LHCb News

LCG-France Meeting

November 18^{th} 2021

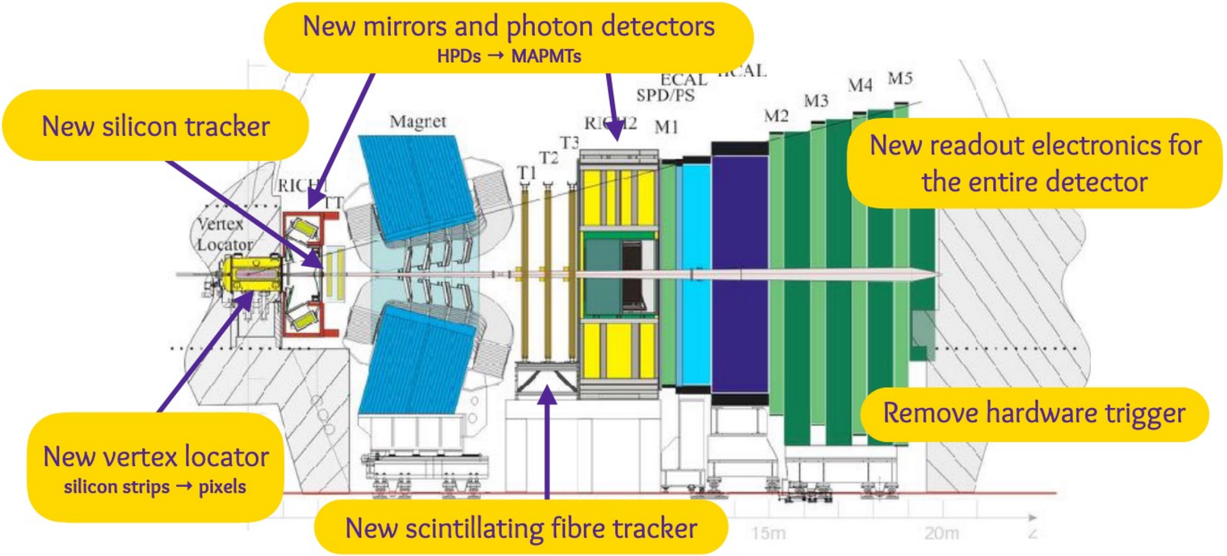
Andrei Tsaregorodtsev, CPPM

Outline

- LHCb Run3 detector upgrade
- Run 3 Computing Model
- Current operations
- Status of requests/pledges 2022-2023 and beyond
- Conclusions

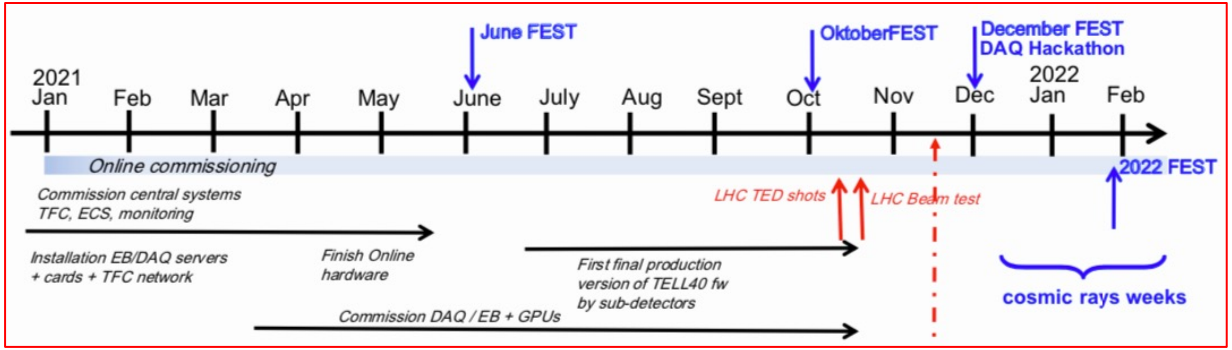
CERN-LHCC-2012-007

The upgraded LHCb detector for Run 3



- This is a new detector !
  - Major detector modifications, new tracker
  - 100% new RO electronics, DAQ

Run3 roadmap



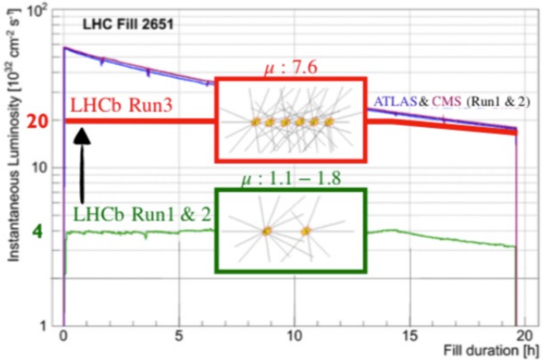
Global activities:

- Online and sub-detector commissioning during LHC beam test
- Commissioning weeks with cosmic ray tests, integrating newly installed detectors while their stand-alone commissioning progresses
- Full Experiment System Test (FEST)
    - Simulated samples injected in the online system
    - Full dataflow run in commissioning weeks

Run 3 conditions for LHCb

Luminosity increase: x5

- More interaction vertices per collision of proton bunches, more tracks, more signal
- Beauty and charm signal rates: 1-10MHz
- Almost all events will have a *b* or *c* hadron in Run 3



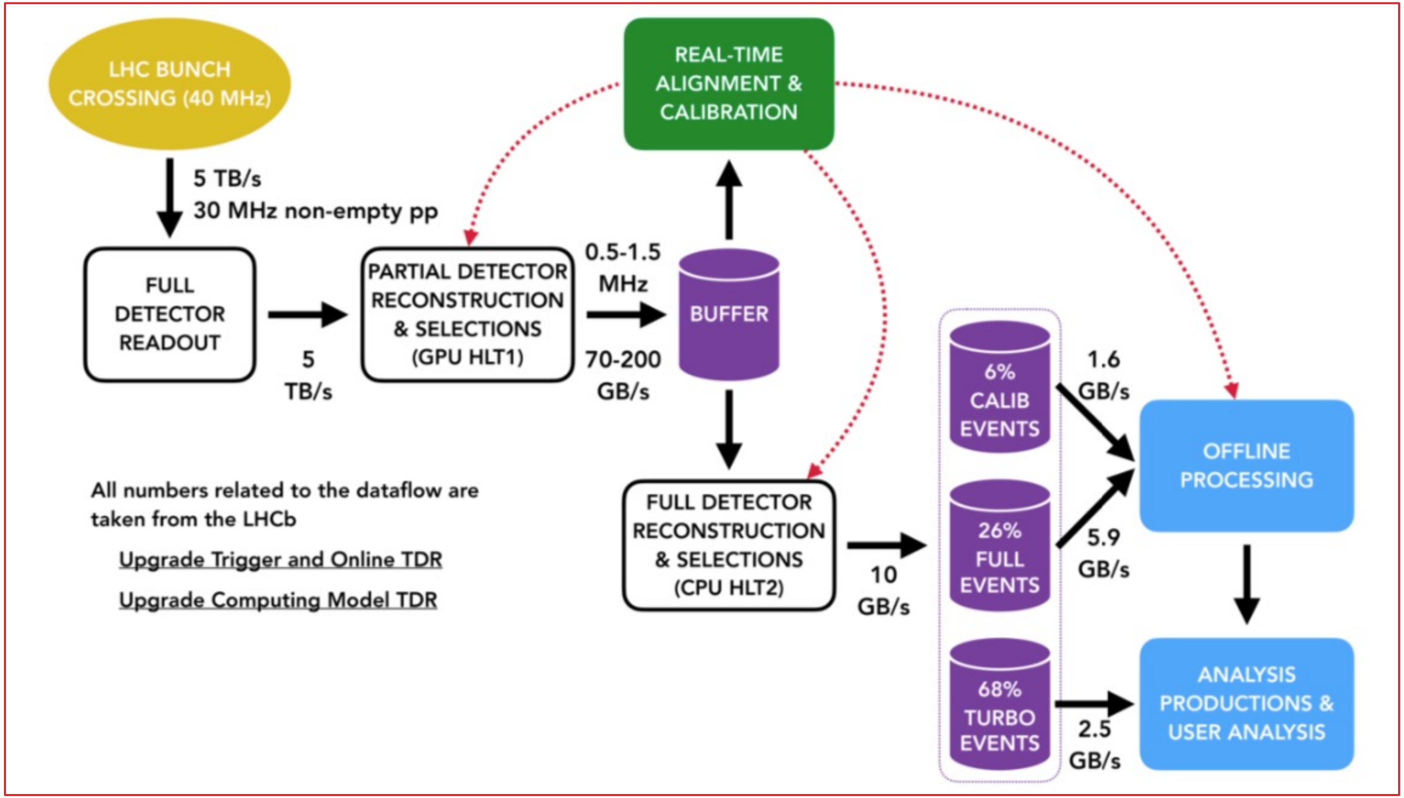
- Hardware trigger is no more an option
  - No simple local criteria
  - Track reconstruction is needed for event selection
    - Discover event topology as early as possible
- Full software trigger is required

The MHz signal era

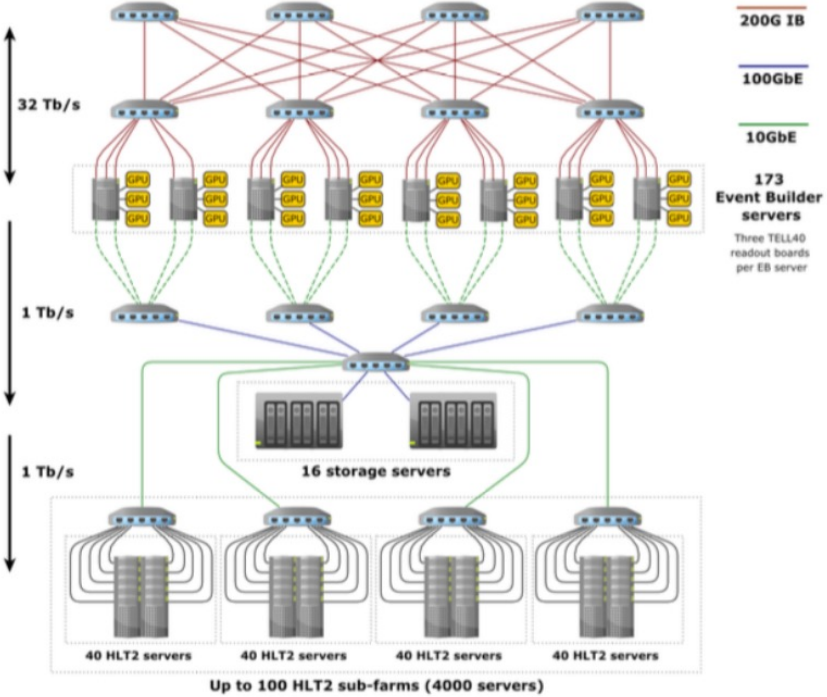
"From a needle in a haystack to an haystack of needles"

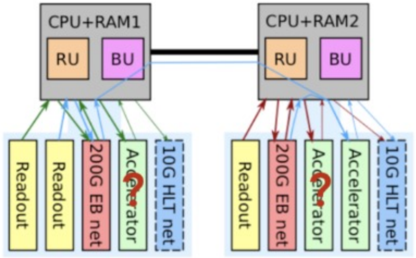
Run3 Computing Model

LHCb Run 3 Data Flow

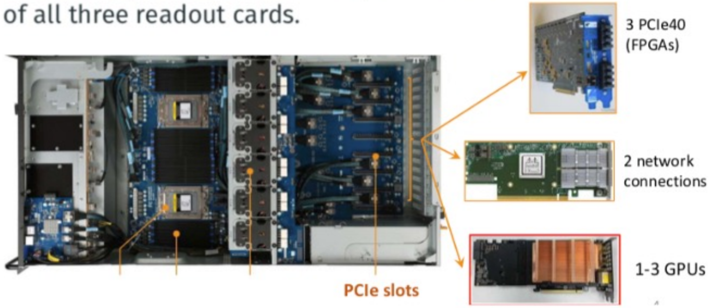


LHCb Run3 HLT practical implementation



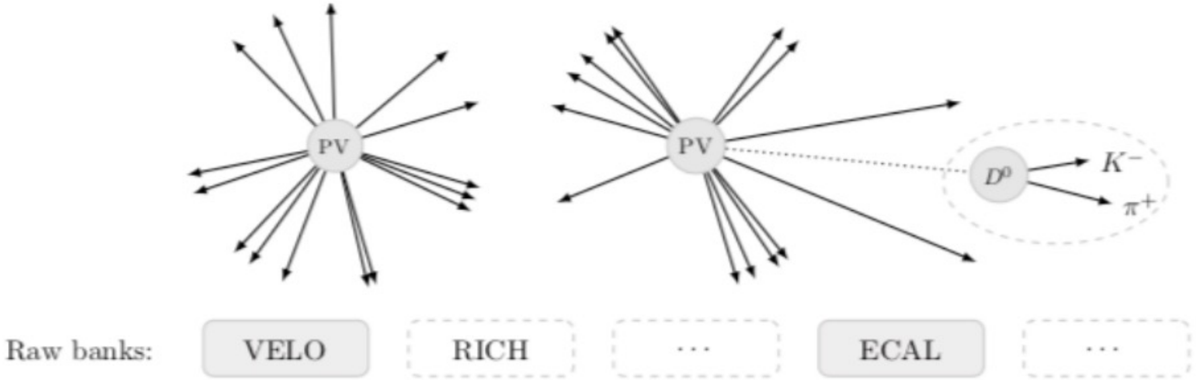


GPU-equipped event builder PC, with traffic of all three readout cards.



Persistency model

- Selective persistency: write out only the "interesting" part of the event.



- Turbo stream:
  - Miminum output: only HLT2 signal candidates
  - Optionally: (parts of) pp vertex (e.g. "cone" around candidate for spectroscopy searches)

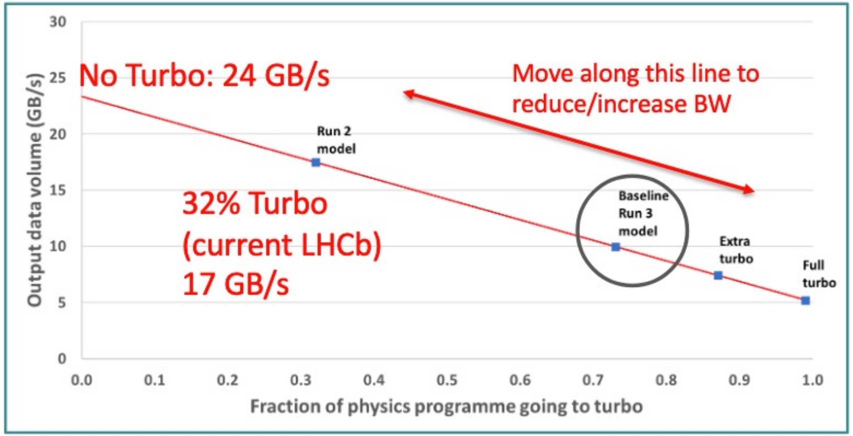
Limitations: cannot refit tracks and PVs offline, rerun flavour tagging etc.

Advantage: Event size O(10) smaller than RAW

- FULL stream: all reconstructed objects in the event
  - Optionally adding selected RAW banks
- TurCal stream: HLT2 candidates and RAW banks
  - Used for offline calibration and performance measurement

Output rates

- Moving a larger fraction of the physics program to Turbo decreases the output bandwidth
  - Turbo events – 16% of Full size events
- Baseline assumes 73% of the physicis selections on Turbo
  - Correponds to the output bandwidth of 10GB/s

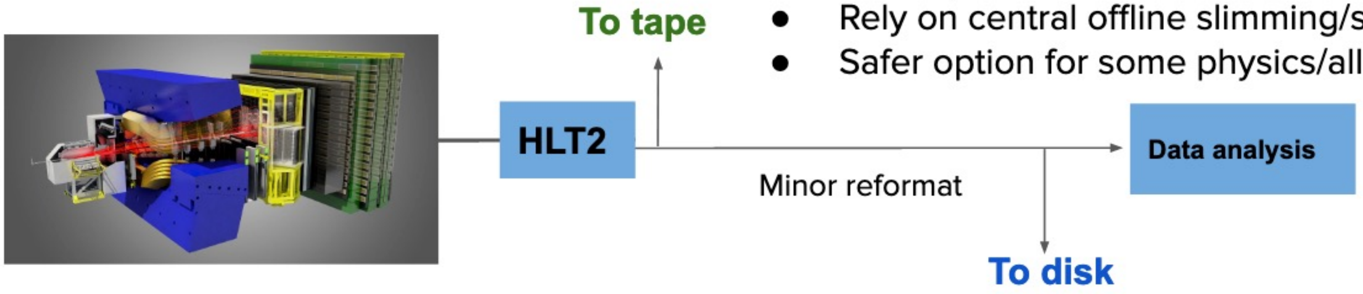


Data flow evolution Upgrade

Cannot save all HLT output straight to **disk**!

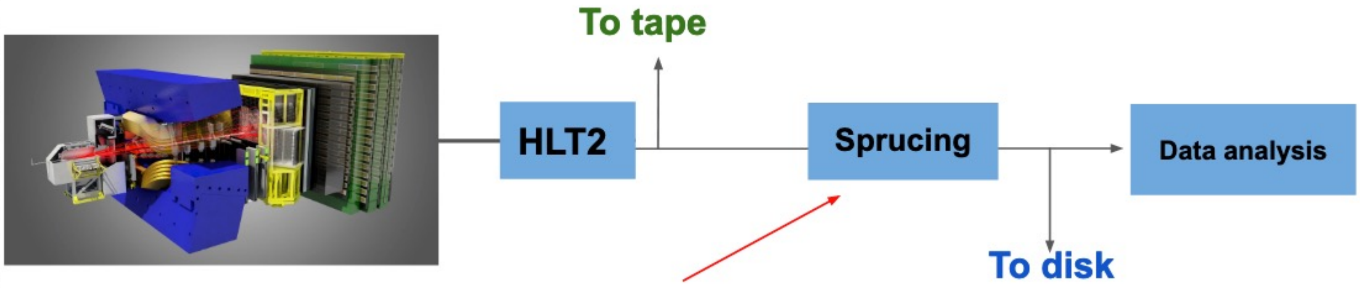
- Utilise cheap **tape storage** for bulk of bandwidth (full stream)
- Rely on central offline slimming/skimming
- Safer option for some physics/allows data mining

Default model
Turbo



Default model - >70% of physics

Sprucing model
FULL and TURCAL



Use cases - topological, inclusive triggers, datamining

A further offline stage of data reduction/selection between tape and disk storage when HLT2 line throughput is too large to go straight to disk. Utilise same selection framework as HLT2

Streaming and filtering in Run3

- Can we 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"
    - Selecting events to store
      - O(10^3) selection lines
    - Selecting a subset of reconstructed objects to store
- Save **3.5 GB/s** to disk!

Throughput to tape

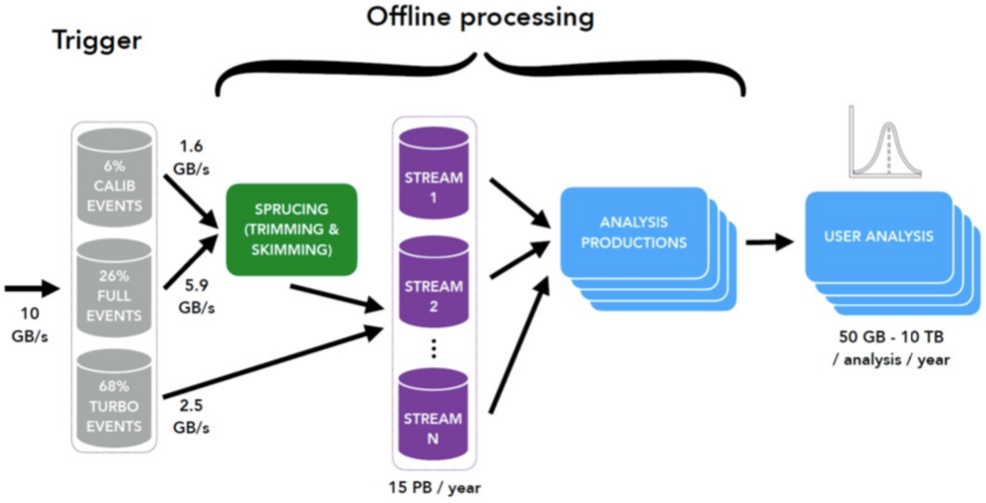
stream	rate fraction	throughput (GB/s)	bandwidth fraction
FULL	26%	5.9	59%
Turbo	68%	2.5	25%
TurCal	6%	1.6	16%
total	100%	10.0	100%

Throughput to disk

stream	throughput (GB/s)	bandwidth fraction
FULL	0.8	22%
Turbo	2.5	72%
TurCal	0.2	6%
total	3.5	100%

“Data Processing and Analysis” (DPA) project
An offline workflow for the 2020s

Very large increase in data volume wrt. Run II brings challenges to offline data processing and analysis
DPA built around 2 main ideas:
- Centralised skimming and trimming (aka Sprucing of significant fraction of HLT2 outputs)
- Centralised analysis productions for physics WGs and users



DPA project
Offline processing/selections/analysis

- WP1 - Sprucing
- WP2 - Analysis Productions
- WP3 - Offline Analysis Tools
- WP4 - Innovative Analysis Techniques
- WP5 - Legacy Software & Data
- WP6 - Analysis Preservation & Open Data

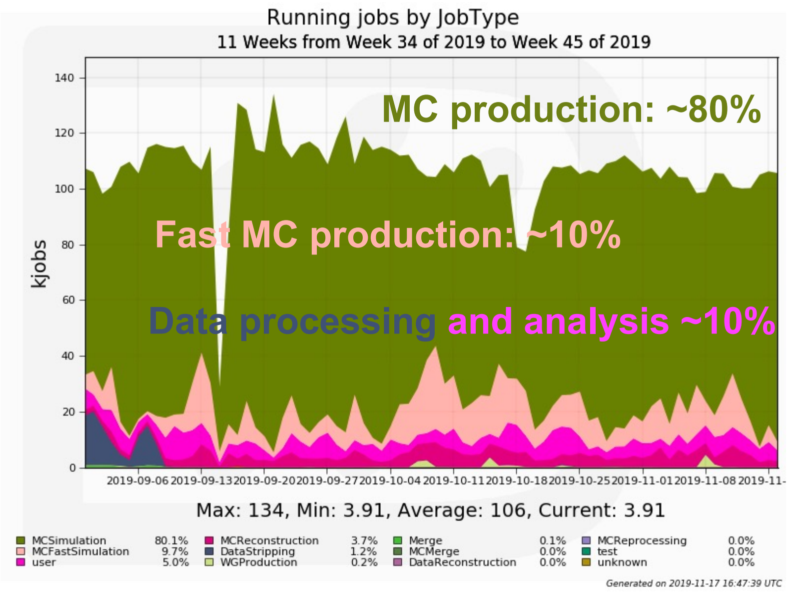
19

V.Gligorov

The model: what about CPU ?

- CPU is dominated by MC production (~90% of CPU power)
- Expected to be the same at the Upgrade

- Baseline simulation numbers:
  - Event timing:
    - Full/fast/parametric simulation: 120/40/2 seconds
  - Sharing full / fast / parametric: 40/40/20
- Aggressive use of faster simulation techniques:
  - Reduce CPU need
  - No effect on tape
  - No effect on disk
  - May not be feasible, strongly linked to analysis

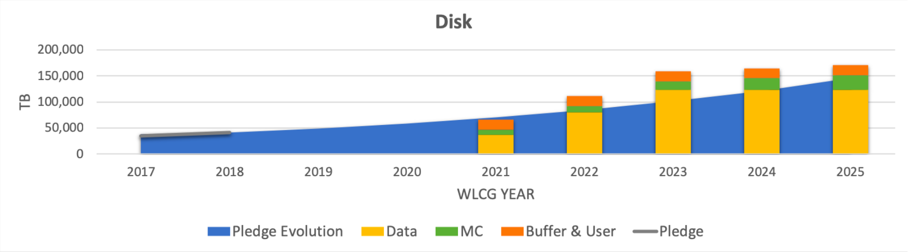


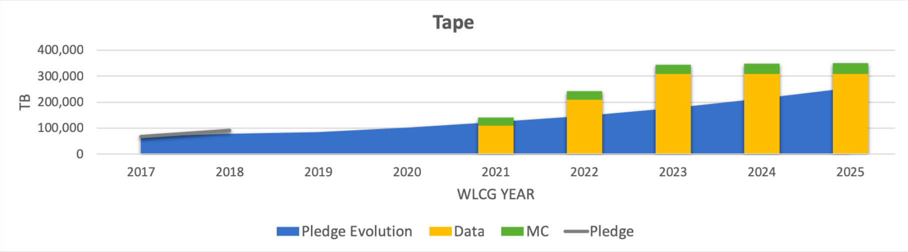
Run 3 Computing model requirements

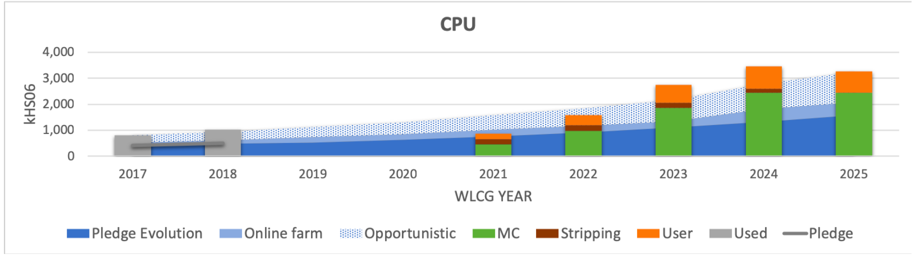
- Assumptions on simulated event volume
  - N. of MC events scales with L_{int}
  - MC production for a data taking years extends over the following 6 years
  - **MC events saved in MDST format (x40 size reduction!)**
- Assumptions on replicas

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

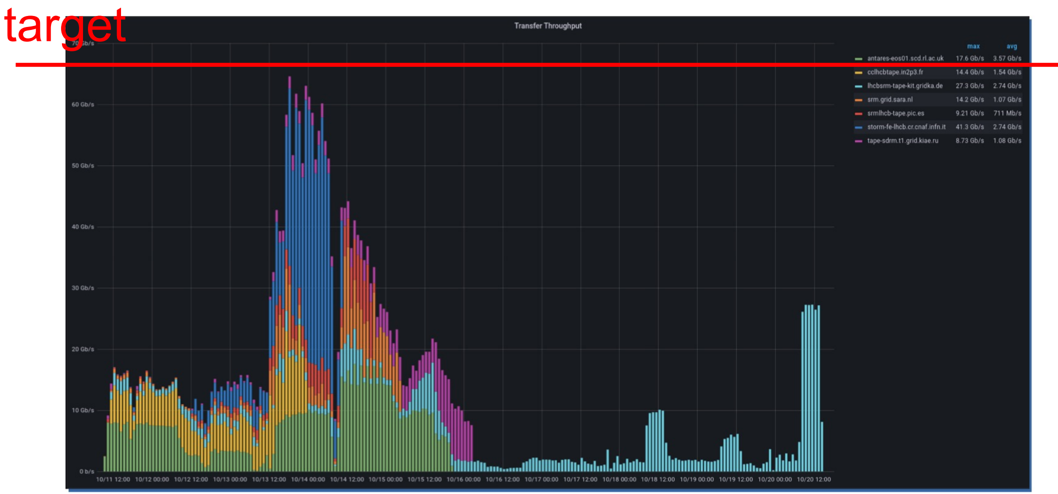






WLCG tape challenge

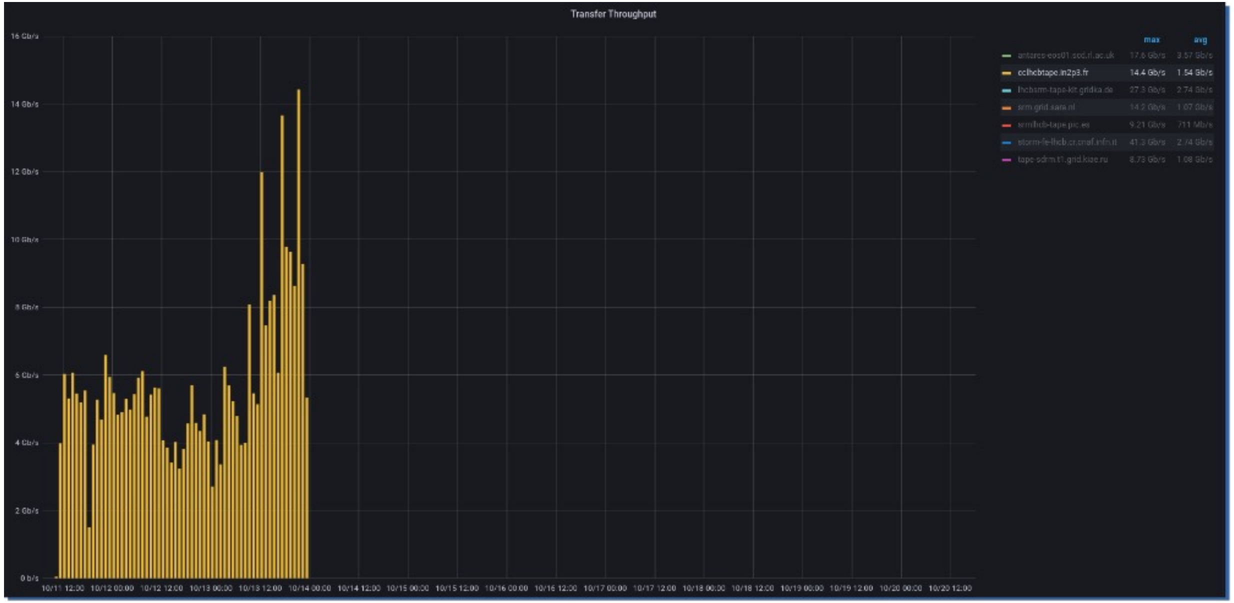
WLCG tape challenge



- Some details here
- EOS -> T1 write tests
  - Real staging activities in parallel
- DIRAC scaled perfectly
- Met average rate, close to peak rate
  - Issues: FTS settings, number of EOS gridftp gateways, sites configuration
  - the main bottleneck (EOS gridftp gateways) should disappear by the start of Run3
    - Moving to (SRM +) HTTPs
- Not a complete success but
  - Good reminder of the FTS tuning we have to do
  - Highlighted the importance of monitoring
    - efforts required in DIRAC
  - Gave ideas to further optimize the data export from P8

Site	Expected Speed (GB/s)	Average Speed (GB/s)	Max Speed (GB/s)	Duration (hours)
CNAF	2.24	1.07	5.16	72
IN2P3	1.26	0.70	1.8	61
NLT1	0.88	0.33	1.77	90
RRC-KI	0.88	0.27	1.09	112
PIC	0.58	0.24	1.15	82
RAL	2.92	0.93	2.2	106
Gridka	2.24	0.35	3.41	220

WLCG tape challenge: CC/IN2P3

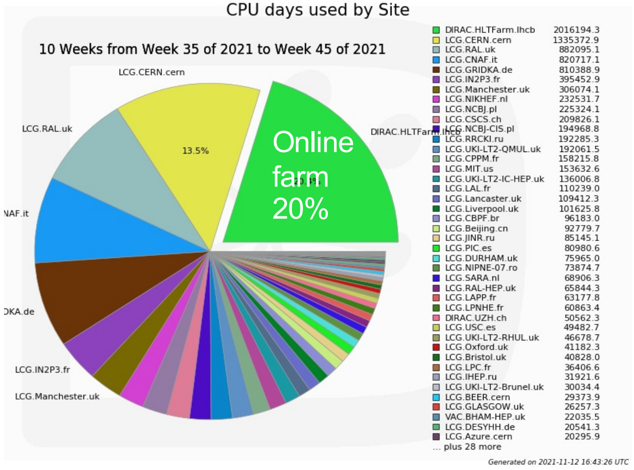


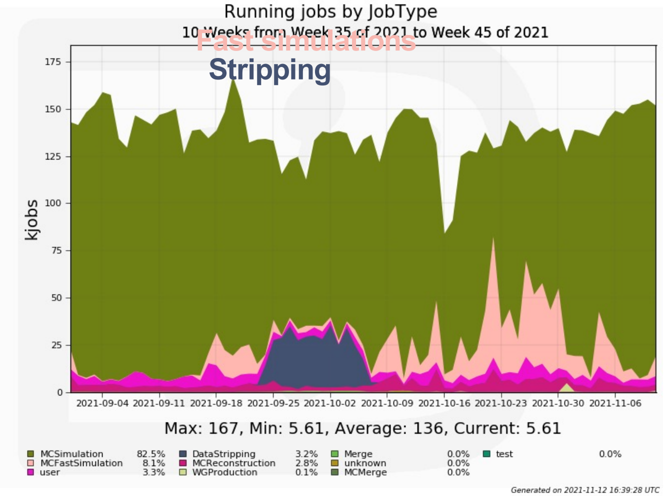
- Immediate start
- Jumps in the throughput
- Target: 1.26 GB/s; average 0.70 GB/s; peak 1.80 GB/s

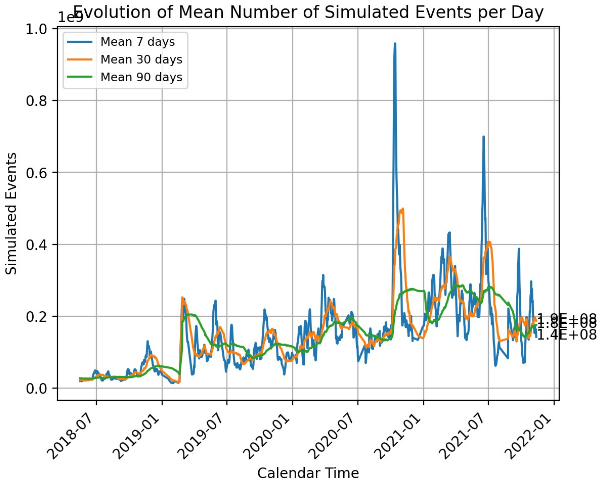
Current Operations

Distributed computing operations

- Computing work dominated by MC production (94%)
- Fast:detailed simulation = 50:50
- Simulating about 180 million events per day
- Incremental stripping of 2018 data recently completed

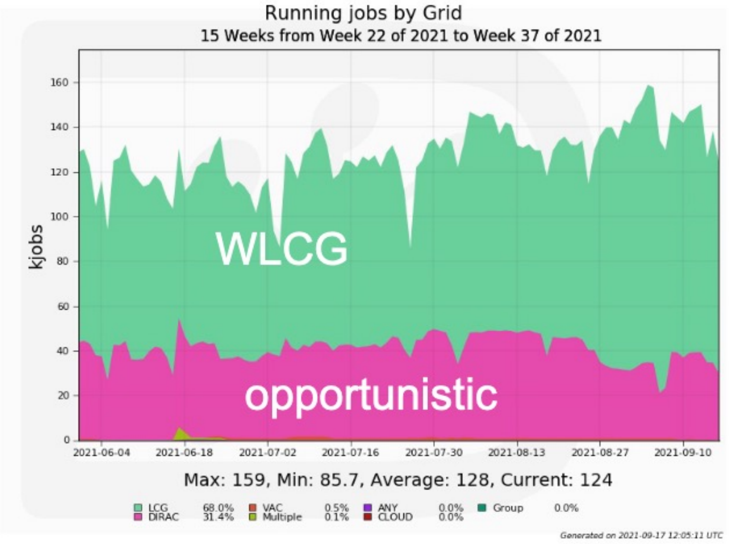


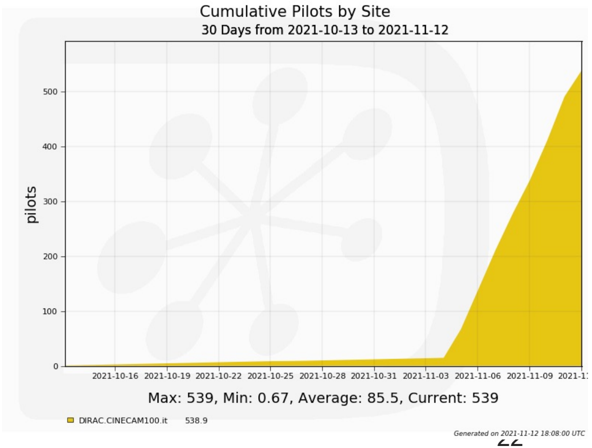




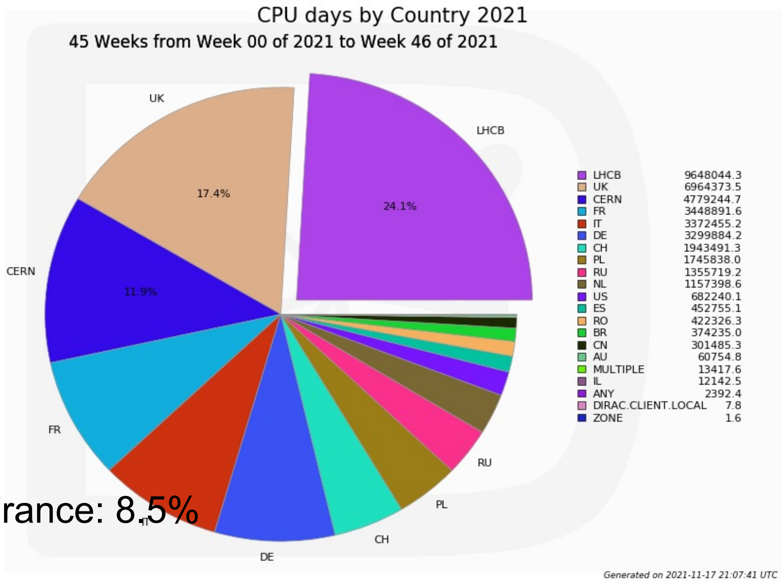
Opportunistic resources

- HLT farm 20%
- Non-pledging sites 10%
- HPCs
  - NERSC, CSCS, SDumont now in production
  - Barcelona Supercomputing Center (BSC), still not in production
    - Installation and configuration of ARC CE
  - CINECA/Marconi100
    - GPU + Power9: difficult to use in normal production workflows, no full software build
    - Some user jobs run locally, very limited CPU consumption
    - DIRAC configured for grid-like access, pilots sent but no matching jobs yet
  - O(1000) computing slots in total
    - Not a lot !

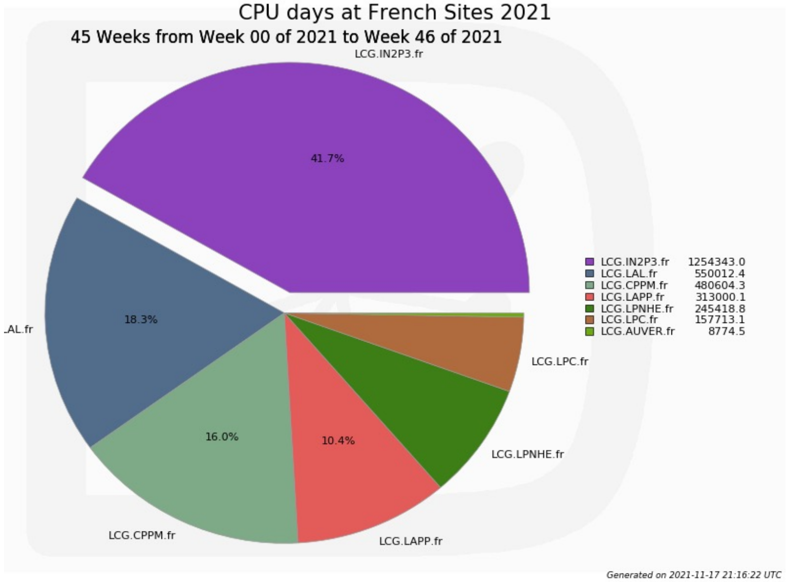




French contributions in 2021



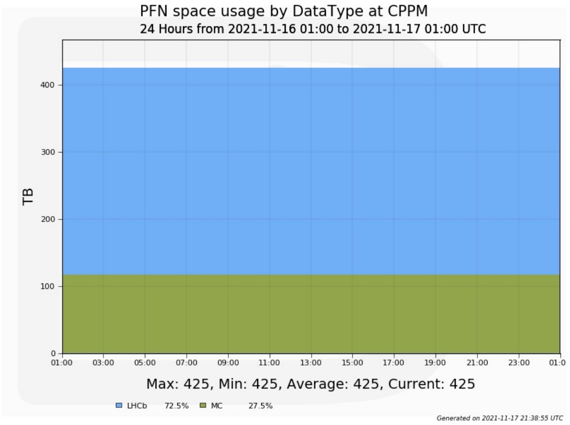
France: 8.5%

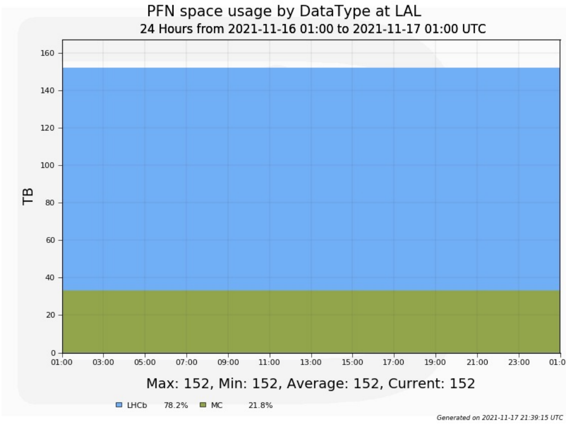


- No particular comments with respect to the French sites functioning
  - Some occasional problems with running pilots at Condor/CC – solved by Vanessa

Disk space usage at T2D's 2021

- CPPM
  - Pledged 600TBs, used 425 TBs
  - Occupancy 71%
- LAL
  - Pledged 383TBs, used 152 TBs
  - Occupancy 40%
- LHCb T2D policy
  - T2D introduced to allow countries without T1's to contribute storage resources
  - No special use of storages at T2's compared to T1 storage - what matters is T1+T2 disk storage
  - But more attention to SEs at T2 sites due to less operational overheads
    - Single person responsible for data management





Requests and pledges

2022 pledges situation

Tier	Pledge Type	Year	LHCb Required	LHCb Pledged	LHCb Balance
0	Tape	2022	81000	81000	0 %
0	Disk	2022	26500	26500	0 %
0	CPU	2022	189000	189000	0 %
1	Tape	2022	139000	116337	-16 %
1	Disk	2022	52900	47783	-10 %
1	CPU	2022	622000	514531	-17 %
2	CPU	2022	345000	332640	-4 %
2	Disk	2022	10200	6941	-32 %
Tier	Pledge Type	Year	LHCb Required	LHCb Pledged	LHCb Balance

~10% lower pledges at Tier1s – significantly less disk at Tier2s
Reality check needed vs. e.g. LHC planning and LHCb readiness

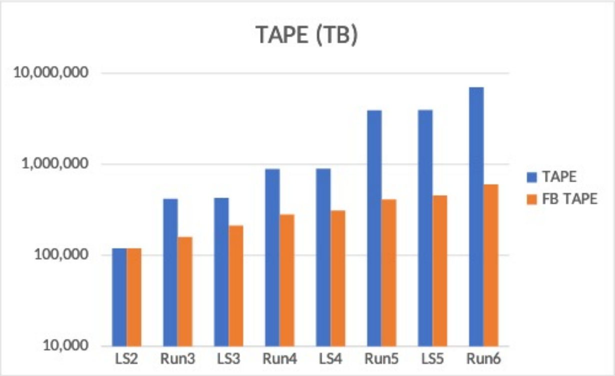
2023 preliminary requests shown at the C-RRB

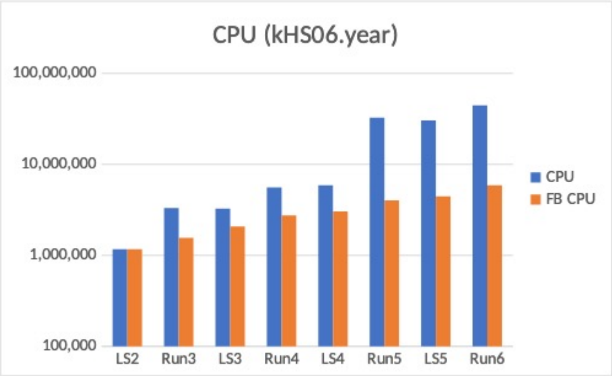
		LHCb-PUB-2021-002		THIS DOCUMENT	
LHCb		**2022**		**2023 (prelim.)**	
		Request	**2022 req. / 2021 CRSG**	**Request**	**2023 req. / 2022 CRSG**
WLCG CPU	**Tier-0**	189	108%	361	190%
	Tier-1	622	108%	1185	191%
	Tier-2	345	107%	657	190%
	HLT	50	100%	50	100%
	Sum	1206	108%	2252	187%
Others		50	100%	50	100%
Total		**1,256**	**107%**	**2,302**	**183%**
Disk	**Tier-0**	26.5	141%	42.8	162%
	Tier-1	52.9	141%	85.6	162%
	Tier-2	10.2	141%	16.5	162%
	Total	**89.6**	**141%**	**144.9**	**162%**
Tape	**Tier-0**	81	184%	132	164%
	Tier-1	139	184%	228	164%
	Total	**219.9**	**184%**	**360.5**	**164%**

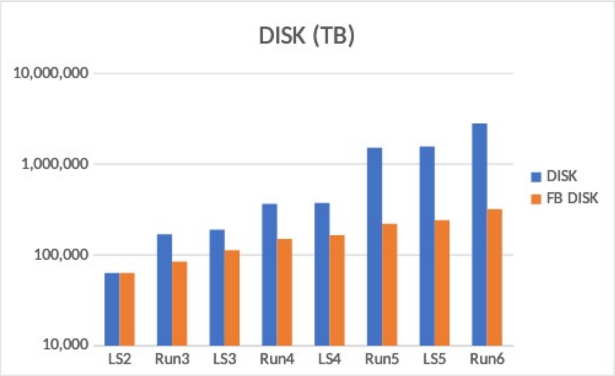
Upgrade I and II computing model assumptions

Model assumptions		
	Upgrade I	Upgrade II
Peak L ($cm^{-2}s^{-1}$)	2×10^{33}	1.5×10^{34}
Yearly integrated luminosity (fb^{-1})	10	50
Logical bandwidth to tape (GB/s)	10	50
Logical bandwidth to disk (GB/s)	3.5	17.5
Running time (s)	5×10^6	
Trigger rate fraction (%)	26 / 68 / 6 Full / Turbo / TurCal	
Ratio Turbo/Full event size	16.7%	
Ratio full/fast/param. MC	40:40:20	
CPU work per event full/fast/param. MC (HS06.s)	1200 / 400 / 20	
Number of simulated events	4.8×10^9/ fb^{-1}/year	
Data replicas on tape	2 (1 for derived data)	
Data replicas on disk	2 (Turbo); 3 (Full, TurCal)	
MC replicas on tape	1 (MDST)	
MC replicas on disk	0.3 (MDST, 30% of the total dataset)	

Resources required for Run 4,5,6







New resources can not be acquired in a scheme where funding is flat and performance increase by 10% each year.

Mitigation strategies

- Similar to ATLAS and CMS, huge R&D effort of the HEP community
  - Simulation
    - GEANT4 running on GPU
    - Calorimeter cluster simulation using ML techniques and/or shower libraries
    - …
  - Reduce storage requirements
    - nanoAOD format
    - Lossless data compression
    - Improves data placement
    - …
  - Skilled manpower is the key for the success !

Conclusions

- Run3 is a huge challenge for the new LHCb detector, trigger, DAQ and offline processing
- Smooth ongoing offline computing operations dominated by the MC production
- Pledges for the coming years are below the LHCb requests
- Ongoing effort to optimise the MC software, data production procedures, onvolve new opportunistic resources including HPCs