Concept for a Si-W ECAL for the ILC based on MAPS

P.D.Dauncey, Imperial College London

14 December 2004

Draft 0.2

Abstract

A conceptual design for an electromagnetic calorimeter for the International Linear Collider is outlined. The calorimeter is a sampling tungsten calorimeter, with silicon wafers forming the sensitive detectors. Unlike other such Si-W designs, the use of Monolithic Active Pixel Sensors is proposed.

Contents

1 Introduction

1.1 Existing ECAL concepts

Several designs for ECALs for the ILC have been proposed. Most are for sampling calorimetry, where the conversion layers are iron, lead or tungsten and the sensitive layers are silicon or scintillator. (Note, there is one proposal for an ECAL made from lead-tungstenate [1] although this does not have widespread support.) Otherwise, a silicon-tungsten ECAL is widely acknowledged to be the best performing ECAL for PFLOW, albeit with some degradation in EM shower resolution, although the cost is very high. Other proposals are mainly driven by the desire to reduce the cost, particularly for detectors with a large tracking volume and hence ECAL inner radius.

The TESLA TDR [2] ECAL is an example of a silicon-tungsten calorimeter. This had 40 layers of silicon and tungsten with a total ECAL thickness of 23 cm. The barrel has an inner radius of 1.68 m and an outer radius of 1.91 m and is 5.4 m long. The planar endcaps cover both ends of the barrel. The total number of radiation lengths is $24X_0$, comprised of 30 layers of $0.4X_0$ and the final 10 layers of $1.2X_0$.

The silicon wafers proposed were $1 \times 1\,\mathrm{cm}^2$ diode pads. These were reverse biased such that a charged particle passing through the wafer would cause electron-hole pairs to be created and hence the resulting charge could be detected in a Very Front End (VFE) preamplifier chip. The analogue signal from each pad was then digitised and readout out. The total ECAL consisted of 32 million pads. It is likely that the 40 layers may not be needed to achieve the required hadronic jet pattern recognition, and so the ECAL considered here will be assumed to have less layers and hence a lower number of channels.

The mechanical structure of the ECAL was that every other sheet of tungsten was imbedded into a carbon-fibre structure, with gaps ("alveoli") between these layers. Into the gaps, slabs of the remaining tungsten sheets were slid, with a layer of silicon wafers on either side mounted on a PCB. In the concept being discussed in this report, the basic structure will be very similar, except the silicon diode pad wafers will be replaced with Monolithic Active Pixel Sensors (MAPS).

1.2 ICL timing structure and physics rates

The timing structure of the ILC is not yet defined. It is generally assumed it will not be too dissimilar from the TELSA timing. The TESLA timing was different for the 500 GeV and 800 GeV options. At 500 GeV, the bunch collided within trains every 337 ns, with a train consisting of 2820 bunches and hence lasting 0.95 ms. The period of these trains was comparatively slow at 5 Hz, meaning there was a duty factor of around 0.5%. The 800 GeV option was more challenging, with collisions every 176 ns and 4886 bunches in a train of length 0.86 ms. The train period was slightly slower, at 4 Hz, giving a duty factor of 0.3%.

The assumptions used in the following to give concrete examples are taken at the higher end of the above numbers. A machine with a crossing time of 200 ns is assumed, with 5000 crossings per train, giving a train length of 1 ms. A train period of 5 Hz is then used, giving a duty factor of 0.5%.

The physics rates relevant to the data volume of the ECAL at the ILC are dominated by two-photon production of hadronic events, the so-called "mini-jet" events. The TESLA study calculates around 200 such events per bunch train resulting in an average of around 100 ECAL silicon pads being hit per beam crossing. However, the production of e^+e^- pairs from photons produced by beam-beam interactions gives an even larger average rate of around 5000 pads per beam crossing and this clearly dominates. Assuming approximately equivalent luminosity per crossing, then with the above timing assumptions, this would result in an average of 2.5×10^7 pads per train being hit.

2 Diode pad vs MAPS concepts

The concept for the MAPS ECAL rests heavily on the similar diode pad design. Hence, the latter is described in some detail here.

The optimisation of the detector design between tracking, calorimetry and cost requirements is an ongoing study. Hence, the size of the ECAL is not yet fixed. Here, it is assumed that the ECAL will have an average radius of around 1.6 m and will be around 4 m long. This gives an average layer area in the barrel of around 40 m^2. Assuming of order 30 layers, this gives a total area of around 1200 m^2 in the barrel. The endcaps need to have a radius at least equal to the ECAL barrel out radius, which will be around 2 m. This gives an endcap area per layer of 12 m^2 and with two endcaps of 30 layers each, this is an area of around 750 m^2 (neglecting the hole in the centre for the beam line). Hence, the total ECAL silicon area is around 2000 m^2 or $2 \times 10^7\,\mathrm{cm}^2$. With $1 \times 1\,\mathrm{cm}^2$ pads, this would be around 20 million channels.

Assuming first beam in 2015, then since the ECAL will need three to four years to build, then construction needs to start in 2012. This requires prototyping of a "final" design to start in 2010, giving five years in which to define this design. Also, all the major choices will need to be made on proven technology at this time, i.e. 2010, so it is likely there are only five, not ten, years of technological progress from now which can be used.

2.1 Diode pad concept

The basic unit of a diode pad ECAL detector is the silicon wafer. Here, we will assume each wafer has an array of 16×16 diode pads, each around $1 \times 1\,\mathrm{cm}^2$. This requires a wafer of diameter $16\sqrt{2} \sim 23\,\mathrm{cm}$ or around 9 inches, which should be feasible and economical as standard within five years. The total number of wafers needed is then around $2 \times 10^7/256 \sim 8 \times 10^4$.

The complete thickness of the wafer is used to collect the particle signal from dE/dx in the wafer. Hence, it is likely that the wafers would be kept relatively thick, namely around 500 μm.

The wafers need to be mounted onto a PCB for electrical connections and mechanical support. Each PCB would be electrically independent. The feasible size of the PCB is a major influence on the design. Ideally, PCBs long enough to fill an octant of the ECAL would be used, as per the TESLA design. With a radius of 1.6 m, this would require a PCB around 1.3 m long. Such a PCB would be able to fit eight 16 cm wafers along its length. Assuming it could also fit a pair of wafers in width, the resulting PCB would be 32 cm wide and would hold 16 wafers, for a total of 4096 channels per PCB. The wafers would be fastened with conductive glue to the PCB. Around 5000 such PCBs would be needed. A module would then consist of a tungsten sheet sandwiched between a pair of these PCBs. A diagram of such a PCB is shown in figure 1, which also shows the assumed readout interface of an FPGA and optical fibre RX/TX.

The section through the PCB is shown in the left side of figure 2.

The thickness of the tungsten sheet will probably be varied with depth in the ECAL with the thinnest at the front. The TESLA design has the first layers with a $0.4X_0$ thickness which is around 1.4 mm. The resulting module is shown in the left side of figure 3.

The total number of modules needed would be around 1500 for the barrel and 1000 for the endcap.

2.2 MAPS concept

The main functional difference between a MAPS and a silicon diode pad wafer is that there is readout circuitry integrated into the MAPS so that this is not required externally. The most obvious thing to then do is to divide the MAPS wafer into $1 \times 1\,\mathrm{cm}^2$ pads and implement the functionality of one channel of the VFE ASIC chip into each pad. This would give a MAPS wafer which is functionally very similar to the diode wafer and VFE chip combination.

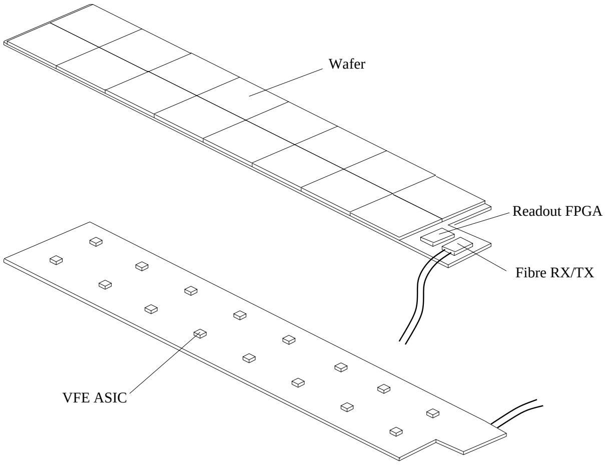

Figure 1: Diode pad PCB, with 16 wafers and 16 VFE ASICs. The MAPS option looks identical except for the absence of the VFE ASICs.

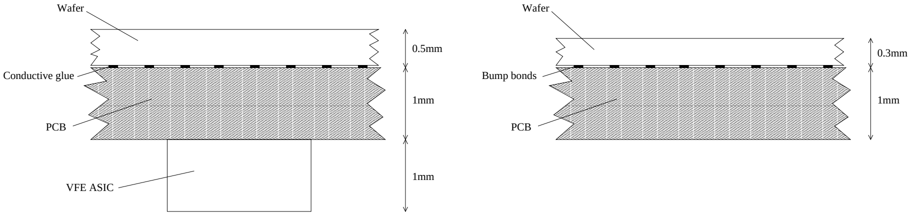

Figure 2: PCB section for (left) diode pads and (right) MAPS. (Not to scale.)

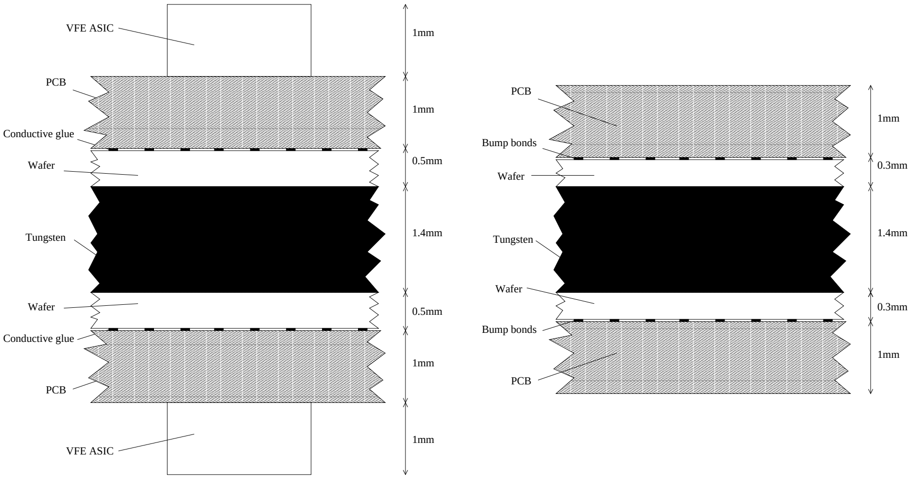

Figure 3: Module section for (left) diode pads and (right) MAPS. (Not to scale.)

This is actually not the proposed use for MAPS, but serves to illustrate some of the advantages which would arise from using MAPS, even if the measured quantities were the same.

With a MAPS rather than pad wafer plus VFE chip, the mechanical structure can be made very similar to the pad wafer case, except there is no VFE ASIC mounted on the opposite side of the PCB. Furthermore, since MAPS do not use the full silicon wafer thickness to produce the signal (it arises only in the epitaxial layer), then they can be thinner; a standard $300\,\mu$m thickness is assumed. The resulting PCB section is shown in the right hand side of figure 2. This makes the whole module structure thinner than for the pad wafer case, as shown in the right hand side of figure 3. Note, the density of connections on the wafers would mean that bump-bonding of the wafer to the PCB may be needed, rather than conductive glue (although see Section 2.5.4).

2.3 Potential advantages

The MAPS design leads directly to several advantages compared to the pad wafer design.

2.3.1 Effective Molière radius

The usual quantitative measure of the width of an EM shower is the Molière radius, defined as the radius of a cylinder which on average contains 90% of the shower energy; for a block of pure tungsten this is 9 mm. In practise, gaps between the tungsten layers, which contain comparatively small amount of material, allow the particles to spread out further. Also, a coarsely segmented detecting element will effectively smear the radius further. The effective Molière radius is is a combination of all these effects and can clearly be minimised by reducing the inter-layer gaps and keeping the granularity high.

With MAPS, the gaps between layers can be reduced due to no longer needing a VFE chip. This directly improves the separability of nearby showers as there is less space for the particles to diverge between the tungsten layers. Clearly, reducing the average PCB thickness from 2.5 mm to 1.3 mm, as shown in figure 2, will help in this respect.

Reducing the gap thickness mainly comes from removing the VFE ASICs and so requires effectively nothing to be placed on the other side of the PCB from the wafers. If simple termination resistors or decoupling capacitors are needed, they may need to be embedded into the PCB to be flush with the surface. One issue is how to join together PCBs if a single board of the right size cannot be manufactured (see Section 2.5.5).

Also, note that the $1 \times 1\,\mathrm{cm}^2$ pads are at the limit of an acceptable granularity; the SLAC/Oregon/Brookhaven design has them at half the size for this reason. The later sections show how the MAPS can improve on the granularity.

2.3.2 Cooling, power cycling and stability

While the power consumption of a MAPS is not yet well known (see Section 2.5.6), the cooling issues compared to a pad diode system can be discussed. With a MAPS system, then the pairs of PCBs in each module would clearly be oriented so the MAPS wafers are facing the tungsten. Depending on the uniformity of the wafer mounting, some thermally conducting foam or grease could be used to ensure a good thermal contact with the module tungsten sheet. This would allow the tungsten to act as a heat sink for the MAPS during the bunch train power cycle.

For diode pads, the main heat source is the VFE chip, which is estimated to produce 4 mW of heat per channel, or 1W per chip (or equivalently per wafer under the assumptions here). It is not yet clear if the VFE chips would be mounted on the outside of the module (as shown in figure 3) or inside, where they are thermal contact with the tungsten sheet. However, even in the latter case, the chip has a limited surface area for thermal contact with the tungsten of order $1\,\mathrm{cm}^2$, compared with around 250 cm^2 for the MAPS. This must make the temperature fluctuations during a bunch train much larger for the VFE chips. This has consequent implications for pedestal (and threshold) stability over the bunch train; these may vary with temperature and so would drift during the time of the train.

If cooling pipes are required, the mechanical design and structure should be simpler when coupling to a flat surface (i.e. the MAPS wafers) rather than the protruding lumped heat sources of the VFE ASICs.

2.3.3 "Single event" disruption

There is a concern that a large EM shower passing directly through a sensitive part of the readout circuit would cause the equivalent of a "single event upset" and corrupt either analogue levels or digital values. This is actively under investigation and the expected rate of such disruption is not yet known. To investigate this for MAPS will require a beam test of sensors in electron showers.

The MAPS has the circuitry spread out over a much larger area than the VFE chip. This means that any shower, which has a width of order the Molière radius of 9 mm, will only hit a small part of the readout circuit, rather than showering into the VFE chip containing the readout for the whole wafer. Hence, while every shower will pass through circuitry, unless SEU is a huge effect, it will only affect a small region of the readout. In contrast, a shower through the VFE ASIC will be rarer, but would affect data from the whole wafer.

Showers into the readout FPGA at the end of the PCB may also be a cause of such problems, but this will be common to both designs. The FPGA will be purely digital and so majority logic might be usable to protect from such upsets.

2.3.4 Cost and number of layers

Cost is one of the major problems with any silicon-tungsten ECAL design. The cost of producing 2×10^7 cm^2 of silicon between 2012 and 2015 is clearly extremely uncertain. Values as low as \$1-2/cm^2 have been mentioned, which would still mean the silicon alone would cost \$20-40 million. The cost is also uncertain due to the uncertainty in the size of the ECAL and the number of layers needed. These factors could easily produce a factor of two difference in silicon area and hence further cost uncertainty.

The comparative cost of diode pad wafers and MAPS wafers is possibly easier to estimate. The high resistivity silicon required for diode pads is rather unusual for commercial processes, whereas MAPS are based on very widely used CMOS processing. MAPS are almost bound to be significantly more complex than diode pad wafers, with 20-30 masks being needed for the former, while the latter may need only 5 masks or less. This makes a major difference to the NRE during prototyping. However, the final ECAL cost is clearly totally dominated by the linear wafer production cost.

Current prices today are that the CMS silicon tracker (which uses high resistivity silicon) is paying around \$10/cm^2, whereas MAPS can be produced for around \$1 thousand for a 15×15 cm^2 area wafer, giving a cost of around \$4.5/cm^2. With the (unrealistic) assumption of no reduction in cost over the next eight years, this would result in an ECAL based on MAPS with a total silicon wafer cost of around \$90 million, compared with \$200 million for diode pads.

There is also the issue of the total detector cost. As described above, the inter-tungsten sheet width will be reduced by around 1.2 mm per layer, resulting in an overall ECAL thickness reduction of around 3.6 cm. This would be 10-20% of the total thickness, allowing the components outside the ECAL to be reduced in radius correspondingly. In particular, the detector solenoidal magnet cost is estimated to change by around \$2 million per cm radius and the HCAL would be more compact too. The thinner ECAL would therefore imply substantial saving in the costs of other systems.

2.3.5 Assembly

The construction of such a large detector will be non-trivial and the scaling up of the assembly needs to be considered as part of the design. The industrialisation of the assembly will require reducing it to a minimum number of steps, with as many as possible being standard processes.

The diode pad design has more steps for the PCB contruction as it requires the VFE ASICs to be connected and the PCB is also double-sided. Furthermore, it requires the wafer bias ground to be connected to the outer side of the wafer, which is currently foreseen to be done using foil wrapped around the assembled module.

The MAPS design has components on only one side of the PCB. It is thought that no ground or bias will be needed on the side of the wafer facing out from the tungsten. This should make assembly significantly simpler.

2.4 Improving the MAPS information

The above compares an ECAL made from diode pad wafers to one made from MAPS, where the information extracted from the silicon is the same in both cases. However, this ignors the new possibilities presented by MAPS and these are discussed here.

The basic idea is to improve the quality of the information read out. This could lead to either a better hadronic jet resolution, a smaller number of layers being needed for the same hadronic jet resolution (and hence cost reduction), or both.

The concept is to not use $1 \times 1\,\mathrm{cm}^2$ pads, but very finely subdivide the wafer into small pixels. The pixel size is chosen so that the probability of more than one charged particle going through each pixel in an event is very small, even in the core of a high energy EM shower. With such a small pixel size, then the pixel signal can be simply discriminated to look for a MIP or not. This allows only binary readout per pixel and hence keeps the readout data volume managable. There are a very large number of pixels in the ECAL and the data are assumed to be read out during the dead period between bunch trains.

The pixel size is determined by the density of charged particles in the core of an EM shower. This is around 100 particles per mm^2 for the shower energies at a ILC. (This sets the maximum signal in the diode pad option with $1 \times 1\,\mathrm{cm}^2$ pads to be around 10^4 MIPs.) This density is equivalent to one particle per $100 \times 100\,\mu\mathrm{m}^2$, which indicates the required pixel size is of this order.

In the following a pixel size of $50 \times 50\,\mu\mathrm{m}^2$ is assumed. This is a factor of 200 smaller in each direction than a $1 \times 1\,\mathrm{cm}^2$ pad and so the number of pixels is larger than the number of pads by a factor of 4×10^4, resulting in around 10^7 pixels per wafer and hence around 8×10^{11} pixels for the whole ECAL. The number of particles passing through such a pixel in a shower can be estimated by assuming the number is Poisson distributed at the level of the pixel size, with a mean given by the above density. With a pixel size of length $100a\,\mu$m, then the mean number in the core of a shower would be a^2. The probability of no particles is then e^{-a^2} and of one is $a^2e^{-a^2}$. For pixels of size $50 \times 50\,\mu\mathrm{m}^2$, then $a = 0.5$ so for no particles, the probability is 0.779, for one is 0.195 and hence more than one is 0.026. Whether 2.6% is sufficiently small needs to be answered by simulation; if 1% was required, then $(1 + a^2)e^{-a^2} \geq 0.99$ which gives $a \leq 0.385$ implying a pixel size of around $39 \times 39\,\mu\mathrm{m}^2$; this would mean subdividing the $1 \times 1\,\mathrm{cm}^2$ pad into a 256×256 pixel array, resulting in 1.3×10^{12} pixels in total.

For storage of the data during a bunch train, the area of the pixel can be used to implement a RAM buffer. Assuming at most half the total available area could be used for memory and that, with a sub-micron process, each memory bit would occupy an effective area of $2 \times 2\,\mu\mathrm{m}^2$, then the total storage which could be implemented would be around 300 bits or around 40 bytes. A total of 32 bytes is assumed in the following. Note, with the $39 \times 39\,\mu\mathrm{m}^2$ pixels, only 25 bytes would be possible.

There is a requirement to be sensitive to isolated charged hadrons, i.e. single MIP tracks, which will hit only one pixel in each layer. Hence, no rejection of isolated pixel hits can be done before data from many layers of the ECAL are considered together.

It should be possible to configure the MAPS wafer to mask out any of the pixels if they become noisy. This prevents them from firing the discriminator and so generating huge amounts of data.

There are two options which are proposed for the use of fine-pixilated MAPS wafers in an ECAL. The first option has the pixel output bits summed over an area not dissimilar to the diode pad size. The second reports out the individual pixel locations. These are discussed in the following two sections.

However, the next subsection first discusses some of the issues common to both proposals.

2.5 Outstanding common issues

2.5.1 Crosstalk

The charge collected in the pixels is generated in the epitaxial layer and diffuses to the contact points where it is sensed by the surface circuitry. If charge from one particle is spread over many pixels, then the accuracy of counting the number of MIPs seen will be degraded; this is what is meant by crosstalk here.

Studies will be needed to stop or minimise crosstalk. Previous studies using electric fields to control the charge motion were not very successful. The easiest option would be to thin the epitaxial layer, possibly to $\sim 5\,\mu\mathrm{m}$, so as to get a favourable pixel width/depth ratio. However, this will be at the cost of lower signal/noise.

There are questions which need to be answered by simulation to guide the MAPS design:

- What is the effect on the EM shower energy resolution of charge sharing near the pixel boundaries?
- Would a 50:50 division of the charge be better seen as two pixels firing (i.e. a low threshold) or neither firing (i.e. a high threshold)?
- What rate of charge sharing can be tolerated before impacting EFLOW pattern recognition in hadronic jets?

2.5.2 Threshold

The comparator threshold will need to be adjustable as the optimal level will depend on both the S/N achieved and the actual background hit rate seen at the ILC. It is assumed that it would be unfeasible to set a threshold per pixel so the level ideally would be common to a wafer. The MAPS would include an on-sensor DAC which can be programmed to set this threshold level. The issues are then how uniform will the threshold be across a wafer and wafer-to-wafer and how stable is the threshold with time, temperature, etc?

2.5.3 Noise

As discussed in Section 4.2, a signal/noise of around 10^{-6} per sample is desirable. This means there would be one noise hit per 1000×1000 pixel area for each beam crossing, which corresponds to an area of $5 \times 5\,\mathrm{cm}^2$. Note, a MIP $\sim 80e^-/\mu\mathrm{m}$, so $15\,\mu\mathrm{m}$ is equivalent to $1200e^-$ signal. Noise values between $20e^-$ and $40e^-$ have been achieved.

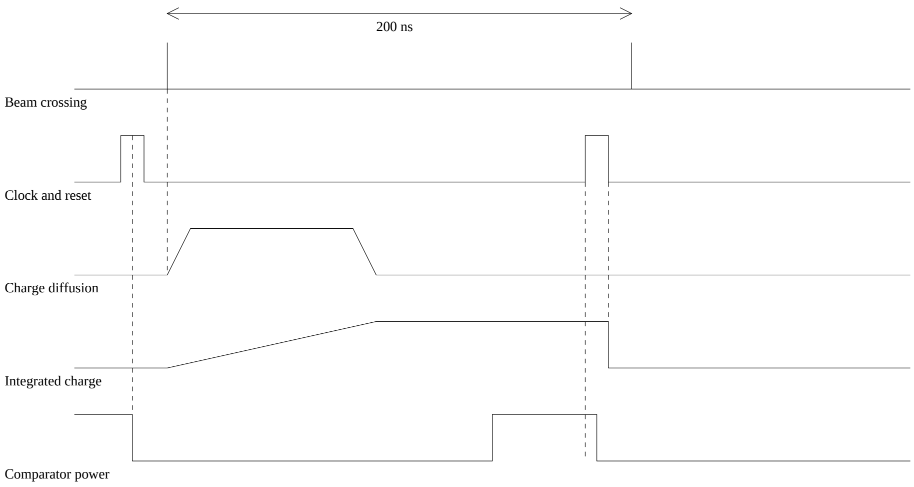

Figure 4: Sample clocking to avoid digital switching noise during the charge collection time.

The noise in a MAPS sensor is usually dominated by reset noise and hence is (to a large extent) unaffected by the epitaxial layer thickness. Hence, the signal/noise $\propto$ epitaxial thickness. This clearly pushes for a thicker epitaxial layer, but obviously reducing the noise is also desirable.

Issues which need to be investigated are

- The possibility to have a “floating” comparator, where the comparator is biased from the reset level itself. This would reduce the noise due to the reset significantly.
- Correlated double sampling is a well-known technique, but it is not clear if this would be possible in the $\sim 200\,\mathrm{ns}$ available between bunch crossings.
- Soft reset or no reset schemes, where the charge is able to leak off with a reasonable RC time constant. This leads to image lag but this may be acceptable given the ILC event rate. This is an issue for simulation.
- A very fast shaping time and feed resulting waveform into either a digital counter or a time-above-threshold integrator. This means no reset is required, and hence there is no reset noise.

Even though dominated by reset noise, it would be sensible to avoid any switching noise by clocking the wafer only outside of integration time, as shown in figure 4.

A major question for simulation is the requirement on the noise rate. The 10^{-6} figure is chosen so that the noise gives a hit rate around the size of the beam-induced background. However, if the DAQ can handle higher rates, then it is important to study at what noise rates the particle flow jet reconstruction and pattern recognition break down. A rough estimate can be obtained by requiring a low probability for a noise hit in an layer within the area which a track seen in the preceding layers would project into. As the distance between layers is only of order one cm, then the area consistent with the track projection is only a few mm^2. Taking e.g. $5 \times 5\,\mathrm{mm}^2$, then this contains a 100×100 pixel array, which is 10^4 pixels. Hence, this would indicate a rate up to of order 10^{-4} would be tolerable for physics.

The ILC will produce a low rate of events in time, but a localised high density of hits within an event in space. Hence, if the data rates prove too high to be dealt with for every beam crossing, then it is generally better to gang in time, not space.

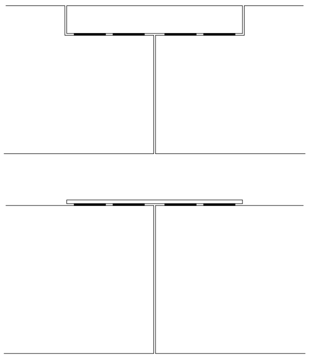

Figure 5: Inter-PCB bonding concepts; (top) embedded PCB piece (bot) thin kapton connecting bridge.

2.5.4 Dead area of wafer

The dead area of diode pads is due to the 1 mm guard ring around the wafer which for a $16 \times 16\,\mathrm{cm}^2$ wafer contributes a 2.5% dead region. The inter-pad 0.1 mm gap contributes another 4%.

MAPS could have around 64 I/O pins, which would fit along one edge only. This would give a density of 2.5 mm per connection. Assuming this global circuitry is around 1 mm wide, then this contributes a 0.6% dead region.

What is the percentage of the sensitive MAPS wafer area lost due to inter-pixel gaps and the sensor I/O pads? Could the I/O connections be placed on top of an insulating layer on the top surface? This would make the sensor appear like a standard ball grid array component, and hence allow cheap standard solder paste connections, as well as reducing the dead area.

There may be some charge absorption in the memory and other circuitry N^+ structures. How can this be minimised? Does this lead to an effective fractional loss of charge everywhere (and so is a signal/noise issue) or does it depend strongly on position within the pixel (and so is an efficiency issue)?

2.5.5 PCB ganging

The ECAL designs (both diode pad and MAPS) call for 1.5 m PCBs to be manufactured and populated with silicon wafers. It is not clear how many manufacturers can make such PCBs. If they cannot be fabricated in one piece, then two (or more) smaller PCBs would need to be ganged together.

The main issue for MAPS is then that to take advantage of the removal of the VFE ASIC (making the PCB thinner), no extra material must be added to the non-wafer side of the PCB for this join. This is less of an issue for the diode pads, as the VFE ASICs already occupy ~ 1 mm in height, allowing a connector of this size without any increase in gap thickness.

Figure 5 shows some possibilities in terms of embedded PCB pieces or thin PCB/kapton connections.

Some issues are:

- If a PCB of at least 1.5 m length can be made, then what yield could be expected? What is the risk of a MAPS sensor failure after assembly, given the large area of each PCB? Can the MAPS reliably be replace after mounting on the PCB?

- If thin inter-PCB connections are straightforward, would it be better to make multiple PCBs, one per MAPS wafer, to allow each wafer to be individually replaced if it goes faulty?

- What thickness of PCB is needed given the density of signals? What density is needed at the join between PCBs?

2.5.6 Power

A comparator (discriminator) might be around $1\,\mu$A from a 2.5 V supply, which is $2.5\,\mu$W. Therefore, a MAPS wafer with 10 million pixels would consume 25 W for the comparators alone while powered on. This would average out to 0.13 W given the machine duty cycle.

The comparators could be rapidly switched so as to come on only when the integrated charge is expected to have built up. The switching could feasibly be done within 10 ns or less, which would mean a reduction by one or possibly two orders of magnitude in power might be possible.

2.5.7 Stitching and large detector areas

There is little incentive to invest R&D in stitching now, given that companies seem to be moving in this direction anyway.

Longer term, the issues are whether it be cheaper to make large area sensors with no dicing, and what yields would we expect? Would it be better to try to have one large sensor per wafer or multiple electrically independent ones? Would we need to dice the wafer even for the latter?

2.5.8 Cosmics in situ

How wide in time is the efficient period for a MIP to pass through the wafer when acquiring charge and hence what is the effective duty cycle for cosmics? Should we try to clock continuously if we want to detect cosmics? Would this be helped with a time-over-threshold method?

3 Particle counter option

3.1 Overview of concept

The basic idea is to sum the number of pixel bits set for each beam crossing over an area of order the original pad size. The summed areas are then equivalent to diode pads but the value obtained is then based on the number of particles which went through the wafer, not the energy deposited by the particles in the silicon. This should be a better measure of the ECAL energy. The area summed over does not have to be the same as for the diode pad option, but is assumed to be so for convenience here.

With this assumption, then each MAPS wafer is assumed to have a 16×16 pad array, with each $1 \times 1\,\text{cm}^2$ pad containing a 200×200 pixel array. It is likely the pixels within the 200×200 array will need to be laid out individually and not using step-and-repeat, due to the combined pad functionality. However, the pads should then be identical, so the 16×16 pad array could be step-and-repeated.

The sum is done for every bunch crossing separately, resulting in 5000 values per bunch train.

3.2 Data rates

Every non-zero value would need to be read out. From beam interactions, there will be 2.5×10^7 non-zero values per bunch train from the whole ECAL, which is effectively one hit per pad per train. The train corresponds to $5000 \times 4 \times 10^4 = 2 \times 10^8$ pixel discriminations per pad and so,

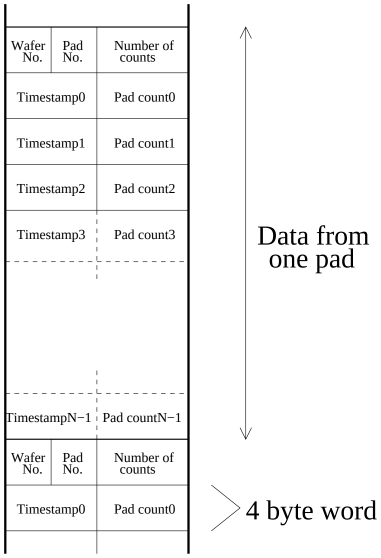

Figure 6: Example format for particle counter option. The wafer number has a range 0-15 and so can fit into a byte. The pad number within a wafer has a range 0-255 and so also fits into a byte. The number of sums (counts above zero) cannot exceed the number of beam crossings and so is in the range 0-5000, requiring two bytes, with an average of around $N \sim 200$. Each timestamp is also in the range 0-5000, fitting into two bytes and the corresponding pad count of pixels above threshold has a range 0-40000, again fitting into two bytes.

with a noise rate above threshold at the 10^{-6} level, this would be 200 hits per pad per train, which therefore completely dominates.

Hence, for each pad, 200 of the 5000 bunch crossings would give a non-zero value in the pixel sum and hence have to be read out. This is a 4% noise, however, all the noise gives a single hit, i.e. a sum of value one, to a very good approximation, and hence should not degrade the resolution or pattern recognition.

Each sum can be up to $4 \times 10^4 \sim 2^{16}$ so two bytes are needed per pad per bunch crossing, or 10 kBytes per pad. For 4×10^4 pixels, each capable of storing 32 bytes, the total memory storage area available is around 1 MByte per pad, which is easily sufficient.

This is an average raw data volume of 400 bytes per pad. The values would also need timestamps in the range up to 5000, which requires 13 bits, i.e. 2 bytes. The pad number would be sent out first; with 200 hits per pads per bunch train, then it would be very unlikely for a pad to have no hits above threshold so this arrangement would mean each hit above threshold does not duplicate the pad number.

Hence, the timestamps effectively double the data volume to 800 bytes per pad, 200 kBytes per wafer, 3.2 MBytes per PCB, 6.5 MBytes per module and 16 GBytes in total for the ECAL. An example of such a format is shown in figure 6.

The data would need to be transfered from the wafer to the FPGA at the end of the PCB within the inter-train dead period. At 5 Hz, this is a minimum rate of 1 MByte/s from each wafer. An example of how this could be achieved would be to use the 5 MHz clock, which is available for the beam crossings, to also clock out the data. Using eight differential data signals from the wafer, requiring 16 I/O pads, then the data rate would be 5 MBytes/s. Then the 200 kBytes from each wafer could be transfered to the FPGA in 0.04 s, easily within the inter-train time of 0.2 s.

These FPGA data would also have to be transmitted off the PCB within the inter-train dead period. At 5 Hz, this is a minimum rate of 16 MBytes/s from each PCB, which is orders of magnitude below what is standard on a fibre. Hence, the data can be sent out well within the dead period and so the PCB powered off for the rest of the time.

3.3 Outstanding issues

3.3.1 Choice of area to sum over

Can the area summed over be made configurable, or even event-by-event adaptable depending on the hit density for that beam crossing?

3.3.2 Crosstalk requirement

Is the requirement from simulation for crosstalk reduced in this option? If summing over a large number of pixels, the number of pixels hit is less sensitive to double counting or edge inefficiencies.

4 Pixel readout option

4.1 Overview of concept

This option is conceptually simpler; instead of summing over the bits set, the individual timestamps for when each pixel was above threshold are read out. Here, every pixel does an identical task and so it is likely a step-and-repeat can be used for the pixel layout.

There will need to be a restricted number of timestamps which can be stored per pixel. The timestamps can have values up to 5000, which would effectively require two bytes. Hence a maximum of 16 timestamps would be feasible.

Physically, the timestamp memory could be implemented as a clock counter continuously filling memory locations; the first gets frozen if the comparator bit is set for a beam crossing and the other continue to count, etc.

4.2 Data rates

How the one hit per pad per train due to beam interactions translates into pixels hit is not certain. An estimate would be that each shower which hits a pad gives a MIP going through on average 100 pixels in that pad. This means one hit per pad per train is equivalent to $1/400 = 0.0025$ hits per pixels per train. The noise rate of 10^{-6} per pixel per beam crossing results in a rate of 0.005 hits per pixels per train. Hence, the beam interactions and the noise are roughly the same size now. This is why the nominal rate to aim for is chosen to be 10^{-6}; it is not worth doing better as the beam background rate would then dominate.

In the following, a total rate of 0.01 hits per pixels per train is assumed. Hence, each pixel normally would not have a hit above threshold within a bunch train. Only pixels which have at least one beam crossing where a hit occured would be read out.

The pixel timestamps would require 13 bits, or in practise two bytes. Within a PCB there are 1.6×10^8 pixels, so each wafer and pixel label takes 28 bits, or in practise four bytes. An example of such a format is shown in figure 7.

Hence, each pixel hit requires 6 bytes on average. For a train, this is 600 kBytes per wafer, 10 MBytes per PCB, 20 MBytes per module and 50 GBytes total for the ECAL, all a factor of three higher than for the counter option.

The rate needed to transfer the data from each wafer during the inter-bunch period is 3 MBytes/s. Using the same assumptions as for the counter option, i.e. eight differential data

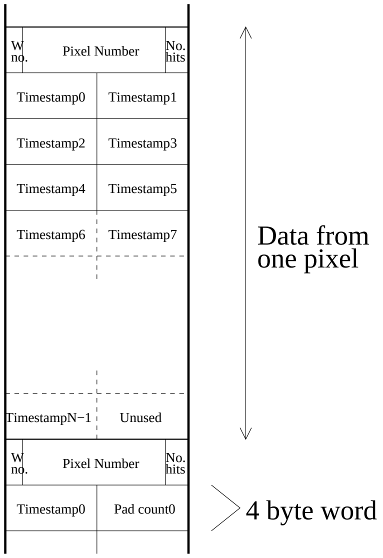

Figure 7: Example format for pixel readout option. The wafer number has a range 0-15 and so can fit into four bits. The pixel number within a wafer has a range 0-10^7 and so fits into three bytes. The number of hits cannot exceed the maximum number which can be stored, here assumed to be 16 and so is in the range 0-15, requiring four bits. The average will be one hit for those pixels which are hit at all. Each timestamp is in the range 0-5000, fitting into two bytes.

signals at 5 MHz, then the 600 kBytes from each wafer could be transfered to the FPGA in 0.12 s, still within the 0.2 s allowed.

The rate from each PCB needed at a 5 Hz train rate is 50 MBytes/s, still well below any fibre limit.

The masking of hot channels should be possible. However, since the number of hits is limited to 16, a hot channel would only ship out 32 bytes maximum anyway. The rate of bad (dead as well as hot) channels could be around per mille, so if half hot, half dead, then unmasked rate out per wafer per train is 32 bytes $\times 0.001/2 \times 10^7 = 160$ kBytes, so this would dominate. Hence, it is still worth having a configurable mask to remove such channels.

4.3 Outstanding issues

4.3.1 Noise requirement

Is there an extra requirement from pattern recognition (simulation) for noise confusion in next layer? As outlined in Sec. 2.5.3, a first estimate for pattern regonition would be that the noise rate could be orders of magnitude higher, but a detailed simulation study is needed to check this.

4.3.2 Data output

Can the pixel data gather be done efficiently during the wafer readout or does it have to be done in advance?

References

[1] R.-Y.Zhu, “Comments on LC Calorimetry,” Internation Conference on Linear Colliders LCWS02, Jeju Island, August 2002, and proceedings thereof.

[2] “TESLA: The superconducting Electron-Positron Linear Collider with an integrated X-Ray Laser Laboratory,” Technical Design Report, DESY 2001-011, March 2001.