LHCb update

Ben Couturier, CERN

Concezio Bozzi, INFN Ferrara

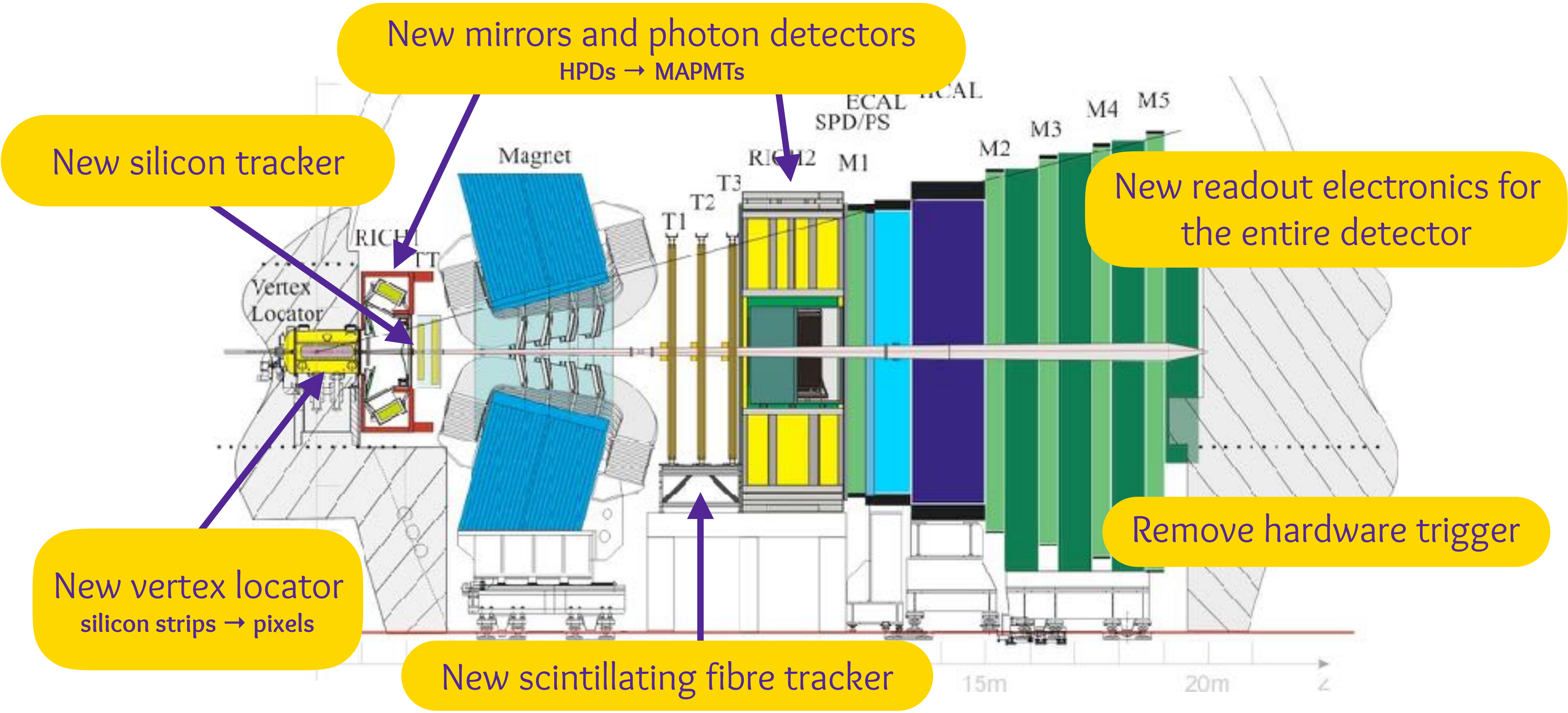
WLCG workshop

Lancaster, November 7th 2022

Overview

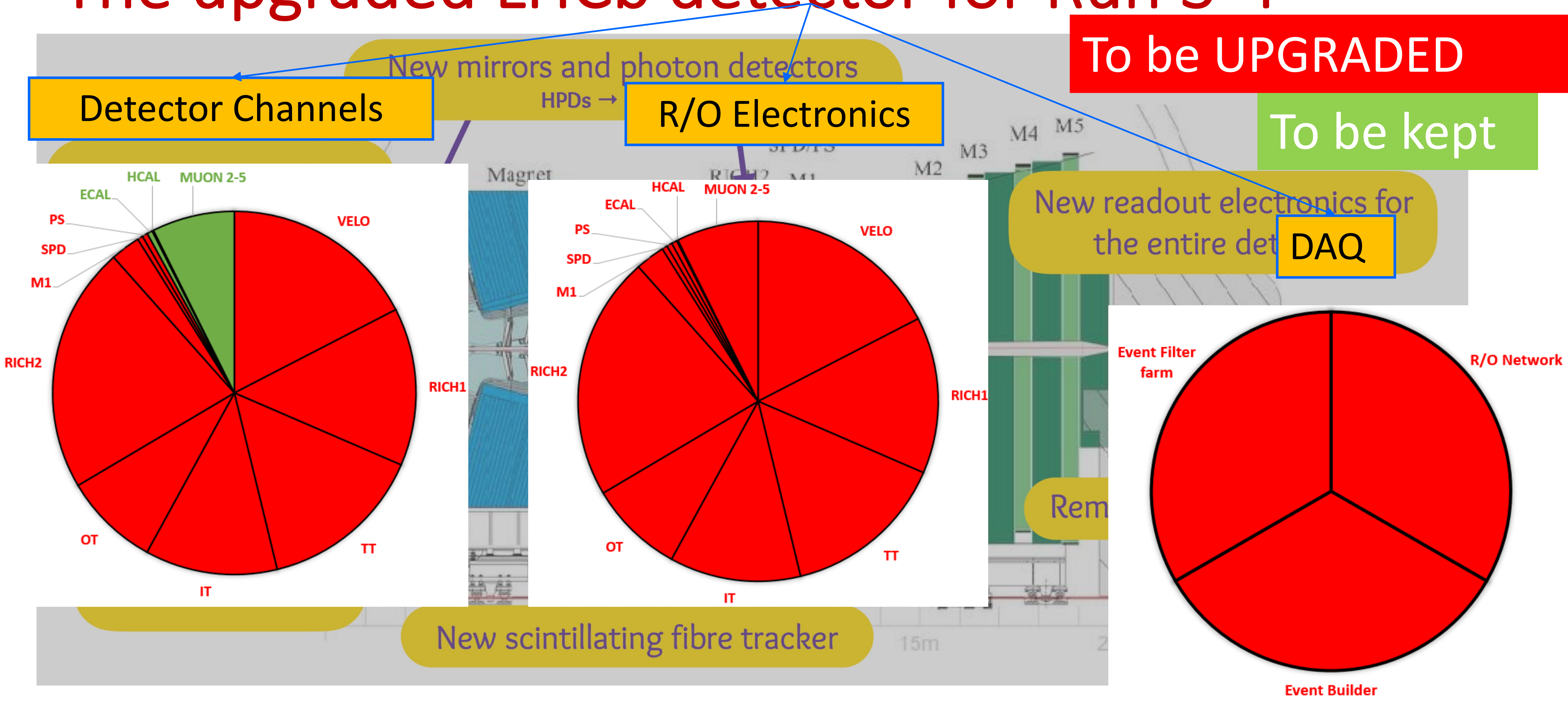
- Run3 + Run4 computing model
- Resource usage
- HPC status
- Analysis facilities

The upgraded LHCb detector for Run 3-4



CERN-LHCC-2012-007

The upgraded LHCb detector for Run 3-4



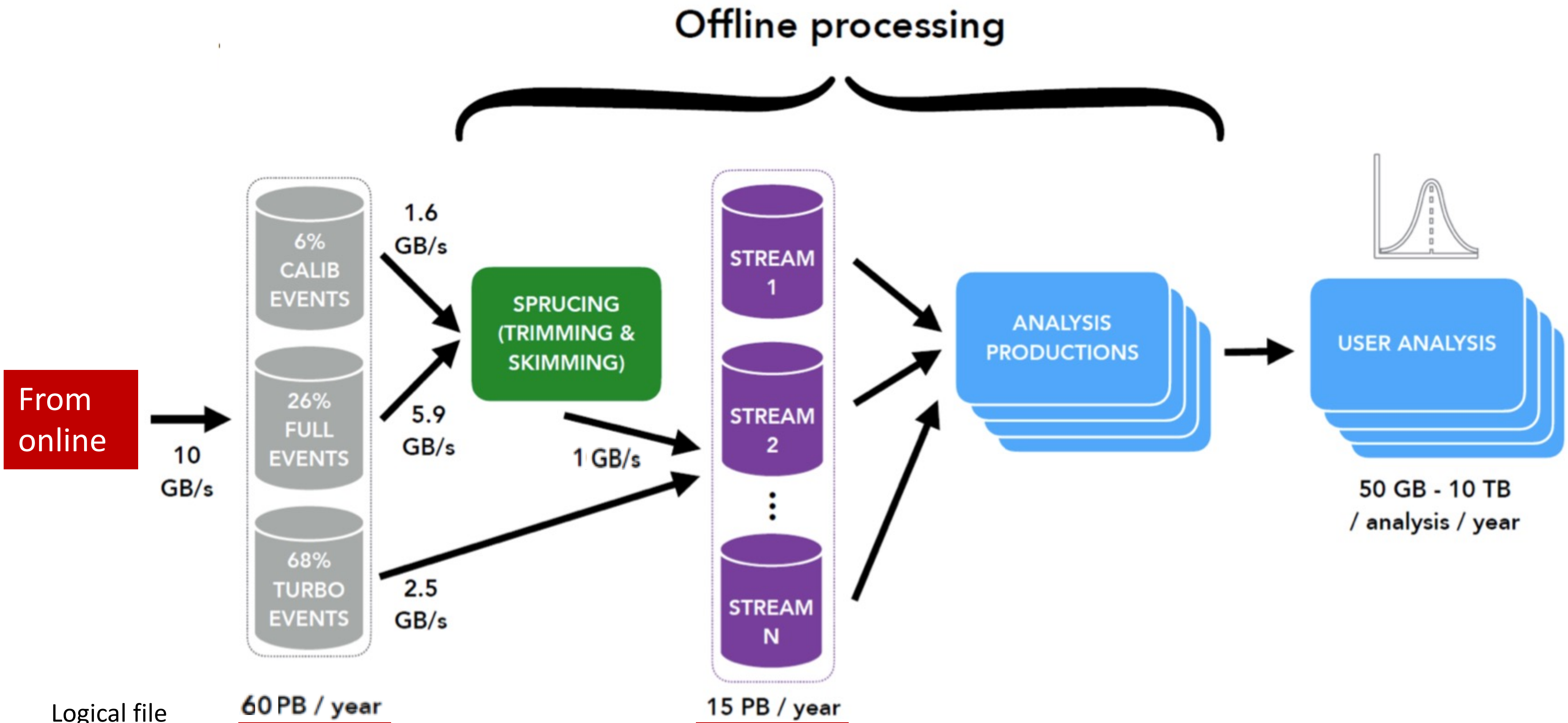
A big challenge in data handling

- Major expansion of LHCb physics programme through:
  - 5-fold increase in instantaneous luminosity
    - 4×10^{32} to 2×10^{33} $cm^{-2}s^{-1}$
  - Full software trigger at 30MHz inelastic collision rate
    - Factor 2 increase in trigger selection efficiency
- Order of magnitude increase in physics event rate to storage
- Pile-up increase
    - Factor 3 increase in average event size
- 30x increase in throughput from the upgraded detector
  - Without corresponding jump in offline computing resources
- Full software trigger and selective persistency to mitigate throughput from online to offline
  - Nevertheless, from ~0.65GB/s (Run2) to 10GB/s (Run3-4)

Data streams and dataflow

- Data from the LHCb detector organised in 3 streams; in all cases; events are reconstructed online at the HLT farm
  - **FULL:** «classic» stream, where information from the entire event is persisted in DST format and input to offline «sprucing» i.e. «slimming and skimming» for subsequent physics analysis
  - **TURCAL:** calibration stream, with both reconstruction output and (some) RAW banks. To be «spruced» offline and used for performance studies.
  - **TURBO:** introduced in Run2, implements selective persistency thus saving selected info that can range from a couple of tracks to the entire event contents. Data ready to be analysed, no further processing needed
- Sprucing is performed at T0 and T1s, concurrently with data taking and during winter shudown («re-sprucing»)
  - T0 for LHCb is equivalent to any other T1 from processing PoV
- Further processing (e.g. tupling) done in centralised Analysis Productions
- Additional analysis steps done on user / local resources

Data streams and dataflow



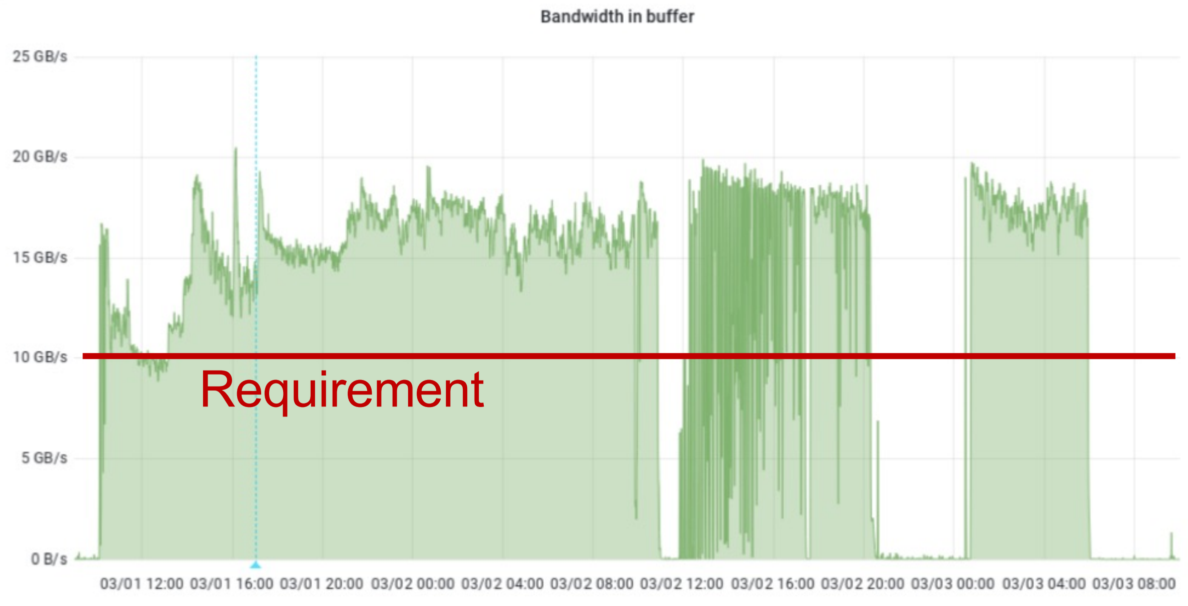
Logical file sizes:

to tape

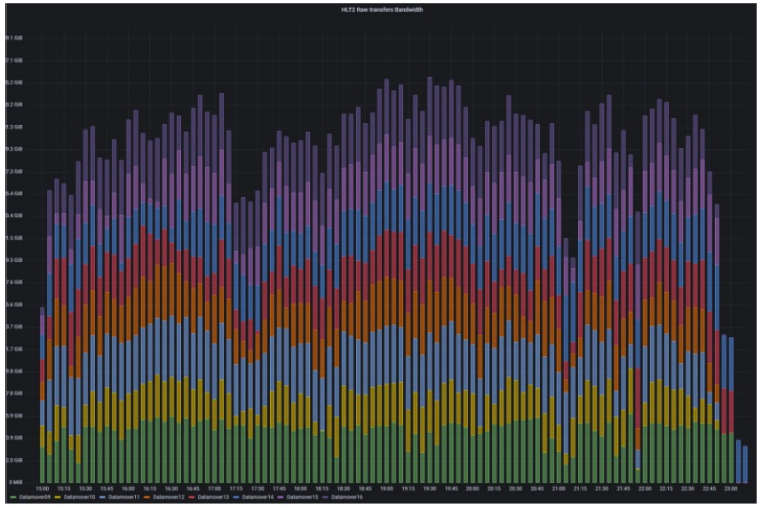
to disk

Data challenges

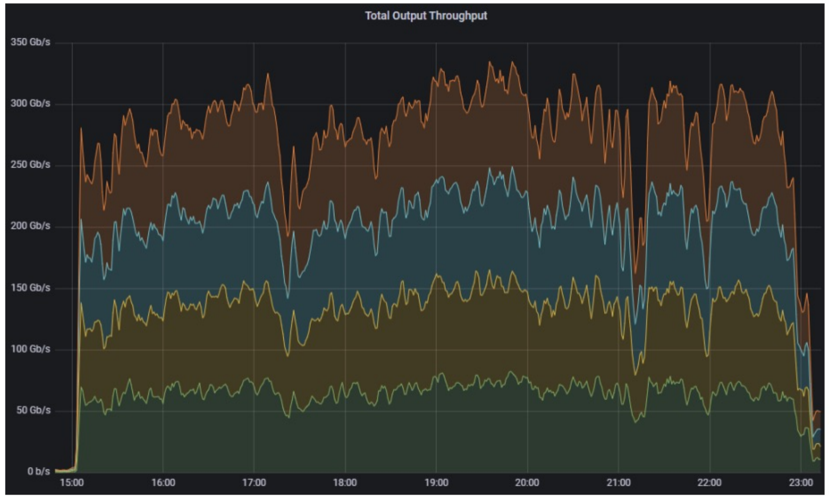
- Large-scale tests for data export from P8 to EOS/CTA performed
  - February 2022: Throughput exceeding target (16 GB/s > 10GB/s)



- Deployment of 4*100Gb/s links from point 8 in summer 2022, giving ~4x over requirement

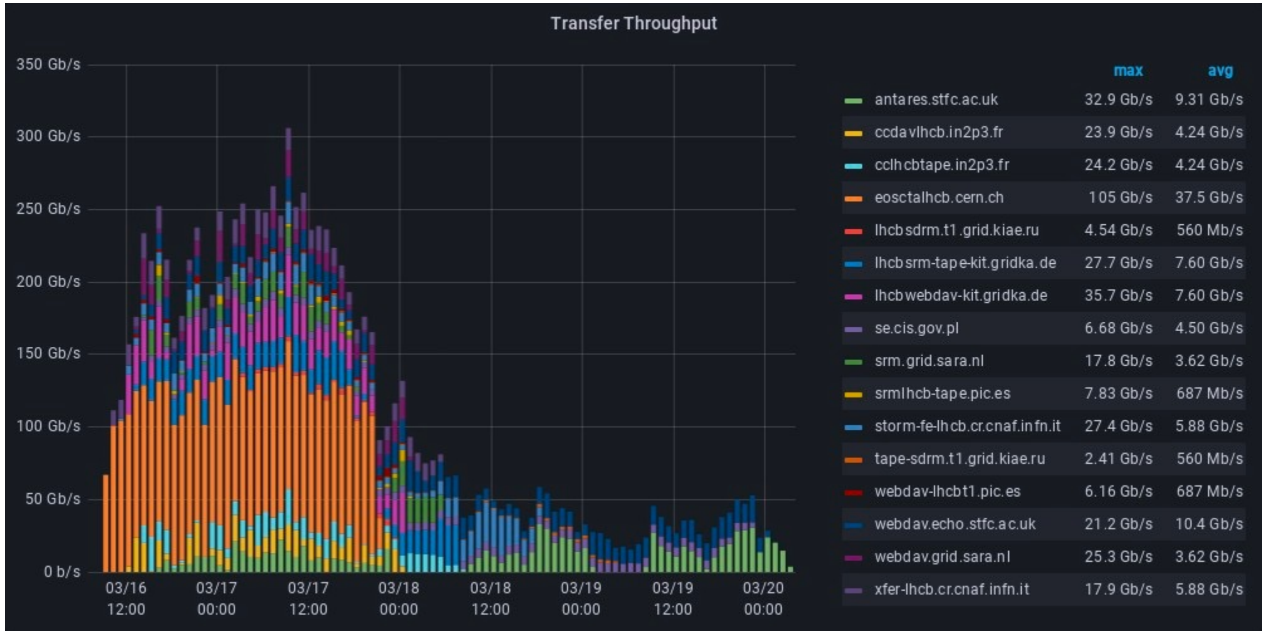


Data transfers from the 8 movers



Corresponding data links traffic

Tape challenges



Write tests:
CERN disk → T1 disk → T1 tape

Site	expected Speed (GB/s)
CERN	11
CNAF	1.72
GRIDKA	2.23
IN2P3	1.25
NCBJ	1.32
PIC	0.2
RAL	2.96
RRCKI	0.25
SARA	1.07

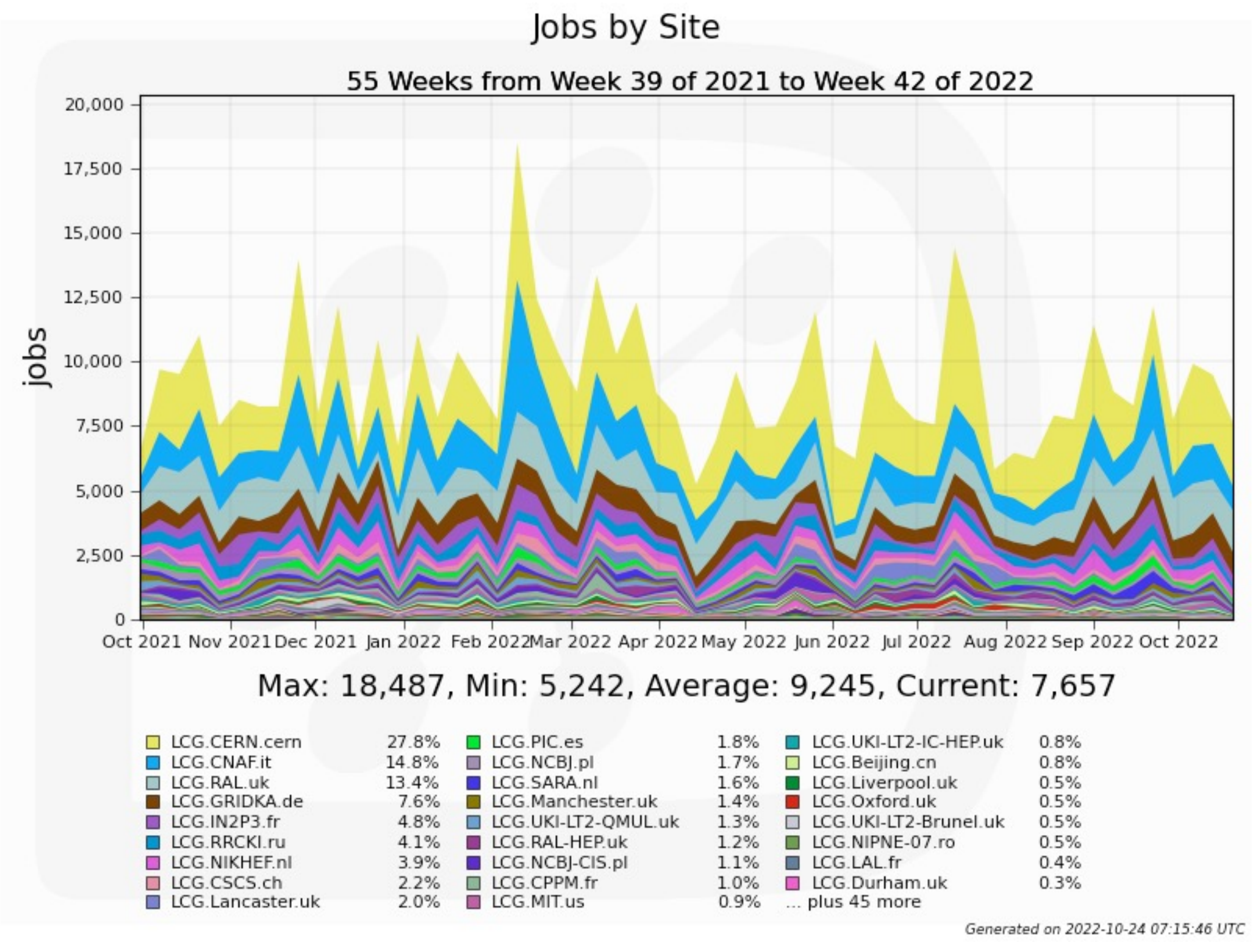
Read tests
T1 tape → T1 disk

Site	expected Speed (GB/s)
CERN	1.90
CNAF	1.35
GRIDKA	1.36
IN2P3	0.98
NCBJ	0.91
PIC	0.17
RAL	1.93
RRCKI	0.21
SARA	0.74

- Both write and read tests OK
  - Requirements exceeded in most sites
- A couple of sites needed following up
  - Tests to be repeated at RAL and NCBJ
- No ”stress test” with real data so far
  - 2022 is a commissioning year for LHCb new detectors and software trigger

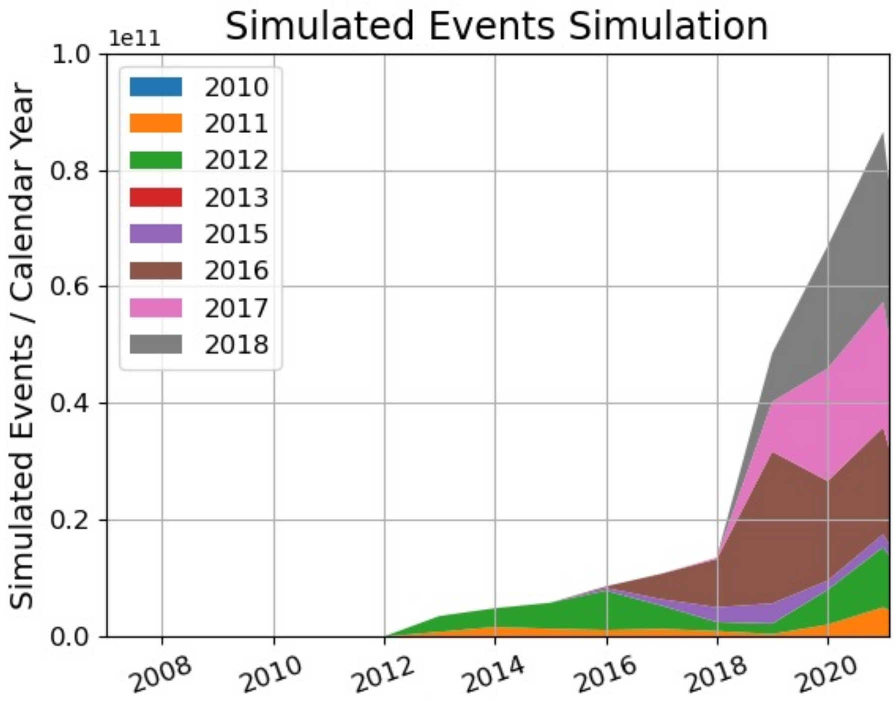
Data distribution for physics analysis

- Data distribution model quite simple
- User jobs run where data is
  - Mostly at Tier0 and Tier1s
- Number of sites with data relatively small
  - 1 T0, 7 T1s, 14 T2-Ds
- Well-balanced CPU and disk resources
  - Grid user jobs are given the highest priority anyway
- No need for caches, pre-placement, etc
- Little impact on WAN other than dataset replication (2 copies)



Monte Carlo simulation

- No input data required. Starting from random seed!
  - Pile-up significantly smaller than GPDs
- Simulation dominates (95%) CPU work, runs everywhere
  - Improvements in simulation and introduction of fast simulation significantly decrease CPU work per event
- Simulation reconstruction is heavily filtered
  - E.g. 80B events simulated in 2021 but only 11B stored, corresponding to 2PB logical volume added
- Simulation is continuously running, with a given data-taking year being simulated for the following N years

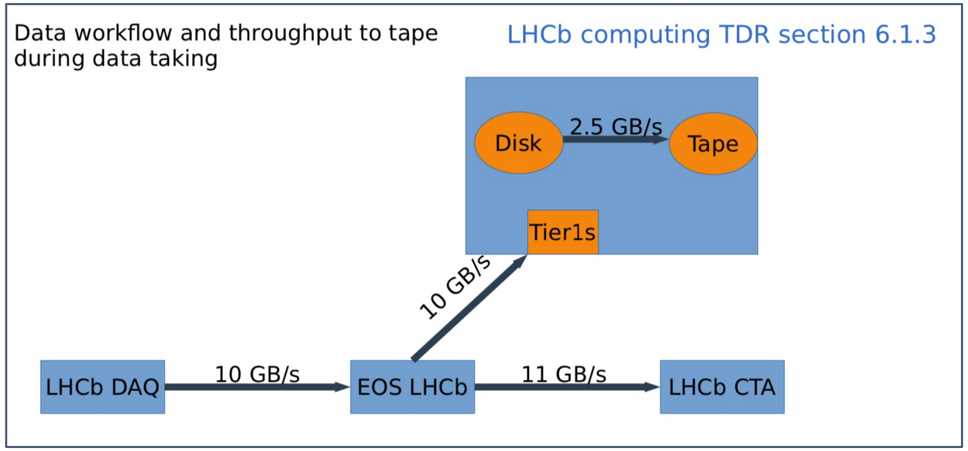


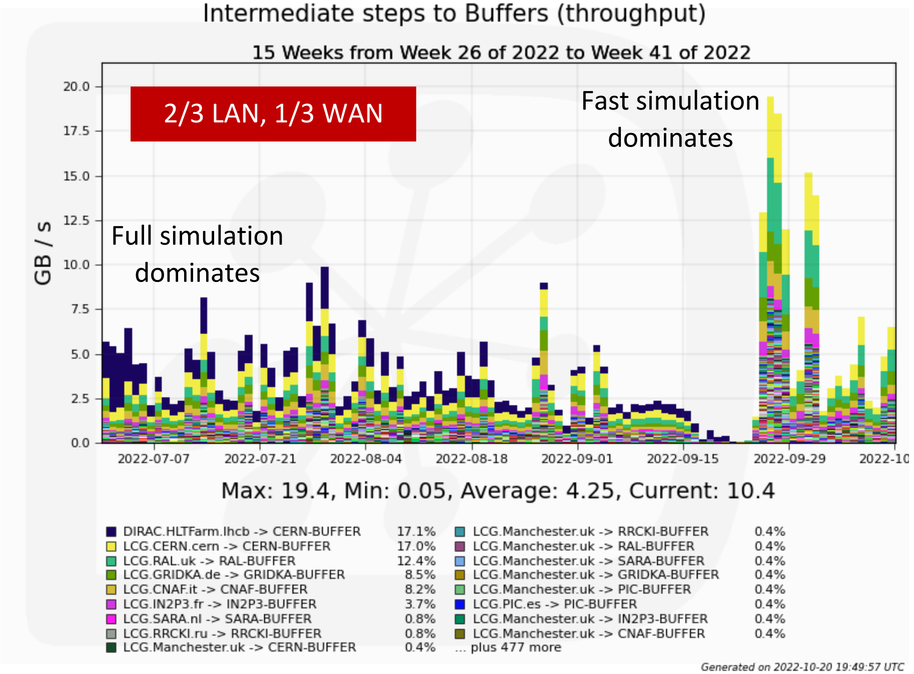
Year	Simulated events (10^9)	Stored events (10^9)	Ratio	CPU work kHS06.y	CPU per event kHS06.s	LFS TB
2017	10.3	4.2	40.3%	817	2.50	640
2018	12.0	3.0	25.3%	1009	2.65	550
2019	45.0	6.9	15.2%	1290	0.90	1110
2020	67.0	16.8	31.7%	1357	0.81	2010
2021	80.0	11.1	13.9%	1815	0.72	2030

More details given at [LHCONE/LHCOPN workshop, Oct 25 2022]

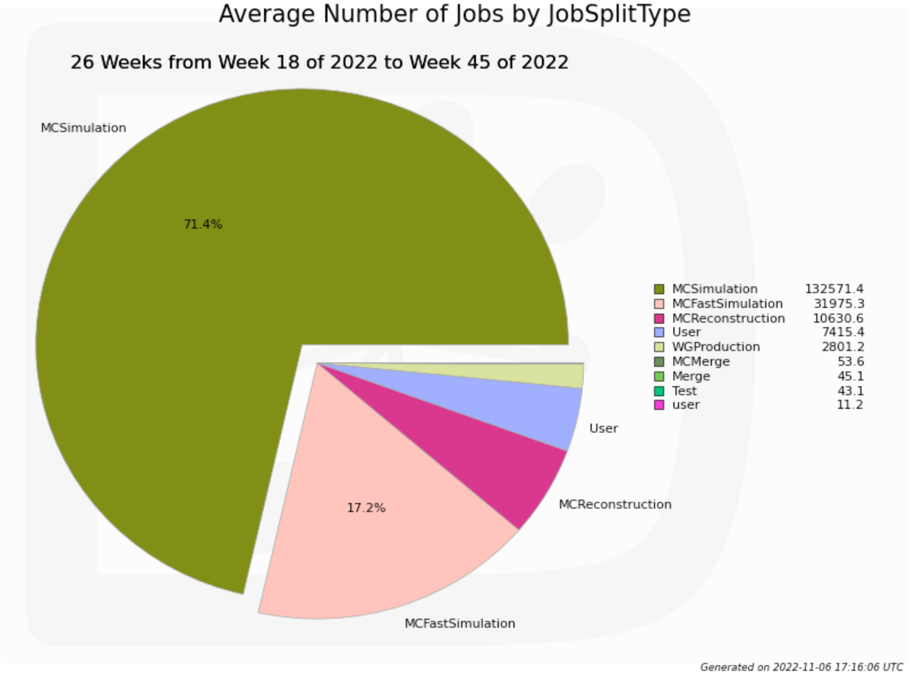
Network

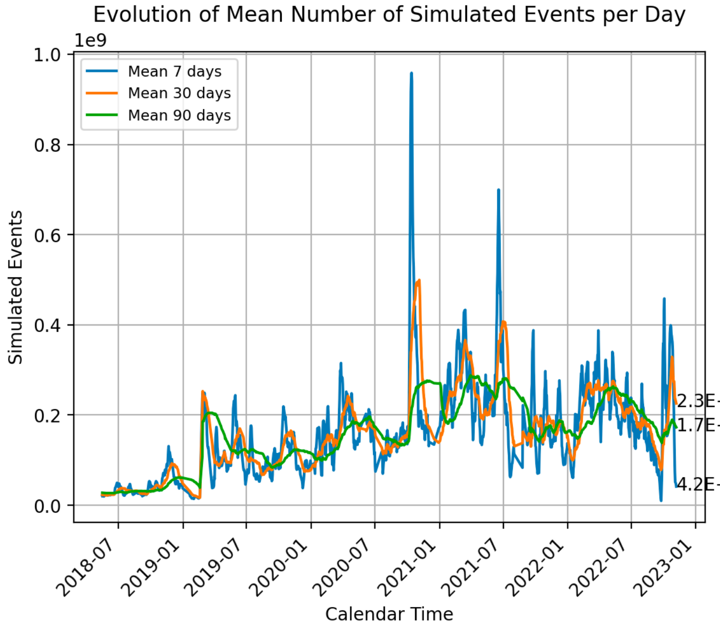
- LHCb will increase network usage in Run3 and beyond
  - Dominated (one order of magnitude!) by real data coming from the detector
  - A factor two expected for simulation
    - Fast simulation requires more BW
- Fast and reliable network is at the basis of our successful computing operations and ultimately of the physics productivity of LHCb
- In general:
  - we favour LAN over WAN
  - when running on a Tier2, we favour the national network before going abroad.



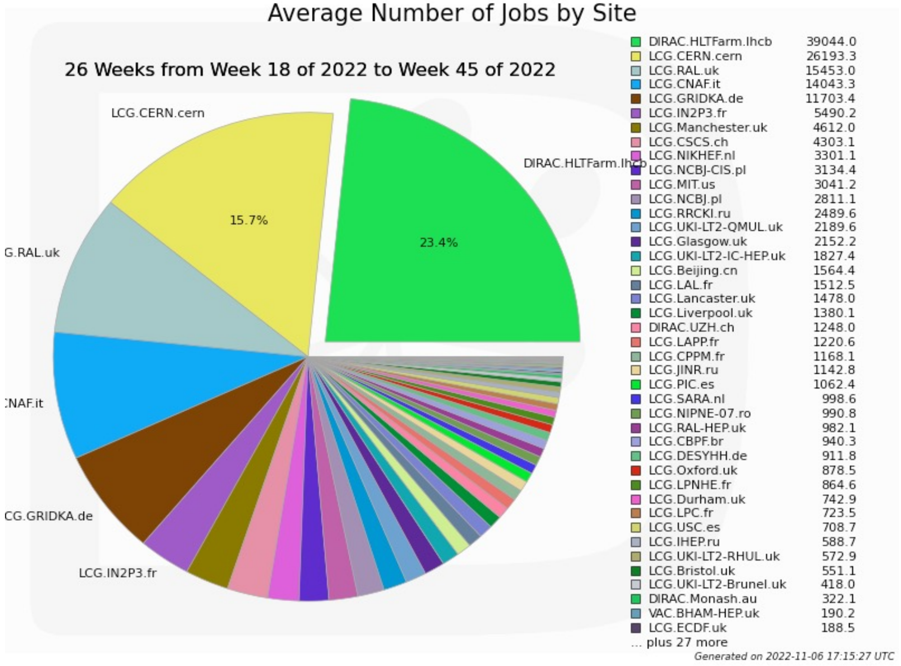


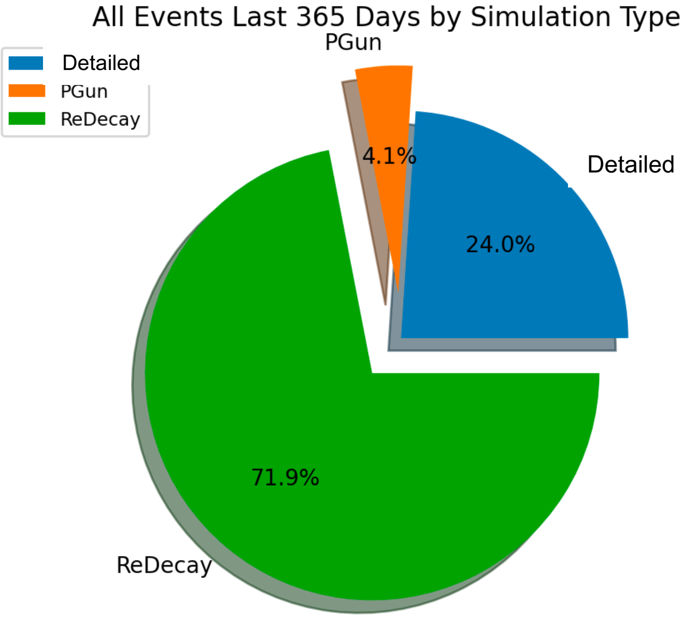
Distributed computing operations





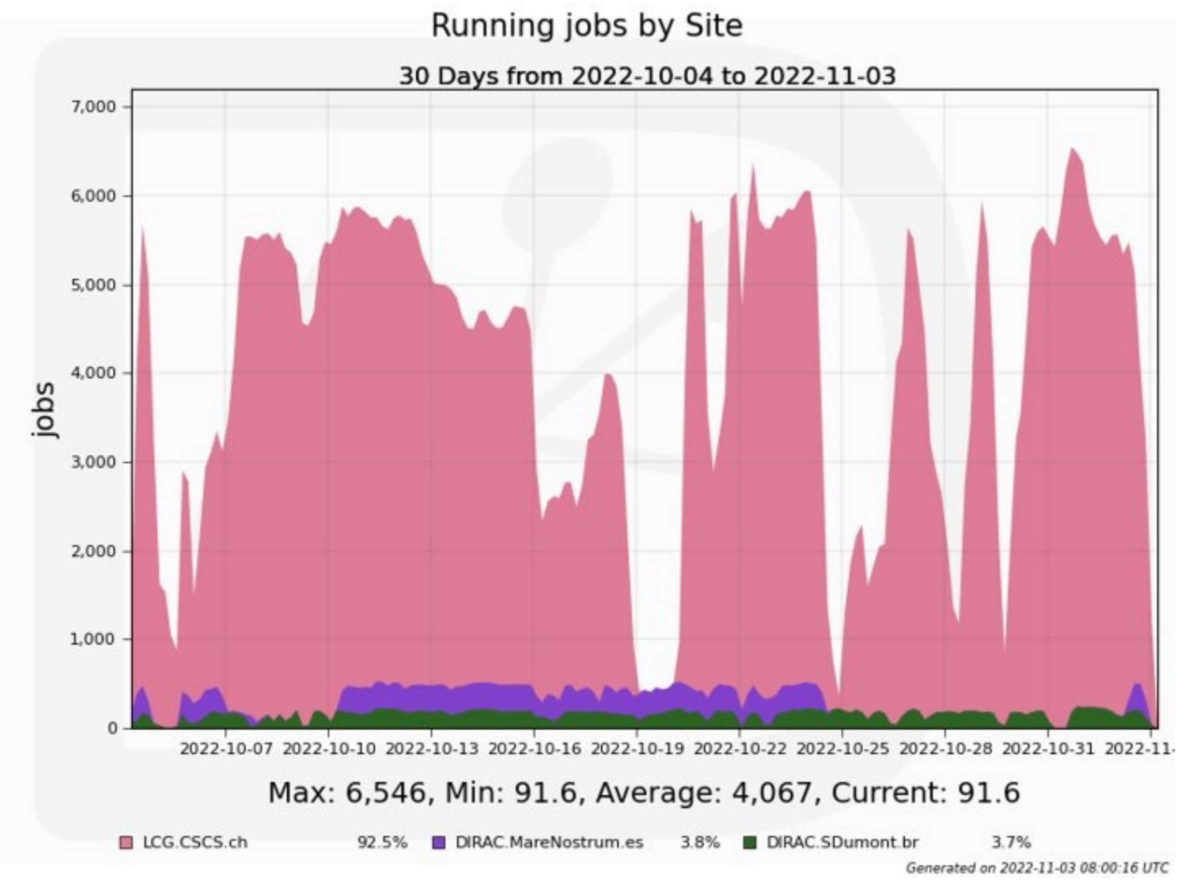
- Computing work dominated by MC production (97%)
- Simulating about 170 million events per day in the last three months
- Only 1/4 of events produced with detailed simulation in the last 365 days
- Strong contribution of HLT farm





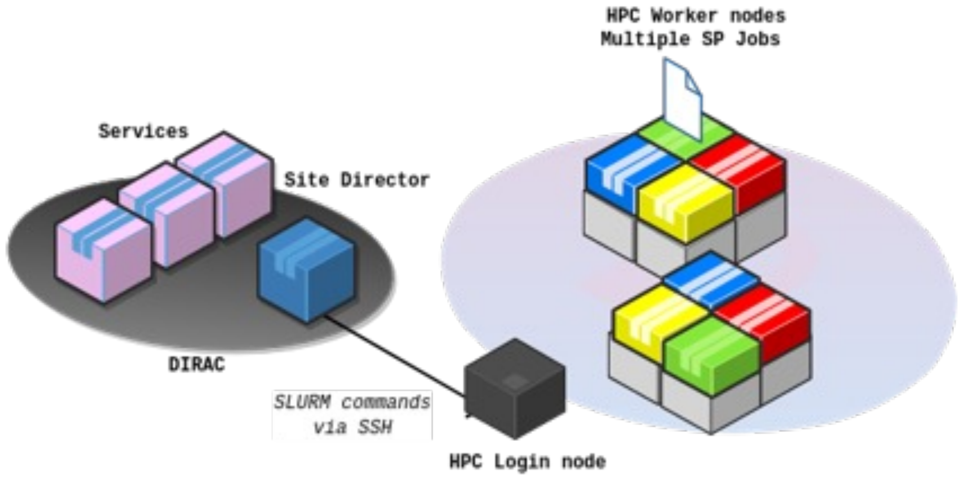
Progress on HPC

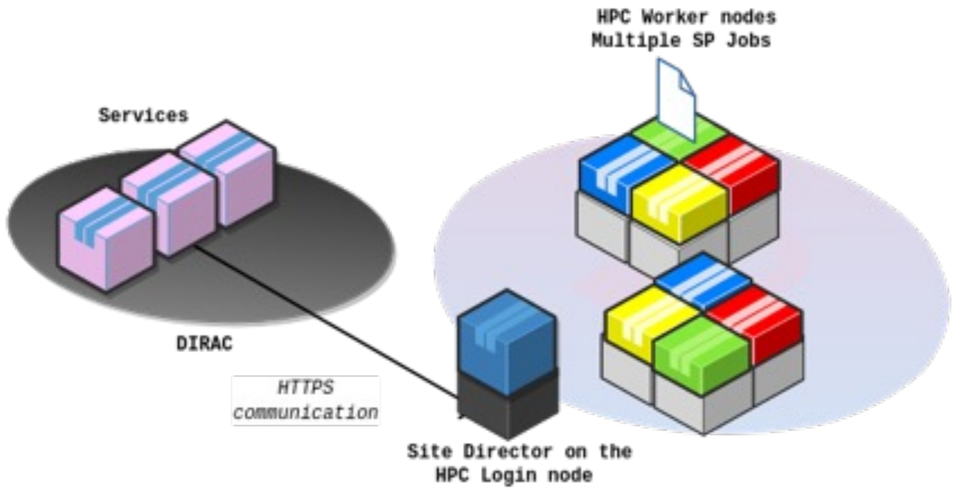
- Mostly used to process Monte Carlo simulation tasks (Gauss)
- Barcelona Supercomputing Center (MareNostrum) in production
  - Currently limited to a few hundred jobs
  - Request granted for 2022Q4
- SDumont.br is saturated by its institutional stakeholders.
- Ongoing efforts on procuring resources and preparing LHCb SW stack to use them
  - Thanks to CERN/IT!



Latest DIRAC developments to support HPCs

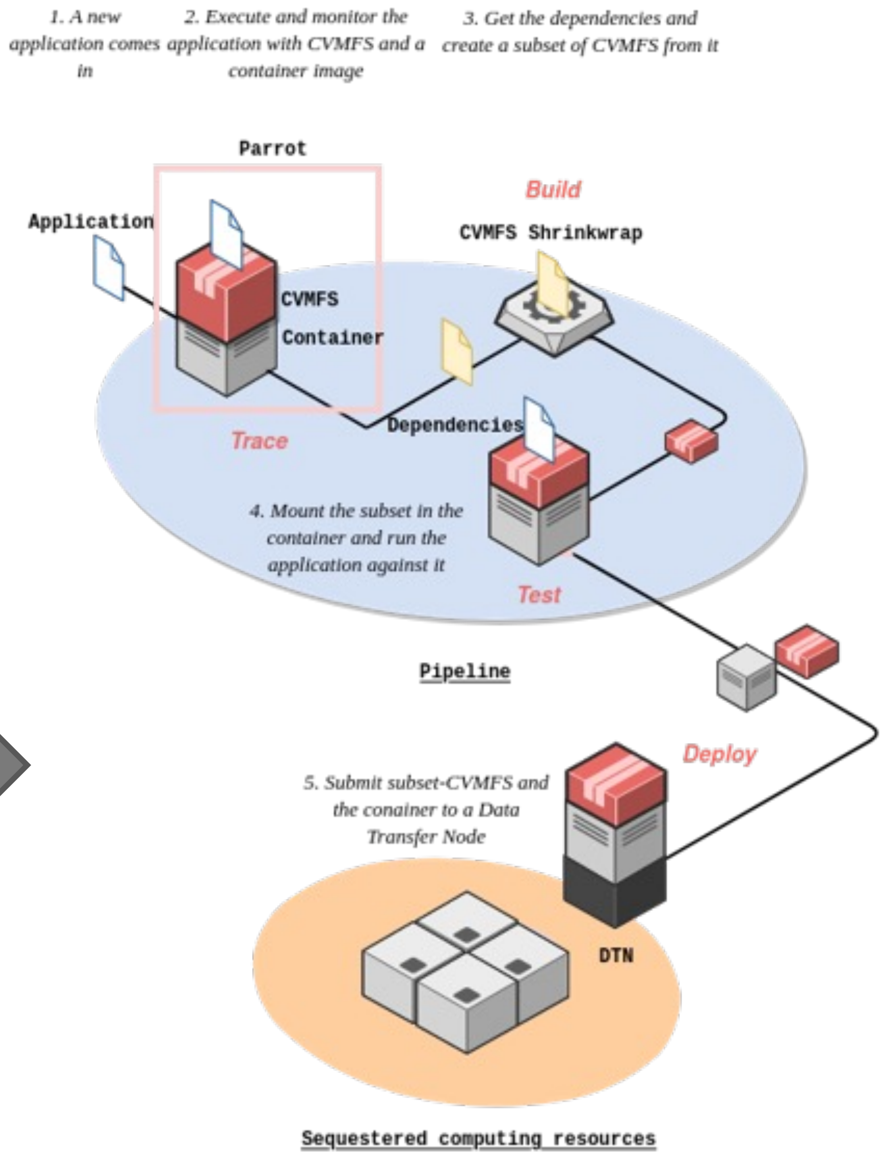
- HPCs with external connectivity:
  - Support AREX (ARC) services
    - Keep leveraging CSCS computing resources
  - Support multi-node allocations
    - Useful when a limited number of large allocations is available
    - Only work with resources orchestrated via SLURM
  - Test a Site Director installation on a HPC edge node
    - Useful when no CE & SSH connection is unstable
    - Experimented but not applied in production





Latest DIRAC developments to support HPCs

- HPCs with no external connectivity:
  - Implement an agent to push jobs via an ARC instance
    - Works similarly to a Pilot-Job but outside the HPC
    - Not scalable because of the current structure of the jobs
  - Implement a generic CI pipeline to extract and deploy a subset of CVMFS in a container to the HPC
    - Used for months in Mare Nostrum, no major issues so far

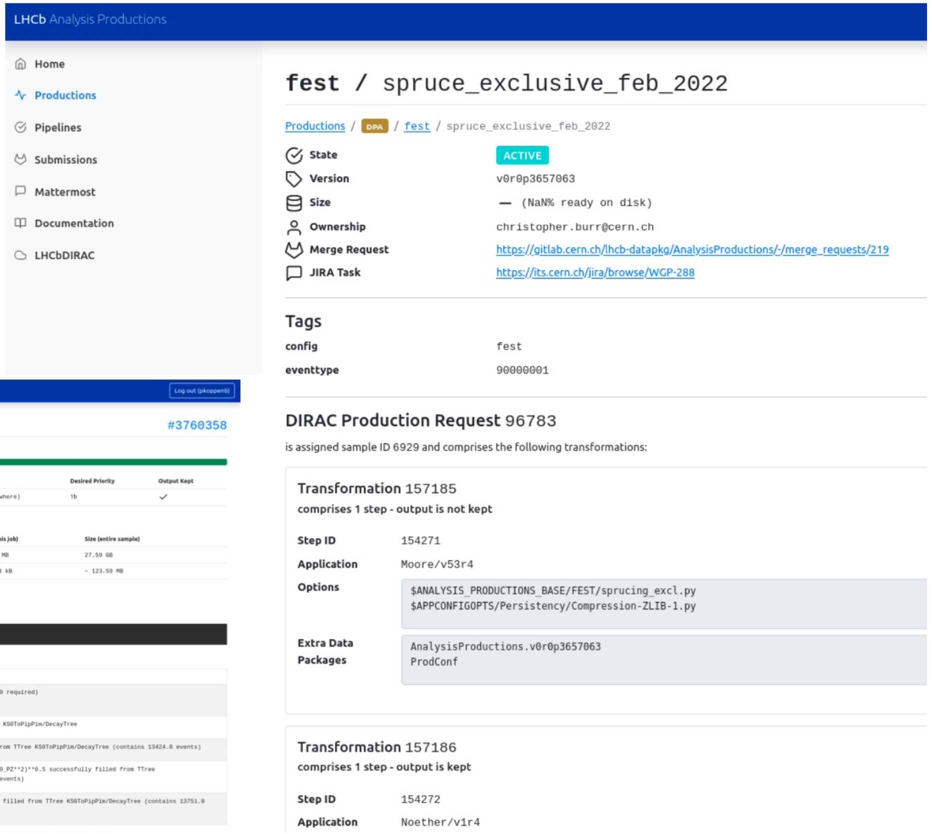


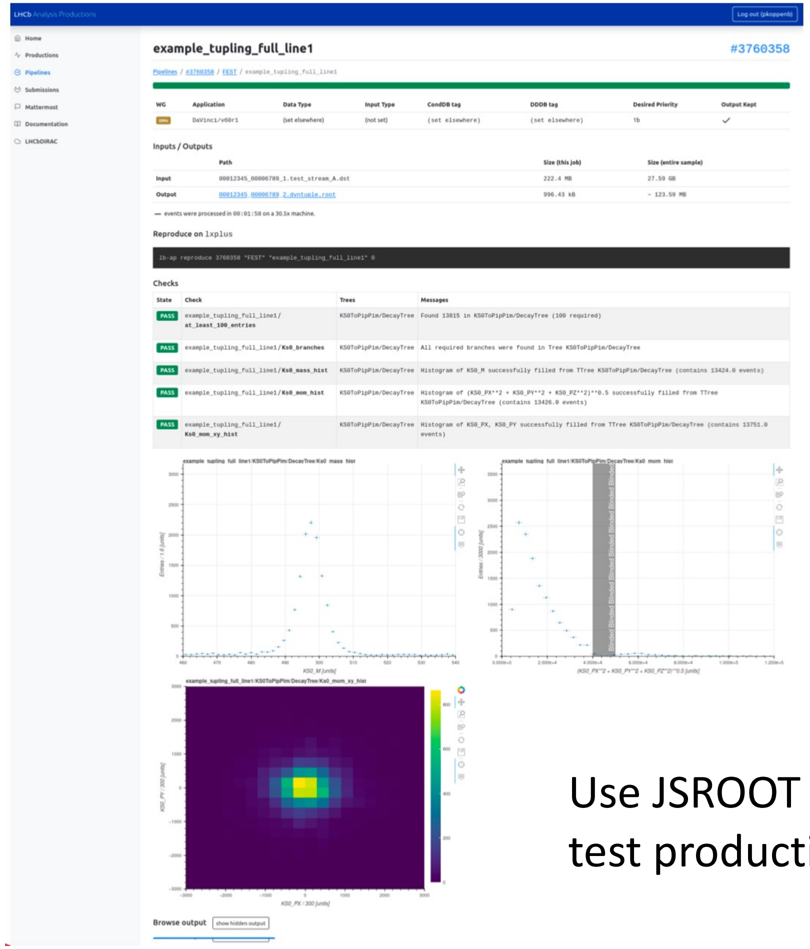
DIRAC news

- Rolling out major DIRAC release (v8.0)
  - First DIRAC release dropping py2 support → fully py3.9
  - Enhanced Monitoring
  - Adds as "technology preview" OAuth/OIDC token-based authN/Z
- Support for token-based authN/Z is being tested for specific use cases
  - Interaction (sending DIRAC pilots) to Computing Elements (HTCondorCE specifically) is in advanced testing
  - Interaction with IAM also being tested
- In development:
  - moving all DIRAC services to HTTP, and later decommissioning of in-house solutions
  - support for Python 3.10 (and 3.11)
  - Better monitoring, especially for pilots

Analysis productions

- Support user processing of data and simulation using the DIRAC transformation system
  - User do not need to monitor GRID jobs
  - Job details / configuration / logs automatically preserved in LHCb bookkeeping / EOS
  - Automated error interpretation / advice
  - Intuitive web interface for requesting / testing / browsing outputs





Use JSROOT for allowing the output of test productions to be browsed.

GDB, Dec 2021
HSF AF kick-off meeting, March 2022

Analysis facilities

- Innovative analysis techniques are being explored e.g.
  - Usage of GPU resources in analyses
    - DNN for jet tagging, Zfit and likelihood inference, DNN for ultra-fast simulation, amplitude analyses, etc.
  - Analyses usually done on local facilities. Resource availability drives implementation choices
  - Quantum Computing applications to HEP, e.g. in jet tagging
- Given the progress in HSF, the Snowmass papers and the proposed prototypes, LHCb is starting to
  - Collect use cases, available and used resources, code developed, etc.
  - Identify the user needs
  - Proceed with a structured activity that may lead to
    - Different AF configurations, depending on site (e.g. HLT1 GPUs at CERN, availabilities in different countries…)
    - Definition and identification of mandatory LHCb-specific requests

Summary

- Run3 + Run4 computing model
  - 30x larger data volume from detector mitigated by aggressive triggering strategy, filtering, selective persistency
  - Network utilisation one order of magnitude larger than Run2
    - Still small wrt other LHC VOs
- Resource usage
  - CPU dominated by simulation production
  - Fast simulation significantly mitigates requirements
- HPC status
  - Usage still limited
  - Gradually overcoming site limitations
  - Proactively seeking for more resources and building on non-x86 architectures
- Analysis facilities
  - Bottom-up approach, collecting use cases towards a more structured activity

backup

Run3 Computing model in a nutshell

- LHCb Upgrade computing model accommodates a trigger output BW of 10 GB/s
  - Massive usage of novel event selection (Turbo) and event size reduction (selective persistence) techniques
  - Save the full bandwidth on cheap storage
  - Reduce by more than a factor of 2 disk requirements using the above techniques
- CPU needs dominated by MC production
  - Massive use of faster simulation techniques
- In summary:
  - Substantial reduction of expensive resources
  - Maintain the full breadth of the physics programme
  - Flexible: incorporate future technology advancements

LHCb Run3 Computing Model assumptions							
L (cm^{-2} s^{-1})	2×10^{33}						
Pileup	6						
Running time (s)	5×10^6 (2.5×10^6 in 2021)						
Integrated luminosity	10 fb^{-1} (5 fb^{-1} in 2021)						
Trigger rate fraction (%)	26 / 68 / 6 Full/Turbo/TurCal						
Logical bandwidth to tape (GB/s)	10 (5.9 / 2.5 / 1.6 Full/Turbo/TurCal)						
Logical bandwidth to disk (GB/s)	3.5 (0.8 / 2.5 / 0.2 Full/Turbo/TurCal)						
Ratio Turbo/FULL event size	16.7%						
Ratio full/fast/param. MC	40:40:20						
HS06.s per event for full/fast/param. MC [a]	1200 / 400 / 20						
Number or MC events[b]	2.3×10^9 / 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)						
Resource requirements							
WLCG Year	Disk (PB)		Tape (PB)		CPU (kHS06)		
2021	66	1.1	142	1.5	863	1.4	
2022	111	1.7	243	1.7	1579	1.8	
2023	159	1.4	345	1.4	2753	1.7	
2024	165	1.0	348	1.0	3467	1.3	
2025	171	1.0	351	1.0	3267	0.9	

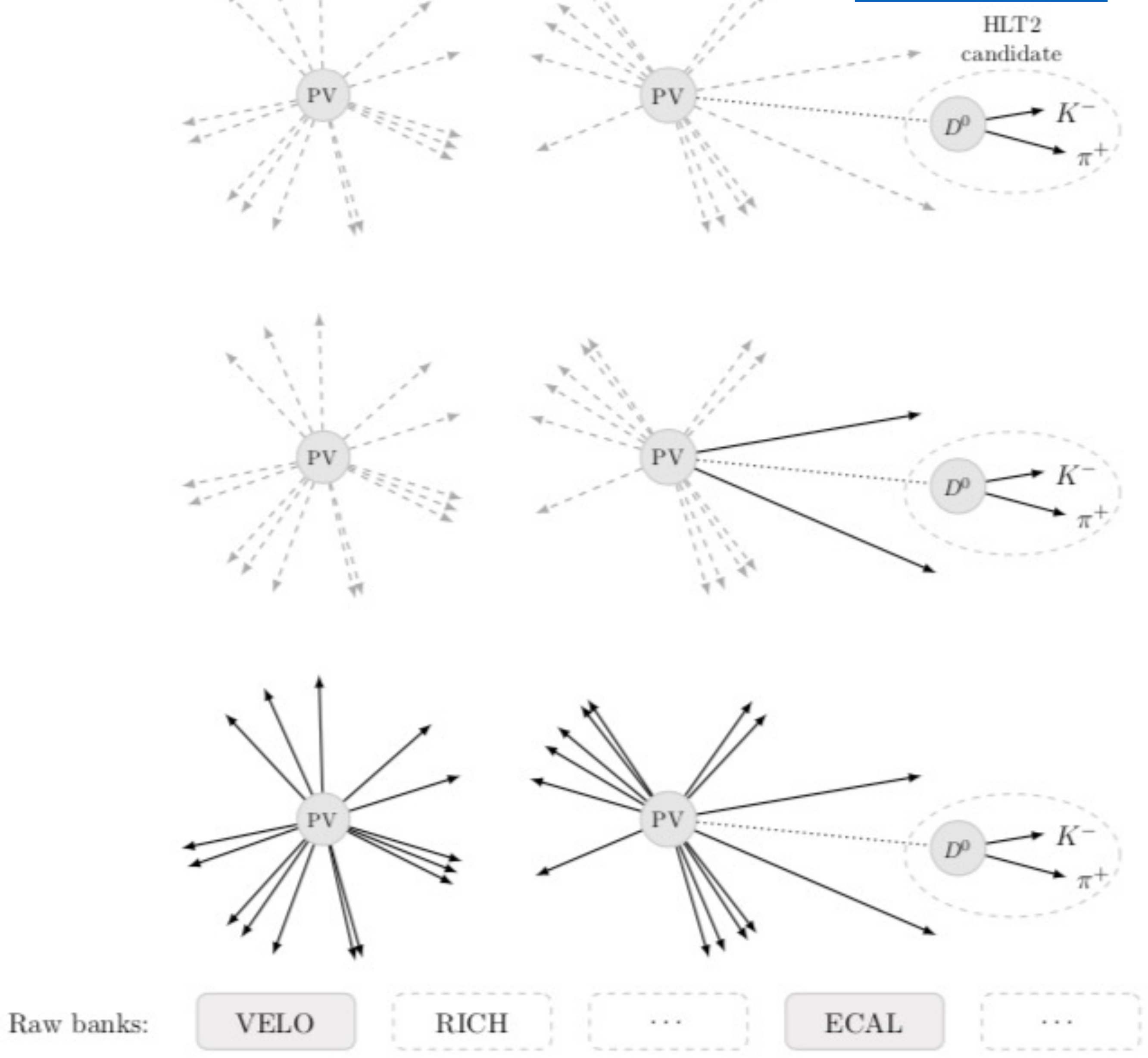
[a] corresponding to 120, 40, 2s on a 10HS06 computing core

[b] simulation of year N starts in year N+1

LHCb-TDR-018

Data persistency

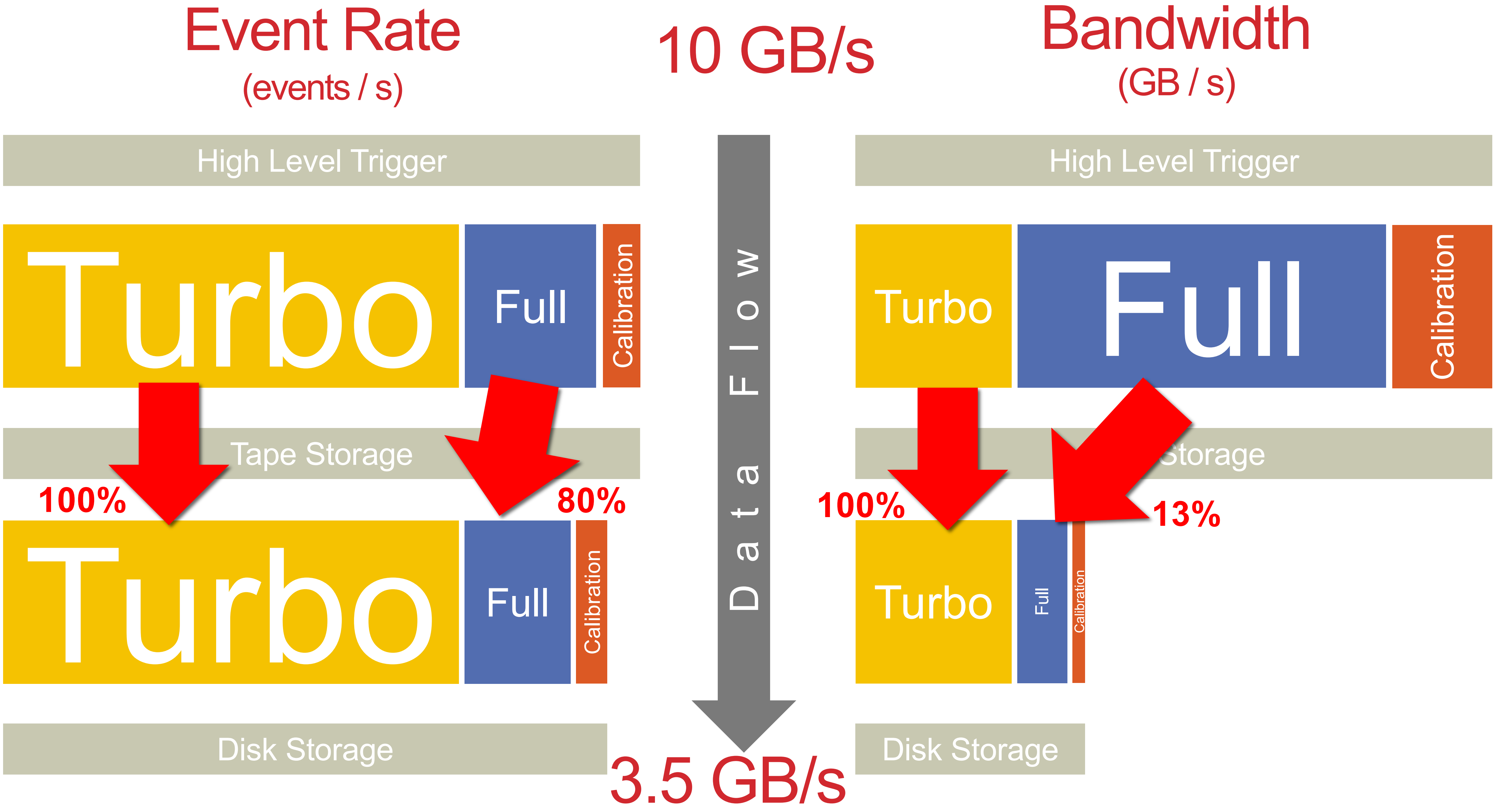
- Different levels of persistency:
  - FULL and TURCAL: the full event is persisted
  - TURBO: selective persistency, ranging from candidate firing the trigger to the entire event, optionally including some RAW subdetector data banks



HLT output bandwidth

- Due to selective persistency, emphasis has shifted from trigger rate (Hz) to bandwidth (bytes/s)
  - save less information and give more rate for a given bandwidth!
- About 60% of the physics selections on FULL in Run2 are migrating to TURBO in Run3
  - Massive migration, not trivial!
- Logical bandwidth to tape: 10 GB/s
- Logical bandwidth to disk reduced to 3.5GB/s by sprucing FULL and TURCAL more aggressively (select substantial fraction but slim by factor 6)
- This gives requirements of O(100PB) tape and O(50PB) disk per data taking year

		Logical Throughput to tape		Logical Throughput to disk	
stream	rate fraction	throughput (GB/s)	bandwidth fraction	throughput (GB/s)	bandwidth fraction
FULL	26%	5.9	59%	0.8	22%
Turbo	68%	2.5	25%	2.5	72%
TurCal	6%	1.6	16%	0.2	6%
total	100%	10.0	100%	3.5	100%


Event Rate
(events / s)
10 GB/s
Bandwidth
(GB / s)
High Level Trigger
Turbo
Full
Calibration
Tape Storage
100%
80%
Disk Storage
Data Flow
3.5 GB/s
High Level Trigger
Turbo
Full
Calibration
Storage
100%
13%
Turbo
Full
Calibration
Disk Storage

Data Processing Workflow per Data Taking Year