The SiD Particle Flow Algorithm (v3, Jan 21)

M.J. Charles[1,2], U. Mallik[1], T.J. Kim[1], R. Cassell[3]

1- University of Iowa, Iowa City, Iowa 52242, USA
2- Oxford University, Oxford, United Kingdom
3- Stanford Linear Accelerator Center, Stanford, California 94309, USA

January 21, 2009

Abstract

A PFA has been developed for the SiD detector concept at a future Linear Collider. The algorithm is descibed in detail and the performance of the version of the algorithm, as used in the SiD LOI, is presented for a number of physics processes with two hadronic jets.

1 Introduction

Reconstruction in SiD is based on the Particle Flow concept in which calorimeter energy deposits from individual particles are separated, allowing the energy of each particle to be measured in the optimal subsystem for that particle (the silicon tracker for charged particles, the EM calorimeter for photons, both calorimeters for neutral hadrons). In the limit of perfect separation, the contribution to the jet energy resolution from charged particles is negligible and only neutral hadrons need to have their energy measured in the hadronic calorimeter, leading to a jet energy resolution of roughly[1] $20\%/\sqrt{E}$ [2]. In practice, this limit is difficult to achieve. Degradation of the resolution due to imperfect separation of energy deposits is generically referred to as confusion, and is the most important effect for well-contained, high-energy jets in the SiD acceptance. A particle-flow algorithm (PFA) has been developed and tuned for SiD in the `org.lcsim` software framework with the goal of minimizing the confusion and therefore the resolution. A snapshot of the PFA has been used for the analysis and benchmarking results reported in this LOI; development is still in progress and performance is expected to continue improving in future versions.

A deliberate effort has been made to keep the code as modular as possible. Different components communicate with one another by reading and writing named objects in standard formats to the event-level data store. This makes the flow of information clear, and allows one component to be substituted for another.

2 Algorithm description

For each event, the SiD PFA takes as inputs the energy deposits in the calorimeters and muon system and the set of tracks found in the tracking system (as described in Reference [3]). The PFA then performs the reconstruction in a series of steps, described below. The general strategy for

[1] E is in units of GeV throughout.

pattern-recognition in the calorimeters is (a) to identify and set aside the easiest, most distinctive showers first, taking maximum advantage of the information, and (b) to recognize common classes of mistakes made earlier in the algorithm and correct for them. The PFA produces as output a collection of reconstructed particles suitable for use in a physics analysis. The first step is to prepare and validate the input. The track reconstruction and calorimeter hit digitization packages are run, and any data which are unphysical or unmeasureable—such as calorimeter hits below an energy threshold or occuring more than 100 ns after the primary interaction—are removed.

The second step is to reconstruct electrons, muons, and photons, since these have distinctive signatures in the calorimeters. Muons are identified by extrapolating tracks through the ECAL and HCAL and requiring them to connect to a MIP stub in the muon system. Electromagnetic showers in the ECAL are reconstructed and identified with a dedicated "photon-finder" clustering algorithm. If the shower is connected to a track whose momentum matches the shower energy, it is taken to be an electron; if it is not connected to any track then it is taken to be a photon; and if it is connnected to a track with the wrong momentum then it is flagged as potentially mis-reconstructed. The latter can occur if the calorimeter deposits of a charged particle and a photon overlap, or if part of a hadronic shower is misidentified as a photon.

The third step is to reconstruct MIP segments in the calorimeters. Hadrons often travel a significant distance before showering and leave a distinctive signature of isolated hits. We find these by propagating each non-leptonic track through the calorimeter, layer by layer, until we can no longer find isolated or semi-isolated hits, either because the MIP segment has ended (typically with a hadronic shower) or because it has overlapped with the shower of another particle.

After setting aside the identified electrons, muons, photons, and MIP segments, the remaining hits are expected to be from hadronic showers. We now use a series of clustering algorithms to find the main structure of these showers. We begin with the DirectedTree clusterer, which groups hits around local maxima in hit density and is quite effective at matching peripheral elements of showers to the correct shower core. This serves as a useful guide in cases where there is little topological information, and in particular plays a large role in the fuzzy clustering step described later. However, the DirectedTree clusters are relatively coarse-grained and do not have the purity we need. We therefore make additional clustering passes, looking for substructure within the DirectedTree clusters: track segments, or dense clumps of hits. This substructure will then form the skeletons of the hadronic showers, together with MIP segments found earlier (and a number of special cases such as DirectedTree clusters with no identified substructure).

We assemble the skeletons of the hadronic showers with an iterative algorithm. We begin with the non-leptonic tracks, each of which is connected to a "seed" cluster in the ECAL—often, but not always, a MIP segment. Starting with the seed, we add clusters to the skeleton. The clusters to add are chosen based on a score which describes how well-connected a pair of clusters is. The way the score is calculated depends on the kind of clusters involved—for example, for a pair of MIP/track segments in the calorimeter we use a likelihood selector taking as inputs the distance of closest approach of the extrapolated track segments, the proximity of the hits in the clusters to the point of closest approach, and whether the point of closest approach is in the calorimeter. We build up the shower recursively, adding clusters which have high scores to be connected to the seed, then looking for further clusters which have high scores to be connected to the ones just added, and so on. We stop when there are no more clusters with high enough scores to add, or when the energy of the shower would become too large compared to the momentum of the track (by default we require $E - p < \sigma$). We then move on to the next track and begin the same process again—except that the clusters we just assigned are no longer available.

After attempting to reconstruct a shower for each track, we look for common mistakes. The reconstruction is fairly conservative by default, and sometimes misses parts of a shower for one of two reasons. Firstly, the score connecting the cluster to the rest of the shower may be too low to pick it up; we deal with this by loosening the score requirement if the cluster energy is too low ($E < |p| - \sigma$). Secondly, if the shower energy has an upward fluctuation it may be prevented from adding all of its clusters; we deal with this by loosening the requirement on $E - p$ in cases where the shower was prevented from picking up a cluster for this reason and the cluster was not subsequently assigned to another shower and the current shower does not already have a high energy compared to the track momentum. When loosening these requirements, we make a relatively small change at each iteration to avoid over-compensating.

In each iteration, the tracks are considered in order of increasing momentum: the reason for this is that lower-momentum tracks have smaller showers which are generally easier to reconstruct, so the risk of incorrectly adding clusters from another shower is reduced. However, in some cases two or more showers are badly overlapped and we are unable to separate them. In this case we group them together for the purposes of shower reconstruction, adding clusters that connect to any of them and requiring that the combined shower energy balance the sum of the tracks' scalar momenta.

After the last clustering iteration, we make final attempts to identify and correct mistakes in the charged hadronic showers. We look for showers whose energy is too low compared to the track momentum and for clusters that were not assigned to any shower, and attempt to match the two. We also look for unassigned clusters downstream of a shower whose energy is too low—these can be caused by secondary neutrals.

In addition to the skeletons of the hadronic showers, we have a large number of individual hits and small clusters whose association is not clear. These typically come from secondary photons or from soft neutrons displaced from the detector material during a hadronic shower. The most likely source is the nearest shower, but since there is little pointing information and secondary neutrals can sometimes travel a significant distance, this may not be correct. We handle this case with a fuzzy clustering technique: the energy in these small fragment clusters is split between any nearby showers which could have contributed in a probabilistic way, favouring closer showers over distant ones. Where possible, this sharing uses information from the DirectedTree clustering pass: fragments which are inside a DirectedTree envelope cluster are shared preferentially with showers inside the same DirectedTree cluster. This energy sharing is handled implicitly throughout the shower-building process, so that associated fragments are taken into account when testing the energy of a shower during reconstruction.

At this point, the only remaining clusters should be from neutral hadrons. We apply a simplified version of the charged hadron reconstruction to these clusters—since there is no track, we cannot make energy-momentum comparisons and there is no need to iterate—and form neutral hadron showers. (Some of these may be misidentified photons; we look for special cases such as when the cluster was flagged earlier as a photon-MIP overlap and move them to the photon list instead.)

The final step is to produce a list of reconstructed particles suitable for physics analysis. This list contains electrons, muons, charged hadrons, neutral hadrons, and photons. The momentum of each charged particles is taken from its track fit; together with the appropriate mass hypothesis (e, μ, π) this defines the four-vector. The energy and direction of neutral particles are computed from the calorimeter energy deposits, and the four-vector is again defined assuming the appropriate mass hypothesis (K_L or γ). We also consider tracks which were not matched to energy deposits in the calorimeter as a special case. If the track lies outside the calorimeter acceptance then we

assume the particle was real and missed, and therefore add it to the output with the pion mass hypothesis. If the track is inside the calorimeter acceptance then the most likely explanations are that the track-cluster matching failed or that the track decayed or interacted before reaching the calorimeter; in either case the energy of the particle reached the calorimeter and will likely already have been included (e.g. as a neutral hadron), so to avoid double-counting we do not put an additional particle with the track's three-momentum in the output.

3 Performance

The true test of performance is the sensitivity to key physics observables—this is discussed in detail in Reference [5]. However, for the purposes of studying and optimizing a PFA it is helpful to look at specific physics processes which are simple to analyse and depend primarily on the quality of the PFA output. We use two such processes:

- $e^+e^- \to q\bar{q}$ at $\sqrt{s} = 100, 200, 360, 500$ GeV, for $q = u, d, s$. Beamstrahlung and bremstrahlung in the initial state are disabled so that the collision energy $E_{\rm CM}$ is the same as $\sqrt{s}$. The figure of merit is the event energy sum residual $\Delta E_{\rm CM}$, i.e. the signed difference between the reconstructed and true values of $E_{\rm CM}$. Plots of the residual distribution are shown in Figure 1. Under the simplifying assumption that the invariant mass of two jets with energies E_1 and E_2 and opening angle θ_{12} is given by $m_{12}^2 = 2E_1E_2(1 - \cos\theta_{12})$, the resolution of energy sum residuals is equal to the resolution of the dijet mass for jets of the same energy.

- $e^+e^- \to Z(q\bar{q})Z(\nu\bar{\nu})$ at $\sqrt{s} = 500$ GeV, for $q = u, d, s$. The figure of merit is the dijet mass residual ΔM, the signed difference between the reconstructed and true values of $m_{q\bar{q}}$. Plots of the residual distribution are shown in Figure 2.

In both cases, the figure of merit depends upon the quality of hadronic jet reconstruction but does not require jet-finding or corrections for primary neutrinos.

Table 1 shows the measured resolutions[2] for the `sid02` detector. The resolution is quoted separately for the barrel ($0 < |\cos(\theta)| < 0.8$) and endcap ($0.8 < |\cos(\theta)| < 0.95$) regions of polar angle. There are several effects at work:

- The calorimetric component of the resolution function is expected to scale as $\sqrt{E}$, i.e. slower than linear. When this dominates, the fractional resolution ($\sigma_{\Delta E_{\rm CM}}/E_{\rm CM}$) decreases as the energy goes up.

- The confusion component of the resolution function will increase as the jet energy goes up and pattern-recognition becomes harder. The energy-dependence is not known from first principles, but it is likely to be at least linear if not faster.

- At high energies, leakage of energy out of the back of the calorimeter becomes important. The impact on the resolution has a strong angular dependence, since the effective depth of the calorimeter varies with $\cos\theta$; this is illustrated in Figure 3. This effect is modest for jet energies of 180 GeV but becomes dominant by 250 GeV. It is partially mitigated in the endcap region by using the muon system as a tail-catcher; this depends strongly on the longitudinal

[2]Resolutions are quoted in terms of rms_{90}, the RMS of the contiguous block of 90% of events with smallest RMS. Similarly, μ_{90} is defined to be the mean of these events. Note that for a Gaussian distribution, the rms_{90} is approximately 78% of the full RMS.

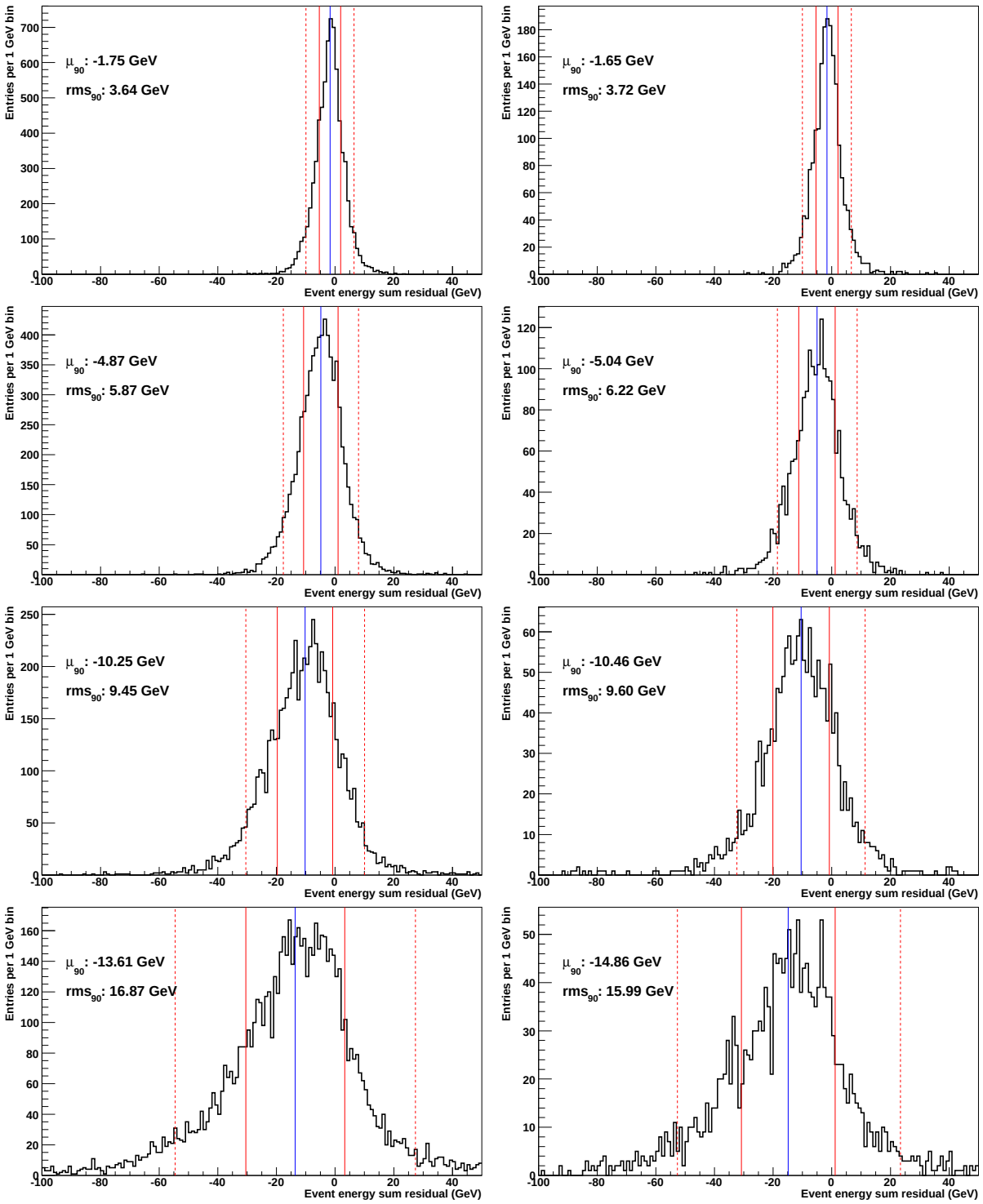

Figure 1: Energy sum residuals for $e^+e^- \to q\bar{q}$ events at $\sqrt{s} = 100, 200, 360, 500$ GeV (top to bottom), shown for the barrel (left) and endcap (right) regions of polar angle. The dashed lines indicate the 90% of events with smallest RMS, and the solid lines indicate the mean and RMS of those events.

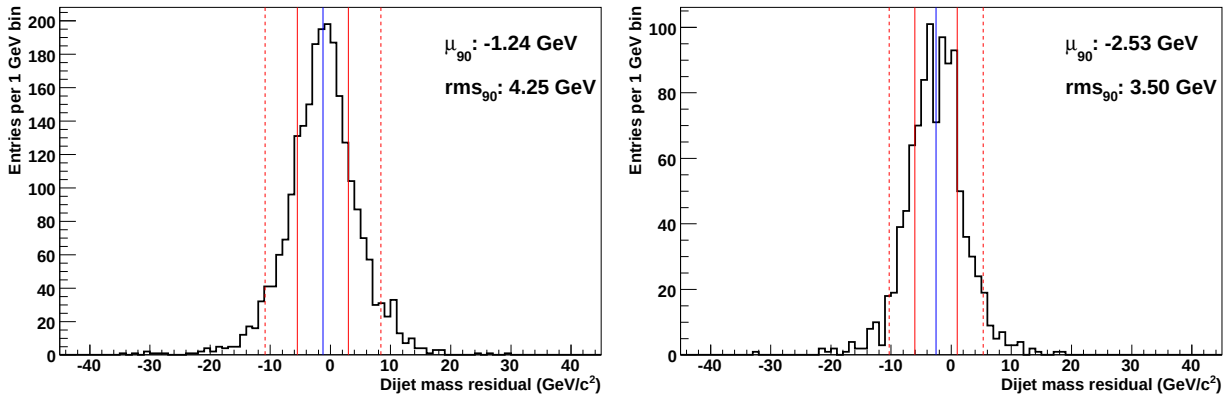

Figure 2: Dijet mass residuals for $e^+e^- \to Z(q\bar{q})Z(\nu\bar{\nu})$ events at $\sqrt{s} = 500$ GeV, shown for the barrel (left) and endcap (right) regions of polar angle. The dashed lines indicate the 90% of events with smallest RMS, and the solid lines indicate the mean and RMS of those events.

	Resolution (real tracking)		Resolution (cheat tracking)	
Process	Barrel	Endcap	Barrel	Endcap
$e^+e^- \to q\bar{q}$, $\sqrt{s} = 100$ GeV	3.7%	3.8%	3.4%	3.5%
$e^+e^- \to q\bar{q}$, $\sqrt{s} = 200$ GeV	3.0%	3.2%	2.8%	3.0%
$e^+e^- \to q\bar{q}$, $\sqrt{s} = 360$ GeV	2.7%	2.7%	2.6%	2.6%
$e^+e^- \to q\bar{q}$, $\sqrt{s} = 500$ GeV	3.5%	3.3%	3.5%	3.4%
$e^+e^- \to Z(q\bar{q})Z(\nu\bar{\nu})$	4.7%	3.9%	4.2%	3.7%

Table 1: PFA performance for `sid02`. For the $e^+e^- \to q\bar{q}$ processes, the rms_{90} of the energy sum residuals is quoted as a fraction of $\sqrt{s}$, and for the $e^+e^- \to ZZ$ process the rms_{90} of the dijet mass residuals is quoted as a fraction of m_Z. Resolutions are quoted for the LOI production snapshot and do not include subsequent improvements.

segmentation in the muon system and was found to be much more effective with an absorber thickness of 5 cm than 20 cm.

- For the ZZ events, the requirement that both jets lie in the angular region of interest constrains the kinematics of the decay.

- The dijet mass resolution measured in $e^+e^- \to ZZ$ events is observed to be larger than the resolution seen in $e^+e^- \to q\bar{q}$ events, even when the jet energy is comparable. This may be due to non-linearity in the energy response: the $e^+e^- \to q\bar{q}$ events have mono-energetic jets by construction and so a non-linear response would simply shift the mean of the energy sum distribution, whereas the jets in $e^+e^- \to ZZ$ events can be quite asymmetric and therefore the dijet mass residual distribution would be broadened by such an effect.

The resolutions in Table 1 are larger than those seen when running the PandoraPFA algorithm on the ILD detector design [6]. Understanding this difference is not straightforward: the performance of a PFA and the design of the detector on which it runs are coupled and it is not meaningful to take either in isolation. It is also technically very difficult to run one PFA on the other detector. However, a work-around has been developed: by starting from the LDC00Sc detector and adjust-

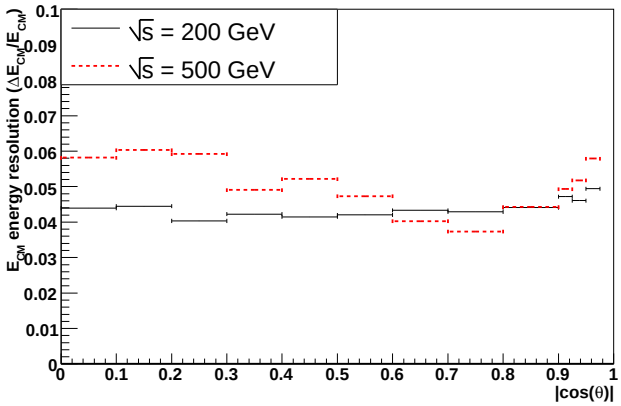

Figure 3: Resolution as a function of angle for $e^+e^- \to q\bar{q}$ for $\sqrt{s} = 200, 500$ GeV. For 200 GeV (solid), leakage is not significant and the angular distribution is roughly flat, rising slowly towards $|\cos\theta| = 0.975$ as the effects of acceptance and tracking become important. For 500 GeV (dashed), leakage has a major impact—this can be seen from how the resolution varies in the barrel between $|\cos\theta| = 0$ where the calorimeter thickness is minimized to $|\cos\theta| = 0.8$ where it is greatest. For both energies, the resolution is very bad for $|\cos\theta| > 0.975$ due to acceptance losses.

ing the calorimeter geometry and layering, we can produce "SiDish" detectors which have similar dimensions to `sid02` and run PandoraPFA v2.01[3] on them [7]. Note that the SiDish detectors still use the same detector technology as LDC00Sc, though: a TPC tracker and iron/scintillator HCAL, unlike SiD's silicon tracker and iron/RPC HCAL.

The $e^+e^- \to q\bar{q}$ event energy sum resolution in the barrel region ($0.0 < |\cos\theta| < 0.7$) found when running PandoraPFA on SiDish detectors resembling `sid02` is 3.1% for $\sqrt{s} = 90$ GeV and 2.8% for $\sqrt{s} = 200$ GeV, superior to the performance we find in Table 1 (3.7% and 3.0%, respectively). Part of this difference is due to the difference between `sid02` and the SiDish detectors. In previous studies comparing SiD detectors with scintillator and RPC instrumentation of the HCAL, we found that the scintillator variant had better performance by about 10% relative (0.3% absolute). Likewise, the use of a TPC tracker gives more complete information for decays and interactions inside the tracking system (e.g. for $K_S \to \pi^+\pi^-$); we can place an upper bound on this of 0.3% for $\sqrt{s} = 100$ GeV and 0.2% for $\sqrt{s} = 200$ GeV from studies with cheat tracking. These effects are sufficient to explain most of the observed performance difference between PandoraPFA and the SiD PFA.

4 Conclusions

There has been a great deal of progress in SiD reconstruction since LCWS08. We have switched to full track reconstruction and found that PFA performance in $e^+e^- \to q\bar{q}$ events remains close to that of cheat tracking. The PFA itself has been largely rewritten and gives event energy sum resolutions of order 3.0–3.5% for jet energies up to 250 GeV. The PFA performance was found to be approaching that of the gold standard, PandoraPFA, when running on a comparable detector design for jet energies up to 200 GeV. This is very encouraging for the jet physics prospects at SiD.

Nonetheless, there is a great deal of improvement still to come. A number of code fixes have

[3] PandoraPFA has continued to develop while this study was carried out; at the time of writing the current version is v03-β.

already been made [4] and more substantial revisions such as the integration of calorimeter-assisted tracking [8] are planned. At the broadest level, the two principal challenges are: (1) to understand the impact of leakage in high-energy jets on the physics potential of the detector, and to reduce it by adapting the algorithm and detector design if needed; and (2) to improve the reconstruction algorithm, and in particular to reduce the dijet mass resolution seen in $e^+e^- \to ZZ$ events. The confusion term still dominates the resolution for the range of jet energies likely to be used in physics analyses at a 0.5 TeV or even 1 TeV collider: we have plenty of room for improvement.

5 Acknowledgments

Funding bodies etc

References

[1] `http://ilcagenda.linearcollider.org/contributionDisplay.py?contribId=143&sessionId=23&confId=2628`

[2] R. Cassell, *SiD: Separating Detector Performance from PFA Confusion*, LCWS08, Chicago.

[3] R. Partridge, *SiD Track Reconstruction*, LCWS08, Chicago.

[4] T.J. Kim, *Implementation of PFA and Muon Identification*, LCWS08, Chicago.

[5] T. Barklow, *Status of SiD Benchmarking*, LCWS08, Chicago.

[6] D. Ward, *PFA Progress and ILD Detector Optimization*, LCWS08, Chicago.

[7] M. Stanitzki, *Detector Optimization for SiD*, LCWS08, Chicago.

[8] D. Onoprienko, *Integrated Tracking-Clustering Algorithm*, LCWS08, Chicago.