LHCb news

A.Tsaregorodtsev,
CPPM-IN2P3-CNRS, Marseille
Journées LCG-France, 11 Dec 2019

Outline

- Run3 Computing Model update
- Resources needs
- Some Production System developments
- Current activities
- Conclusions

Streams and event sizes in Run 2

- Trigger output is saved in 3 different streams using different file format

Stream	Content	File format
FULL	Full event information	RDST
Turbo	Selected event information	MDST
Calibration	Full event information + raw banks	RAW or RDST

- Run2 event sizes and rates
  - Event size Turbo/FULL ~0.5

stream	event size (kB)	event rate (kHz)	rate fraction	throughput (GB/s)	bandwidth fraction
FULL	70	7.0	65%	0.49	75%
Turbo	35	3.1	29%	0.11	17%
TurCal	85	0.6	6%	0.05	8%
total	61	10.8	100%	0.65	100%

Extrapolation of Run2 rates to Run3 conditions

- With the upgrade conditions
  - Luminosity $4*10^{32}cm^{-2}s^{-1}$ to $2x10^{33}cm^{-2}s^{-1}$
  - HLT efficiency increase because of removal of L0 hardware trigger
  - Raw event size increase due to pileup, according to simulation
- Without any changes the HLT output rate would increase in Run3 to 17.4 GB/s

	Run 2 (GB/s)	Lumi	No L0	Raw size	Run 3 (GB/s)
Full	0.49	x5	x2	x3	14.7
Turbo	0.11	x5	x2	x1	1.1
Calibration	0.05	x5	x2	x3	1.6
Total	0.66				17.4

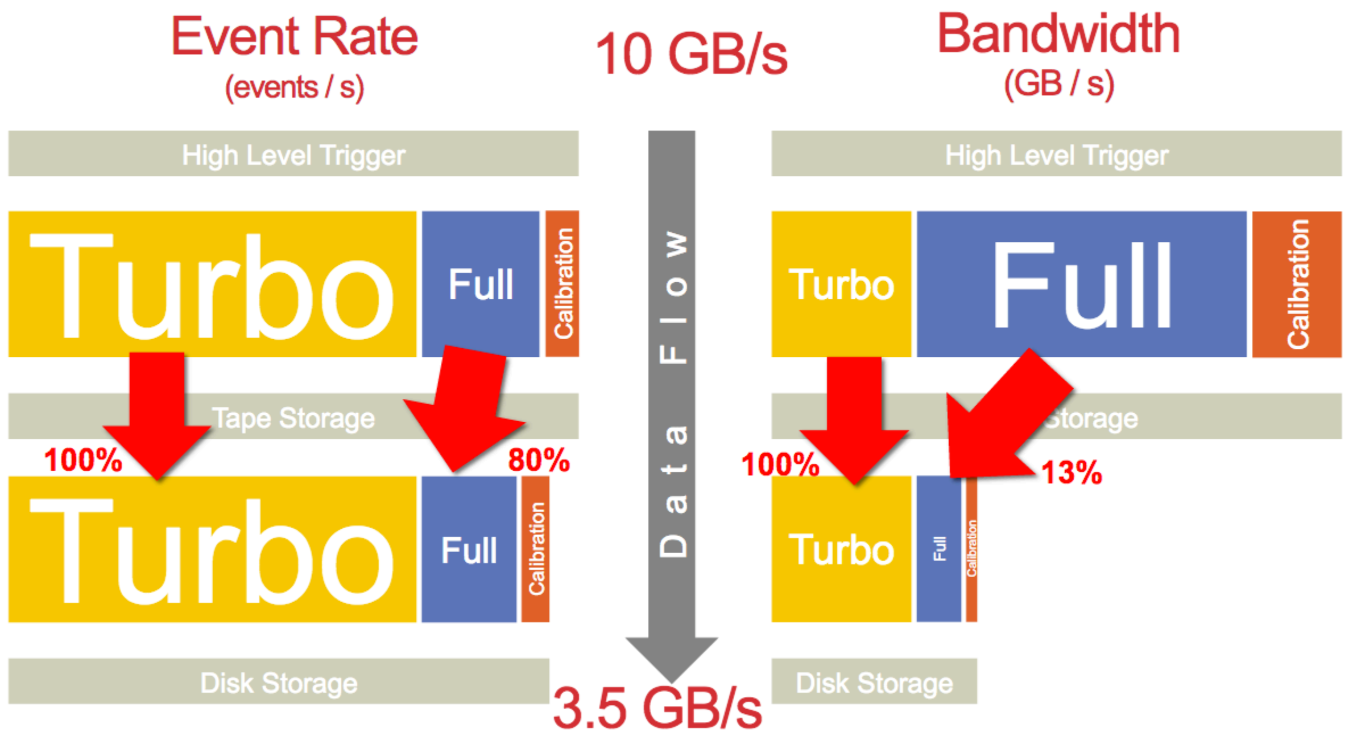
- How to cope with that ?

Optimizing the HLT output bandwidth

- Need to optimize the bandwidth to achieve 10GB/s to tape
- Moving a larger fraction of the physics program to Turbo decreases the output bandwidth
- Making Turbo events considerably smaller (16 % of Full size)
  - Some selections need to stay in Full
    - Keep some flexibility, recover from eventual errors, develop new analysis ideas
- For the baseline model assume 60% of the physics selections currently on FULL stream migrating to Turbo

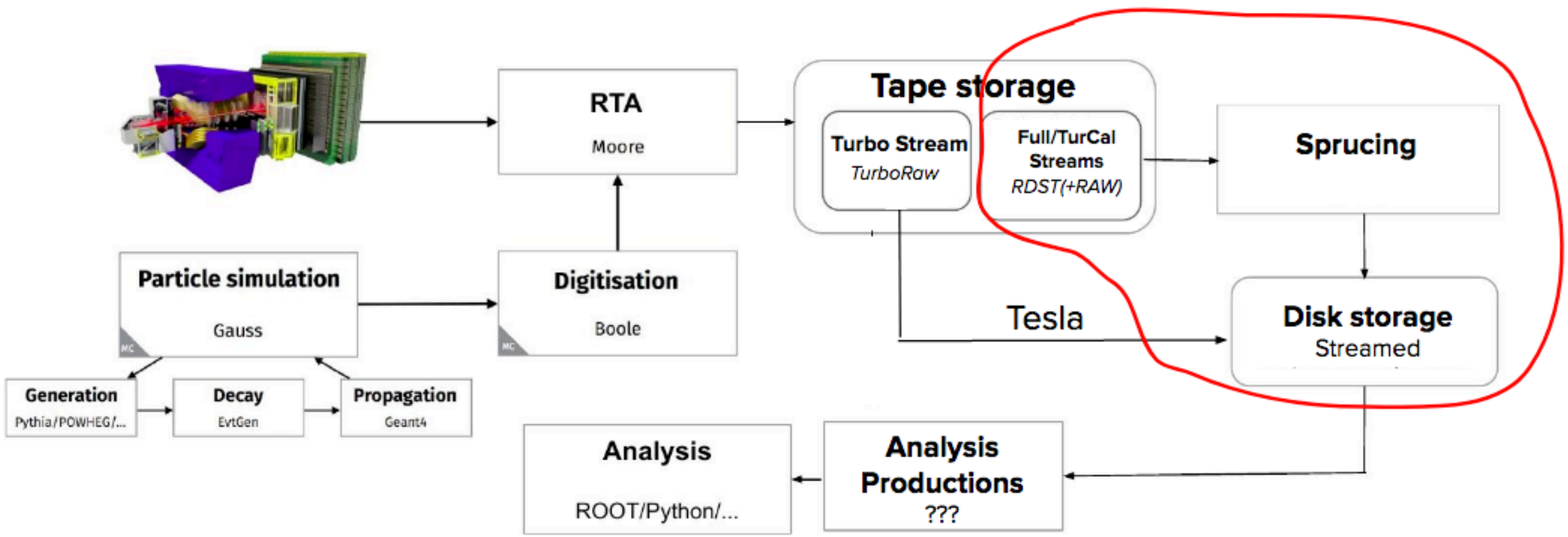
Streaming and filtering in Run3

- How to fit 10 GB/s in a reasonable amount of storage resources ?
  - 10 GB/s to tape
  - Reduce by ~1/6 FULL and Calibration data volume with “sprucing”
  - Save 3.5 GB/s to disk



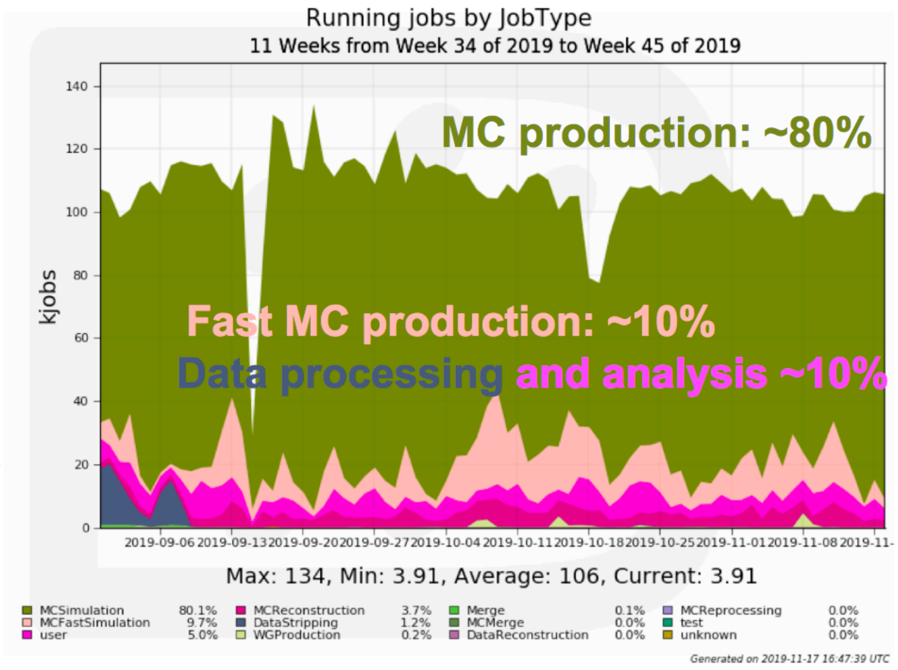
Reducing disk storage requirements

Sprucing proposal - selection, streaming and formatting all data consistently utilising the same applications



Run3 Simulation

- CPU needs dominated by MC Simulation
  - Number of needed MC events scale with luminosity
    - As seen in Run2
      MC events/fb^{-1}/year = 2.3×10^9
- Assume the same scaling for Upgrade



Run3 event volume

Assumptions on simulated event volume

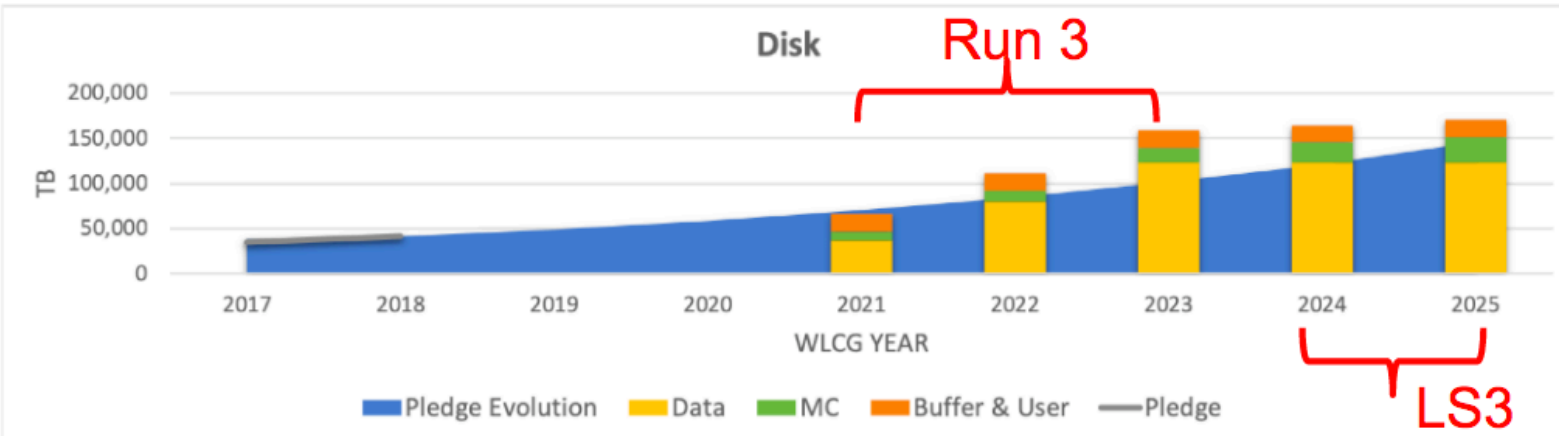
- MC events saved in MDST format (x40 size reduction!)
- MC production for a data taking years extends over the following 6 years

- Assumption on replicas

stream	tape	disk
FULL	2× RDST + 1× MDST	3× MDST
Turbo	1× TurboRaw + 1× MDST	2× MDST
TurCal	2× RDST + 1× MDST	3× MDST
Simulation	1× MDST	1 × MDST (30% data set only)

- All Run 1 + 2 data will be reduced in the end to 1 replica
- The first year of LHC Run 3 (2021) is considered a “commissioning year” with half the luminosity delivered

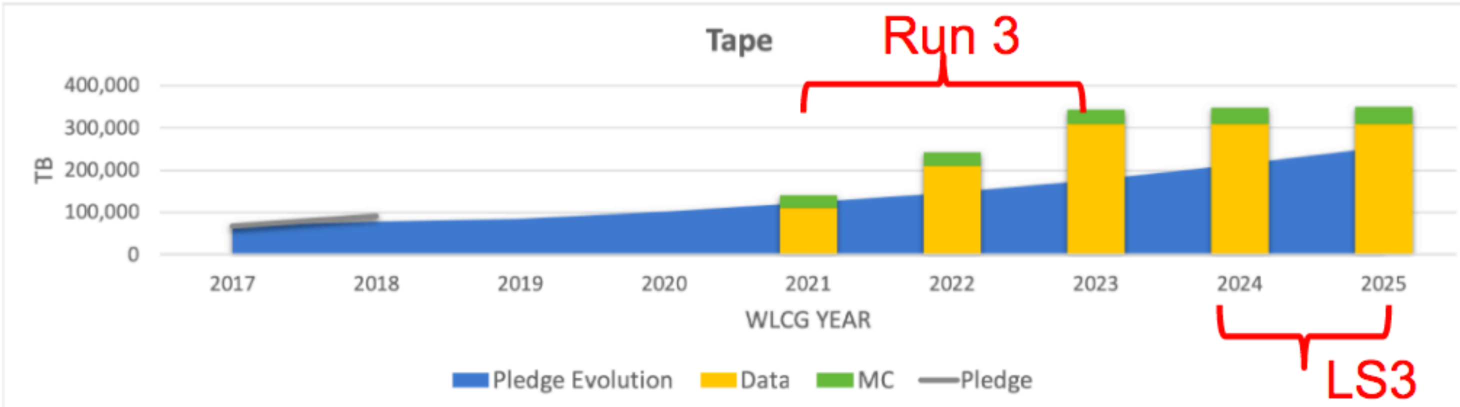
Run3 Model: disk requirements



	WLCG Year	Disk	
		PB	Yearly Growth
Run 3	2021(*)	66	1.1
	2022	111	1.7
	2023	159	1.4
LS 3	2024	165	1.0
	2025	171	1.0
Average end of Run 3			1.4
Average end of LS 3			1.2

- Pledge evolution assumes a "constant budget" model (+20% more every year)
- Max deviation from this model ~**1.6**
- In line with the model by the end of LS3
- 3.5 factor reduction compared to the assumptions in summer 2018 !

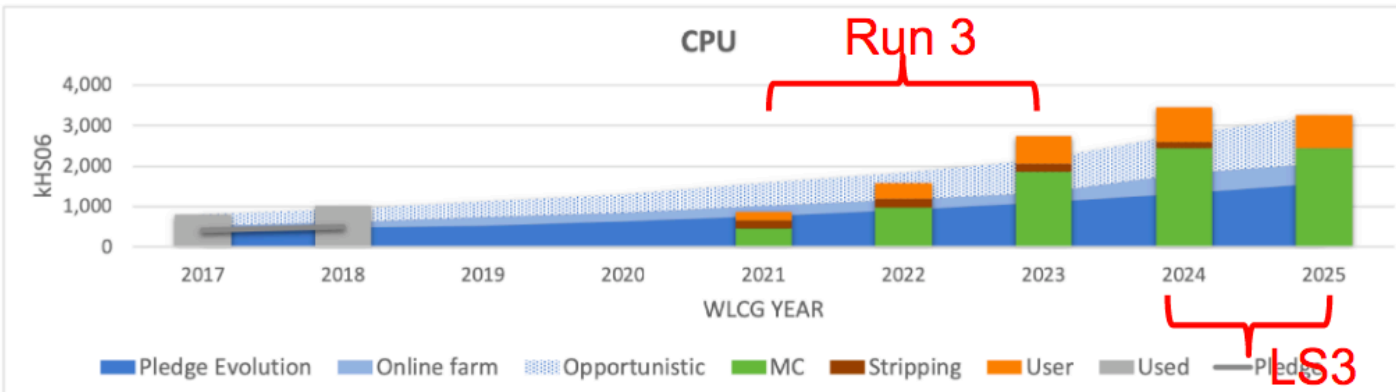
Run3 Model: tape requirements



	WLCG Year	Tape	
		PB	Yearly Growth
Run 3	2021(*)	142	1.5
	2022	243	1.7
	2023	345	1.4
LS 3	2024	348	1.0
	2025	351	1.0
Average end of Run 3			1.5
Average end of LS 3			1.3

- Pledge evolution assumes a "constant budget" model (+20% more every year)
- Max deviation from this model ~**1.9**
- In line with the model by the end of LS3

Run3 Model: CPU requirements



	WLCG Year	CPU	
		kHS06	Yearly Growth
Run 3	2021(*)	863	1.4
	2022	1.579	1.8
	2023	2.753	1.7
LS 3	2024	3.476	1.3
	2025	3.276	0.9
Average end of Run 3			1.6
Average end of LS 3			1.4

- Pledge evolution assumes a "constant budget" model (+20% more every year)
- Max deviation from this model ~**2.5**
- Plan to use opportunistic resources, which are however not granted
- Online farm used opportunistically when idle

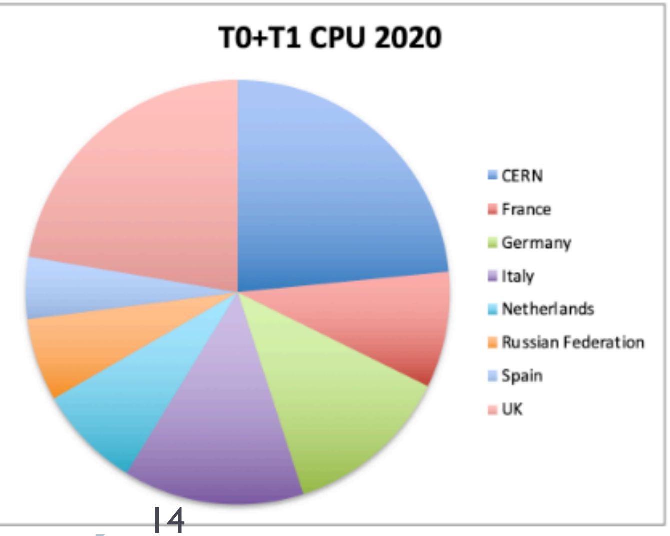
Possible solutions to reduce resources requirements

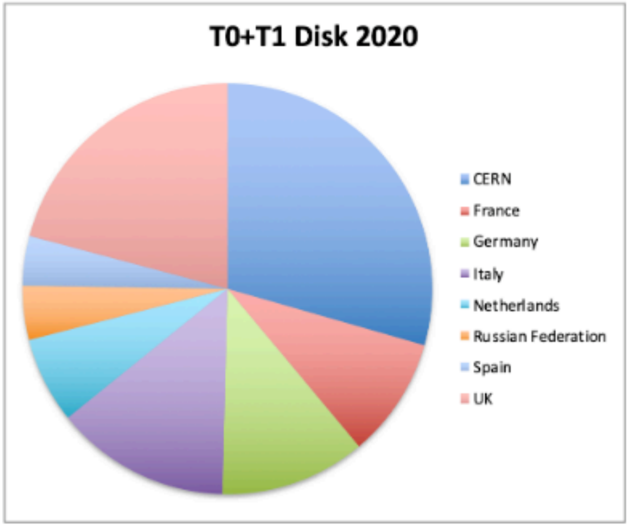
- Agressive use of faster simulation techniques
  - Baseline:
    - Full/fast/parametric simulation: 120/40/2 seconds
    - Sharing full/fast/parametric: 40/40/20
  - Needs a lot of developments on fast MC techniques
  - Changing sharing will reduce CPU needs but no effect on tape/disk
- Agressive use of Turbo to reduce HLT bandwidth
  - Helps to save tape but can have impact on the physics reach
- Data parking to save disk storage
  - Impact on operations (tape throughput, intelligent staging)

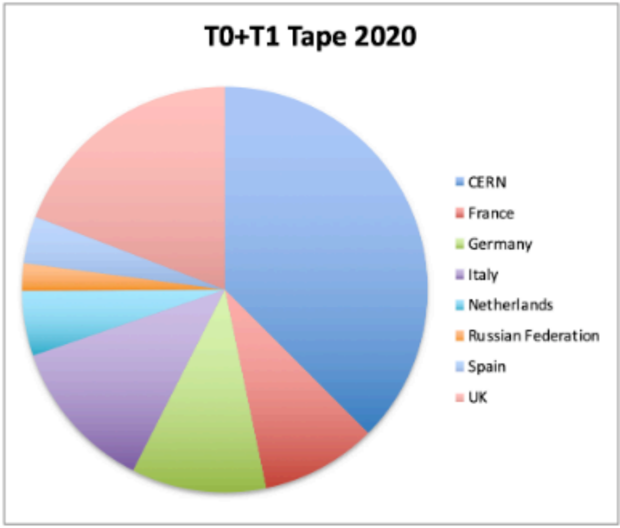
2020 T0+T1 pledges

- Generally covering requests
  - Slightly lower in CPU and disk
  - T1 France contribution on the level of ~15-16%

2020 T0+T1	CPU HS06	Disk Tbytes	Tape Tbytes
CERN	98000	17200	36100
France	47200	4650	8170
Germany	54780	5545	9270
Italy	55760	6868	13362
Netherlands	26203	2645	4725
Russian Fed	16400	2300	3000
Spain	13120	1328	2220
UK	81300	8370	15270
Total	**392763**	**48906**	**92117**
Requested	**426000**	**50400**	**91600**
Difference	**-7.8%**	**-3.0%**	**0.6%**



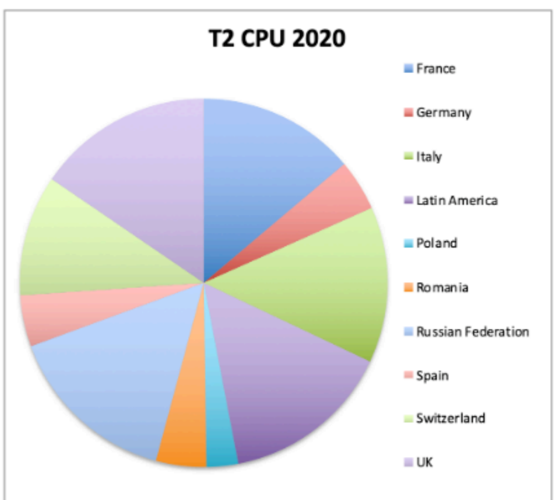


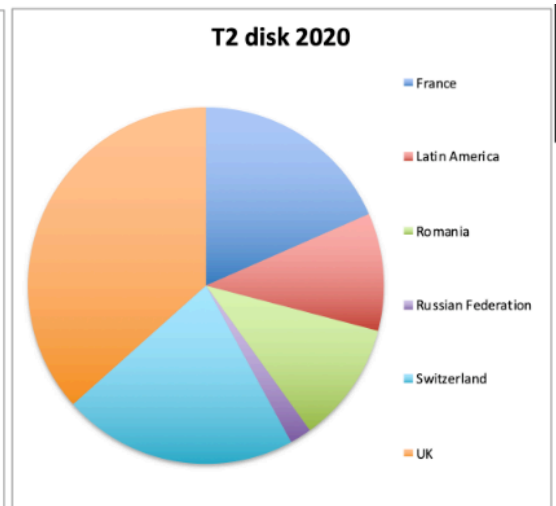


2020 T2 pledges

- With respect to requests: sligthly lower CPU, half disk
  - No demand for the new T2-D’s
  - French contribution
    - ~17% CPU
    - ~23% Disk

2020	CPU	Disk
Tier2	HS06	Tbytes
France	29905	902
Germany	10600	21
Italy	31450	0
Latin America	1000	0
Poland	7400	0
Romania	6900	400
Russian Federation	18212	65
Spain	7000	1
Switzerland	32000	1080
UK	31636	1500
Total	**176103**	**3969**
Requested	**185000**	**7200**
Difference	**-4.8%**	**-44.9%**





Offline computing requests for 2021

- Same model as in LHCb Upgrade Computing Model TDR
  - Instantaneous luminosity: 1×10^{33}
  - Integrated luminosity:
    - $3 fb^{-1}$ baseline,
    - $7 fb^{-1}$ contingency
- Detailed LHC planning for end of LS2 and Run3 being discussed

CPU Power (kHS06)	2020	2021
Tier 0	98	112
Tier 1	328	367
Tier 2	185	205
Total WLCG	**611**	**684**
HLT farm	10	50
Yandex	10	50
Total non-WLCG	**20**	**100**
Grand total	**631**	**784**

Disk (PB)	2020	2021
Tier0	17.2	20.7
Tier1	33.2	41.4
Tier2	7.2	8.0
Total	**57.6**	**70.1**

Tape (PB)	2020	2021 (baseline)	2021 (contingency)
Tier0	36.1	56	85
Tier1	55.5	96	147
Total	**91.6**	**152**	**232**

Outcome of C-RRB October 2019

LHCb Recommendations

LHCb-1 C-RSG finds that the LHCb 2021 estimates conform to the needs resulting from the upgrade LHCb computing model. The C-RSG notes that some work is still needed in the commissioning of the software trigger and the parametric MC simulation.

LHCb-2 C-RSG notes that 60 PB increase in tape storage for 2021, while CPU and disk increases are 10 to 20%. For 2022 and 2023, LHCb predicts 100 PB/year of tape and increases of 70-80% per year in CPU and disk. No increase in computing resources is foreseen for the LS3 period (2024 and 2025). The C-RSG encourages funding agencies to consider multi-year funding in order to smooth out this Run 3 profile.

LHCb-3 C-RSG requests LHCb to estimate computing resources needed for the heavy ion run in 2020 and include the corresponding requests in the next scrutiny round.

LHCb-4 C-RSG recommends LHCb continue investing in workload management system and application software to enable HPC opportunistic resources.

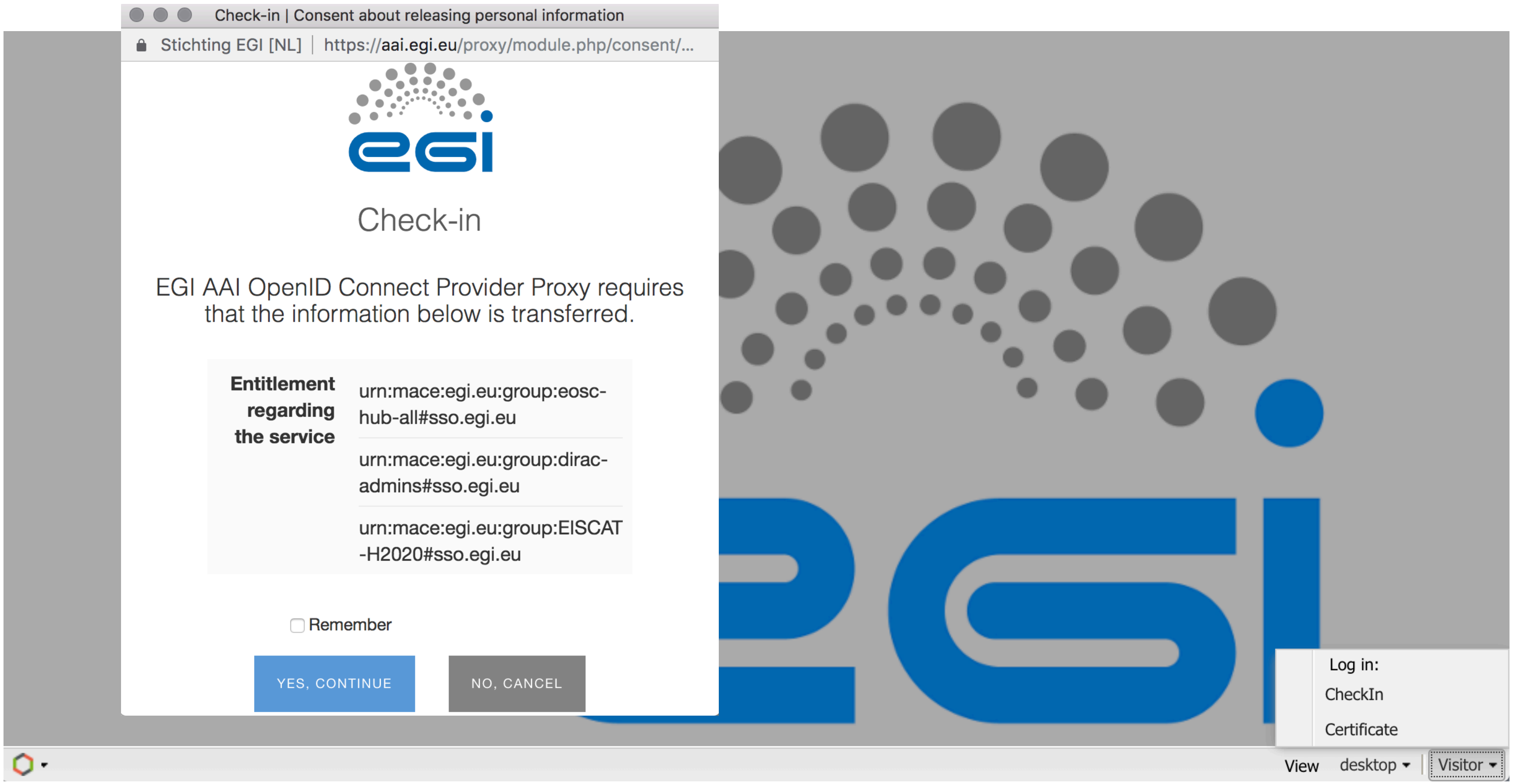
LHCb-5 C-RSG encourages the ongoing work in organized analysis to reduce storage and CPU usage resulting from individual user analyses.

Developments: AuthN/AuthZ

- VOMS will stay for a while but will not be the only AuthN/Z provider for long
- INDIGO AIM is the VOMS replacement
  - Chosen by WLCG
  - DIRAC and LHCb will have to interface to it
    - in 2020
- Auth2/OIDC support in DIRAC is developed for the EGI Workload Manager service
  - Using EGI Check-In AuthN/Z service
  - Can be easily adapted to LHCb
    - INDIGO AIM ?
    - CERN SSO ?

Developments: AuthN/AuthZ

- Web Portal authentication



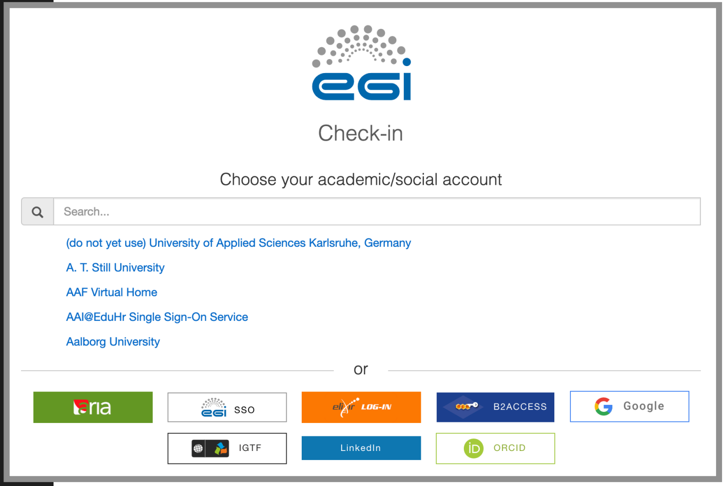
Developments: AuthN/AuthZ

CLI Authentication

```
[[dirac@ce-emi pro]$ python DIRAC/FrameworkSystem/scripts/dirac-proxy-init.py -O CheckIn -g training_user -q
 OAuth authentification from CheckIn.
 Use link to authentication..
 https://ce-emi.bitp.kiev.ua:9943/oauth2/oauth?getlink=MZ7XnO4iyMYTx9Vw2wkpBbHrm3Gz8f

[ Waiting 3.0 minutes when you authenticated..    ..*      $ [3~

 Proxy generated:
 subject      : /DC=org/DC=ugrid/O=people/O=BITP/CN=Andrey Litovchenko/CN=3461819742
 issuer       : /DC=org/DC=ugrid/O=people/O=BITP/CN=Andrey Litovchenko
 identity     : /DC=org/DC=ugrid/O=people/O=BITP/CN=Andrey Litovchenko
 timeleft     : 23:59:59
 DIRAC group  : training_user
 rfc          : True
 path         : /tmp/x509up_u3310
 username     : alitov
```

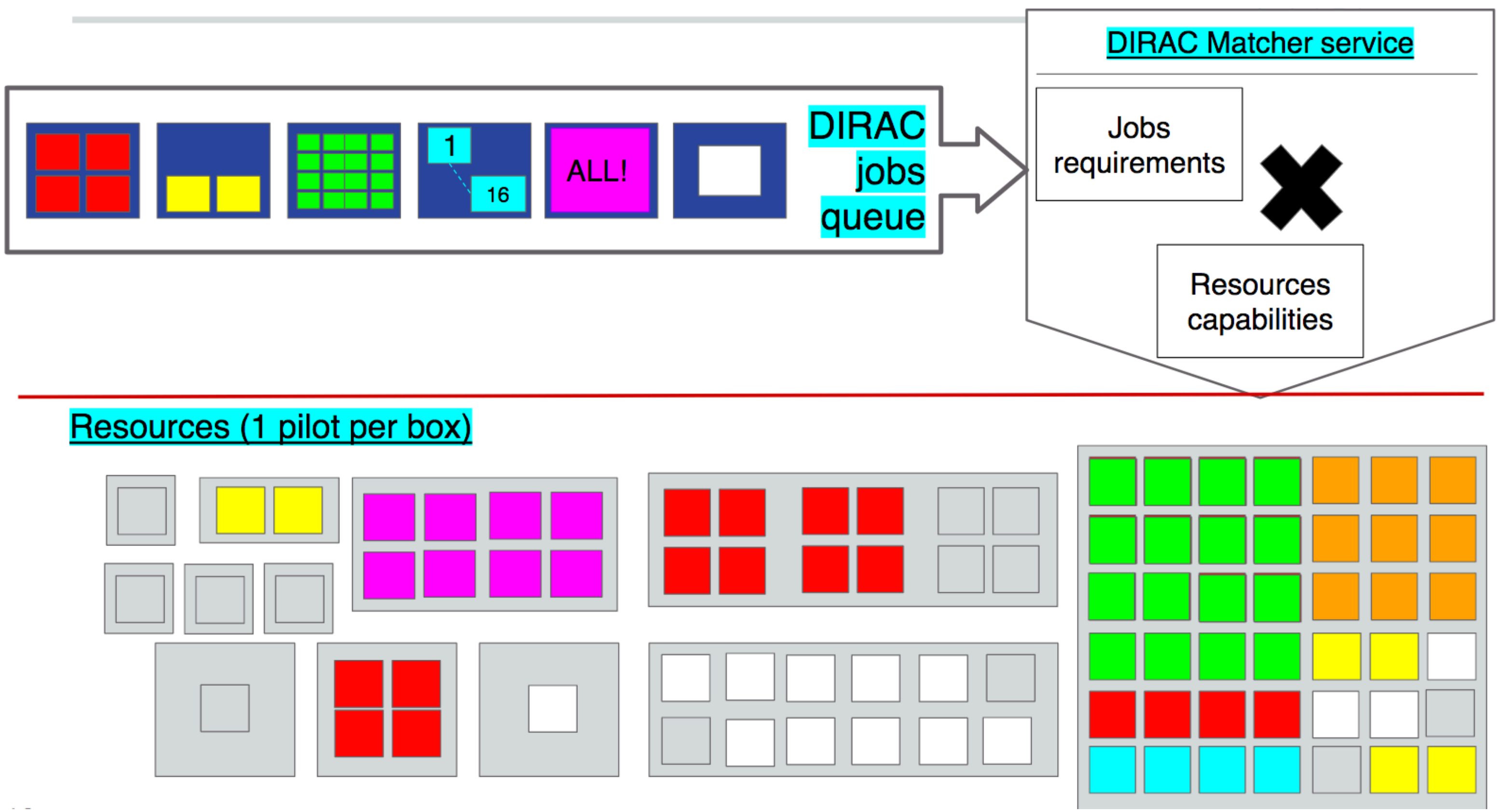


Developments: MP Jobs

- LHCb has access to several HPC centers
  - CSCS, CINECA/Marconi, Santos Dumont
- Example Marconi A-2 at CINECA node
  - 68 processors XeonPhi 7250
  - 272 logical processors
  - 96 GB RAM
    - 350MB RAM per logical processor !
  - Node outbound connectivity available
  - CVMFS available
- Fat node
  - DIRAC needs to partition the node for optimal memory and throughput

Developments: MP Jobs

- Using DIRAC PoolComputingElement – running a small batch system inside the pilot on the worker node
  - Matching parallel or SP jobs to fully exploit the node



Developments: Python3

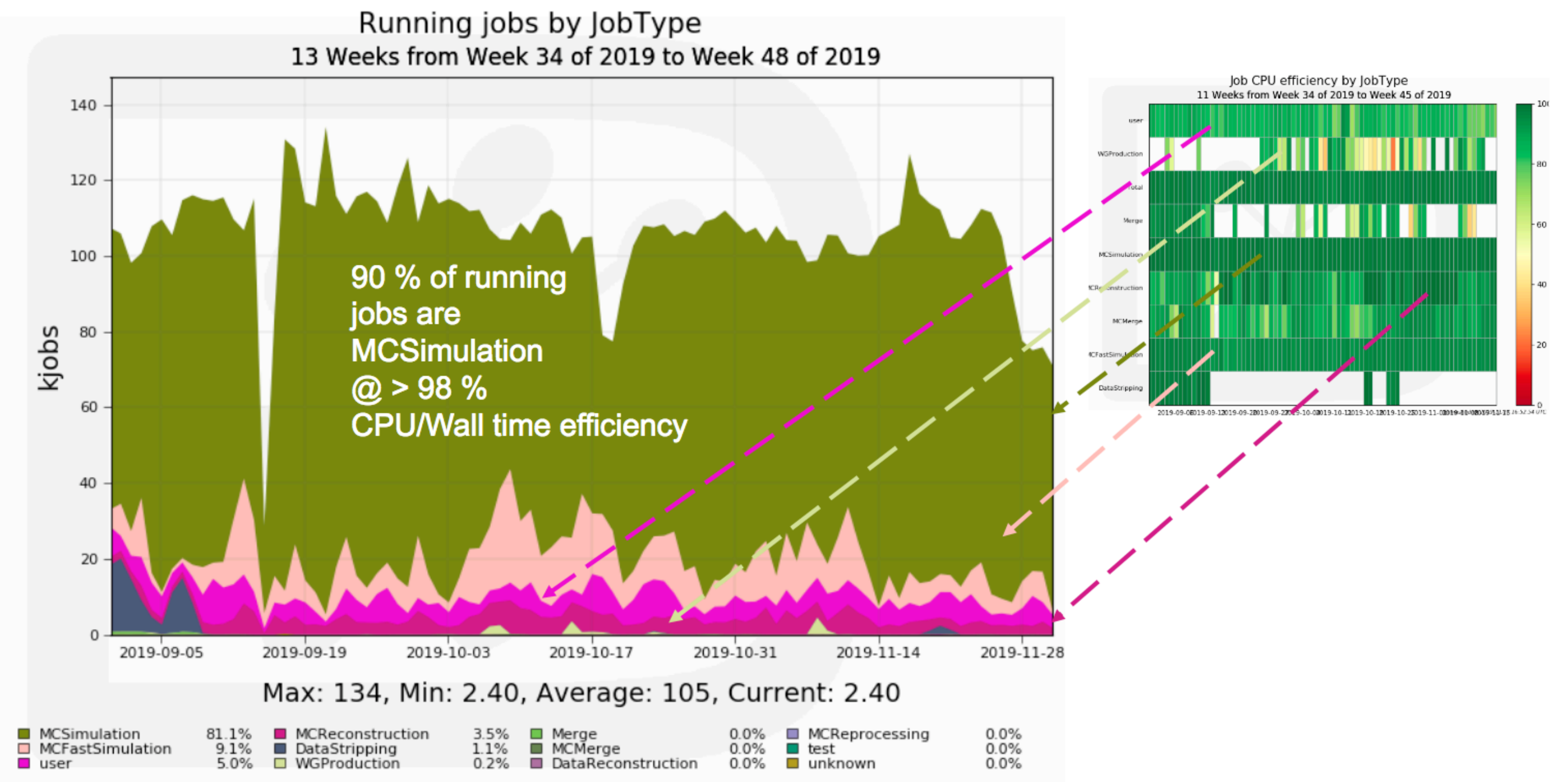
- (LHCb)DIRAC is written in Python2
  - ~400K lines of code
- DIRACOS shipped with DIRAC is containing Python2 (currently 2.7.13)
  - Can rely on Python2 forever, but…
  - Some dependency software will not support Python 2 soon
- Some codes must run with the OS Python
  - Pilots, DIRAC installation scripts
  - SLC7, CC7, CC8
- Work in progress
  - Progressively make code Python 2 and 3 compatible through 2020
  - Drop Python 2 in the longer term

Developments: Containers

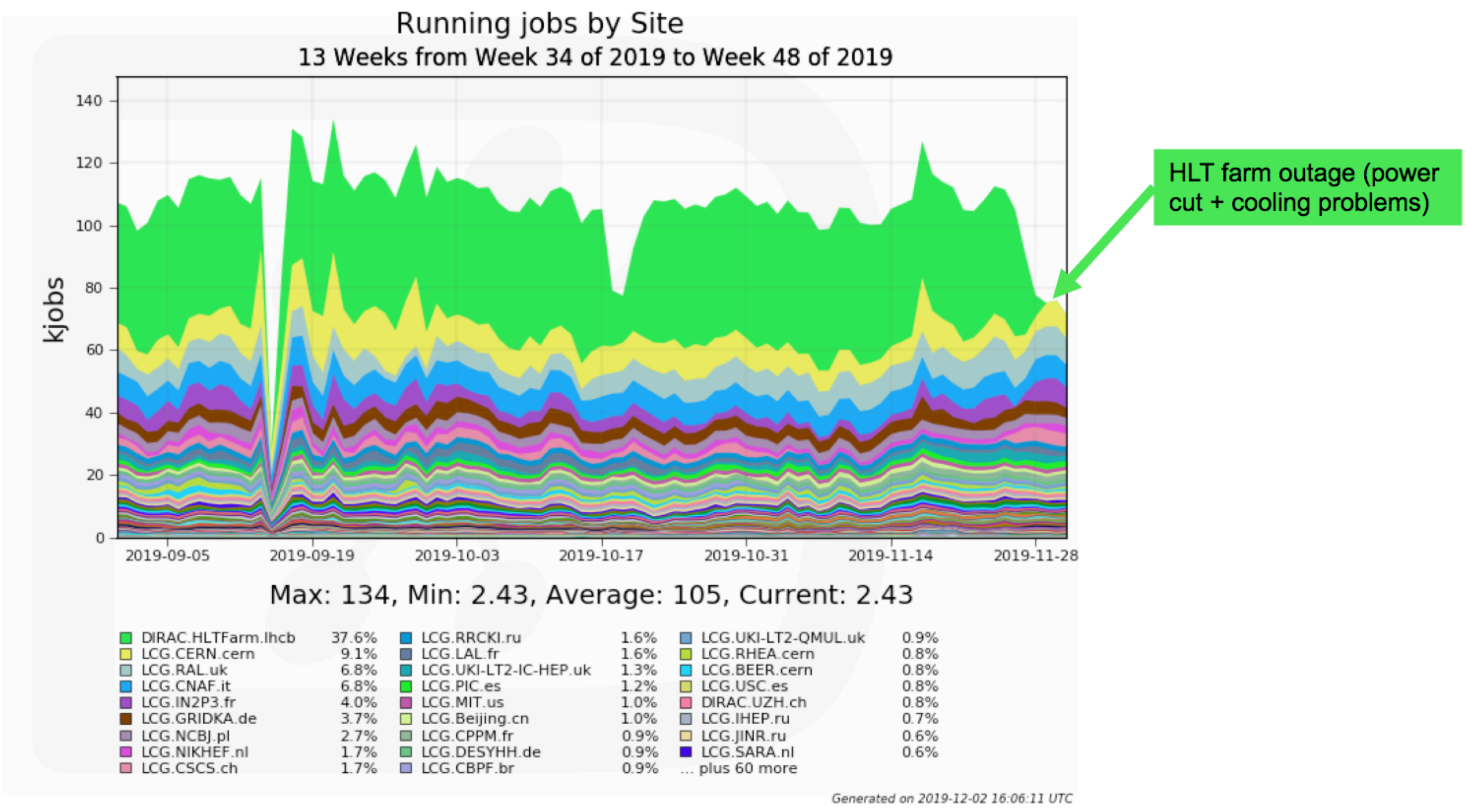
- For software preservation
  - E.g. running SLC5 compiled legacy trigger code on CC7 nodes
- For user analysis
  - Ganga is planning to encapsulate user applications in containers
- Payload isolation
  - glexec -> Singularity
    - Using SingularityComputingElement of DIRAC

- LHCb asked all the T1 and T2-D sites to provide Singularity
  - Running Singularity from CVMFS requires user namespace mode
  - Other T2's will be asked to provide Singularity also

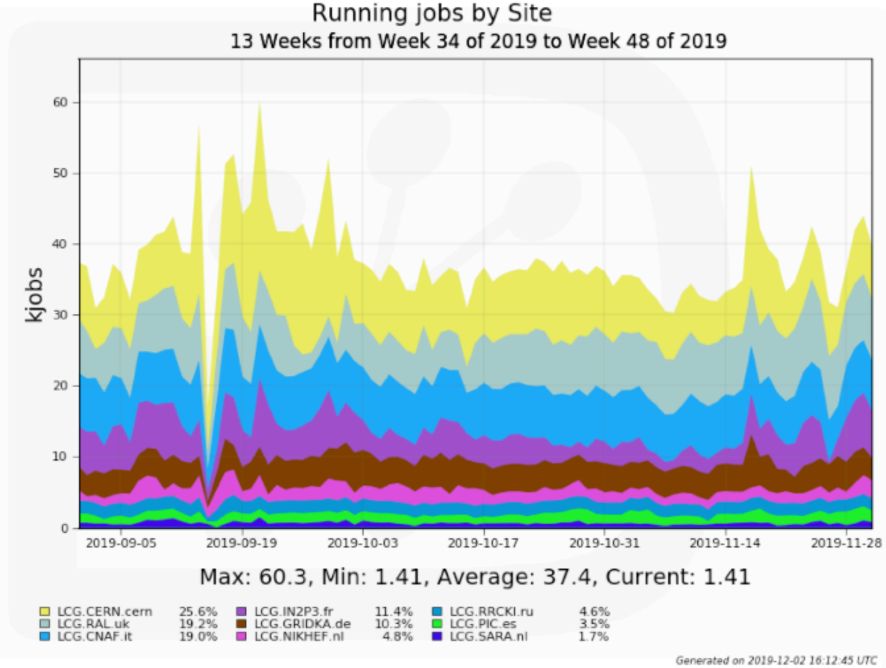
Current activities

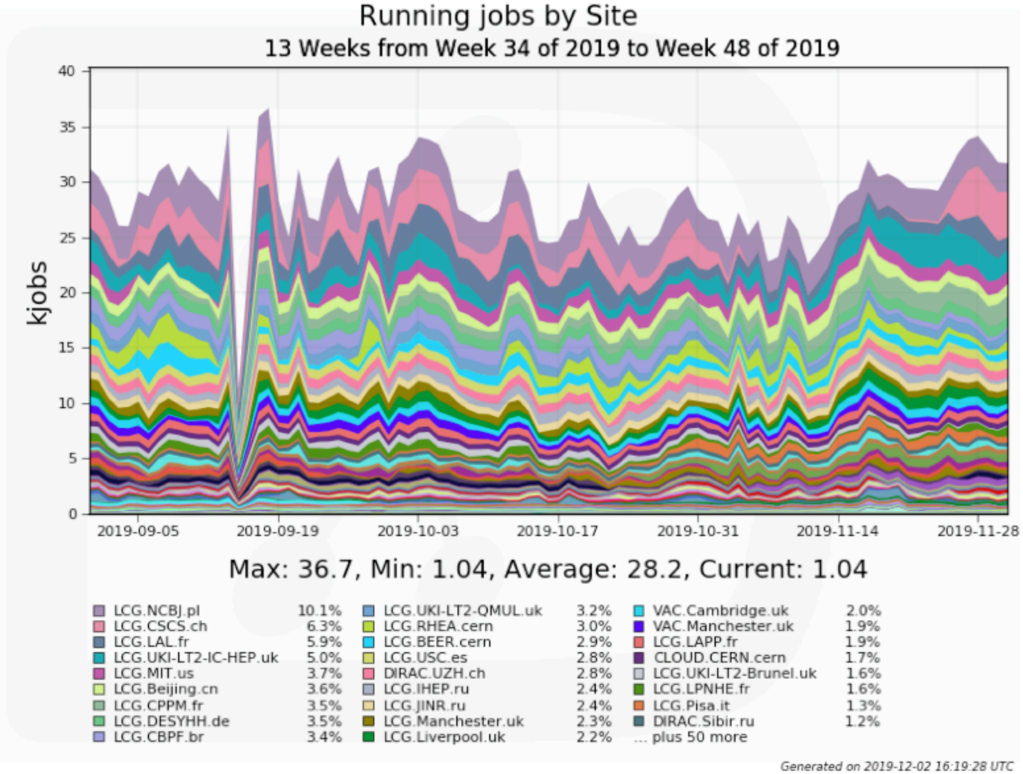


Current activities



Current activities



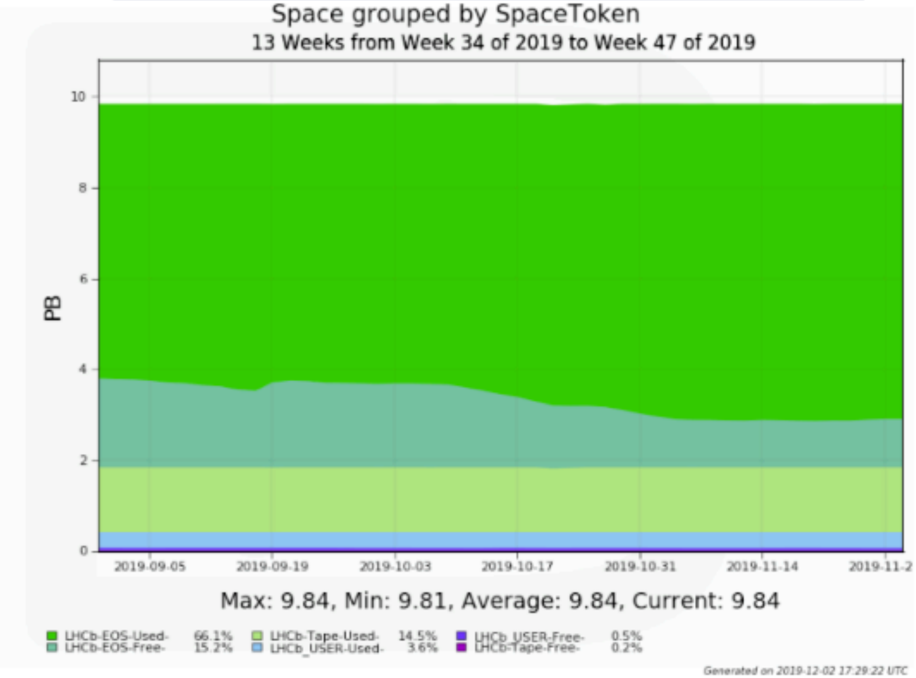


T0+T1 sites

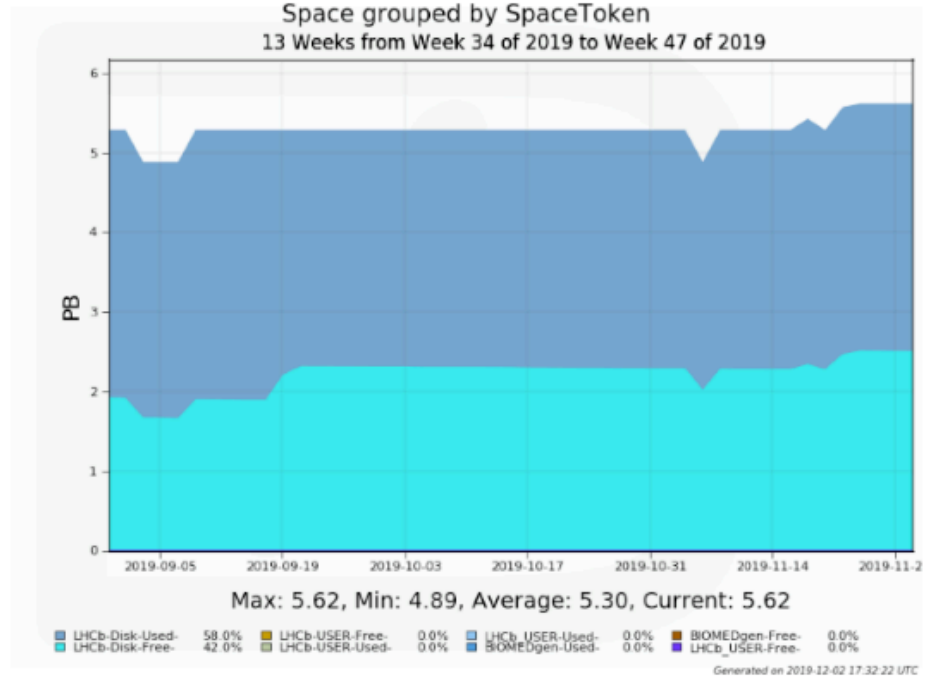
All the rest (HLT farm excluded)

Storage usage (by space token)

CERN (84% full)

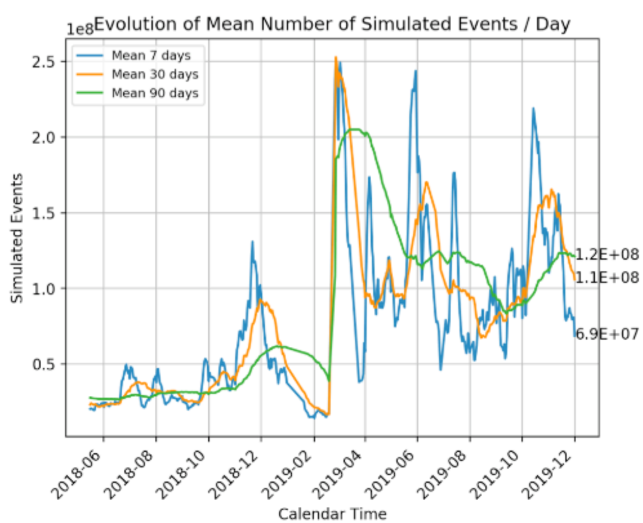


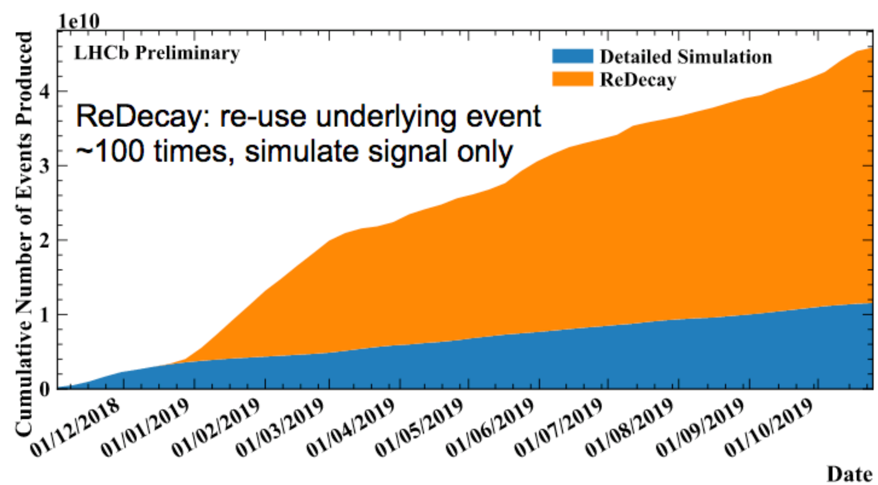
Tier 2D sites (58% full)



Current activities

- Legacy stripping campaigns for all Run1 and Run2 data under way
- Simulation is using 90% of the computing power
  - "fast" simulation used to produce 80% of events in last year





Conclusions

- LHCb Computing model for the Run 3 Upgrade is updated to reduce the use of expensive resources
  - trigger output bandwidth of 10 GB/s to tape/3.5 GB/s to disk
- CPU needs for Run 3 are dominated by MC production
  - Massive use of faster simulation techniques
- Developments are ongoing to accommodate advancements in software and technologies (python, AuthN/Z, HPC, containers, etc)
- Smooth running of LHCb Computing project, most of the computing resources is for the MC production currently
- Smooth running of the french sites (T1, T2, T2-D)