

# Status of ILD new 250 GeV common MC sample production

LCWS2021

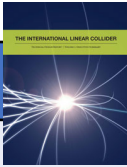

2021 Mar. 15

NDU Hiroaki Ono, KEK Akiya Miyamoto



# ILD new 250 GeV sample production

Full set of common MC samples were produced for several studies (DBD, IDR)  
 New full set of high statistics 250 GeV MC samples are requested for Physics study

	DBD (2013) 	IDR (2019) 	mc-2020 (2020)
Aim	Physics study	Detector Opt.	Physics study
Ecm	250 GeV (250 fb <sup>-1</sup> ) 350, 500 GeV, 1 TeV	500 GeV	250 GeV (1 ab <sup>-1</sup> )
Large cross section SM	40~100 fb <sup>-1</sup>		1 to 5 ab <sup>-1</sup>
Beam param	TDR_ws	TDR_ws	250-SetA
GEN sample	Whizard 1.95 stdhep	Re-use DBD sample	Whizard 2.8.5 slcio
Detector SIM	Mokka	DDSim	DDSim
ILCSoft	v01-16	v02-01	v02-02
Detector model		Hybrid CAL L5/S5	Hybrid CAL L5



Use ILCDirac for mass production

# Sample requests from ILD Physics WG

- Requested 1 or 5  $ab^{-1}$  statistics, at least 10 k events for small cross section channel
- Generator samples were produced by ILD Generator group

Beam param : 250-SetA

Generator : Whizard : 2.8.5

Process	Statistics	
$2f\_l, 2f\_h$	$eL.pR/eR.pL$ $5 ab^{-1}$	$eL.pL/eR.pR$ $1 ab^{-1}$
all $4f$		
all $6f$	10 k	
$e\gamma/\gamma e/\gamma\gamma$ process ( $3f, 5f, aa\_2f, aa\_4f$ )	$1 ab^{-1}$	
$2f\_bhabhag(ee\gamma)$	$1 ab^{-1}$	
$h \rightarrow inclusive$	$1 ab^{-1}$	
$h \rightarrow each\ decay$ (5x9 channels)	100 K	
$Z \rightarrow qq, Zh \rightarrow vvqq$ for LCFI	50 k	
$Z \rightarrow qq$ (91 GeV) for LCFI	50 k	

# Resource requirement estimation (on KEKCC)

Process type	M Events	KEK CPU days		Data size(GB)		DST
		Sim	Rec	Sim	Rec	
higgs inclusive	7.0	1.6	0.6	22.5	26.6	0.6
higgs exclusive	29.4	5.5	1.8	76.8	88.3	2.0
2f_Z_hadronic	992.0	196.3	62.6	2,704.6	3,051.5	72.4
2f_Z_leptonic	188.5	12.6	4.5	168.8	216.9	4.6
3f_Z_hadronic	18.5	6.6	2.4	163.1	251.9	6.0
3f_Z_leptonic	205.9	10.1	3.6	174.7	246.7	5.6
3f_others	0.340	0.018	0.008	0.215	0.384	0.012
4f_large(sw_sl,ww_sl/h,zzorww_h)	281.7	58.2	20.2	943.2	1,082.7	25.8
4f_rest	155.9	23.3	6.1	285.0	337.7	6.8
5f	0.066	0.007	0.002	0.069	0.100	0.003
aa_2f_Z_hadronic.bBB	42.3	1.6	0.8	22.0	48.5	1.6
aa_2f_leptonic_eB_pB	233.4	3.6	2.5	35.9	126.3	5.2
aa_4f	0.130	0.002	0.002	0.025	0.094	0.004
<b>Total</b>	<b>2,155.2</b>	<b>319.4</b>	<b>105.1</b>	<b>4,596.9</b>	<b>5,477.8</b>	<b>130.7</b>

CPU : ~ 424 k KEK CPU days, ~ 8.5 M HS06 days

Storage : SIM = 4.6 PB, REC = 5.5 PB, DST = 130 TB

Only SIM and DST files are kept (SIM remove for large Xsec)

REC : 10% or 500 files are kept

Available capacities

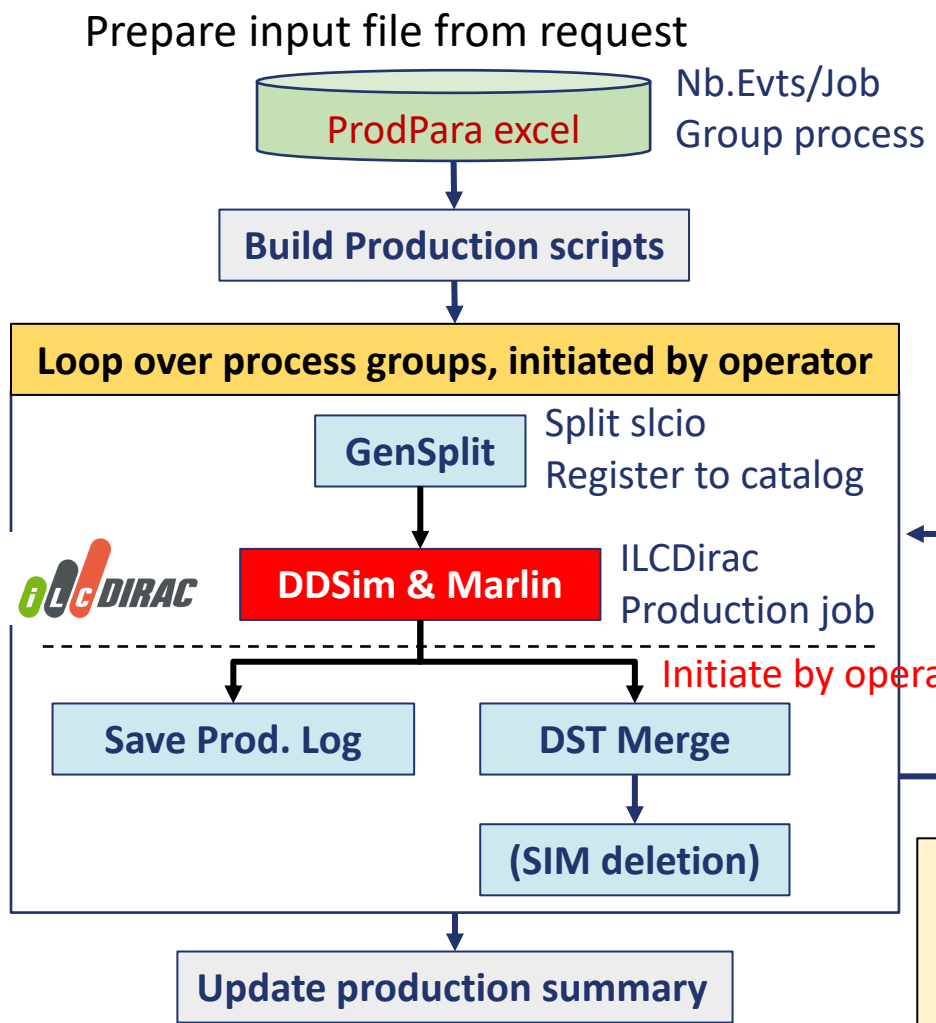
KEK tape : 600 TB (1.3 PB in total)

KEK disk : 300 TB

DESY disk : 300 TB (→ 600TB )



# ILD MC production workflow (ildprod)



**Operator task**

- **Prepare** production scripts  
Processes, # of Evt/job, # of files to process, etc.
- **Start** each production, adjust prod. parameters for better through put.
- **Terminate** ILCDirac production step when >99% files are processed
- **Monitor** the production comes to DONE status.

Manual intervention if error happens, stuck

**Cron task**

- Show progresses on web
- Add jobs controlled with param.
- Initiate sub-steps of DST merge and retrieve logs
- Monitor error

**Elog** Record progress in each step  
<https://ild.ngt.ndu.ac.jp/elog/dbd-prod/>

**ildprod new feature**

- Automatic retry if jobs failed with error
- Set gensplit speed, # of running jobs (SIM/REC)

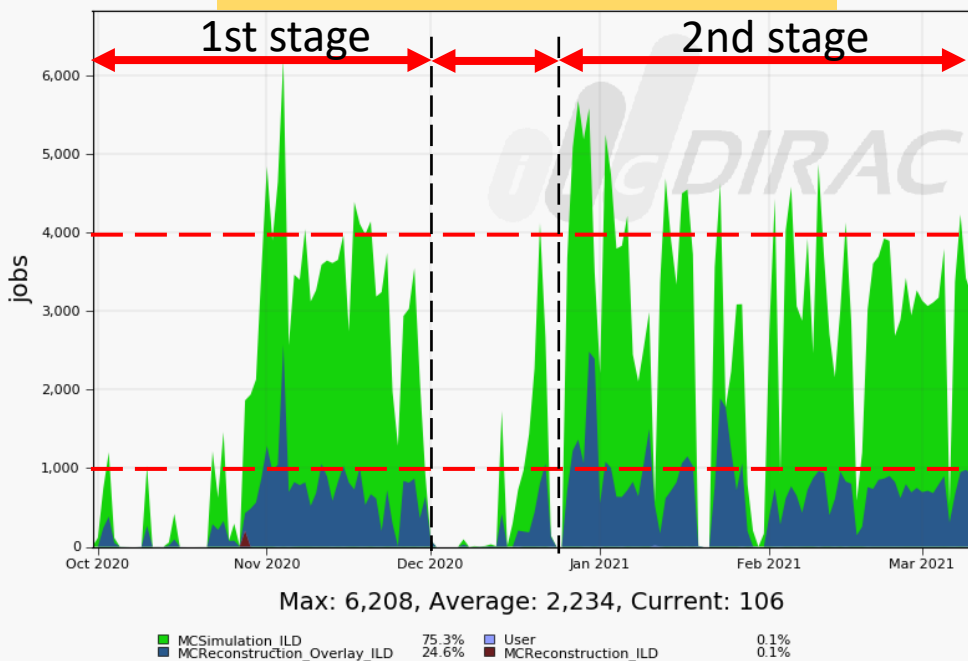
<https://ild.ngt.ndu.ac.jp/mc-prod/prodmon/prodsum-mc2020.html>



# Production progress situation

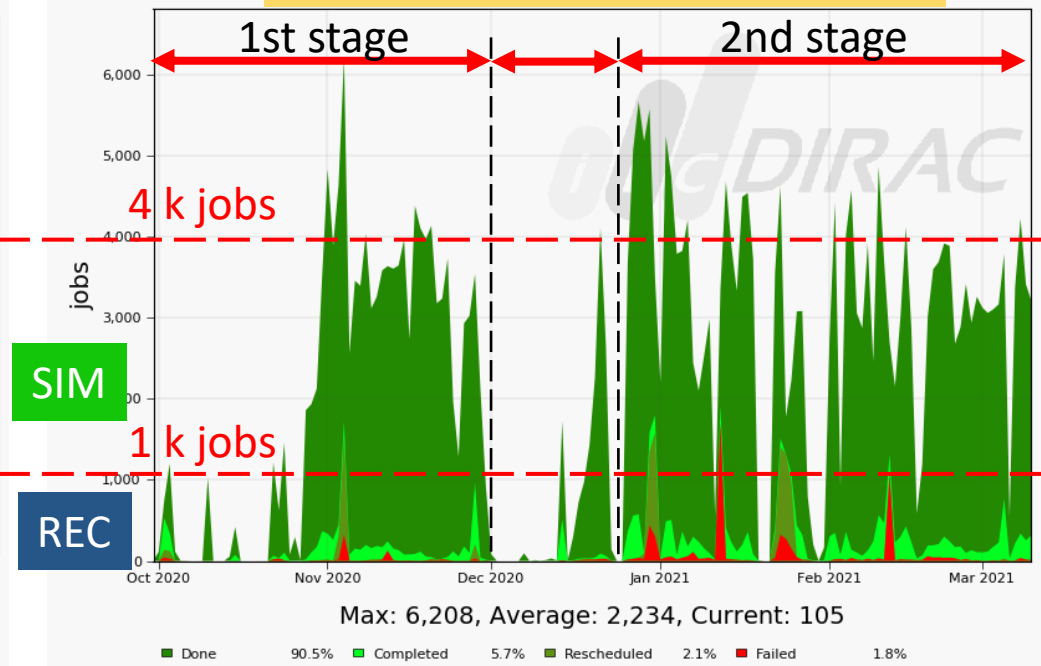
SIM : 3 ~ 4 k jobs continuously run  
 REC : Limit ~1.5 k jobs due to BG overlay access IO issue  
 (Increase failure if REC jobs run too much)

# of running jobs per JobType



Generated on 2021-03-11 02:43:24 UTC

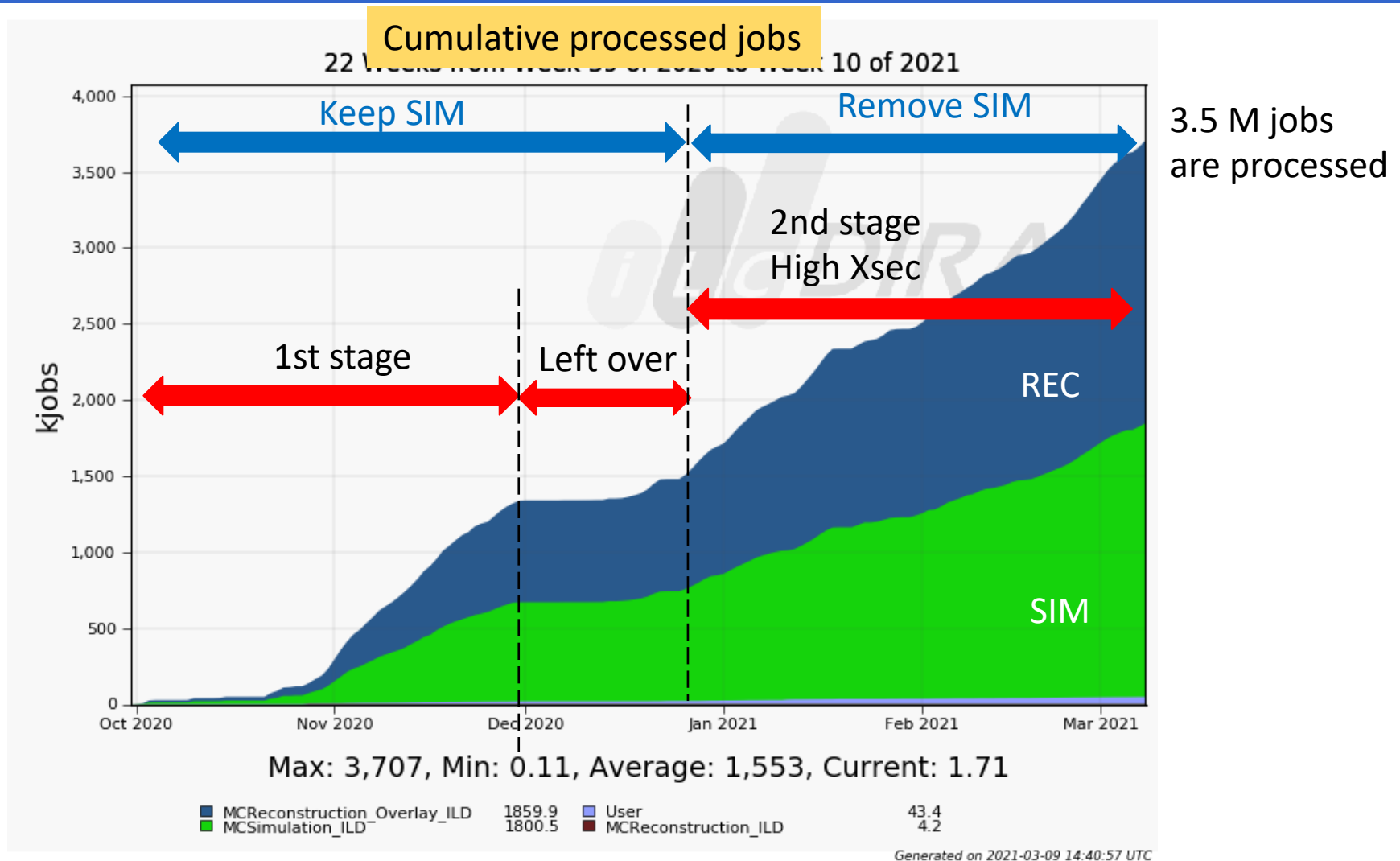
# of running jobs per Final status



Generated on 2021-03-11 02:42:38 UTC



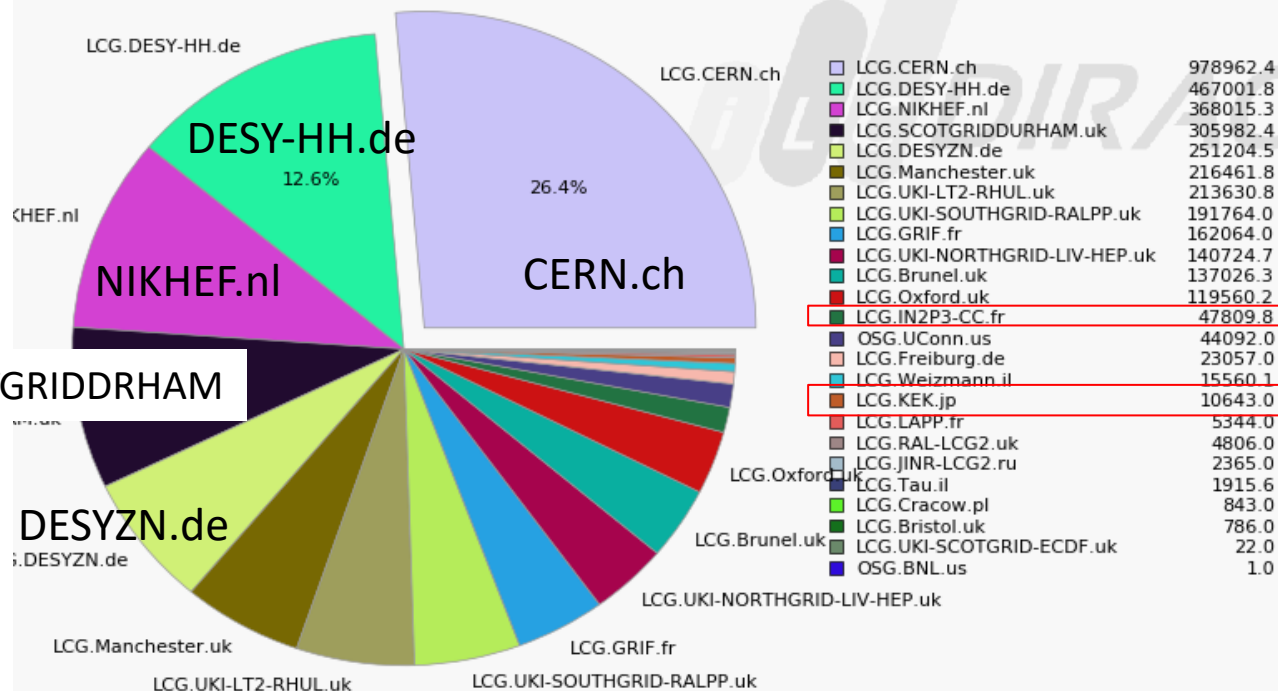
# Production progress situation



# Job processed sites on ILCDIRAC

## Production jobs processed sites

22 Weeks from Week 39 of 2020 to Week 10 of 2021



**IN2P3 issue**  
 - Expire certificate  
 - File transfer failure to DESY

IN2P3 Fixed at Feb. 2021

**KEK**  
**KEK site issue**  
 Miss configuration after renewal  
 Mismatch with ILCDirac param.  
 ILCDIRAC-933

Largest fraction of jobs are processed at CERN, DESY and European sites  
 → BG overlay files are replicated to CERN, DESY, KEK SE but mostly retrieved from KEK-DISK

File access IO issue happened → JIRA LCDIRAC-941 : Update to select better SOURCE-SE



# Major issues during the production

Issues are notified via JIRA ticket and thanks to the ILCDIRAC admin for quick fix

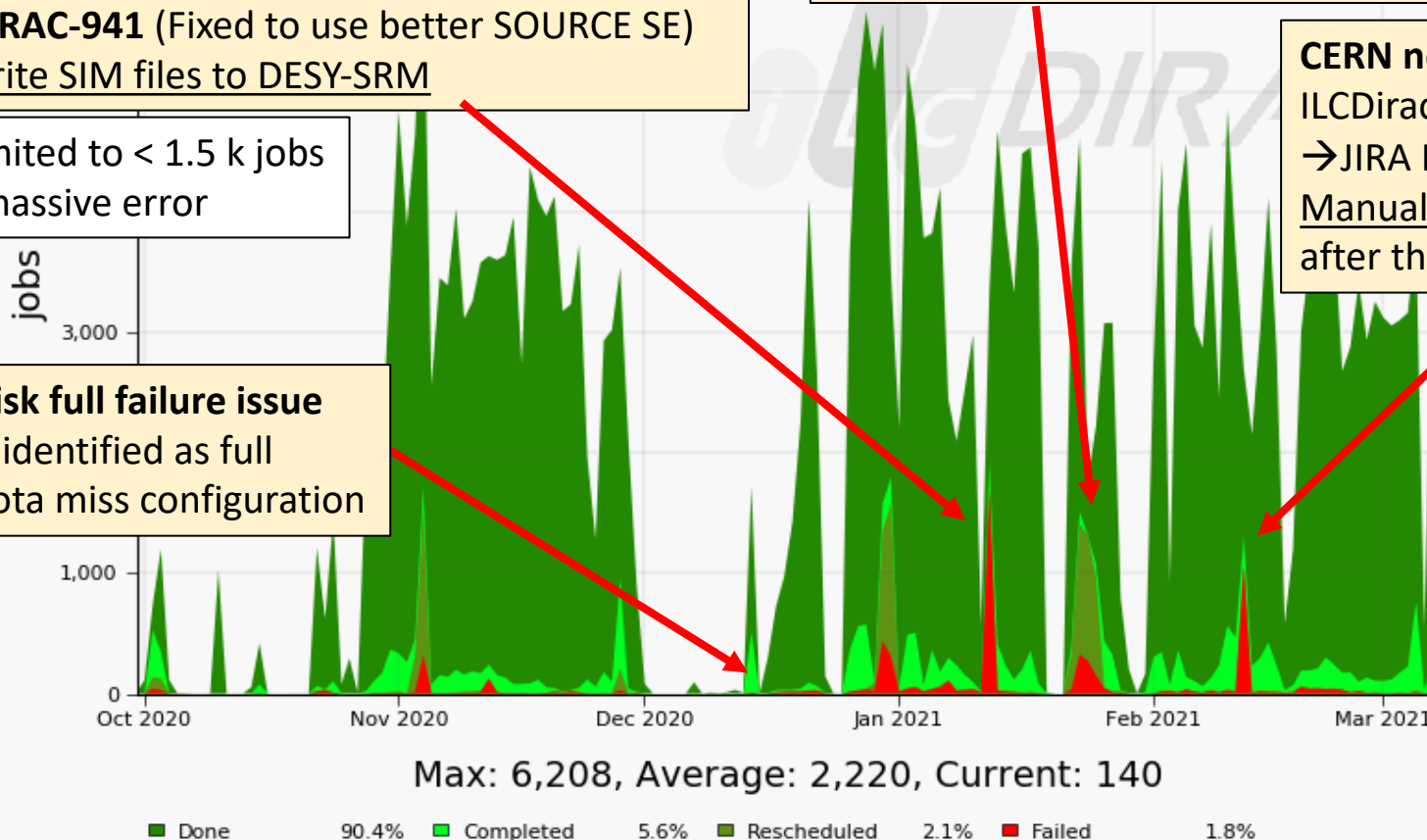
**BG Overlay file retrieve issue (Job killed as stalling)**  
All jobs access to KEK-DISK to retrieve BG overlay file even though stored also at CERN-SRM, DESY-SRM  
→ JIRA ILCDIRAC-941 (Fixed to use better SOURCE SE)  
→ Stick to write SIM files to DESY-SRM

**SIM/REC file loss by miss operation**  
→ Reproduce BG overlay files (Increase events/file)

REC jobs : limited to < 1.5 k jobs preventing massive error

**CERN network down**  
ILCDirac is not accessible  
→ JIRA ILCDIRAC-942  
Manual recovery work after the down

**KEK-SRM disk full failure issue**  
KEK-SRM is identified as full by TAPE quota miss configuration



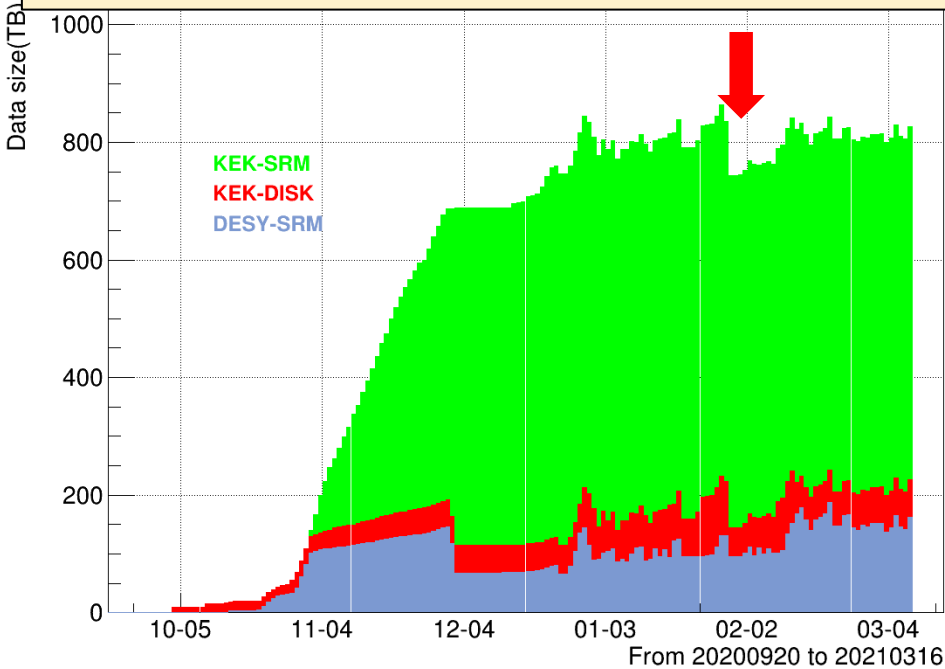
Generated on 2021-03-09 14:27:51 UTC



# SIM/REC file loss by miss operation

## Disk usage

File loss happened by miss operation Jan. 29



## ILCDirac WEB portal

Transformation Monitor [Untitled 1] × Accounting [Untitled 2] × Job Monitor [Untitled 4] ×

Start Stop Flush Complete Clean Items per page: 200 Page 1 of 2

ID ↓	Type	Name	Files	Process
15276	MCSimulation...	ILD-Opt_sim_250_2f_hadronic_eR_pL_I5_v02_2021...	17750	99.9
15275	MCRReconstru...	ILD-Opt_dstov_250_2f_hadronic_eR_pL_I5_o1_v02...	17731	99.9
15274	MCSimulation...	ILD-Opt_sim_250_2f_hadronic_eR_pL_I5_v02_2020...	17750	99.8
15273	MCRReconstru...	ILD-Opt_dstov_250_2f_hadronic_eL_pR_I5_o1_v02...	31985	99.9
15272	MCSimulation...	ILD-Opt_sim_250_2f_hadronic_eL_pR_I5_v02_2020...	32000	99.9
15120	MCRReconstru...	ILD-Op... i1_v02_202...	0	0
15119	MCRReconstru...	ILD-Op... i1_v02_202...	0	0
15118	MCSimulation...	ILD-Op... 2_20201030...	0	0
15114	MCRReconstru...	ILD-Op... o1_v02_20...	0	0
15113	MCRReconstru...	ILD-Op... o1_v02_20...	0	0
15112	MCSimulation...	ILD-Op... J2_2020102...	0	0
15111	MCRReconstru...	ILD-Opt_dstov_250_4f_VWV_semileptonic_I5_o1_v0...	0	0
15110	MCRReconstru...	ILD-Opt_recov_250_4f_VWV_semileptonic_I5_o1_v0...	0	0
15109	MCSimulation...	ILD-Opt_sim_250_4f_VWV_semileptonic_I5_v02_202...	0	0
15108	MCRReconstru...	ILD-Opt_dstov_250_4f_singleZnuu_leptonic_I5_o1...	0	0
15107	MCRReconstru...	ILD-Opt_recov_250_4f_singleZnuu_leptonic_I5_o1...	0	0
15106	MCSimulation...	ILD-Opt_sim_250_4f_singleZnuu_leptonic_I5_v02...	0	0
15105	MCRReconstru...	ILD-Opt_dstov_91_flavortag_I5_o1_v02_20201028_3	0	0
15104	MCRReconstru...	ILD-Opt_recov_91_flavortag_I5_o1_v02_20201028_2	0	0

**Clean** instead of **Complete**

- SIM and REC files are **cleaned** by miss operation on DIRAC (Including BG overlay files)
- Merged-DST files are **not affected** → Make copy on off-grid disk (50 TB)

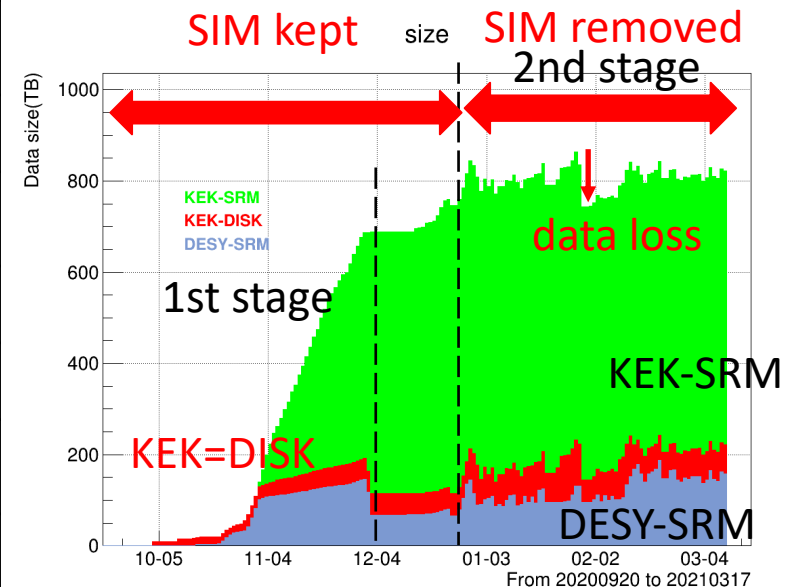
After reproducing aa\_lowpt BG overlay files, production has been resumed



# SE Storage resource usage

Resource usage as of 2021-03-10 on `/ilc/prod/ilc/mc-2020`

	Data Size (TB)			
mc-2020	Total	DESY-SRM	KEK-DISK	KEK-SRM
SIM	640.4	32.6	7.1	599.8
REC	34.1	33.7	0.04	0.4
DST-Merged	100.4	50.0	50.3	---
Gen	11.7	5.8	5.9	---
<b>Total</b>	<b>823.4</b>	<b>157.4</b>	<b>64.9</b>	<b>600.2</b>



<https://ild.ngt.ndu.ac.jp/mc-prod/prodmon/mcprod.html>

DESY-SRM : Disk space (mc-2020 300 TB → 600 TB in preparation)  
 KEK-DISK : DISK-only resource (300 TB in total)  
 KEK-SRM : TAPE storage (1.3 PB in total, 80% used)

# Production summary

Processes	1st stage production	2nd stage production
Higgs inclusive decay	Each 500 k events $> 1 \text{ ab}^{-1}$	<i>unnecessary</i>
Higgs exclusive decay	Each 100 k or 500k events	<i>unnecessary</i>
$2f_{\text{hadronic}}$	$100 \text{ fb}^{-1}$	$+0.9 \text{ ab}^{-1}$ <b>Done</b> , $+4\text{ab}^{-1}$ <b>ToDo</b>
$2f_{\text{leptonic}}$	$2.5 \text{ ab}^{-1}$	$+2.5 \text{ ab}^{-1}$ <b>ToDo</b>
$4f_{\text{mid-xsec}}$	1 or $5 \text{ ab}^{-1}$ in most channel	<i>unnecessary</i>
$4f_{\text{high-xsec}}$	$500 \text{ fb}^{-1}$	$4f_{\text{sw\_sl}}$ : $+4.5 \text{ ab}^{-1}$ <b>Done</b> $4f_{\text{sze\_l}}$ : $+4.5 \text{ ab}^{-1}$ <b>Done</b> $4f_{\text{ww\_sl}}$ : $+4.5 \text{ ab}^{-1}$ <b>Done</b> $4f_{\text{ww\_h}}$ : $+0.5 \text{ ab}^{-1}$ <b>Done</b> , $+4\text{ab}^{-1}$ <b>ToDo</b> $4f_{\text{zzorww\_h}}$ : $+0.5 \text{ ab}^{-1}$ <b>Done</b> , $+4\text{ab}^{-1}$ <b>ToDo</b>
$e\gamma/\gamma e/\gamma\gamma(3f, 5f, aa_{2f}, aa_{4f})$	$1 \text{ ab}^{-1}$	<i>unnecessary</i>
$2f_{\text{ee}\gamma}$	<b>Generator files are not ready</b>	
$6f$	<b>Generator files are not ready</b>	
LCFI $Z \rightarrow qq, Zh \rightarrow vvqq$	Each 500 k events	<i>unnecessary</i>

At least  $1 \text{ ab}^{-1}$  samples were produced if generator samples are available



# Summary

- New 250 GeV production has been processed with ILCDirac
  - Produced at least  $1\text{ab}^{-1}$  of all channels except  $2f_{ee\gamma}$  and  $6f$ , consuming  $\sim 1/2$  of total CPU time required.
  - Production of remaining channels are in progress. It will take 2 or 3 more months to complete with a current production speed ( $2f_{ee\gamma}$  and  $6f$  not counted)
- Issues in production system have been solved time to time and ILCDirac works rather fine now, though
  - data processing speed is currently limited by IO performance
- Additional production will follow to produce
  - Samples reconstructed with SDHCAL option (ILD\_I5\_o2)
  - Dedicated samples for detector, PID, jet clustering studies etc





# Planned productions

Producing following samples (1 to 5  $\text{ab}^{-1}$  high statistics samples)

physics processes

- Higgs Inclusive decay channels : nnh, eeh, mumuh/tautauh, qqh
- Higgs exclusive decay channels : Separate with decay channels

BG samples

- 2f, 4f, 3f, 5f, 6f, aa\_2f, aa\_4f

Calibration for LCFIPlus

- 91 GeV  $Z \rightarrow qq$
- 250 GeV  $Z \rightarrow qq$

Last discussion with physics convener

Process pol.	eL.pR	eR.pL	eL.pL	eR.pR
2f_l, 2f_h	5 $\text{ab}^{-1}$	5 $\text{ab}^{-1}$	1 $\text{ab}^{-1}$	1 $\text{ab}^{-1}$
all 4f				
all 6f	10K	10K	10K	10K
2f_bhabhag	1 $\text{ab}^{-1}$	1 $\text{ab}^{-1}$	1 $\text{ab}^{-1}$	1 $\text{ab}^{-1}$
H $\rightarrow$ inclusive	1 $\text{ab}^{-1}$	1 $\text{ab}^{-1}$	1 $\text{ab}^{-1}$	1 $\text{ab}^{-1}$
H $\rightarrow$ each mode (5x9 channels)	100K	100K	10K	10K

All processed on ILCDIRAC  
with latest ILDProd tool  
(REC keep fraction is configurable)



# Software conditions

- ILCSOFT/ILDConfig v02-02 for CentOS7
- Simulation
  - ILD\_L5\_v02 only ( not S5 ), hybrid calorimeter
  - With 7 mrad crossing angle
  - IP smearing and offset depending on initial particle, defined in ILDConfig
- Reconstruction
  - Options
    - o1 model (Si-ECAL + AHCAL) : Full statistics
    - o2 (SDHAL) : requested by R&D group to do partially after validation
    - o3 (ScECAL): needs and validation not clear.
  - Background overlay
    - Pairs : selected-pairs to every bunches.  
Pre-simulated file: 500 files/200 BXs per file.

■ aa\_lowpt :

	#BG/Bx	NbFiles	NbEvt/File
eW.pW	0.126	3458	50
eB.pW	0.297	3633	100
eW.pB	0.297	3649	100
eB.pB	0.830	4090	200

# Detector model for 250 GeV production

Current setup of the ILCSoft detector model

	Large	Small	ECAL	HCAL
SIM	ILD_I5_v05	ILD_s5_v05	Both	Both
REC	ILD_I5_o1_v02	ILD_s5_o1_v02	Analog	SiW
	ILD_I5_o2_v02	ILD_s5_o2_v02	Semi-digital	SiW
	ILD_I5_o3_v02	ILD_s5_o3_v02	Analog	ScintillatorW
	ILD_I5_o4_v02	ILD_s5_o4_v02	Semi-digital	ScintillatorW

All hybrid  
Default o1  
Reprocess  
o2, o3 by  
request

I5/s5\_v02 SIM detector model includes all the detector setup (Hybrid)

→ **Large ILD\_I5\_v05** is used for the 250 GeV production

Each detector model study can be processed from the same SIM

→ A scheme to re-process SIM has been prepared in ILDProd