Computing for the ILD experiment

Akiya Miyamoto[1*]

[1]High Energy Accelerator Research Organization (KEK),
1-1 Oho, Tsukuba, Ibaraki, 305-0801 Japan

Wednesday 1st July, 2015

Abstract

Computing resources necessary for the ILD experiment are estimated based on the experience of Monte Calro productions for ILC TDR and the US snowmass study. All standard model processes were generated and simulated for these studies at the centre of mass energy of 250, 350 and 500 GeV. The CPU times used and the produced data size were used to estimate the ILD data taking at ILC assuming the H20 running scenario with the push-pull operation with SiD were assumed. The raw data size was based on the estimation studied for the ILC TDR.

Contents

*e-mail: akiya.miyamoto@kek.jp

1 Introduction

The computing cost for ILC experiments was not described in the ILC TDR because it was hard to estimate reliably developments of computing technology in more than ten years future. On the other hand, it is requested to estimate all resources for all period of the project in order to push forward the ILC project. In this document, a first attempt to meet this request is presented based on ILD software tools and recent experiences of Monte Carlo productions for DBD and the Snowmass study. Present day software and hardware are assumed for this estimation with many assumptions. Therefore, the estimation will subject to many changes.

The estimation described in this document is based on discussions at various occasions such as AWLC14, ILD Workshop@Oshu, a Fuse meeting, private communications, etc.

2 A model for ILD data processing

A possible ILD data processing model is shown in Figure 1. The front end electronics (F.E.) of ILD collects data and sent to the on-line computer system at the control room near IP. The purpose of the on-line computer system is to build a train data collecting sub-detector data from ILD, then send them to the main computer for data storage. The another role of the on-line computer is to select a few data for reconstruction and monitor data in order to ensure quality of data taken. It should also provide a temporary data storage for the case when data connection to the main computer is broken.

The main computer system receives data from the on-line computer and write them to a tape system as a raw data (**RD**). No data filtering will be applied before writing the raw data in order to maximize the data safety.

Subsequently, a real time data processing will be made to select events of interest for a train data. The procedure will include preliminary reconstruction, identify bunches of interest, apply calibration and alignments, background hit rejection, full event reconstruction and event classification. The reconstructed data are stored as Online Processed Data (**OPD**). In this real time processing, it is important to remove beam induced background hits as much as possible while keeping low pt low multiplicity events as much as possible. The reduction of data size helps data storage and transfer to other collaborating institutes in later stage. Same time, physics events with clear signature are written separately as Fast Physics Data(**FPD**). Calibration data (**CD**) are prepared in advance and used for the on-line data processing. CD is improved after on-line data processing and used by Off-line data processing to create higher quality data.

Raw data is store at the main campus and one copy of raw data will be stored for redundancy at sites other than ILC laboratory. FPD will be small enough to be replicated to many collaborating institutes. ORD contains raw data after background hit removal and reconstructed objects. It will be smaller than Raw data, but significantly larger than FPD. When high quality calibration data are created, OPD data are re-processed by GRID based computing and resultant data are stored as Off-line Reconstructed Data(**ORD**). It also produce condensed data sample as **DST**). In parallel, a large statistics Monte Carlo data production are produced.

3 ILD raw data size at various CM energies

ILD raw data size was estimated in TDR, which is shown in Fig. 2.

The ILD data size was dominated by those produced by pair background. It was estimated at 500 GeV. In order to evaluate the data size at different energy points, the number of simulator hits (`SimTrackerHit` and `SimCalorimeterhit`) per bunch crossing were studied using the samples simulated by ILD_o1_v05 model with Mokka.

The Mokka simulation files used for this study is shown in Table 1. 100 bunches of pair background event samples at the centre of mass energy of 250, 350 and 500 GeV were studied. Note that the magnetic field map, `fieldX02`, was used for the detector solenoid including the anti-DID field.

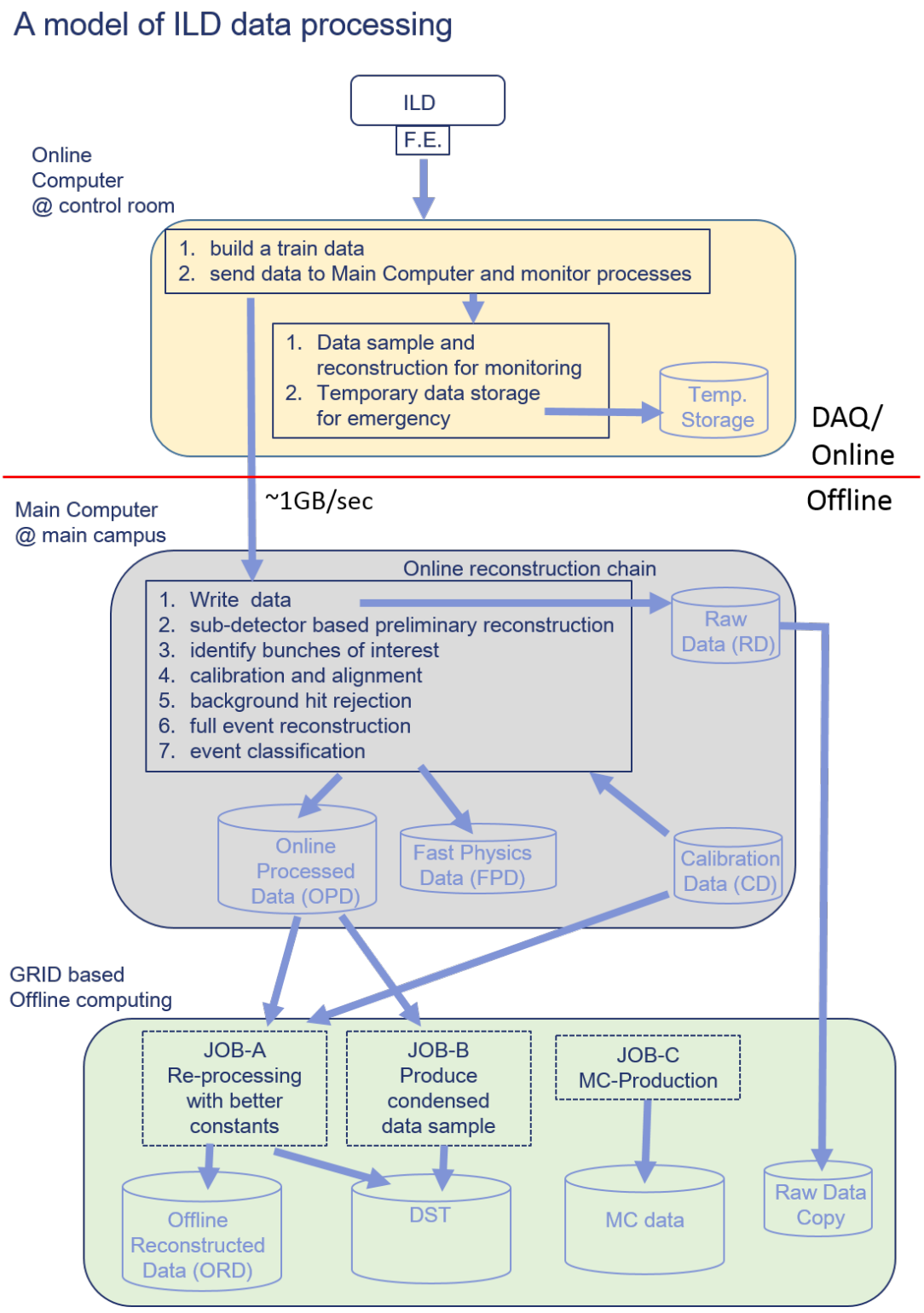

Figure 1: *An example of ILD data processing model*

Sub-detector	Channels [10^6]	Beam induced [Hits/BX]	Noise [Hits/BX]	Data volume per train [MB]
VTX (CPS)	300	1700	1.2	< 100
VTX (FPCCD)	4200	1700	1200	135
TPC	2	216	2000	12
FTD	1	260	0.3	2
SIT	1	11	0.3	6
SET	5	1		1
ETD	4			7
SiECAL	100	444	29	3
ScECAL	10	44	40	
AHCAL	8	18000	640	1
SDHCAL	70	28000	70	
MUON	0.1		8	≤ 1
LumiCal	0.2			4
BeamCal	0.04			126**

Figure 2: *ILD data size in TDR(table.xxx)*

Energy	file-path
250	`mc-dbd/ild/sim/250-TDR_ws/eepairs/ILD_o1_v05/v01-14-01-p00_fieldX02/` `sv01-14-01-p00_fieldX02.mILD_o1_v05.E250-TDR_ws.PBeamstr-pairs.I270000.`*NNNN*`.slcio`
350	`mc-dbd/ild/sim/350-TDR_ws/eepairs/ILD_o1_v05/v01-14-01-p00_fieldX02/` `sv01-14-01-p00_fieldX02.mILD_o1_v05.E350-TDR_ws.PBeamstr-pairs.I260000.`*NNNN*`.slcio`
500	`mc-dbd/ild/sim/500-TDR_ws/eepairs/ILD_o1_v05/v01-14-01-p00_fieldX02/` `sv01-14-01-p00_fieldX02.mILD_o1_v05.E500-TDR_ws.PBeamstr-pairs.I230000.`*NNNN*`.slcio`

Table 1: *Pair background samples used for background hit data estimation. NNNN is a bunch number, which is 1 to the maximum number of bunch in a pulse. In this study, 100 bunches are used.*

Table III-5.4. Pair induced backgrounds in the subdetectors for nominal 500 GeV and 1 TeV collision energy beam parameters [380]. The numbers for the ECAL and the HCAL are summed over barrel and endcaps. For the vertex detecor, the double-layer option has been chosen for this simulation. The TPC hits are the digitised hits that would be written to the data acquisition system. The errors represent the RMS of the hit number fluctuations of ≈ 100 bunch crossing (BX) simulations.

Sub-detector	Units	Layer	500 GeV	1000 GeV
VTX-DL	hits/cm^2/BX	1	6.320 ± 1.763	11.774 ± 0.992
		2	4.009 ± 1.176	7.479 ± 0.747
		3	0.250 ± 0.109	0.431 ± 0.128
		4	0.212 ± 0.094	0.360 ± 0.108
		5	0.048 ± 0.031	0.091 ± 0.044
		6	0.041 ± 0.026	0.082 ± 0.042
SIT	hits/cm^2/BX	1	0.0009 ± 0.0013	0.0016 ± 0.0016
		2	0.0002 ± 0.0003	0.0004 ± 0.0005
FTD	hits/cm^2/BX	1	0.072 ± 0.024	0.145 ± 0.024
		2	0.046 ± 0.017	0.102 ± 0.016
		3	0.025 ± 0.009	0.070 ± 0.009
		4	0.016 ± 0.005	0.046 ± 0.007
		5	0.011 ± 0.004	0.034 ± 0.005
		6	0.007 ± 0.004	0.024 ± 0.006
		7	0.006 ± 0.003	0.022 ± 0.006
SET	hits/BX	1	0.196 ± 0.924	0.588 ± 2.406
		2	0.239 ± 1.036	0.670 ± 2.616
TPC	hits/BX	-	216 ± 302	465 ± 356
ECAL	hits/BX	-	444 ± 118	1487 ± 166
HCAL	hits/BX	-	18049 ± 729	54507 ± 923

Figure 3: *Pair background hits shown in TDR*

From 100 bunches of these files, the number of SimTrackerHits and SimCalorimeterHits in sub-detector collections are counted, resulting Hits per bunch crossing summarized in Table 2. In the Table 2, the number of background hits per bunch crossing relative to those at 500 GeV are shown in columns 5 and 6. The name of the collections included in each sub-detector is shown in Table 3.

	Hits/BX			Relative to 500GeV		
Energy	250	350	500	250	350	500
VXD	807.8	1047.2	1889.8	0.427	0.554	1
TPC	1273.9	1984.5	4048.0	0.315	0.490	1
FTD	84.4	117.4	250.9	0.337	0.468	1
SIT	10.5	14.4	17.6	0.597	0.816	1
SET	0.3	0.8	0.9	0.297	0.824	1
SiECAL	99.0	160.6	321.6	0.308	0.499	1
AHCAL	3419.0	5782.3	18145.6	0.188	0.319	1
Muon	59416.6	61949.2	145783.9	0.408	0.425	1
LumiCAL	104.8	133.6	323.8	0.324	0.412	1
BCAL	172922.7	275519.2	703877.8	0.246	0.391	1
LHCAL	199.2	337.2	1153.1	0.173	0.292	1

Table 2: The number of sim hits created by pair background at the centre of mass energy of 250, 350 and 500 GeV (column 2 to 4) and the ratio relative to 500 GeV (column 5 to 6) for each detector components

In order to estimate the ILD data size, it is necessary to take into account the signal accumulation time of each sub-detector (namely how many bunches are accumulated) and a data size per Sim Hits. In stead of assuming these numbers, we scaled the beam induced background hits shown in Table 4 by the relative pair background hit ratio shown in Tabl 2. Noise hits per BX were kept constant. As a result, we obtain the raw data size at 250 GeV and 350 GeV in Table 4. The total data size at different center of energy is summarized in Table 5. For this table, the number of bunch crossings (BXs) and the pulse rate of H-20 running scenario [1] proposed by the ILC running scenario working group have been used.

The number of bunch crossings (BX) and the pulse rate described in the ILC TDR is assumed for the Table 5. The ILC running scenario working group has revised these parameters [1].

4 Information from MC samples

4.1 CPU time and data size

In order to estimate the CPU time and the data size for simulation, small number of events were simulated using the standard stdhep files generated for DBD and Snowmass studies. The number of events are 50 to 500 depending on processes. The simulation jobs were executed on KEKCC batch servers equipped with Xeon X5670@2.93GHz chips. Their performance has been measured as 14.72 HEP-Spec06 per core.

The results of the simulation are shown in Tables 9∼12 in the Appendix A. As seen in these tables, there were no strong dependences on beam polarization, thus only the case with e^- -80% and e^+ +30% polarization are considered in the following discussion. The CPU days necessary to simulate 1 fb^{-1} of data and its data size are summarized in Table 6.

From Table 6, the event rate per bunch crossing are obtained, which is shown in Table 7. In this table, the cross section of bhabha process, $e^+e^- \to e^+e^-$, was calculated using Grace including beam luminosity

Detector	Collactions
VXD	`VXDCollection`
TPC	`TPCCollection, TPCLowPtCollection, TPCSpacePointCollection`
FTD	`FTD_PIXELCollection, FTD_STRIPCollection`
SIT	`SITCollection`
SET	`SETCollection`
ECAL	`EcalBarrelSiliconCollection, EcalBarrelSiliconPreShowerCollection, EcalEndcapSiliconCollection, EcalEndcapRingCollection, EcalEndcapSiliconPreShowerCollection, EcalEndcapRingPreShowerCollection`
HCAL	`HcalEndCapRingsCollection, HcalBarrelRegCollection, HcalEndCapsCollection`
Muon	`MuonBarrelCollection, MuonEndCapCollection`
LHcal	`LHcalCollection`
LCal	`LumiCalCollection`
BCAL	`BeamCalCollection`

Table 3: `MCParticle` *and* `COILCollection` *were not counted.*

	datasize MB/train	Beam induced [Hits/BX]	Noise [Hits/BX]	byte/Hits	datasize(MB/Train) 250	350
VXD	135	1700	1200	35.13	89.7	99.7
TPC	12	216	2000	4.09	11.2	11.4
FTD	2	260	0.3	5.80	0.7	0.9
SIT	6	11	0.3	400.73	3.6	4.9
SET	1	1	0		0.3	0.8
ETD	7				0.0	0.0
SiECAL	3	444	29	4.79	1.1	1.6
AHCAL	1	18000	640	0.04	0.2	0.3
Muon	1	0	8		1.0	1.0
LumiCAL	4	1	0		1.3	1.6
BCAL	6.3	1	0		1.5	2.5
LHCAL						
Total	178.3	20634	3877.6		110.6	124.9

Table 4: *Data size of sub-detectors scaled from the estimation at 500 GeV presented in the TDR. Beam induced parts were scaled according the ratio of the number of pair background hits to those at 500 GeV. As sub-detector technologies, FPCCD was used for VXD. For calorimeters, SiECAL and AHCAL were used because they were assumed for the simulation of pair background.*

CM Energy	GeV	250	350	500
Number of BXs per train	BXs/train	1312	1312	1312
Pulse rate	Ntrain/sec	10	5	5
Annual Integrated Luminosity	fb^{-1}	120	80	144
Total data size	MB/Train	110.6	124.9	178.3
Data size/sec	MB/sec	1106.1	624.26	891.50
Data size(push-pull)/year	PB/year	8.85	4.99	7.13

Table 5: *Summary of raw data size at different centre of mass energy. The number of bunch crossings (BXs), pulse rate, and the total integrated luminosity are taken from the H-20 running scenario [1] proposed by the ILC running scenario working group. Only the case before the luminosity upgrade is shown. Due to the push-pull scheme of the ILD and SiD data taking, 1 year of ILD data taking is* 0.8×10^7 *sec and the integrated luminosity with the nominal operation condition is shown in the 4-th line.*

Energy	250	350	500
kEvents/fb^{-1}	2380.3	1927.5	2436.8
CPU days/fb^{-1}	228.3	237.2	364.4
Data size (GB)/fb^{-1}	261.5	228	297.5

Table 6: *Summary of the number of events, CPU days, and data size per unit integrate luminosity at the centre of mass energy of 250, 350, and 500 GeV for simulation.*

Energy	250	350	500	GeV
Number of bunches	1312	1312	1312	1/Train
Number of Train	10	5	5	1/sec
Total number of BX per ILD year	104960	52480	52480	M Bxs/1year
Total number of events per ILD year	286	154	351	M events/ILD year
Number of events/BX	0.272	0.294	0.669	%
Bhabha events/BX	0.265	0.180	0.159	%
(Bhabha+signal)/BX	0.537	0.474	0.828	%

Table 7: *The total number of events per 1 year for ILD (*0.8×10^7 *sec) and the event rate per bunch crossing for only signals (second last row) and adding signal and bhabha events (last row).*

spectrum at tree level. The angular range of $|\cos\theta| < 0.996$, which corresponds to the ECAL coverage, was calculated. From this table, we can conclude that less than 1% of bunch crossing contains meaningful events, however, several signal events are included in one train of collision.

4.2 Data size of reconstruction files and DST files

The data size of the reconstruction file is estimated from the ratios of the sizes between the reconstruction file and the simulated files, They are shown in Figure 4 in the case of 350 GeV samples. Each points in the figure represent one process type. The ratio slightly depends on process type but typically they are about 2. This is because detector hit corrections created by simulator hits has almost same size and both hits are kept in the output corrections. This strategy may change at the time of the experiments but for this excessive we keep the same assumption. 250 GeV and 500 GeV data show similar property.

The ratios of the DST file size and the reconstruction file size are shown in Figure 5. In DST files, detector hit collections are removed and only collections for physics analysis are kept, thus the sizes reduced significantly and they are about 3% or less.

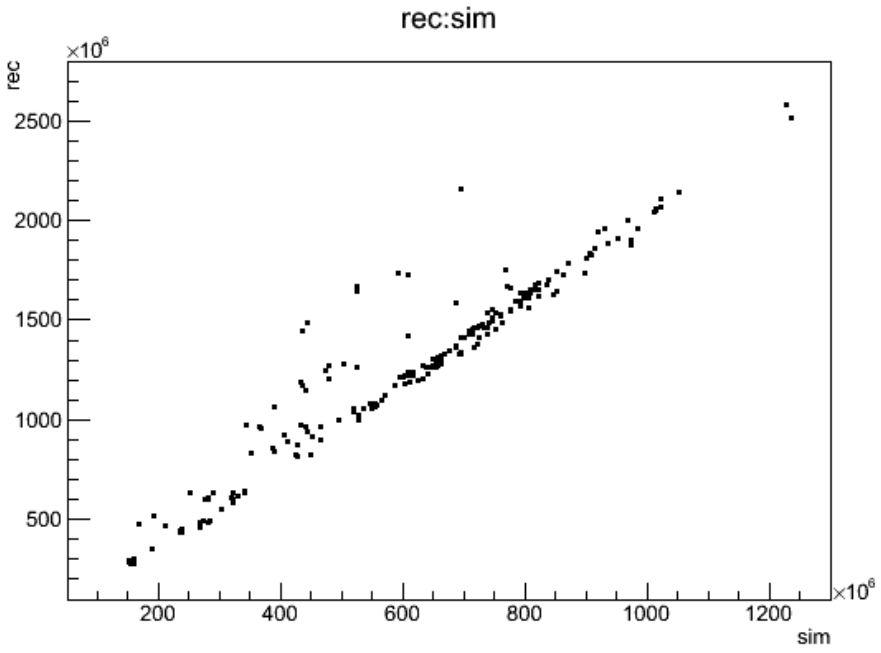

Figure 4: *A typical data size ratio of simulation data and reconstruction data file. The shown is the case for 350 GeV Monte Carlo. Each point corresponding to each process type.*

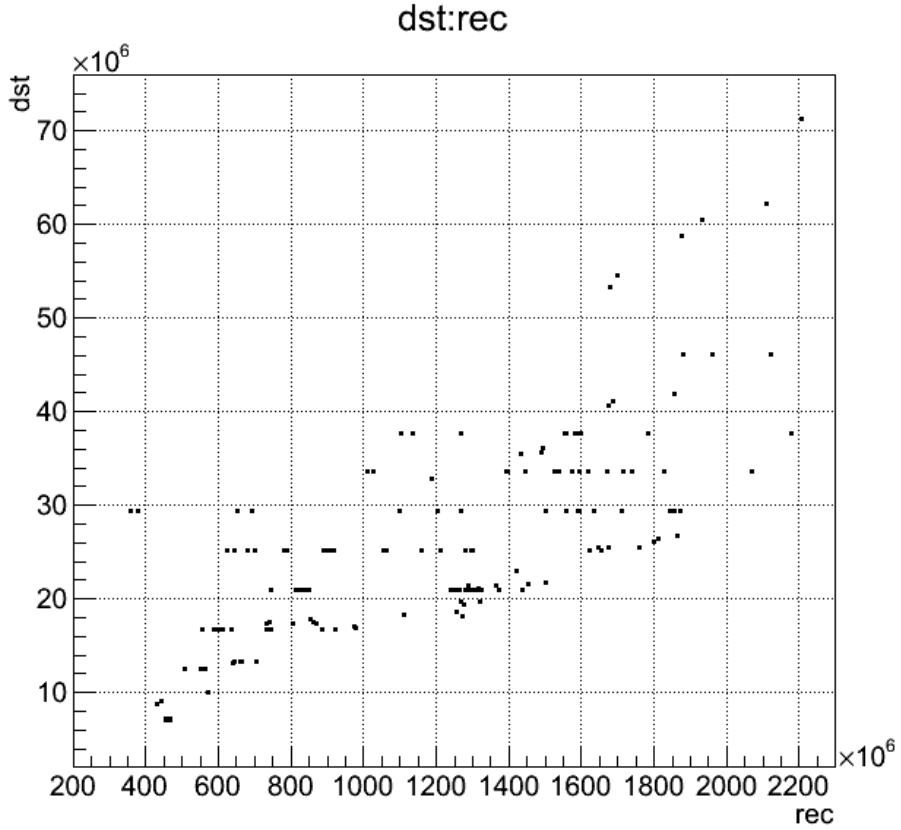

Figure 5: *A typical data size ratio of reconstruction data and dst data file. The shown is the case for 250 GeV Monte Carlo. Each point corresponding to each process type.*

5 Total storage size and CPU requirements

From the estimation of raw data size and MC CPU time, the total resources for ILD are estimated with following assumptions. It is summarized in Table 8. Assumptions used in this table are described below. Note that the numbers shown in this section assumes 1.0×10^7 sec of running per year at design luminosity. The actual data size and CPU resources needed will depend on the running scenario.

5.1 Data size

5.1.1 Online Processed Data (OPD)

The Online Processed Data is produced after selecting meaningful events from raw data. From table 7, only less than 1% of bunch crossings contains meaningful events. The purity of the event selection would be hardly 100% in order to keep high efficiency for meaningful event selection. Therefore, including a margin of factor 2, we assume 2% of raw data events are kept in OPD.

The data size of each event will increase after event processing. In the case of Monte Calro data, the data size after Marlin job increases by factor 2 for reconstructed data and another 3% for DST data. Adding these two factors, we assume the data size of each event increases by a factor 2.03 in On-line Processed Data.

Therefore, the size of OPD is obtained as

$$S_{OPD} = S_{RD} \times 0.02 \times 2.03 \tag{1}$$

where S_{RD} is the data size of raw data mad S_{OPD} is that of OPD.

5.1.2 Off-line Reconstructed Data

In the offline reconstructed data, another reconstructed data is added to OPD. Therefore, the size of ORD is obtained as

$$S_{ORD} = S_{OPD} + S_{RD} = S_{RD} \times 3.03 \tag{2}$$

where S_{OPD} is the size of ORD

5.1.3 MC data

The data size of simulated data are shown in Tables 9-11 in Appendix A for assumed integrated luminosity. Since there are not strong beam polarization dependence, we consider the case for -80% electron and +30% positron polarization. In addition, we assumed that

- we simulate 10 times more statistics than real data.
- A margin of factor 2 for events such as bhabha and two photon processes, which are not counted in Tables 9-11.
- Another factor of 2.03 to include reconstructed data and DST data

As a result, the size of the MC data is calculated by multiplying a factor $10 \times 2 \times 2.03 \times (integrated\ luminosity)$ to the size shown in Tables 9-11

5.1.4 Fast Physics Data and DST

They are small compare to the other data and neglected.

5.1.5 Data replication

We assume that

- One set of raw data is stored in the main computer system at ILC laboratory. One copy of raw data is stored at sites other than ILC laboratory, several sites around the world.
- OPD, ORD and MC DST are copied to 10 major GRID sites for quick access to these data. Their sizes will be small enough to allow many copies.
- MC simulation and reconstruction data are not replicated.

5.2 CPU time

CPU days to run Mokka simulation of all processes considered in DBD and Snowmass studies are shown in Tables 9-9 and summarized in Table 6.

5.2.1 MC data production

The total CPU time for productions of MC samples per year was calculated by the following formula:

$$MC\ CPU\ days = [CPU\ days/fb^{-1}] \times Int\_Lumi \times F_{event} \times F_{lumi} \times (1 + F_{rec}) \quad (3)$$

where Int_Lumi is the annual integrated luminosity, F_{event} is a factor to take into account CPU times for Bhabha and two-photon processes, etc, which are not counted in Table 6, F_{lumi} is a luminosity factor needed for MC samples, and F_{rec} is a CPU time for the reconstruction job relative to that of the simulation job. In the Table 8, $F_{event} = 2$, $F_{lumi} = 10$ and $F_{rec} = 0.2$ are assumed.

5.2.2 Online Processed Data

We calculated the CPU time to process one year of Online Processed Data as follows:

$$OPD\_CPU = [CPU\ days/fb^{-1}] \times Int\_Lumi \times F_{rec} \times F_{event}/F_{sig\_rate} \quad (4)$$

where $CPU\ days/fb^{-1}$ is the one for the signal Monte Carlo samples; F_{rec} and F_{event} are same as those used for the section above; F_{sig_rate} is the fraction of signal and bhabha events to the number of bunch crossing. All bunch crossing data has to be processed in the OPD processing and F_{sig_rate} is introduced to take into account this fact. As shown in Table 7, $F_{sig_{rate}}$ is less than 1%. Here we assumed $F_{sig_{rate}} = 0.01$. This would be equivalent to assume that the CPU to process raw data is same in all bunch crossing disregarding whether it contains signal events or not. This would be correct if the most of CPU time is consumed by reductions of background hits produced by pair background. Further study is required to improve this estimation.

5.2.3 Off-line Reconstructed Data(ORD)

The CPU time to produce Off-line Reconstructed Data was estimated as follows:

$$ORD\_CPU = [CPU\ days/fb^{-1}] \times Int\_Lumi \times F_{rec} \times F_{event} \quad (5)$$

5.2.4 Number of CPU cores

The CPU days to process data can be converted to the number of cores necessary to process in a given time. Except data for monitoring, it is not mandatory to process all data immediately. Note that ILD will collect only 1/4 of year (0.8×10^7 sec per year). Here we assume that it is sufficient to complete processing within 240 days (2/3 of year).

CM Energy	250	350	500	250Up	500Up	GeV
Annual Int. Lumi.	120	80	144	240	288	fb^{-1}
Nb. of Signal + Bhabha / BX	0.53%	0.47%	0.83%	0.53%	0.83%	
Raw data size/sec	1106	624	892	2213	1784	MB/sec
Data size at Campus per ILD year						
Raw Data (RD)	8.85	4.99	7.13	17.71	14.27	PB
On-line Processed Data (OPD)	0.36	0.20	0.29	0.72	0.58	id.
Off-line Reconstructed Data (ORD)	0.54	0.30	0.43	1.07	0.87	id.
MC Data (Sim+REC+DST)	1.27	0.74	1.74	2.55	3.48	id.
Sub Total	11.02	6.24	9.59	22.05	19.19	id.
Data size global per ILD year						
Raw Data (RD)	17.70	9.99	14.26	35.41	28.54	PB
Online Processed Data (OPD)	3.59	2.03	2.90	7.19	5.79	id.
Offline Reconstructed Data (ORD)	5.36	3.03	4.32	10.73	8.65	id.
MC data(Sim+Rec)	1.27	0.74	1.74	2.55	3.48	id.
MC DST with copy	0.19	0.11	0.26	0.38	0.51	id.
Total Data size per ILD year	26.86	15.16	21.76	53.74	43.55	PB
CPU						
MC production	658	455	1259	1315	2519	k CPU days
Online Processed Data	1096	759	2099	2192	4198	id.
Offline reconstructed (*)	11	7.6	21	22	42	id.
Total CPU days	1764	1222	3379	3529	6759	id.
# of cores to process data in 240 days						
all data	7351	5091	14080	14702	28160	cores
online data	4566	3162	8745	9132	17491	id.

Table 8: *ILD data size and CPU requirements per year (0.8×10^7 sec) with the push-pull operation with SiD. H-20 running scenario of the ILC running scenario is assumed. 250Up and 500Up are the case for the operation after the ILC luminosity upgrade.*

6 Evolution of data size for sample running scenarios.

Needs of data storage capacity and CPU cores depends on ILC running scenario. Here, we consider the H20 running scenario proposed by the ILC running scenario working group [1] and approved by the LCC recently. It assumes annual running time of 1.6×10^7 sec shared between ILD and SiD, therefore, the ILD running time is 0.8×10^7 sec. The collision rate at 250 GeV is twice larger than the ILC TDR and twice more bunches per pulse is assumed after the luminosity upgrade. The accumulation of the integrated luminosity at each energy is shown in Figure 6.

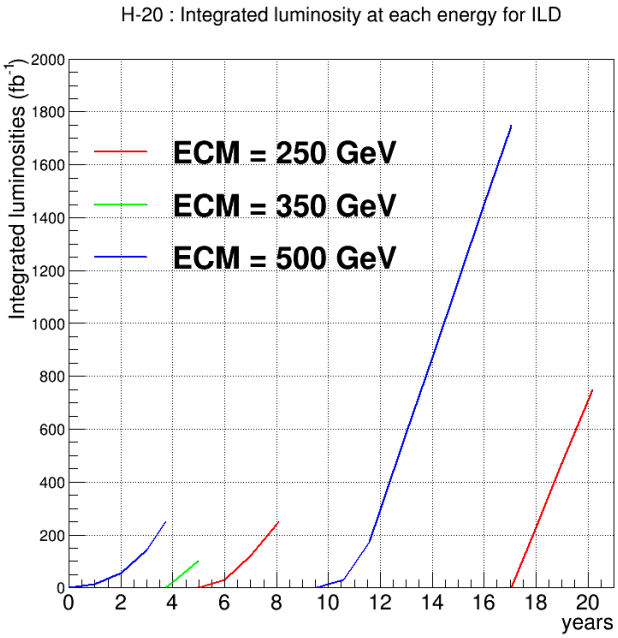

Figure 6: *Accumulated integrated luminosity of ILD at each energy points in the case of the H20 scenario.*

When the energy is fixed, the annual data size is proportionally depends on the annual integrated luminosity. Based on the data size shown in the Table 8, the growth of the accumulated data size was calculated and shown in Figure 7 and Figure 8.

Figure 7 shows that about 50 PB of storage for ILD will be required on the ILC campus before the shut-down of the ILC luminosity upgrade, and the global data size will be about 113 PB. The global data size after about 20 years of running will be about 550 PB. But lessons acquired during the first phase of ILC running will be applied to reduce the needs for storage space.

CPU resources necessary to process data taken by ILD at ILC campus is shown in Figure 9. Note that this resource is for the on-line data processing at campus. The last line of the Table 8 was scaled by the integrated luminosity in each year and the number of cores is converted in terms of HEP-Spec06. The HEP-Spec performance of the CPUs used for this study has been measured to be 14.72 HEP-Spec06/core. Note that the processing of 500 GeV data is most CPU demanding, however, due to the luminosity ramp-up assumption of H-20 scenario, the ILC luminosity reaches the design value only at 4-th year when the machine energy is reduced 350 GeV in the middle of 4-th year. Therefore, the CPU resources could be smaller than full year running at 500 GeV. Such case will happens after the luminosity upgrade.

Before the luminosity upgrade, ILD will need about 110 k HEP-Spec06 CPU cores on Campus and 172 k HEP-Spec06 globally. The need for the Campus computing facility could be reduced if some part of the processing can be done using the GRID computing infrastructure.

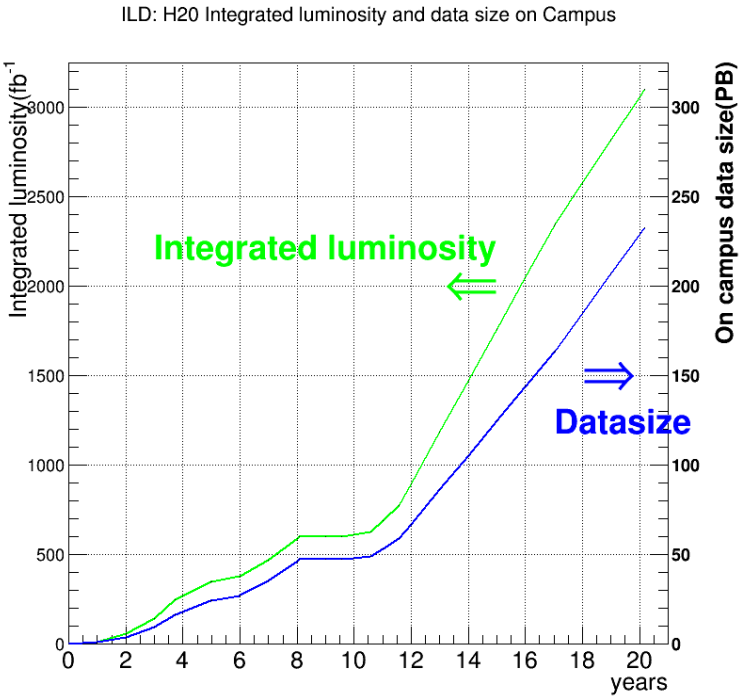

Figure 7: *Year by year growth of the accumulated ILD data at ILC campus. The left axis is the integrated luminosity and the right axis is for the data size.*

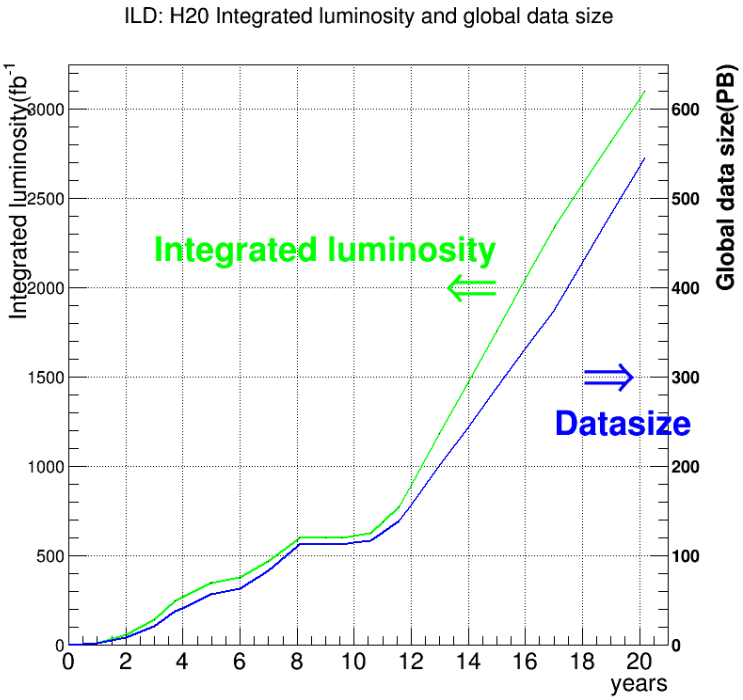

Figure 8: *Year by year growth of the accumulated ILD data the world wide.*

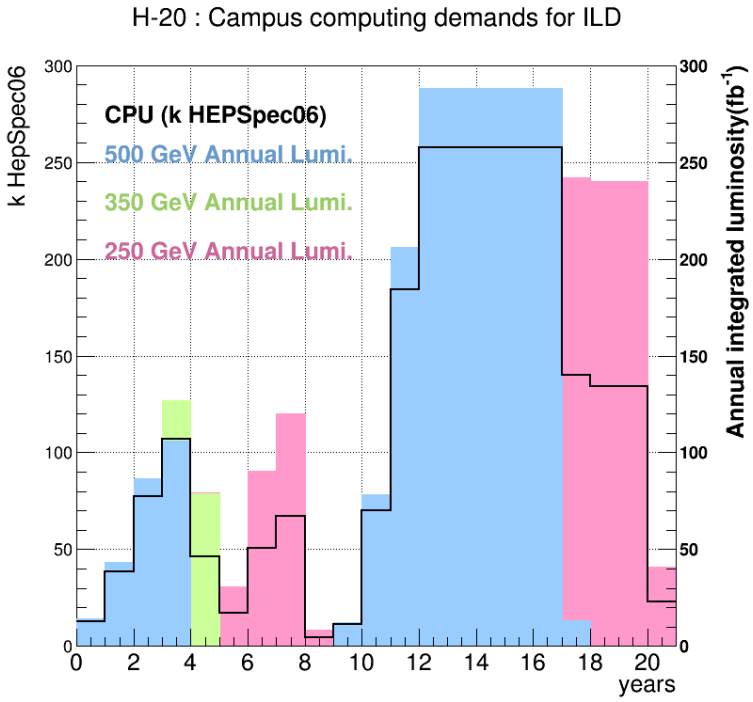

Figure 9: *Year by year needs of CPU resources for ILD on Campus is shown by a black histogram. The annual integrated luminosity taken at 500 GeV, 350 GeV and 250 GeV is shown by coloured histogram, cyan, light green, and pink, respectively.*

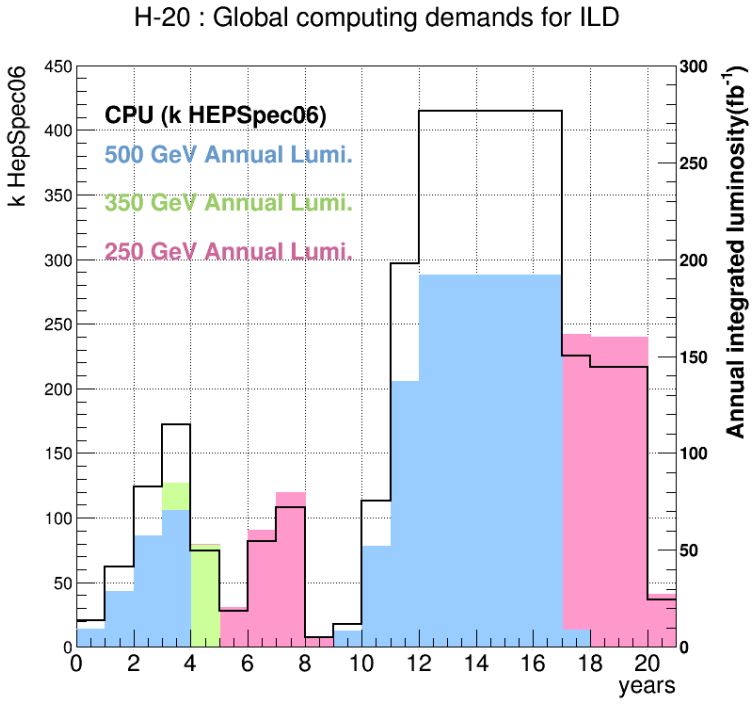

Figure 10: *Year by year needs of CPU resources globally for ILD is shown by a black histogram. The annual integrated luminosity taken at 500 GeV, 350 GeV and 250 GeV is shown by coloured histogram, cyan, light green, and pink, respectively.*

7 Summary

We presented a model of ILD data processing. Adapting the performances of ILD software used for DBD studies and current computing performance, we estimated the data size and CPU resources necessary for ILD experiment applying the H-20 running scenario, which plans to correct about 3100 fb^{-1} ILD data in total at 500 GeV, 350GeV and 250 GeV during the total running period of 21.2 years.

We concluded that the ILD data size at Campus before the luminosity upgrade ($\sim$ 8 years) will be about 50 PB. The data storage model is to store one set of raw data on the ILC campus and one copy of raw data globally off-campus. Including other type of files, the total ILD data size after about 21 years of data taking will be about 540PB.

The CPU resources for the 500 GeV data taking is most demanding. Before the shut-down for the luminosity upgrade, ILD will need about 100k on campus, which will increase to about 250 k HEP-Spec09 for the luminosity updated 500 GeV running. Globally, we will need 60% of more CPUs for MC sample production using GRID.

The results summarized here are based on many assumptions. Especially, the number of CPU cores on site is basically determined by the CPU time necessary to process raw data; the CPU time for calibration, alignment, event reconstruction of data including many backgrounds hits not seriously estimated and further investigations are necessary. Other issues to be improved include (1) efficiency of non-signal event filtering; (2) CPU time and data size of data taken during the luminosity ramp up period; (3) consistency of raw data size estimated in TDR and pair background simulation; (4) How many MC events do we need to produce ?; (5) How many data reprocessing do we need ? ; (6) resources during the construction; (7) Disk size necessary for data processing.

These issues need to be resolved and the estimation should be improved in further study.

Acknowledgments

Author would like to thank J. Strube, F. Gaede, N. Graf, M. Stanitzki and A. Sailer for helpful discussion.

A CPU time and data size for simulation at 250, 350 and 500 GeV

The CPU time and data size for simulation of 250 fb^{-1} for 250 GeV, 350 fb^{-1} for 350 GeV and 500 fb^{-1} for 500 GeV. The data shown in the table are estimated by simulating small number of events.

	250GeV/250fb^{-1},Pol(e-:-80%, e+:+30%)			250GeV/250fb^{-1},Pol(e-:-80%, e+:+30%)		
Process	k events	CPU days	Data size(GB)	k events	CPU days	Data size(GB)
1f	365541.4	36186.5	25259.6	365635.2	36176.8	25452.7
2f	29055.9	5595.8	13267.6	20299.6	3220.0	8724.9
3f	23087.4	2592.3	3405.4	22645.4	2638.3	3250.4
4f	10213.4	2373.2	8614.5	1274.8	127.5	858.7
aa_2f	165274.4	9879.9	13988.3	165274.4	9879.9	13988.3
aa_minj	1818.1	435.1	741.5	1818.1	435.1	741.5
ffh	79.9	22.4	100.7	51.5	13.6	58.2
Total	595070.4	57085.1	65377.6	576999.0	52491.2	53074.7

Table 9: *A summary of number of events (k events), CPU days and data size at 250 GeV obtained from Monte Carlo simulation samples.*

	350GeV/350fb^{-1},Pol(e-:-80%, e+:+30%)			350GeV/350fb^{-1},Pol(e-:-80%, e+:+30%)		
Process	k events	CPU days	Data size(GB)	k events	CPU days	Data size(GB)
1f	355028.8	52965.0	33423.1	355166.6	52816.4	32659.9
2f	18028.8	4693.1	9080.4	13605.7	3030.4	5203.6
3f	32033.8	5045.6	5727.0	31119.1	4802.6	5177.0
4f	7990.8	2386.4	6100.7	1074.5	149.4	652.6
6f	39.4	16.8	58.3	14.0	5.2	21.6
aa_2f	256407.9	16480.1	22809.6	256407.9	16480.1	22809.6
aa_4f	14.9	4.2	7.9	14.9	4.2	7.9
aa_minj	4765.6	1322.1	2174.7	4765.6	1322.1	2174.7
ffh	65.4	22.0	74.6	35.1	12.5	48.3
mixed_5f	10.3	3.3	7.6	7.4	1.9	4.7
mixed_6f	3.3	0.6	1.6	0.9	0.2	0.6
mixed_aa_4f	14.9	4.2	7.0	14.9	4.2	7.0
mixed_aa_minijet	3.4	0.9	2.0	3.4	0.9	2.0
tt	206.8	88.2	342.1	98.9	38.8	167.7
Total	674613.9	83032.5	79816.6	662328.9	78668.8	68937.2

Table 10: A summary of number of events (k events), CPU days and data size at 350 GeV obtained from Monte Carlo simulation samples.

	500GeV/500fb−1,Pol(e-:-80%, e+:+30%)			500GeV/500fb−1,Pol(e-:-80%, e+:+30%)		
Process	k events	CPU days	Data size(GB)	k events	CPU days	Data size(GB)
1f	446288.2	104310.4	48775.9	446424.9	104685.1	48666.5
2f	6599.5	2313.4	4527.7	4405.9	1312.9	2407.9
3f	66730.9	16202.0	13434.6	64136.5	15992.4	12206.4
4f	8111.7	2824.9	5241.4	2791.8	629.7	969.6
5f	35.8	18.2	28.4	24.8	10.4	18.9
6f	196.1	111.5	312.8	91.4	49.8	154.8
aa_2f	669274.0	50391.6	66076.9	669274.0	50391.6	66076.9
aa_4f	52.4	22.9	29.8	52.4	22.9	29.8
aa_minj	21127.6	5990.6	10309.1	21127.6	5990.6	10309.1
ffh	4.5	2.3	4.5	2.9	1.2	2.7
Total	1218420.9	182187.8	148741.2	1208332.3	179086.6	140842.7

Table 11: A summary of number of events (k events), CPU days and data size at 500 GeV obtained from Monte Carlo simulation samples.

B Subprocesses used for the cpu time and data size of simulation

process	sub-processes
1f	ae_ea, ea_ea
2f	2f_z_bhabhag, 2f_z_h, 2f_z_l
3f	ae_eee, ae_ell, ae_evv, ae_exx, ae_eyy, ae_lvv, ae_vxy, ea_eee, ea_ell, ea_evv, ea_exx, ea_eyy, ea_lvv, ea_vxy
4f	4f_sw_l, 4f_sw_sl, 4f_sze_l, 4f_sze_sl, 4f_szeorsw_l, 4f_sznu_l, 4f_sznu_sl, 4f_ww_h, 4f_ww_l, 4f_ww_sl, 4f_zz_h, 4f_zz_l, 4f_zz_sl, 4f_zzorww_h, 4f_zzorww_l
5f	ae_eeeee, ae_eeell, ae_eeevv, ae_eeexx, ae_eeeyy, ae_eelvv, ae_eevxy, ae_elevv, ae_ellll, ae_ellvv, ae_ellxx, ae_ellyy, ae_elvxy, ae_evexy, ae_evlxy, ae_evvvv, ae_evvxx, ae_evvyy, ae_exxxx, ae_exxyy, ae_eyyyy, ae_lllvv, ae_llvxy, ae_lvvvv, ae_lvvxx, ae_lvvyy, ae_vvvxy, ae_vxxxy, ae_vxyyy, ea_eeeee, ea_eeell, ea_eeevv, ea_eeexx, ea_eeeyy, ea_eelvv, ea_eevxy, ea_elevv, ea_ellll, ea_ellvv, ea_ellxx, ea_ellyy, ea_elvxy, ea_evexy, ea_evlxy, ea_evvvv, ea_evvxx, ea_evvyy, ea_exxxx, ea_exxyy, ea_eyyyy, ea_lllvv, ea_llvxy, ea_lvvvv, ea_lvvxx, ea_lvvyy, ea_vvvxy, ea_vxxxy, ea_vxyyy
6f	eeeeee, eeeell, eeeexx, eeeeyy, eellxx, eellyy, eeveev, eevelv, eeveyx, eevlev, eevllv, eevlyx, eexyev, eexylv, eexyyx, llllee, llllll, llvelv, llveyx, llvlev, llvllv, llvlyx, llxyev, llxylv, llxyyx, vvveev, vvvelv, vvveyx, vvvlev, vvvllv, vvvlyx, vvvvxx, vvvvyy, vvxyev, vvxylv, vvxyyx, xxveev, xxveyx, xxvlyx, xxxxee, xxxxll, xxxxvv, xxxxxx, xxxylv, xxxyyx, yycyyc, yycyyu, yyuyyc, yyuyyu, yyveev, yyvelv, yyveyx, yyvlev, yyvllv, yyvlyx, yyxyev, yyxylv, yyyyee, yyyyll, yyyyvv, yyyyyy
aa_2f	aa_ee, aa_ll, aa_xx, aa_yy
aa_4f	aa_eeee, aa_eell, aa_eevv, aa_eexx, aa_eeyy, aa_elvv, aa_evxy, aa_levv, aa_llll, aa_llvv, aa_llxx, aa_llyy, aa_lvxy, aa_vexy, aa_vlxy, aa_vvxx, aa_vvyy, aa_xxxx, aa_xxyy, aa_yyyy
aa_had	aaddhad
aa_minj	aamin_04_10_m1, aamin_04_10_m4, aamin_10_20_m1, aamin_10_20_m4, aamin_20_40_m1, aamin_20_40_m4, aamin_40_xx_m1, aamin_40_xx_m4
tt	tt-2l2nbb, tt-6q, tt-en4q, tt-ln4q
ffh	e1e1h, e2e2h, e3e3h, nnh, qqh

mixed_5f	mixed_5f
mixed_6f	mixed_6f
mixed_aa_4f	mixed_aa_4f
mixed_aa_minijet	mixed_aa_minijet

Table 12: The names of process and sub-process used for the study of Monte Carlo samples.

References

[1] T. Barklow, J. Brau, K. Fujii, J. Gao, J. List, N. Walker, K. Yokoya, "ILC Operating Scenarios," arXiv:1506.07830. ILC-NOTE-2015-068.