Status of large scale 250 GeV MC production

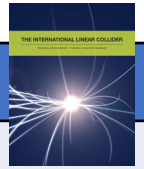
ILD meeting 2021

NDU Hiroaki Ono, KEK Akiya Miyamoto

ILD software group MCProd conveners

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^{-1}) 350, 500 GeV, 1 TeV	500 GeV	250 GeV (1 ab^{-1})
Large cross section SM	40~100 fb^{-1}		1 to 5 ab^{-1}
Beam param	TDR_ws	TDR_ws	250-SetA
GEN sample	Whizard 1.95 stdhep	Re-use DBD sample	Whiard 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

MC-2020 production summary

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

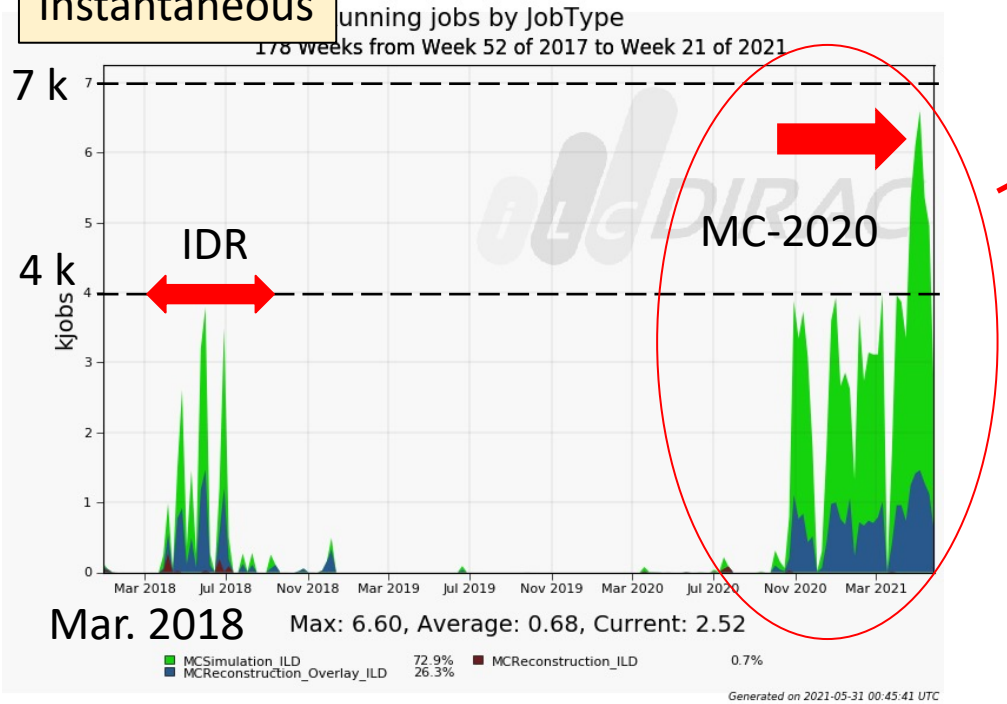
Processes	1st stage production	2nd stage production
Higgs inclusive decay	Each 500 k events > 1 ab^{-1}	*unnecessary*
Higgs exclusive decay	Each 100 k or 500k events	*unnecessary*
$2f$_hadronic	100 fb^{-1}	+2.4 ab^{-1} **Done**, +~2.5 ab^{-1} ToDo
$2f$_leptonic	2.5 ab^{-1}	+2.5 ab^{-1} **Done**
$4f$-mid-xsec	1 or 5 ab^{-1} in most channel	*unnecessary*
$4f$-high-xsec	500 fb^{-1}	4f_sw_sl : +4.5 ab^{-1} **Done** 4f_sze_l : +4.5 ab^{-1} **Done** 4f_ww_sl : +4.5 ab^{-1} **Done** 4f_ww_h : +4.5 ab^{-1} **Done** 4f_zzorww_h : +4.5 ab^{-1} **Done**
eγ/γe/γγ($3f, 5f, aa_2f, aa_4f$)	1 ab^{-1}	*unnecessary*
$2f$_eeγ	Generator files are **not ready**	
$6f$	Generator files are **not ready**	
LCFI $Z \to qq, Zh \to \nu\nu qq$	Each 500 k events	*unnecessary*

Most of production has completed except for $2f \to q\bar{q}$ and GEN not ready samples $ee\gamma, 6f$

Production progress

1st stage : SIM kept, REC partially kept
2nd stage : SIM/REC all removed

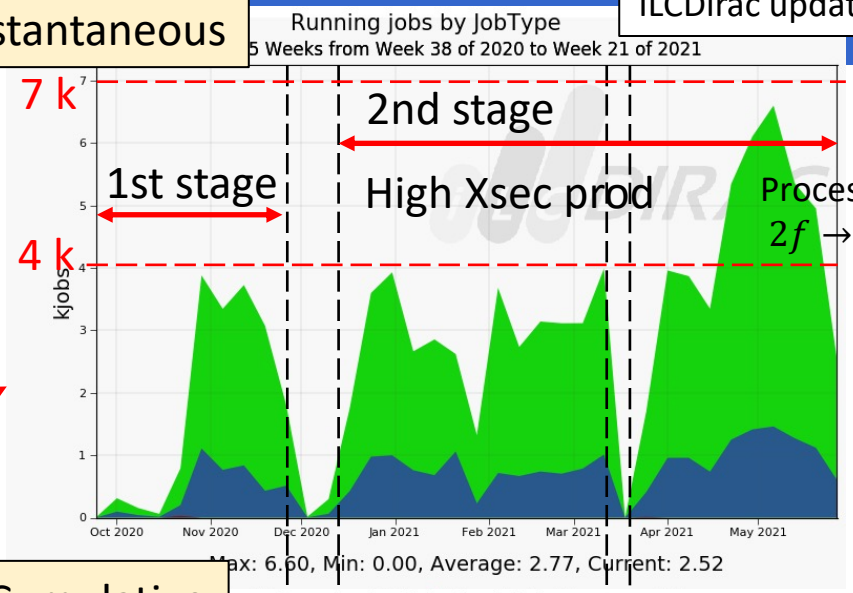
Instantaneous



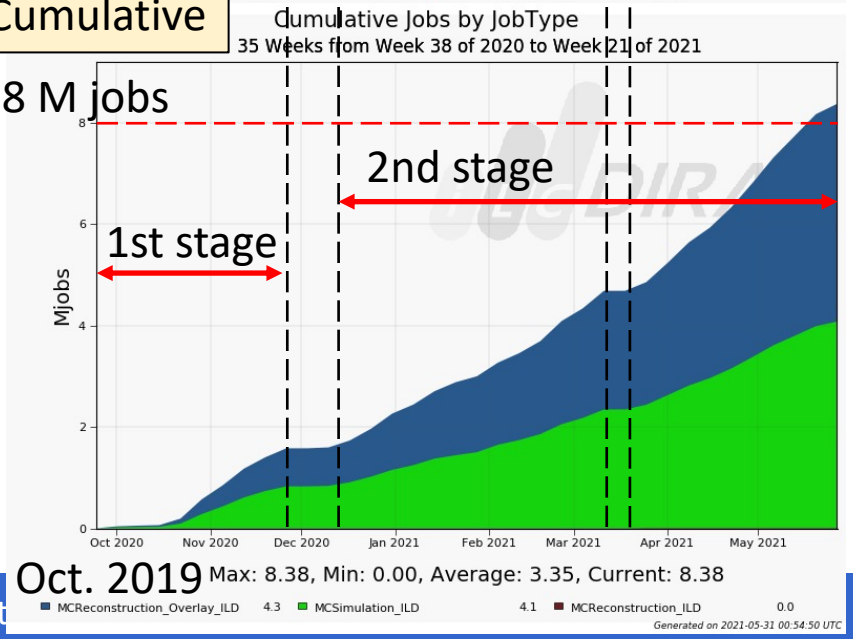
Now 6 to 7 k jobs are concurrently running

ILCDirac update

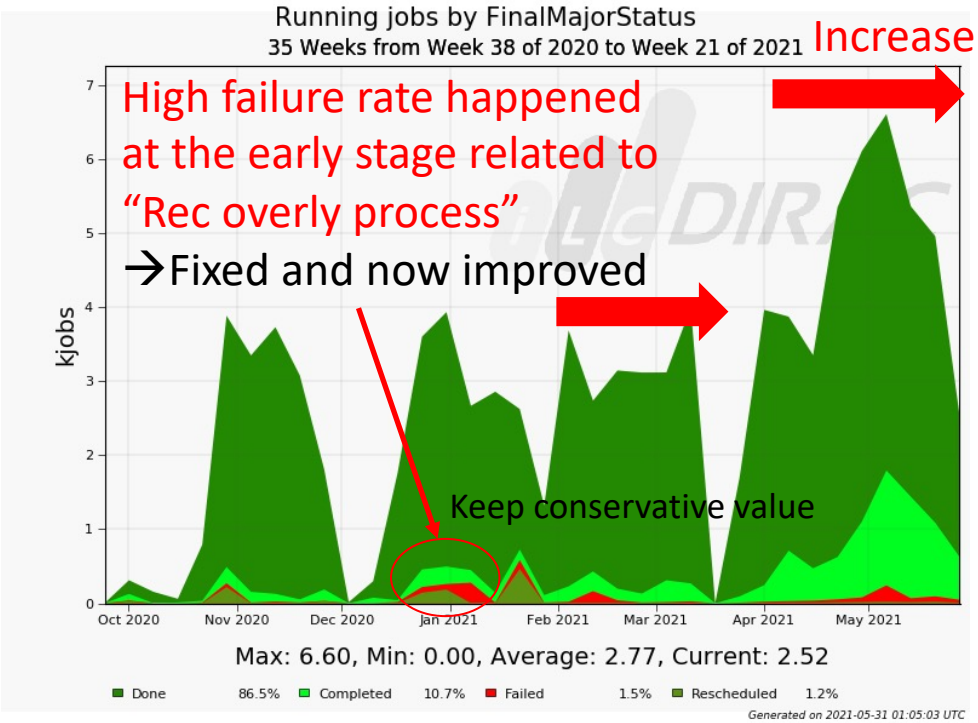
Instantaneous



Cumulative



Failure rate and improvement



Also we have affected by site downtime (Expected or unexpected)

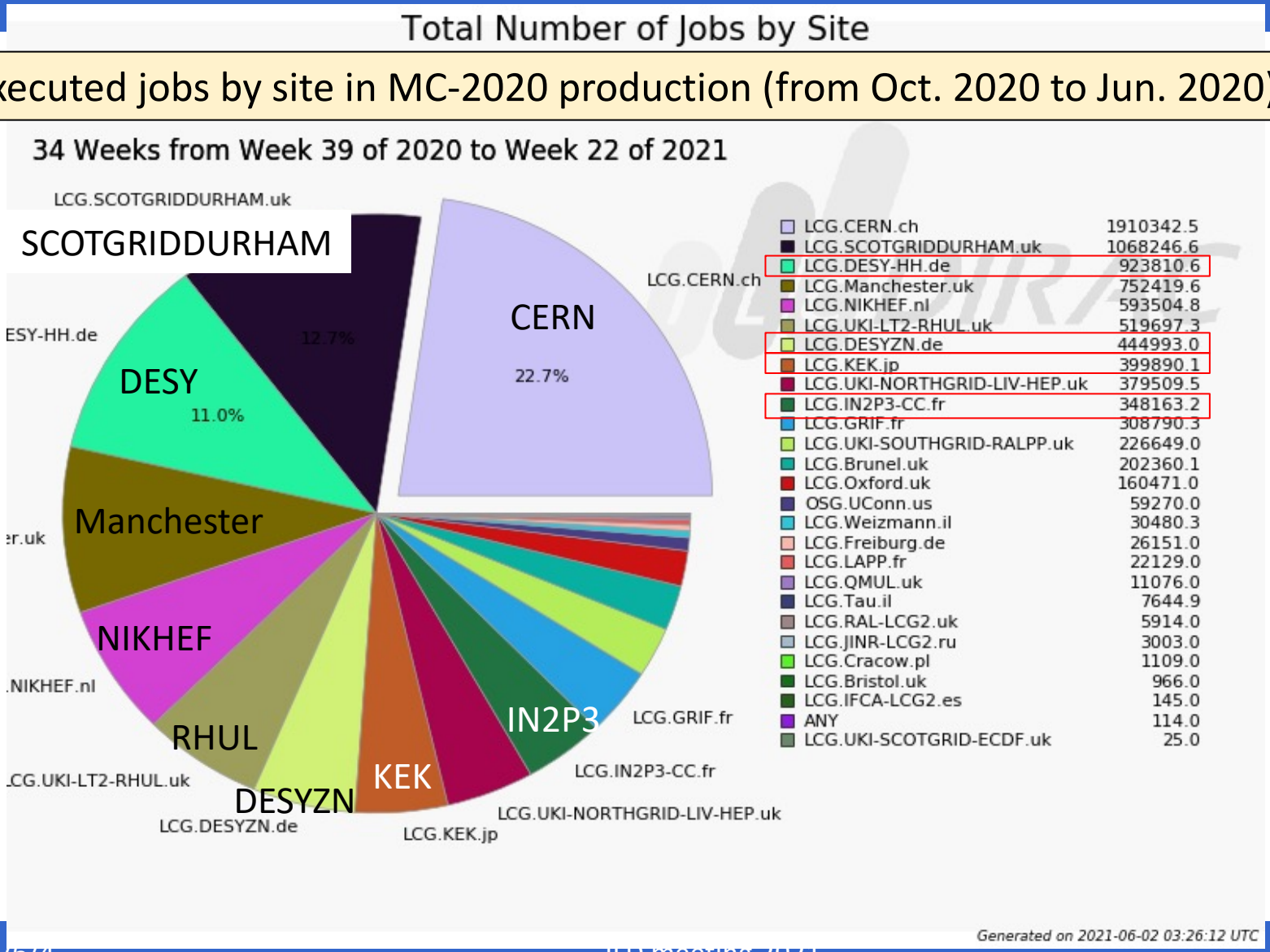
- CERN unexpected network down
- DESY network maintenance
- DESY-SRM situation
- KEKCC situation (GPFS trouble)
- …

- Keep stable production (conservative around ~3 to 4 k jobs)
  →Increase concurrent running jobs up to 7 – 8 k jobs.
- Avoid high failure rate situation → Cause negative feedback by reschedule jobs

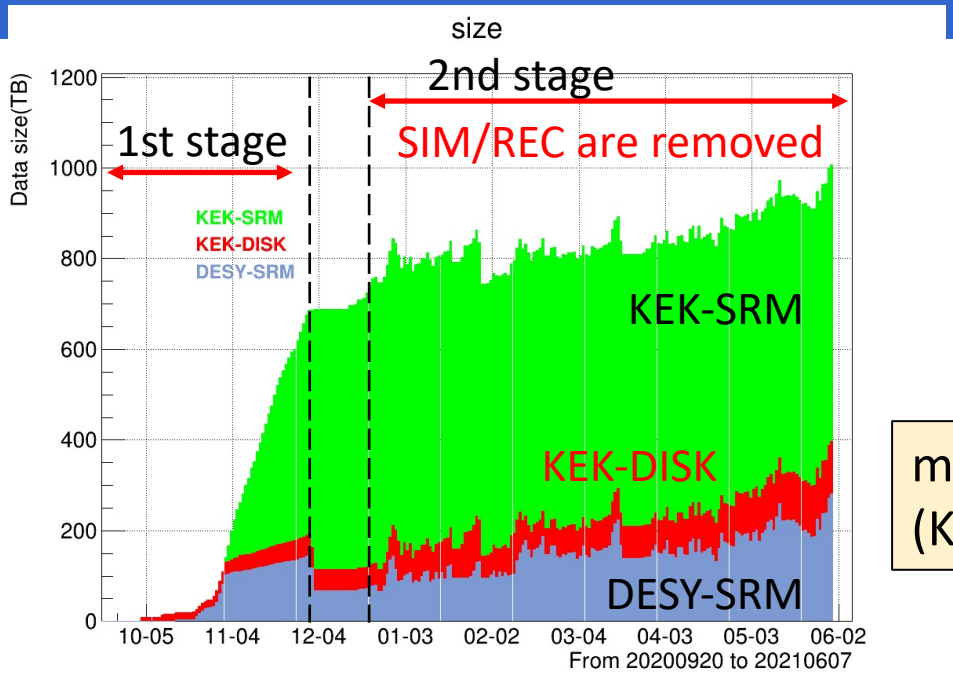
We have already taken 10 month and additional 1 to 2 month need to complete.

CPU resources

of executed jobs by site in MC-2020 production (from Oct. 2020 to Jun. 2020)



Disk space usage



https://ild.ngt.ndu.ac.jp/mc-prod/prodmon/size.png

mc-2020 merged-DST total: 130 TB (expected)
(Kept on KEK-DISK, DESY-SRM disk area)

	DISK area	Used (2021.06.02)	Available
KEK-DISK	300 TB (DISK)	120 TB	~150 TB
KEK-SRM	1.5 PB (TAPE)	1.1 PB (610 TB for MC-2020)	~400 TB
DESY-SRM	600 TB (DISK)+TAPE	280 TB (DISK mc-2020)	~300 TB? (DISK)

Summary of disk space usage
https://ild.ngt.ndu.ac.jp/mc-prod/prodmon/mcprod.html

Next step

- Process remaining channels : $ee\gamma, 6f$
- Re-produce incomplete processes : 4f_WW_l.eL.pR, 4f_WW_sl.eL.pR, 2f_Z_h
  →Their statistics are not enough yet due to various reasons.

- A request of o2 samples for 2f, 4f, higgs_excl has been received.
  - It would require about 6 months for the same statistics as o1 samples.
- No official request of o3 option yet.

- ILCDirac upgrade : Dirac v7r0 (python2)→v7r2 (python3) or later
  - By the end of 2021 at the latest
    → Need an interruption of MC production
  - More disk SEs to the production workflow for more samples.
- Consider to other energy production (500 GeV?)
- Check available resources (DISK spaces) and estimate requirements
- Generator sample situation and prospects (with Whizard2)

BACKUP

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	
$Z \rightarrow qq, Zh \rightarrow \nu\nu qq$ for LCFI	50 k	
$Z \rightarrow qq$ (91 GeV) for LCFI	50 k	

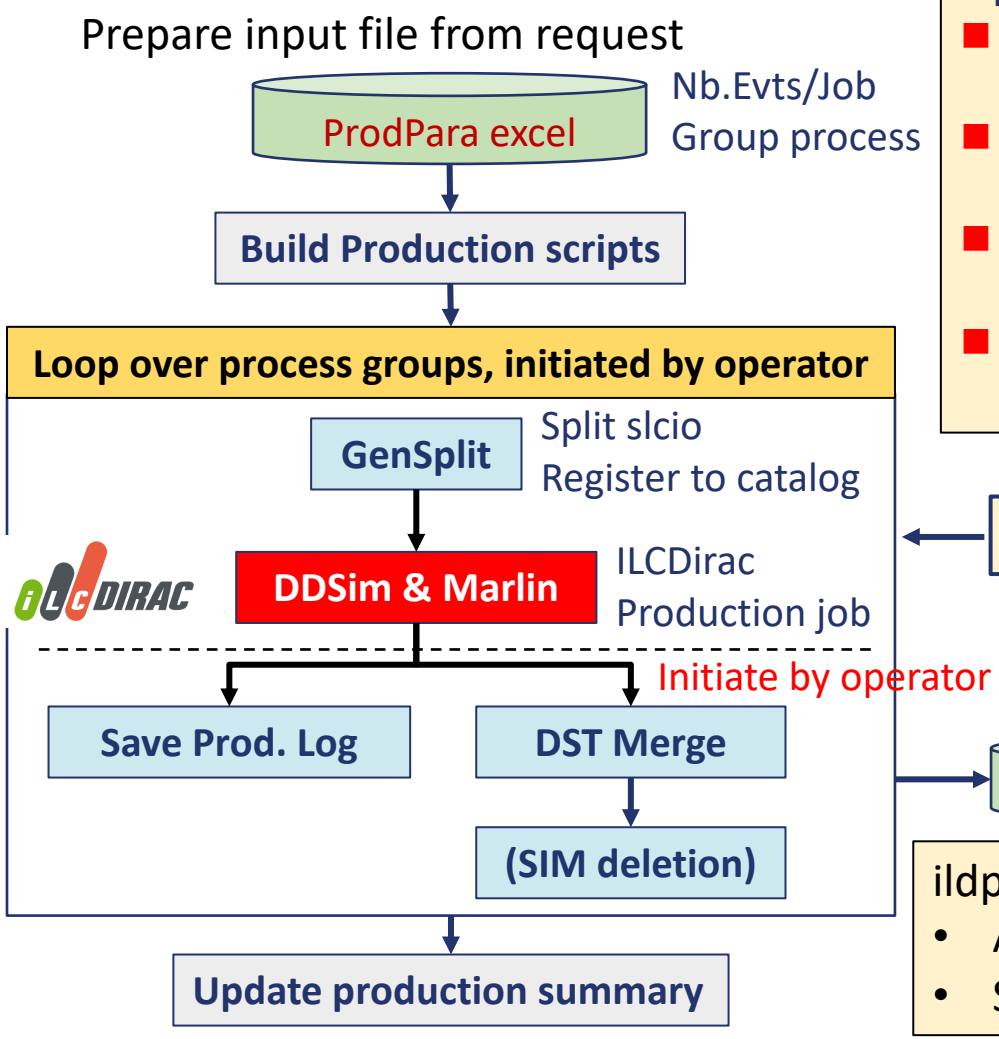
Resource requirement estimation (on KEKCC)

		KEK CPU days		Data size(GB)		
Process type	M Events	Sim	Rec	Sim	Rec	DST
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
Total	2,155.2	319.4	105.1	4,596.9	5,477.8	130.7

Available capacities
KEK tape : 600 TB (1.3 PB in total)
KEK disk : 300 TB
DESY disk : 300 TB (➔ 600TB)

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

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

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_l5_v05	ILD_s5_v05	Both	Both
REC	ILD_l5_o1_v02	ILD_s5_o1_v02	Analog	SiW
	ILD_l5_o2_v02	ILD_s5_o2_v02	Semi-digital	SiW
	ILD_l5_o3_v02	ILD_s5_o3_v02	Analog	ScintillatorW
	ILD_l5_o4_v02	ILD_s5_o4_v02	Semi-digital	ScintillatorW

All hybrid

Default o1

Reprocess o2, o3 by request

l5/s5_v02 SIM detector model includes all the detector setup (Hybrid)
→Large ILD_l5_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