of temperature and of impurity concentration. At room temperature, in lightly doped silicon, we have

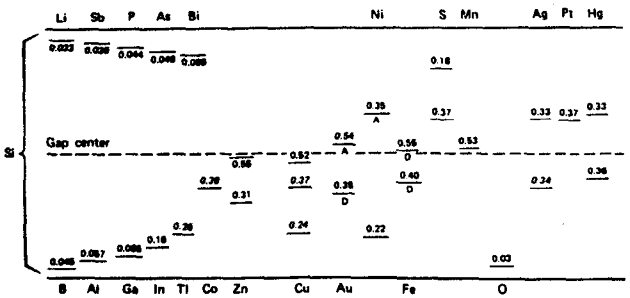

Figure 16 From Reference 23. Energy levels in electron volts of various impurity elements in silicon. Levels of acceptor atoms are measured from the top of the valence band, and levels of donor atoms are measured from the bottom of the conduction band.

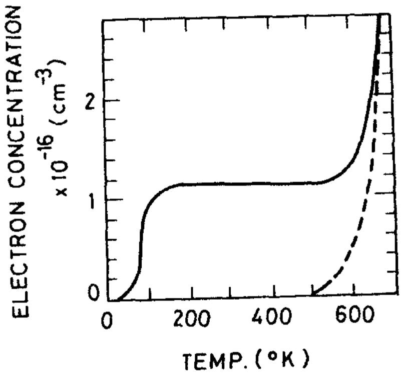

Figure 17 Electron concentration vs. temperature for n-type (arsenic-doped) silicon. The dashed curve shows the concentration for intrinsic material.

electron mobility μ_n = 1350 cm^2 $(V.s)^{-1}$,

hole mobility μ_p = 480 cm^2 $(V.s)^{-1}$,

and the resistivity is given by

$$\rho = \frac{1}{e(\mu_n . n + \mu_p . p)} \qquad (3.6)$$

[e is the charge on the electron and n and p are the electron and hole concentrations].

For pure silicon at room temperature $n_i = p_i = 1.45 \times 10^{10}$ cm^{-3} which gives ρ_i = 235 KΩ cm.

The resistivity as a function of impurity concentration is shown in Figure 18. For reasons which will become clear, we are often concerned in silicon detectors with unusually high resistivity material, beyond the range of these graphs. For example, 20 KΩ cm p-type silicon requires a dopant concentration of 5 x 10^{11} cm^{-3}. Remembering that the crystalline silicon has 5 x 10^{22} atoms per cm^3, this implies an impurity level of 1 in 10^{11} which even in the highly developed art of silicon crystal growing is a major challenge. The resistivity noted above in connection with pure silicon is entirely unattainable in practice.

There is a useful approximation which relates the carrier concentration to the Fermi level. For n-type material, the carrier concentration is dominated by

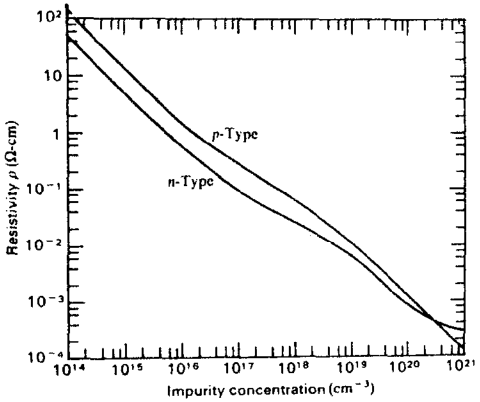

Figure 18 From Reference 24. Resistivity of silicon at room temperature as a function of acceptor or donor impurity concentration.

N_c the effective density of states at the conduction band edge

and $E_c - E_f$ the energy separation between the Fermi level and those available states.

Thus $n = f \cdot g$

becomes $n = N_c \exp\left[-[E_c - E_f]/kT\right]$.

In fact, the extent to which this is an approximation is illustrated by the temperature dependence of the effective density of states, which can be shown to be proportional to $T^{3/2}$. But it is still true that the overall temperature dependence is dominated by the exponential term. Similarly we can define N_v as the effective density of states at the valence band edge, and

$$p = N_v \exp\left[-[E_f - E_v]/kT\right].$$

Now for intrinsic material we have

$$n = p = n_i \quad \text{and} \quad E_f = E_i \ (\sim \text{mid band gap}).$$

Thus in general (intrinsic or doped silicon) we have

$$\left.\begin{aligned} n &= n_i \exp\left[(E_f - E_i)/kT\right] \\ p &= n_i \exp\left[(E_i - E_f)/kT\right] \end{aligned}\right\} . \qquad (3.7)$$

This shows that the deviation of a doped semiconductor from the intrinsic material can be simply represented by the energy separation of the Fermi level from the intrinsic Fermi level.

Notice that

$$np = n_i^2 = N_c N_v \exp\left[-E_g/kT\right], \qquad (3.8)$$

where $E_g = E_c - E_v$.

This is a particular example of the very important law of mass action which applies as much in semiconductor theory as it does in chemistry, and which states that in a sample in thermal equilibrium the increase in electrons (eg. by donor doping) has as a result a decrease in holes (by recombination) such that the np product is constant.

Notice also from (3.8) that the main term in the temperature dependence of the intrinsic carrier concentration is exp $[- E_g/2kT]$ which implies a factor 2 increase in n_1 *for each 8° K increase in* temperature around room temperature.

It is generally valid to think of n-type material as containing only electrons and p-type material as containing only holes. These are referred to as the majority carriers in each case. An important point to be aware of in designing semiconductor detectors is the possibility of growing thin (10 to 50 μm) epitaxial layers of high resistivity material on a low resistivity substrate, with ~ 1 μm transition region between (Figure 19). It is even possible to grow epitaxial layers of alternating conductivity type (n and p), which opens up *some very interesting possibilities* for particle detection.

3.3.2 The np Junction. We now need to introduce a most important fact relating to conducting materials which are electrically in contact with one another and in thermal equilibrium; they all must establish the same Fermi energy. This applies to

metal/semiconductor systems

n-type/p-type systems etc.

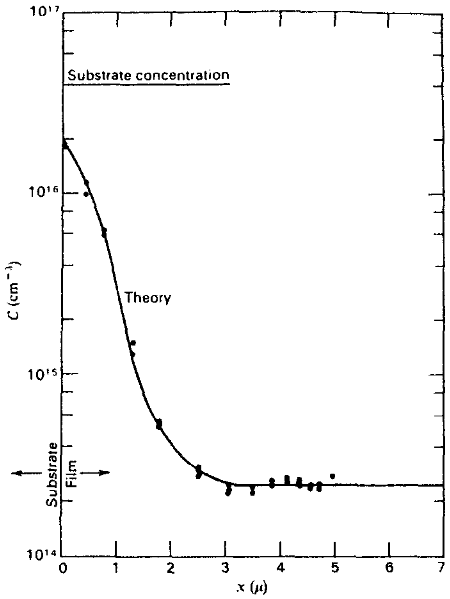

Figure 19 From Reference 24. Impurity distribution after epitaxial growth.

Charges flow from the high to low energy region until this condition is established. For example, at an np junction there is exposed a fixed space charge of ionized donors and acceptors, creating a field which opposes further drift of electrons and holes.

The depletion approximation *says that the semiconductor changes* abruptly from being neutral to being fully depleted. This is far from obvious and in fact there is a finite length (the Debye length, typically ≈ 0.1 μm) over which the transition takes place. But the depletion approximation will be adequate for all the examples we need to consider.

Let us look in some detail at the important case of the np junction. Before contact (Figure 20(a)) the surface energy E_o is equal in both samples; the p-type Fermi level is close to E_v and the sample is densely populated by holes; the n-type Fermi level is close to E_c and the sample is densely populated by electrons.

On contact, the electrons diffuse into the electron-free material to the left, and the holes diffuse to the right. In so doing the electrons leave exposed donor ions (positively charged) in the n-type material, and the holes leave exposed acceptor ions (negatively charged) in the p-type material. This builds up an electric field which eventually just balances the tendency for current flow by diffusion. Once this condition is reached (Figure 20(b)) the Fermi levels in the materials have become equal. The electrical potentials in the two samples (the potential energy at the surface E_o, or at the conduction band edge, E_c) are now unequal.

Intuitively, this can be understood as follows. Initially, the electrons at a particular level in the conduction band of the n-type material see equal-energy levels in the p-type material which are unpopulated, so they diffuse into them. The developing space charge bends the energy bands so that these levels become inaccessible. Eventually, only very high energy electrons in the n-type material see anything other than the empty states of the band gap of the p-type material and conversely for the holes in the p-type material.

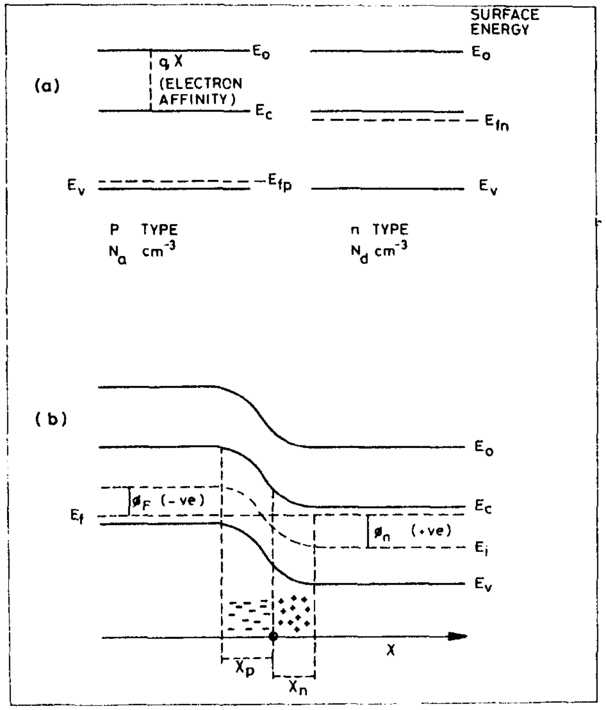

Figure 20 (a) Energy levels in two silicon samples (of p- and n-type) when electrically isolated from one another.
(b) When brought into contact, the Fermi level is constant throughout the material. The band edges bend in accordance with the space charge generated.

Let us develop this quantitatively, adopting a co-ordinate system in which the np junction of Figure 20(b) is at position $x = 0$.

The same x dependence is followed by E_o, E_c, E_i, and E_v. The zero of the electric potential ϕ is arbitrary, so we define

$$\phi = -\frac{1}{e}(E_i - E_f)$$

Thus $\phi = 0$ for intrinsic material
positive for n-type
negative for p-type.

From (3.7) $$\phi_n = \frac{kT}{e}\ln\frac{N_d}{n_i},$$

$$\phi_p = -\frac{kT}{e}\ln\frac{N_a}{n_i}$$

where N_d and N_a are the concentration of donor and acceptor atoms in the n- and p-type material respectively.

The potential barrier $\phi_i = \phi_n - \phi_p = \dfrac{kT}{e}\ln\left(\dfrac{N_d N_a}{n_i^2}\right)$. (3.9)

Notice that the potential barrier falls linearly with temperature since it is sustained by the thermal energy in the system.

We may deduce the electric field strengths near the junction by using Poisson's equation

$$\frac{d^2\phi}{dx^2} = -\frac{e}{\varepsilon_s}\rho(x)$$

where ε_s is the permittivity of silicon $= \varepsilon_r \varepsilon_o$,

ε_o is the permittivity of space $= 8.85 \times 10^{-14}$ F cm^{-1}

$= 55.4$ electron charge/V.µm,

ε_r is the dielectric constant or relative permittivity of silicon $= 11.7$,

$$= -\frac{dE}{dx}$$

where E is the electric field.

For $x > 0$

$$\left.\begin{array}{l} \dfrac{dE}{dx} = -\dfrac{eN_d}{\varepsilon_s} \quad \therefore E(x) = -\dfrac{eN_d}{\varepsilon_s}(x_n - x). \\ \text{For } x < 0 \\ \dfrac{dE}{dx} = +\dfrac{eN_a}{\varepsilon_s} \quad \therefore E(x) = -\dfrac{eN_a}{\varepsilon_s}(x + x_p). \end{array}\right\} \quad (3.10)$$

The undepleted silicon on either side of the junction is field-free. The depleted silicon close to the junction contains an electric field whose strength is maximum at the junction and is directed to the left, i.e. opposing the flow of holes to the right and opposing the flow of electrons to the left.

Requiring continuity of the field strength at x = 0 implies

$$N_a x_p = N_d x_n . \qquad (3.11)$$

Thus, if one wants to make a deep depletion region on one side of the junction (important, as we shall see, for many detectors) we need to have a very low dopant concentration, i.e. very high resistivity material.

The electric field strength varies linearly with x; the electric potential, by integration of (3.10), varies quadratically.

$$\left.\begin{aligned} &\text{For } x_n > x > 0 \quad \phi(x) = \phi_n - \frac{eN_d}{2\varepsilon_s}(x_n - x)^2 . \\ &\text{For } x_p < x < 0 \quad \phi(x) = \phi_p + \frac{eN_a}{2\varepsilon_s}(x + x_p)^2 . \end{aligned}\right\} \qquad (3.12)$$

Requiring continuity of the potential at x = 0 implies

$$x_n + x_p = \left[\frac{2\varepsilon_s}{e}\phi_i\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{\frac{1}{2}} . \qquad (3.13)$$

From (3.9) ϕ_i depends only weakly on N_a and N_d.

If for example $N_a >> N_d$ we have $x_p \approx 0$ and (3.13) gives $x_n \propto \frac{1}{N_d^{\frac{1}{2}}}$.

So a factor 2 increase in resistivity leads to a factor $\sqrt{2}$ increase in depletion depth.

Figure 21 summarises these results on the characteristics of an unbiased np junction, with the inclusion of some typical numerical values based on $N_a = 10^{14}$ cm^{-3} and $N_d = 2 \times 10^{14}$ cm^{-3}. The peak field in this case is about 3 kV/cm. By high doping concentrations and large bias voltages it can easily happen that one approaches the limiting field of about 300 kV/cm at which internal breakdown in the silicon sets in.

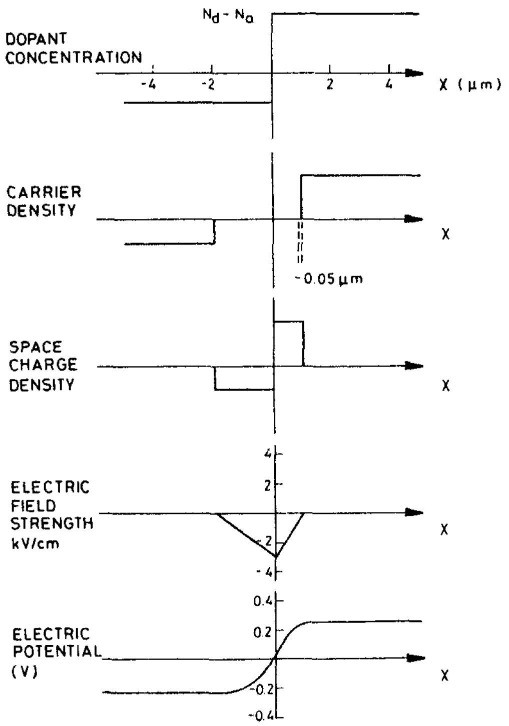

Figure 21 Summary of various quantities across an unbiased np junction.

We now consider the effect of applying a voltage across the junction. Under equilibrium conditions, electron-hole pairs are continually generated by thermal excitation throughout the semiconductor. In the case of zero bias (Figure 22(a)) the electrons and holes generated within the bulk of the semiconductor recombine. Those generated in the depletion region are swept into the undepleted silicon, holes to the left, electrons to the right. This effect would act to reduce the potential barrier and so is compensated by a leakage of majority carriers which diffuse across the barrier in the opposite directions at just the rate needed to cancel the charge generation in the depleted material. The overall effect is of no current flow.

By applying a forward bias (Figure 22(b)) we separate the previously equal Fermi levels by an amount equal to the bias voltage; the system is no longer in thermal equilibrium or this condition could not be maintained. Although there is still an electric field in the depletion region which is directed against the current flow, the depletion region is narrowed and the potential barrier is now inadequate to prevent majority carriers from flooding across it, holes from the left and electrons from the right. Many of these will recombine within the depletion region giving rise to the recombination current. Those which survive are absorbed within one or two diffusion lengths by recombination with the majority carriers on that side of the junction, giving rise to the diffusion current. Beyond these regions there is just a steady flow of majority carriers supplied from the voltage source to keep the current flowing. Notice that in a forward biased junction the current flow is due entirely to electron-hole recombination.

With a reverse bias, we have the situation shown in Figure 22(c). The depletion region is now much wider and electron-hole pairs generated within it are efficiently swept into the undepleted silicon, electrons to the right and holes to the left giving rise to the generation current. Unlike the case of the

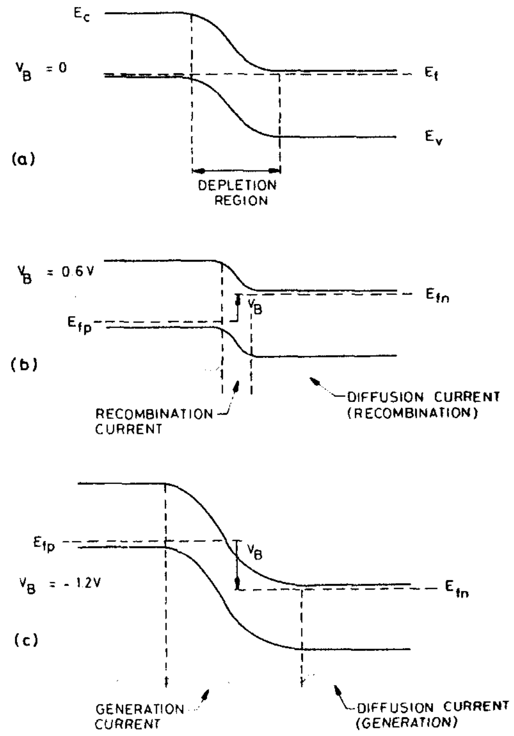

Figure 22 Effect of an applied voltage on the semiconductor in the region of the junction.

unbiased junction, there is now no supply of majority carriers able to overcome the increased potential barrier across the junction. On the contrary, the thermal generation of minority carriers within one or two diffusion lengths of the depletion region leads to some holes generated in the n region reaching this depletion region and then being briskly transported across it, and conversely for electrons generated in the p region. This leads to the so-called diffusion current. In the case of the reverse-biased junction, the current flow is thus due entirely to electron-hole generation. The current flow across reverse-biased junctions is of great importance in determining the noise limits in silicon detectors. An immediate observation is that, since the current arises from thermal generation of electron-hole pairs, the operating temperature will be an important parameter.

Before continuing to discuss this point, it is worth noting that we have finally collected up enough information to calculate the characteristics of a typical particle detector, and it is instructive to do so. Referring to Figure 23, we have a silicon detector made of good quality, high resistivity p-type silicon (ρ = 10 KΩ cm). On the front surface we make a shallow implant of donor atoms and on the back surface we make a highly doped p-type implant to provide a good low-resistivity ground electrode. The terms n^+ and p^+ are conventionally used to represent high doping levels. Now we apply a positive voltage V to the n-type surface with the aim of completely depleting the detector. In this way we shall ensure complete collection of the electrons and holes generated by the passage of a charged particle; with incomplete depletion we would lose signal by recombination. Equation (3.13) applies, with the difference that we replace ϕ_i by V + ϕ_i since the junction is biased in the direction which assists the previously existing depletion voltage.

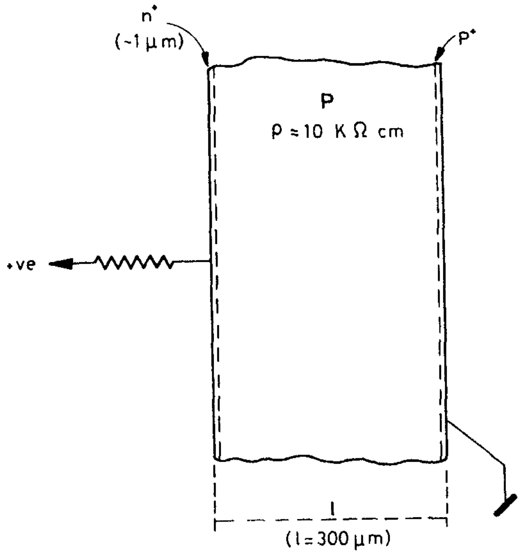

Figure 23 Typical structure used for particle detection. Essentially it is no more than a diode, reverse biased so as to fully deplete the thick p-type layer.

We have

$$x_n + x_p \approx x_p = \left[\frac{2\varepsilon_s}{e} \left(V + \phi_i\right)\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{\frac{1}{2}}$$

$$\approx \left[\frac{2\varepsilon_s}{e} x \frac{V}{N_a}\right]^{\frac{1}{2}} \quad .$$

From Figure 18, we see that $N_a \approx 10^{14} x \frac{150}{\rho}$

and we require $x_p = \ell$

$$\therefore V = \frac{e}{2\varepsilon_s} x \frac{1.5 x 10^{16}}{\rho} x \ell^2$$

$$= \frac{10^{-4}}{2 x 55.4 x 11.7} x \frac{1.5 x 10^{16}}{\rho} \ell^2 x 10^{-8}$$

where ℓ is in μm and ρ in Ω cm

$$\therefore V = \frac{11.6 \ \ell^2}{\rho} \quad .$$

For the above example, V = 105 volts is the potential needed to fully deplete the detector.

Returning to the properties of the reverse biased junction, Figure 24 shows the current/voltage characteristics of a typical silicon junction over a wide temperature range.

At high temperatures the leakage current is dominated by thermal electron-hole generation within approximately one diffusion length of the depletion edge. The diffusion length for minority carriers is

$$L_D = \sqrt{D\tau} \qquad (3.14)$$

where D is the diffusion constant and is related to the mobility μ by $D = \frac{kT}{e} \mu$.

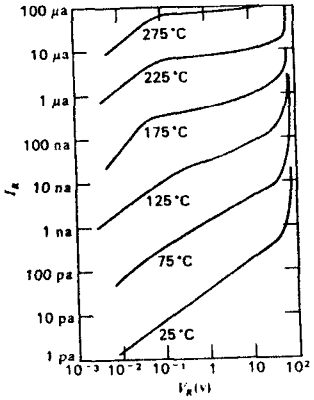

Figure 24 From Reference 24. Current vs voltage in a reverse biased silicon diode at various operating temperatures.

$$\left.\begin{array}{lll}\text{For electrons } D_n & = & 34.6 \text{ cm}^2 \text{ s-1} \\ \text{For holes } D_p & = & 12.3 \text{ cm}^2 \text{ s}^{-1}\end{array}\right\} \text{at room temperature.}$$

The τ is the minority carrier lifetime, and it can vary from ~ 100 ns to > 1 ms depending on the care taken in the silicon processing. This point will be discussed further.

This high temperature leakage current (termed the diffusion current, as previously noted) is almost independent of the reverse bias voltage, but is highly temperature dependent. The temperature dependence stems of course from the thermal generation of the minority carriers.

At lower temperatures (≲ 100° C) the diffusion current becomes negligible and the generation current dominates. This continues to show a similarly fast temperature dependence, but is now also quite voltage dependent, as seen in Figure 24, since the depletion width is proportional to $V^{\frac{1}{2}}$.

The diffusion and generation currents depend on the rate of electron-hole generation, and the diffusion current depends also on the minority carrier lifetime. These quantities are in fact closely related. Direct thermal generation of an electron-hole pair is quite rare in silicon for reasons which depend on the details of the crystal structure. Most generation occurs by means of intermediate generation-recombination centres (impurities and lattice defects) near the band-gap centre. Thus an electron-hole pair may be thermally created in a process where the hole is free in the valence band and the electron is captured by the trapping centre, to be subsequently emitted into the conduction band. These bulk trapping states vary enormously in their density and can be held down to a low level by suitable processing. It is precisely these states which determine the minority carrier lifetime already mentioned. Reducing the density of bulk trapping states does two things. It cuts down the thermal generation of charge carrier pairs in the material, so reducing the concentration of minority carriers available for the generation of current across a reverse-biased junction. It also increases the minority carrier lifetime and so the diffusion length (but only at $\tau^{\frac{1}{2}}$). The first effect vastly outweighs the second, so that a low density of bulk trapping states is highly advantageous in ensuring low leakage current. As we shall see in Section 8, even originally high grade silicon can deteriorate due to the production of bulk trapping states by radiation damage.

Mid band-gap impurities such as gold (see Figure 16) are a particularly serious source of bulk trapping centres. As shown in Figure 25, even in low concentrations, gold atoms strongly reduce the carrier lifetimes, and lead to greatly increased dark current.

These effects are obviously not serious in cases where one is collecting large signals promptly. But in cases of small signals and/or long storage times (such as in a silicon drift chamber, or

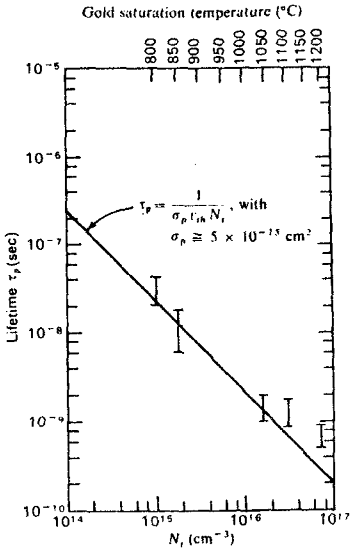

Figure 25 From Reference 24. Effect of gold impurities of various concentrations on the minority carrier (hole) lifetime in n-type silicon.

CCD), particular care is needed. One important design criterion is to keep the stored charges well away from the surface of the silicon, since the surface region is more likely to have picked up some undesirable impurities as well as having a high level of lattice defects.

3.3.3 Electron Transport in Silicon. While the charge generated by an ionizing particle is being transported by the internal field in the detector, there is inevitably the process of diffusion which spreads out the original very fine column of charge during this transportation process. In the case of very highly ionizing particles (such as alphas) the original density of electrons and holes can be so high that space-charge effects are important. In the case of min-I particles, however, such effects are negligible and the time development of the electron and hole charge distributions may be treated by simple diffusion theory.

Consider a local region of charge (electrons or holes), for example a short section of the particle track length within the silicon. Under the influence of the internal field, this will be drifted through the material and at the same time will diffuse radially as shown in Figures 26 and 27. The RMS radius of the charge distribution increases as the square root of time (as (3.14)), with standard deviation $\sigma = \sqrt{2Dt}$. Thus 50% of the charge is contained within a radius of 0.95 $\sqrt{Dt}$. For electrons at room temperature this gives:

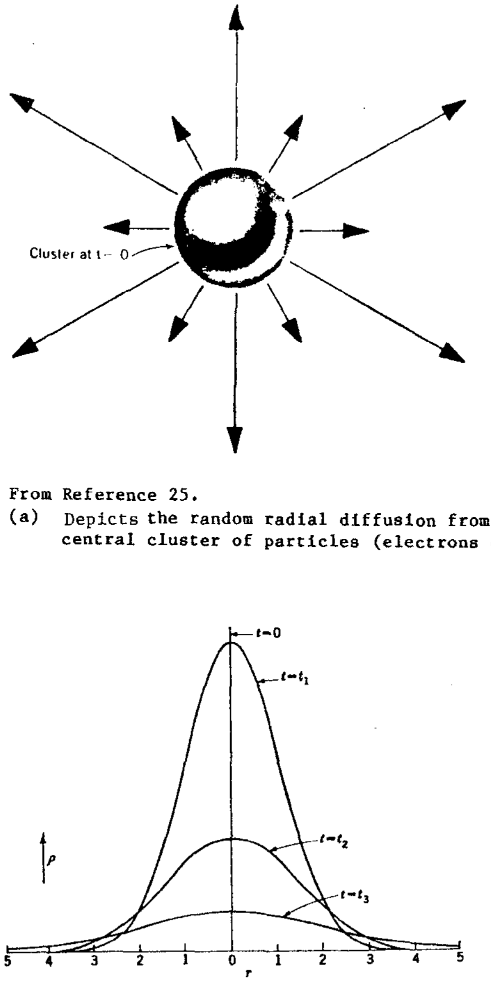

Figure 26 From Reference 25.
(a) Depicts the random radial diffusion from a small central cluster of particles (electrons or holes).

(b) Radial density distribution as a function of time over 3 equal time intervals.

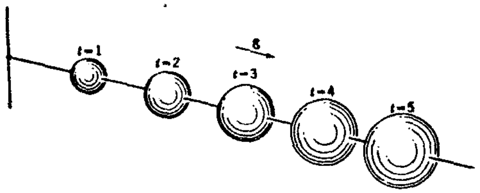

Figure 27 From Reference 25. Combined drift and diffusion of an initially small cluster of particles (electrons or holes) as a function of time over equal time intervals. The drift distance is proportional to t but the radius grows only as $t^{\frac{1}{2}}$.

Time	Radius	Drift Distance
10 ns	6 μm	135 μm
1 μs	60 μm	14 mm
100 μs	0.6 mm	140 cm

The drift distances listed in the third column are obtained by assuming a 'typical' drift field in depleted silicon of 1 kV/cm, and using the fact that the drift velocity $V_d = -\mu_n E$.

Diffusive charge spreading is an attractive option for improving spatial precision beyond the limits of the detector granularity. For example, one might hope to achieve precision of one or two microns from a strip detector with 25 μm strip pitch, by centroid finding on the basis of measured pulse heights in adjacent strips. This depends on achieving a charge radius of $\gtrsim$ 30 μm which (from the above table) implies large drift distances and/or very gentle drift fields. These constraints have different implications for different types of detector but in general, ideas for centroid finding lead to the need for processing high resistivity silicon.

4. MICROSTRIP DETECTORS

These are based on fully depleted silicon slices with the general diode structure described in Section 3.3, of thickness typically 300 μm. The practical lower limit on thickness is set by the need for good signal/noise for min-I particles. Making the detector thinner loses signal charge directly and further reduces the output voltage due to the increased capacitance of the detector.

It is desirable to achieve full depletion with bias voltage $\lesssim$ 100 V in order to avoid large leakage currents (which produce noise) or even internal breakdown. Therefore, these detectors are built on high resistivity silicon (typically $\gtrsim$ 10 KΩ cm). Even so, the mean collection time for electrons is only ~ 4 ns giving a radius for the charge collected at the surface of approximately 4 μm.

One surface is electrically subdivided into conducting strips for charge collection on a pitch of typically 20 μm. Readout can be connected to every strip or to every n^{th} strip, using capacitive charge division between the floating and connected strips to provide the position co-ordinate.

Figure 28 shows one option. Readout is connected to every 5th strip. In this pioneering detector,[26] the diode structure was made by the surface barrier technique. More recently[27] detectors have been built with the diode structure made by separate p^+ implants beneath each one of the aluminium strips. Such detectors have proven to be very efficient. The precision when reading out one strip in 3 (with 20 μm pitch) is found to be $\sigma = 4.5$ μm. This precision degrades by about a factor of 2 if only one strip in 6 is read out.