

# ILD MCProduction for detector optimization

Akiya Miyamoto

29 May 2018

ALCW2018

Simulation and Reconstruction session

# Contents

---

- MC production for optimization
- ILD MC production system using ILCDIRAC
- Status
  - ◆ ILCSoft-v02-00
  - ◆ slightly on ILCSoft-v02-00-01
- Summary

# Optimization production

## ■ ILD detector optimization

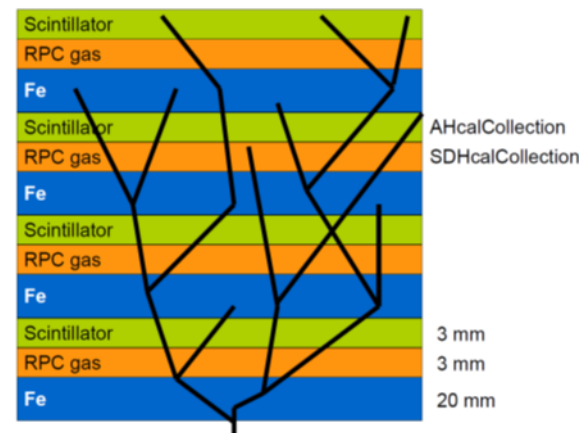
- ◆ Performance study based on a large scale MC samples.
- ◆ Large and small detector geometry
- ◆ Calorimeter options : Hybrid
  - Analog HCAL vs Semi-Digital HCAL

## ■ Software : full update since DBD

- ◆ Mokka → DD4Sim
- ◆ New/improved reconstruction tools are used.

## ■ Samples

- ◆ Calibration samples: single particles, UDS, flavor tag
- ◆ DBD 500 GeV SM samples. Same statistics as existing DBD samples
  - overlay pairs, as well as aa\_lowpt
  - correct IP vertex offset and smearing
- ◆ 250 GeV by Whizard2
  - with the latest ILC beam parameter generated by Whizard2
- ◆ More if allows



# Computing Resource for 500 GeV case

- CPU time and data size was estimated by KEK batch ( ~23HS06/CPU), 50 events/process, 1 Detector(large) and 1 CAL option

Process	Nb.Procs	k Evts	Nb. Jobs	CPU days		Data size(GB)		
				SIM	REC	SIM	REC	DST
uds	12	120	430	30	13	385	379	6
single	94	1,520	589	37	36	382	307	4
higgs	32	951	6,189	322	243	4,367	4,417	72
2f	8	3,780	16,475	1,198	706	14,386	13,740	215
4f	40	11,289	48,219	3,108	1,726	33,369	34,721	475
5f	200	2,029	7,627	520	342	6,175	6,442	116
6f	188	6,907	530,029	2,900	2,564	38,456	40,506	725
aa_4f	80	816	2,344	158	115	1,794	2,042	41
flavortag	5	535	5,329	290	262	3,943	4,142	76
Sum	659	27,947	617,231	8,563	6,007	103,257	106,696	1,732

- Resource needs for all samples.

Total for 500 GeV	CPU(HS06 year)	Storage (TB)
2 Size x 1 Cal	1,882	423
2 Size x 2 Cal	2,657	640

# DIRAC and ILCDirac

---

- **DIRAC** ( Distributed Infrastructure with Remote Agent Control) :High level interface between users and distributed resources



- ◆ Job managements, File catalog, ..
- ◆ Transformation system for productions
- ◆ Written in Python 2
- ◆ Web interface

- **ILCDirac** : An extension for the ILC VO and CALICE VO

- Developed & operated by CLICdp group since 2010.
- Provide simple interface for user jobs
- DIRAC file catalog for file and metadata
- Central system for large scale production

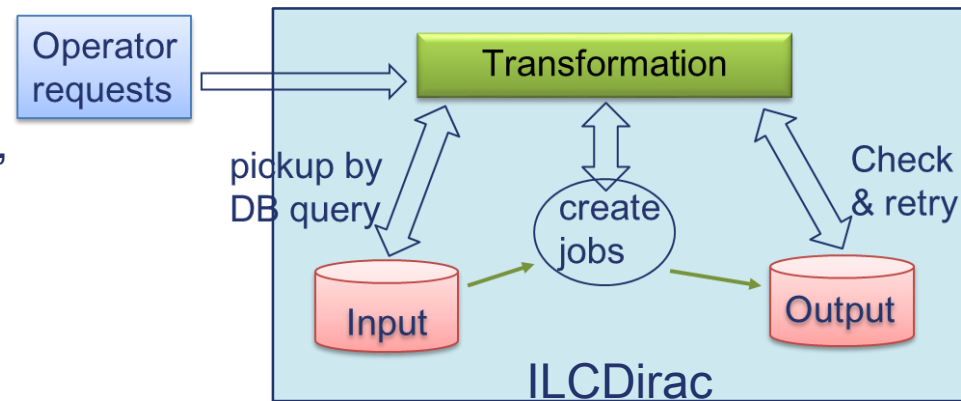


- ILD began to use ILCDirac after DBD
  - ◆ Essential tool for a large scale MC production

# Production with ILCDirac

## ■ Transformation

- ◆ Automatically creates job scripts, submit jobs, retry when failed, and maintain job logs, ...



- ◆ Jobs have to be submitted by the same steering. File names should be known to the server
  - **Input** file : picked up from DB based on query given by the operator request.
  - **Output** file: File name is generated by the transformation server and written to a defined directory.
  - **BKG overlay** : Relevant files are picked randomly and given to the job.
- ◆ *Workflow in the ILCDIRAC server and shared among other groups.*
  - ➔ Hard to modify or implement new workflow, though not impossible.

# ILD specific features

---

## ■ Naming rules for file and directory and meta key matters.

- ◆ Input file search by DIRAC is based on the meta keys attached to the directory and the file. In addition, names are constructed using meta keys.

ex: ...ild/rec/500-TDR\_ws/2f\_Z\_leptonic/ILD\_o1\_v05/v01-16-p05\_500/

rv01-16-p05\_500.sv01-14-01-p00.mILD\_o1\_v05.E500-TDR\_ws.l250108....

→ Modules for ILD naming rule have been prepared.

→ Avoid very long file name. Need to be within 128 characters.

## ■ Generator input ( DBD stdhep files ) are provided as files

- ◆ GenSplitting is mandatory for jobs to be within CPU limit.  
GenSplit is performed at local host, then uploaded to DIRAC
  - stdhep → Lcio conversion
  - Add split number in file name. Used later to merge DST files.
  - Add header information to LCIO header. Consistent with whizard2 files

## ■ Grouping of generator processes.

- ◆ Grouped by output directory, beam nature (  $\gamma_{\text{beam}}$  or not ), lepton or not, ...
- ◆ ~600 Gen. processes are grouped to O(50) groups

# ILD specific features - 2

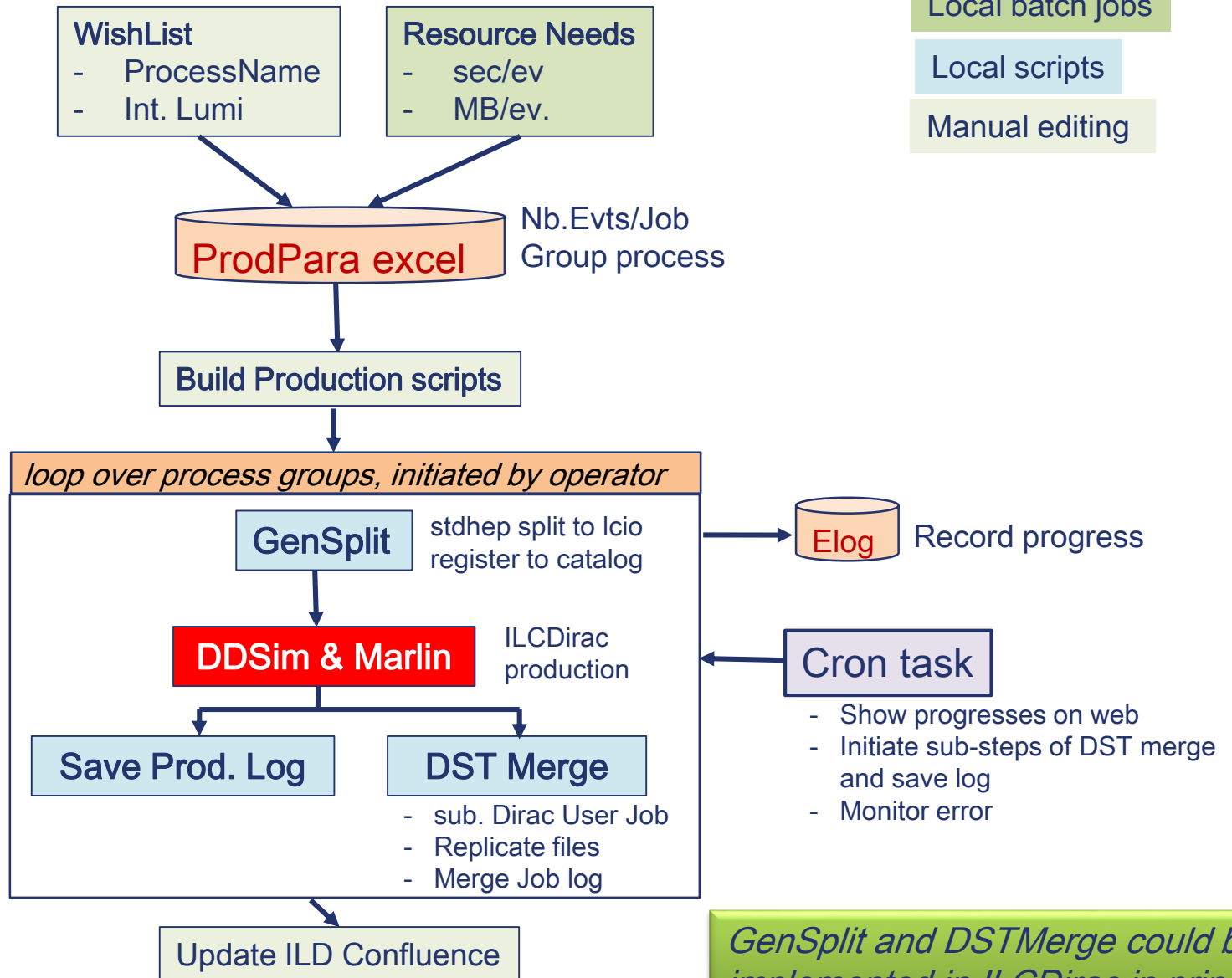
---

- Not all tasks of ILDProduction are provided by ILCDirac.
  - ◆ **DST Merge** :
    - Wait completion of all jobs.
    - Keep the order of gensplit using split number
    - ProcessID aware merging. Note many processIDs by 1 production.
  - ◆ **Log files**: not registered in the Catalog
    - ➔ Copied to the standard location after production.
  - ◆ **Data base** of produced samples
    - ILCDirac concept is to use **DIRAC catalog and meta information**
    - ILD : Providing ELOG server, <https://ild.ngt.ndu.ac.jp/elog/dbd-prod/>
      - ➔ Timely update of ELOG information



# ILD MC production workflow ( v02-00 )

Established since v02-00



Process ILCDirac

Local batch jobs

Local scripts

Manual editing

*GenSplit and DSTMerge could be implemented in ILCDirac in principle*

Searchable entries

Updated by scripts using python interface. manual correction possible

List | Find | Login

ILDConfig: v02-00-01

Full | Summary | Threaded

-- JobStatus -- v02-00-01 -- Ecm -- -- Detector -- -- ProcName -- 15 Entries

ID	Date	SimID	RecID	Ecm	ProcName	ProcID	Pol	ILDConfig	Detector	JobStatus	NEvents	Text
192	2018/05/26 15:29:52	10299,10300	10301,10302(ovl)	500	4f_ZZ_semileptonic	250014,250016	mixed-pol	v02-00-01	ILD_Is5_o1_v02	Production	444373 events will be submitted	2018-05-26 15:29:52 :
191	2018/05/26 12:55:25	10295,10296	10297,10298(ovl)	500	higgs_ffh	106524-106730	mixed-pol	v02-00-01	ILD_Is5_o1_v02	Production	239686 events will be submitted	2018-05-26 12:55:25 :
190	2018/05/26 09:26:34	10291,10292	10293,10294(ovl)	500	4f_singleZsingleWMix_leptonic	250049-250052	mixed-pol	v02-00-01	ILD_Is5_o1_v02	Production	676521 events will be submitted	2018-05-26 09:26:34 :
189	2018/05/26 09:12:44	10287,10288	10289,10290(ovl)	500	4f_singleZnunu_leptonic	250054,250056	mixed-pol	v02-00-01	ILD_Is5_o1_v02	Production	146824 events will be submitted	2018-05-26 09:12:44 :
188	2018/05/25 03:22:09	10273,10274	10275,10276(ovl)	500	6f_ttbar	108659-108676	mixed-pol	v02-00-01	ILD_Is5_o1_v02	Production	1916046 events will be submitted	2018-05-25 03:22:09 :
187	2018/05/24 22:22:05	10269,10270	10271,10272(ovl)	500	6f_ttbar	108655-108669	mixed-pol	v02-00-01	ILD_Is5_o1_v02	Production	2681636 events will be submitted	2018-05-24 22:22:05 :
186	2018/05/24 18:31:44	10265,10266	10267,10268(ovl)	500	4f_singleW_leptonic	250045-250048	mixed-pol	v02-00-01	ILD_Is5_o1_v02	MergeDST:logging	1477239 events will be submitted	2018-05-24 18:31:44 :
185	2018/05/24 17:11:46	10261,10262	10263,10264(ovl)	500	higgs_ffh	106515-108240	mixed-pol	v02-00-01	ILD_Is5_o1_v02	MergeDST:Logging_Done	650002	2018-05-24 17:11:45 :
184	2018/05/24 16:52:05	10257,10258	10259,10260(ovl)	500	higgs_ffh	108161-108164	mixed-pol	v02-00-01	ILD_Is5_o1_v02	Done	160000	2018-05-24 16:52:04 :
183	2018/05/24 12:52:06	10253,10254	10255,10256(ovl)	500	2f_Z_leptonic	250106,250108	mixed-pol	v02-00-01	ILD_Is5_o1_v02	MergeDST:Logging_Done	1430566	2018-05-24 12:52:05 :
182	2018/05/24 06:22:37	10241,10242		500	aa_lowpt	39213	eW.pB	v02-00-01	ILD_Is5_v02	Done	633600	2018-05-24 06:22:37 :

# Details of one Elog record

genmeta dbd-prod2 dbd-prod worklog simCPUSize test-prod

ILD Post-DBD MC production with ILCDIRAC including the optimization

Navigation icons: Home, Back, Forward, Refresh, List, Find, Login

Message ID: 172 Entry time: 2018/05/15 19:09:59

Worker:	A.Miyamoto
SplitID:	500192
SimID:	10129,10130
RecID:	10131,10132(ovl)
Ecm:	500
ProcName:	4f_WW_semileptonic
ProcID:	250018,250020
Pol:	mixed-pol
ILDConfig:	v02-00
Detector:	ILD_Is5_o1_v02
JobStatus:	Done
NEvents:	999820
Log_SimJob:	see attached file, log_upload_list.txt
Log_RecJob:	see attached file, log_upload_list.txt

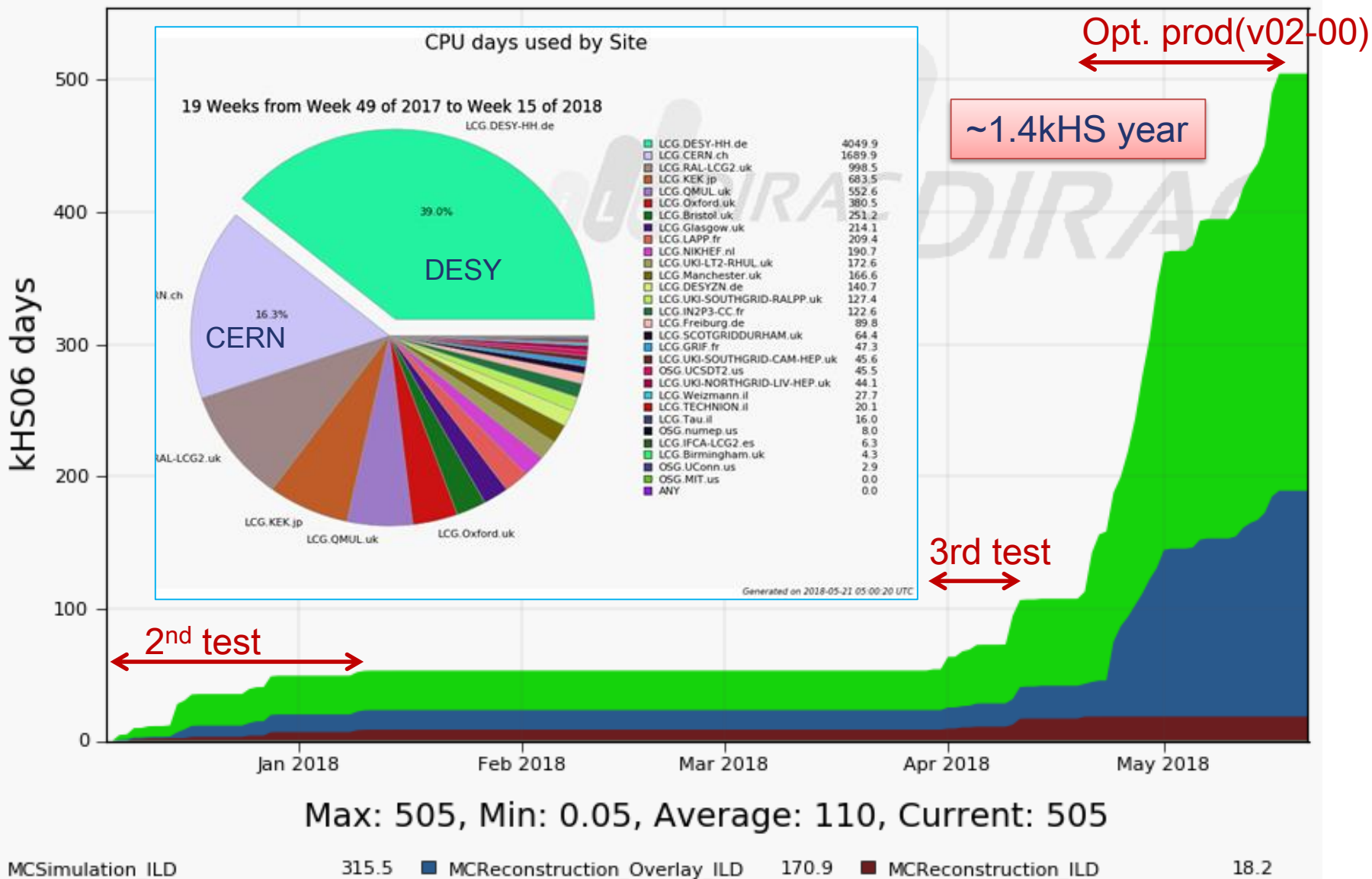
- 2018-05-15 19:09:59 : Production for 4f\_WW\_semileptonic.bWW.1lep has registered.
  - Beam polarizations of input samples : eL.pR,eR.pL
- 2018-05-15 19:11:26 : gen data splitting has initiated and sim/rec production will start/has started already.
- 2018-05-15 19:23:55 : sim production, 10129(ILD\_Is\_v02),10130(ILD\_s5\_v02), was submitted.
- 2018-05-15 19:28:15 : rec ovl production, ,10131(ILD\_s5\_o1\_v02),10132(ILD\_Is\_o1\_v02), was submitted.
- 2018-05-17 16:16:11 : Production completed and dst-merge job has started.
- 2018-05-17 16:30:20 : download/upload log files have started.
- 2018-05-17 18:30:32 : Submitting UserJobs to merge DSTs
- 2018-05-17 18:35:22 : DST merge jobs have completed. Downlodng job outputs.
- 2018-05-17 22:12:04 : log files were uploaded. List of log files will be found in the attachement.
- 2018-05-17 22:22:29 : log files were uploaded. List of log files will be found in the attachement.
- 2018-05-18 05:37:56 : Number of events of DST-Merged files is uploaded to dstm-nevents-summary.txt.
- 2018-05-18 05:38:04 : Start replicating DST-Merged files to KEK-SRM.  
Request 'ILD-500-4f\_WW\_semileptonic-10131', RequestID(s): 4195154
- 2018-05-18 06:30:40 : Replication of DST merged files have completed.
- 2018-05-18 06:32:04 : DST-Merged job log files were uploaded.
- 2018-05-18 06:32:04 : All steps have completed. JobStatus is changed to Done

Attachment 1: input\_genfiles.list 1 kB | Hide | Hide all

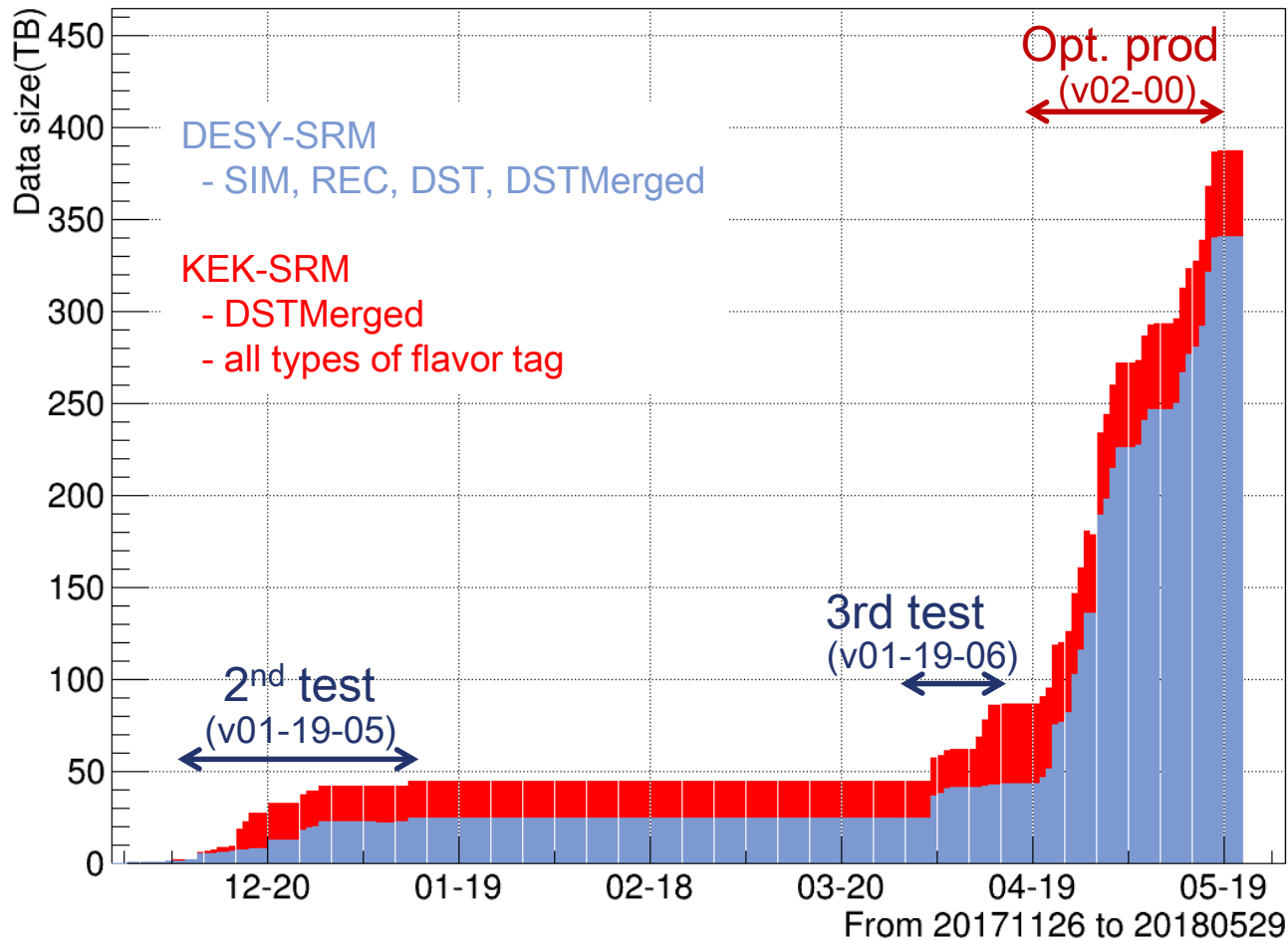
E500-TDR\_ws.P4f\_ww\_s1.Gwhizard-1\_95.eL.pR.I250018.001.stdhеп,nw\_per\_file=200  
E500-TDR\_ws.P4f\_ww\_s1.Gwhizard-1\_95.eL.pR.I250018.002.stdhеп,nw\_per\_file=200  
E500-TDR\_ws.P4f\_ww\_s1.Gwhizard-1\_95.eL.pR.I250018.003.stdhеп,nw\_per\_file=200

# Cumulative Normalized CPU ( Dec. 2017 to mid. May )

## Normalized CPU used by JobType 23 Weeks from Week 49 of 2017 to Week 20 of 2018



# Produced data since Nov. 2017 to mid. May



Production	Nb. phys. procs.	Nb.Trans.	Nb.Evts.
2nd	10	58	
3rd	3	46	0.8M
Opt	32	140	33M

(2 detectors)

- Optimization production : ~ 300TB samples were produced in < 3 weeks.
- 2 detector models ( Large and Small ), Si ECAL + Analog HCAL only
- ~ 60% of planned events ( 500 GeV SM ) were produced.
- Stopped due to DDSim bug in v02-00. v02-00-01 has started since last Thursday.

# Problems in the recent production (v02-00/v02-00-01)

## 1. Job failures

- **DDSim**: by missing libraries(fixed), Geant4 exception, file upload errors  
Failed jobs/Submitted jobs : > 50% worst case, < 1% best case
- **Marlin** : Initially, high failure rate due to error to download bkg files.
  - bkg file sites: DESY only → replicate them to several sites, KEK, CERN,
- Processed files/input files : > 99%, mostly > 99.5% , thanks to auto recover

## 2. Error to upload files

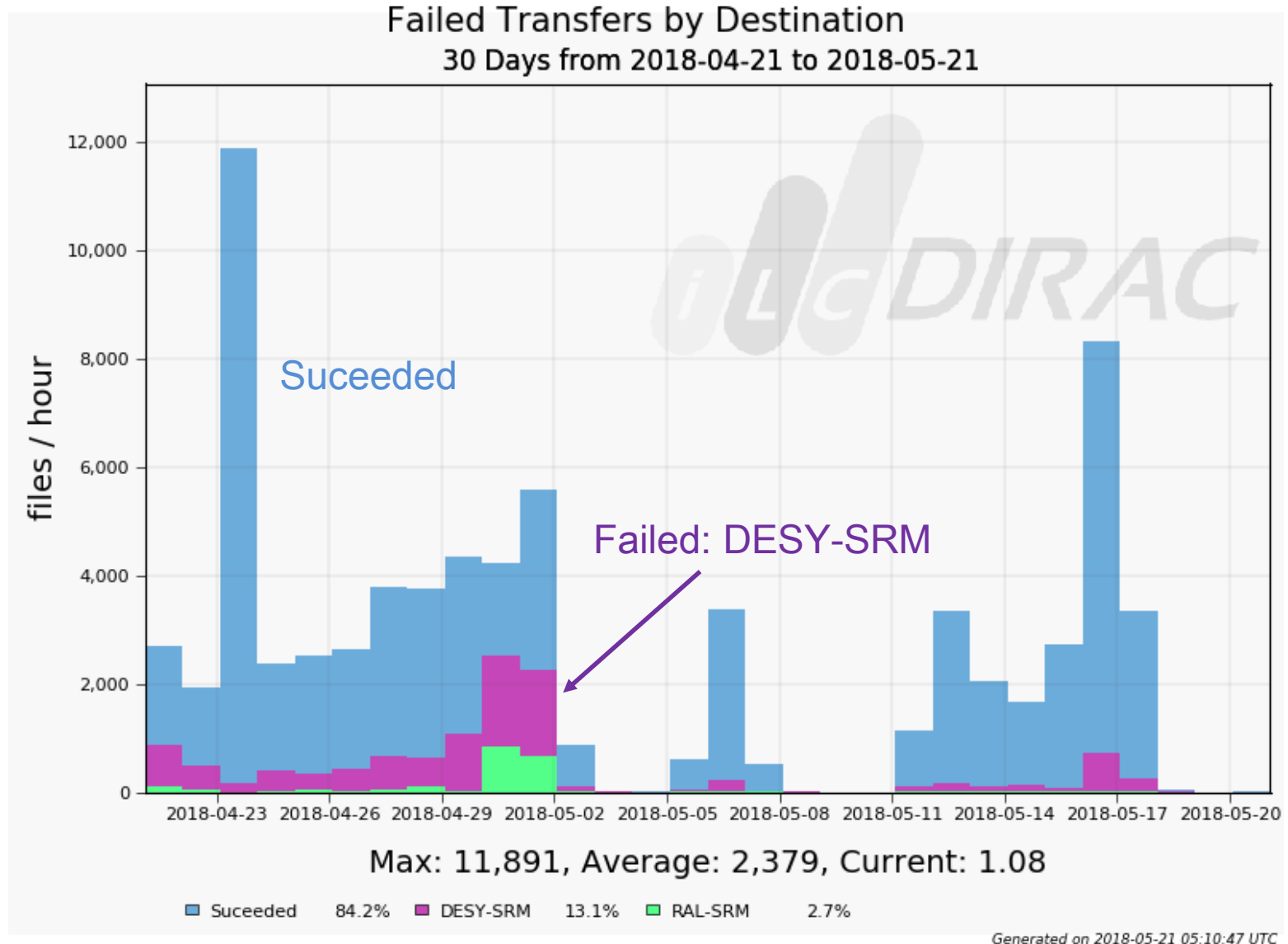
- 2GB/job → 1.2GB/job by reducing nb. events/job
- Failure rate increases when nb. of running jobs increased.
  - Adjusting by a speed of gensplit upload and nbtasks.

## 3. v02-00-01 production

- Noticed that few jobs failed due to Marlin error. Retry jobs were succeed. The cuase is not known
- up to 10% of job failed due to bkg file download.  
Worse if data rate > 1 GB/sec

Quick and better error detection is desirable

# Failed transfer by destination (v02-00)



It seems that error rate of transfer increases if data rate  $> 1$  GB/s  
( 1 jobs/sec )

# Summary

---

- ILD has producing MC samples using ILC DIRAC
  - After three test productions since last summer, the optimization production has finally started last April.
  - ILD production scripts have been improved continuously. Basysitting works have been reduced significantly. A power of ILCDirac is fully utilized and data production speed has increased a lot.
  - About 70% of first 500 GeV sample was produced with v02-00 in less than three weeks effectively. A bug in DDSim was fixed in v02-00-01, and new production has started. Few issues, but running well.
  - It would be desirable to enrich tools to check errors and monitor data quality
-



# BACKUP