ILD MC production for detector optimization

Akiya Miyamoto∗†
High Energy Accelerator Research Organization (KEK),
1-1 Oho, Tsukuba, Ibaraki, 305-0801 Japan,
and

Hiroaki Ono
Nippon Dental University School of Life Dentistry at Niigata,
1-8 Hamaura-cho, Chuo-ku, Niigata 951-8580, Japan

Abstract
A large scale Monte Carlo production has been pursued since spring 2018 for the ILD detector optimization studies based on physics benchmark processes. A production system based on ILCDirac has been developed to produce samples in timely manner. The system and its performance are presented in this talk.

1 Introduction
The basic concept of the ILD detector for International Linear Collider was established in 2009 [1] and it has been used extensively to establish physics case of ILC DBD [2] and thereafter. However, the optimumness of the design was concerned in view of progresses in detector technologies and evolution of ILC accelerator design in recent years. To address the question, the detector optimization study based on physics benchmark processes based on a large scale Monte Carlo production has being conducted.

Two detector geometries, a large one called ILD_{l5_v02} and a small one ILD_{s5_v02}, were selected for comparison and they were implemented in the DD4hep [3] based detector simulation. In the calorimeter simulation, two types of read out for electro-magnetic calorimeter, silicon and scintillator, and two types for hadronic calorimeter read out, analogue and semi-digital, were implemented in order to save CPU time and clear comparison of model difference. The simulated data were reconstructed by Marlin data reconstruction framework.

The initial target of the MC production was to produce 500 GeV standard model samples which were used in the DBD study [2] with the same statistics. Differently from the DBD time, the interaction vertex points were shifted and smeared depending on the initial beam types and low energy $e^+/e^-$ pairs produced by beamstrahlung were overlaid to signal events.

*Talk presented at the International Workshop on Future Linear Colliders (LCWS2018), Arlington, Texas, 22-26 October 2018. C18-10-22.
†e-mail: akiya.miyamoto@kek.jp
together with lowpt hadron events of large cross section by $\gamma\gamma$ interaction. In addition, samples such as single particle events and quark pair events for detector calibration and basics performance studies were produced.

Before launching the production, required resources were estimated by processing small number of events for all processes. It was found that about 0.7M HEPSpec2006 CPU days and about 400 TB of data storage would be required to process samples planned. This was not a huge requirements for recent GRID infrastructure, however, it was also noticed that the reduction of baby sitting works associated with the production of more than 600 physics processes required in order to complete in timely manner.

2 Overview of the ILD production system with ILCDirac

ILD is using DIRAC [4] and ILCDirac [5] for GRID based MC production after DBD era. DIRAC provides a high level interface between users and GRID resources, such as job managements, file catalogue, transformation system for productions, web based interface for monitoring and control. ILCDirac is an extension for the ILC VO, being developed and operated by CLICdp. The common ILC software and production tools are implemented in ILCDirac.

ILD began to use ILCDirac after DBD and ILD MC data produced during DBD and thereafter have been imported to ILCDirac data catalogue.

In DIRAC system, production is conducted by the transformation system. When an operator request is received, it selects input files through database query, creates job scripts, and submit jobs to ILC VO cites. It retries job submissions, in case of job failures, several times because most of the job failure happens due to problematic sites or temporary network problem. Servers for the transformation and workflow are shared among ILCDirac users. The output file name is generated by the transformation workflow module, which was updated to support the ILD file name convention.

In ILD, the file name was constructed by a comma separated list of meta information consists of one character of key followed by meta value. The meta information includes the nominal beam energy and machine parameter, process name, detector model, etc. See Ref. [6] for details. For production purpose, file splitting number of generator files and job numbers were added in the file name for easy tracking of file creation order. As a consequence, file name length got close to 128 characters limit of ILCDirac and meta values had to be decided carefully some times.

In the production, input DBD generator files had to be split (GenSplit) to smaller number of events so as to run one job within allowed CPU time limit and produced data size within a limit. This task was done by a local script. The split files were created in slcio format to add header information such as the cross section of the processes and file sequence number, etc. The DBD generator files were produced with stdhep format which did not allow to store additional header information. In splitting, the number of events for each file were determined so that the reconstruction data size per job be within O(1GB) and CPU time for simulation be less than O(8 hour) depending on input generator processes.

The DBD samples at 500 GeV CM energy includes events not only by $e^+e^-$ collision but also $\gamma\gamma$ and $e\gamma$; final states from 2-fermion to 6-fermion. There were more than 600 processes, some of them could complete within a little CPU time; some consumed quite a lot of CPU time. In order to produce them efficiently, processes were grouped to about 50
groups considering the initial beam nature (IP smearing parameters and type of background to be overlaid). Number of jobs per transformation were also limited to be O(10k). The baby sitting work of production were made based on group.

After the one group of production, DST merge jobs were executed using ILCDirac UserJob tool. The DST files, which have typically several to a few tens of MB, were collected based on the physics process name and ID and merged keeping the order of file sequence in GenSplit.

Overall workflow of the ILD MC production is shown in Figure 1. In the Figure, the task done by ILCDirac production is shown by a red box, tasks done by ILCDirac user job tools are indicated by cyan boxes. Prior to start production, a ProdPara excel file (shown by pink disk) was created based on the production WishList and the Resource Needs estimation. It was used to group the generator processes and build a chain of scripts to perform GenSplit, DDSim and Marlin production, DSTMerge and Saving log for each group. The chain of scripts were monitored by a cron script, which also initiated jobs for DST merge and save log automatically.

ELOG system [7] was used to keep a track of DIRAC transformation and a database for the produced samples. ELOG uses ASCII data records, thus easy to use and flexible. A simple search function provided is also useful. Python interface to ELOG were utilized for automatic logging of production progresses and status. Our ELOG server is available at https://ild.ngt.ndu.ac.jp/elog. This server is also used for the database of generator samples produced since DBD era. Files of all generator meta data in json format is also provided.
3 Production results

The optimization MC production was launched late April 2018, using ILCSoft version v02-00. Unfortunately, the problem in v02-00 was found and new production was launched again late May 2018 with v02-00-01 ILCSoft. Since then, the production proceeded smoothly and most of the planned events could be produced within a month. Production of left over samples and additional samples continued time to time since then. This could be seen in Figure 2 which shows the cumulative normalized CPU used since April 2018.

![Figure 2: Cumulative normalized CPU in unit of Mega HEPSpecInt2006 units, since beginning of the optimization production.](image)

The size and the number of data produced this period is shown in Figure 3.

![Figure 3: Produced data size (left) and number of replicas (right) since middle of April 2018](image)
production, all produced data, namely simulated, reconstructed, DST, and DSTMerged data were kept and stored at DESY. DST-merged files and small part of files produced for calibration were replicated to KEK. Erroneous data produced by ILCSoft v02-00 were removed in September to save about 300TB of storage space.

Since DESY was the primary storage, its performance affected error rates of production. The rate of the all transfer and failed transfer to DESY monitored by ILCDirac are shown in Figure 4. It is seen that the data transfer rate exceeded 1.5GB/sec at peak and the error rates correlated to the transfer rate. In order to avoid high error rate, the number of jobs in the ILCDirac was restricted to less than about 5000 by limiting the number of gensplit jobs running in parallel or the number of tasks assigned to the transformation. Former method limited the number of new input files for the simulation jobs, thus restricted the job submission by transformation.

![Throughput by Destination](image1)

![Failed Transfers by Destination](image2)

**Figure 4:** Day by day records of throughput by destination (left) and failed transfer by destination (right).

### 4 Summary

A tool for ILD MC production has been developed using ILCDirac. It has been used for the production of MC samples for ILD detector optimization. About 1 PB of data has been produced since May 2018 and met the needs for the ILD detector optimization studies. Produced samples were mostly 500 GeV samples with ILD large and small geometries. Relatively small amount of samples of other energies, 91 GeV and 1000 GeV, and those for detector calibration and basic performance studies were produced as well. The ILD production tools worked smoothly and baby sitting works, which had been the major limitation in previous test productions (summer 2017-winter 2018), were reduced significantly. Currently, the data transfer rate is the major limiting factor of the production rate. The refinement of the production scripts are in progress, aiming for the use in coming 250 GeV production.
Acknowledgements

The authors would like to thank Andre Sailer and Marco Petrič for supporting ILCDirac for the MC production, Frank Gaede and ILD Software group members for discussion and comments during the course of this work.

References

[1] ILD Concept Group Collaboration, T. Abe et al., “The international Large Detector Letter of Intent,” arXiv:1006.3396 [hep-ex]

[2] H. Abramowicz et al., “The International Linear Collider Technical Design Report - Volume 4: Detectors,” arXiv:1306.6329 [physics.ins-det]

[3] M. Frank, F. Gaede, M. Petric, and A. Sailer, “Aidasoft/dd4hep,” Oct., 2018. https://doi.org/10.5281/zenodo.592244 webpage: http://dd4hep.cern.ch/.

[4] A. Tsaregorodtsev, M. Bargiotti, N. Brook, A. C. Ramo, G. Castellani, P. Charpentier, C. Cioffi, J. Closier, R. G. Daz, G. Kuznetsov, Y. Y. Li, R. Nandakumar, S. Paterson, R. Santinelli, A. C. Smith, M. S. Miguelez, and S. G. Jimenez, “DIRAC: a community grid solution,” J. Phys.: Conf. Ser. 119 (2008) 062048. https://cds.cern.ch/record/1176117

[5] on behalf of The CLIC Detector and Physics Study (CLICdp) Collaboration, C. Grefe, S. Poss, A. Sailer, and A. Tsaregorodtsev, “ILCDIRAC, a DIRAC extension for the Linear Collider community,” J. Phys.: Conf. Ser. 513 no. CLICdp-Conf-2013-003, (Nov, 2013) 032077.5. https://cds.cern.ch/record/1626585

[6] The naming convention of ILD file names and directories is available at https://confluence.desy.de/display/ILD/Conventions+of+file+names+and+directory+names

[7] “The ELOG Home Page.” https://elog.psi.ch/elog/index.html

[8] The latest code will be found at https://gitlab.cern.ch/amiyamot/ildprod