LHCb update

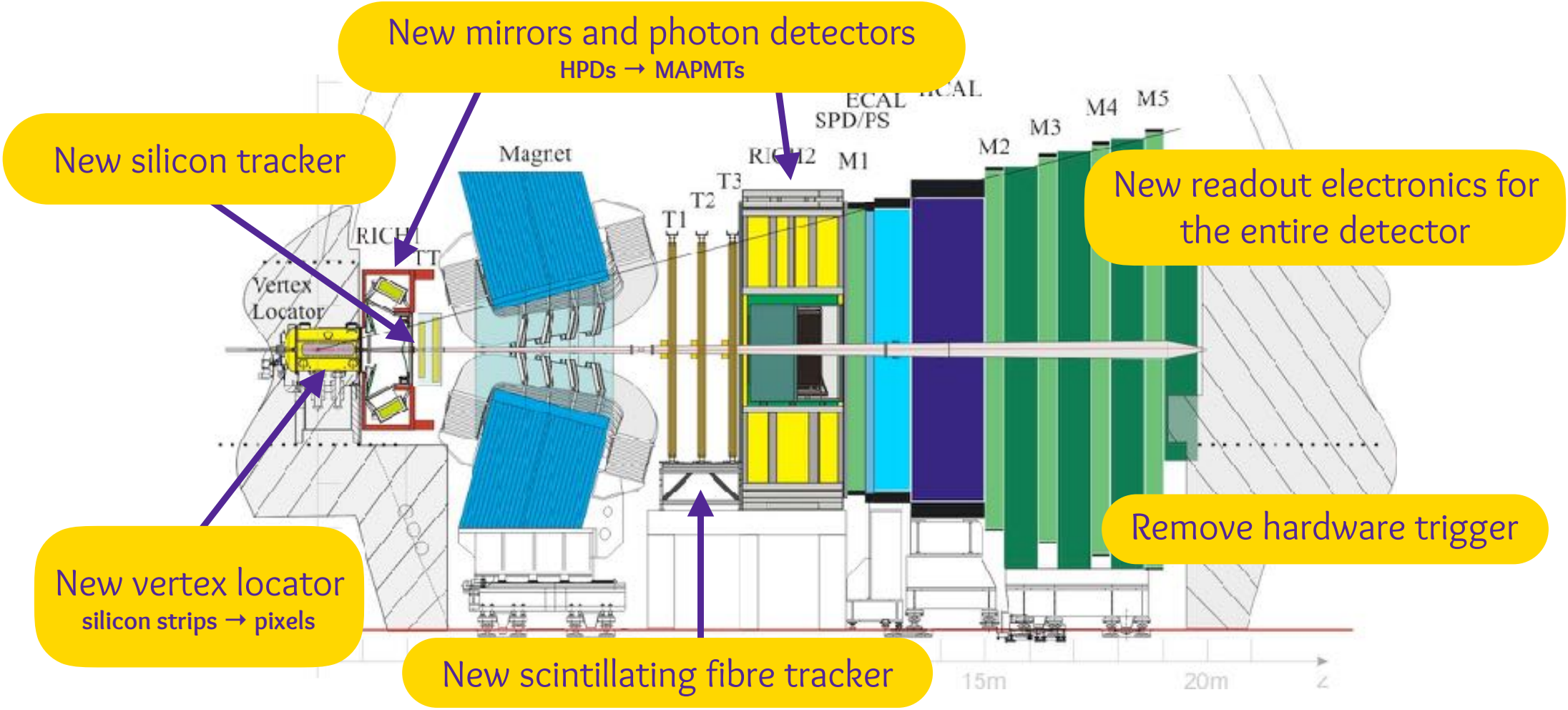
Concezio Bozzi, INFN Ferrara

Christophe Haen, CERN

LHCONE/LHCOPN workshop

CERN October 25th 2022

The upgraded LHCb detector for Run 3-4



The upgraded LHCb detector for Run 3-4

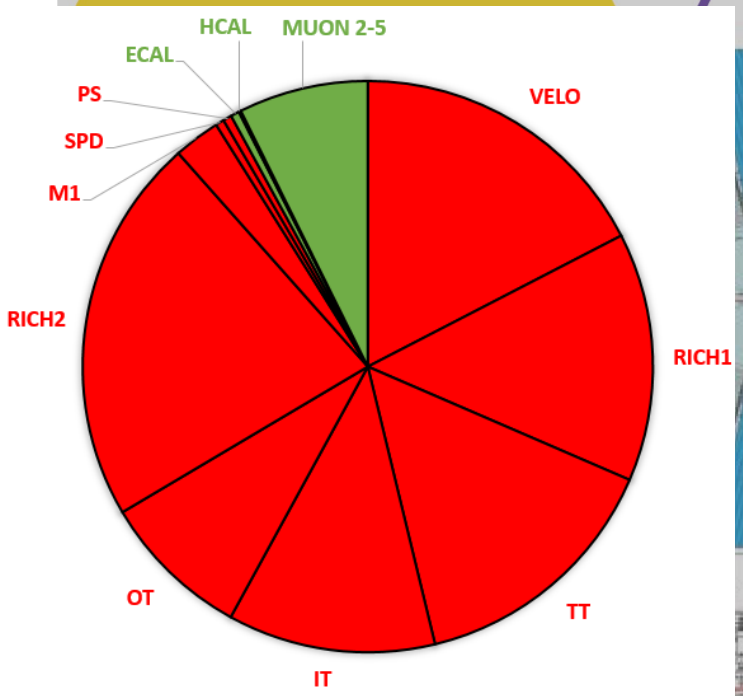
To be UPGRADED

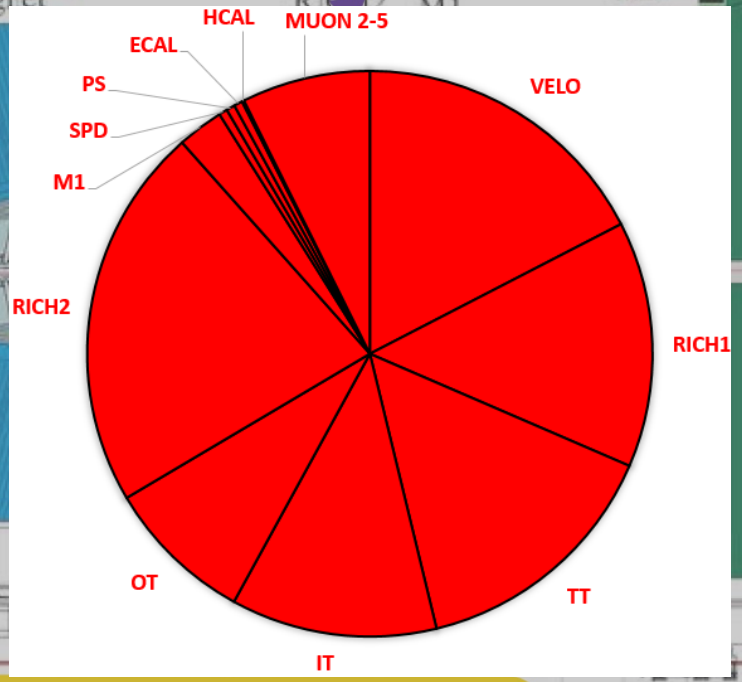
To be kept

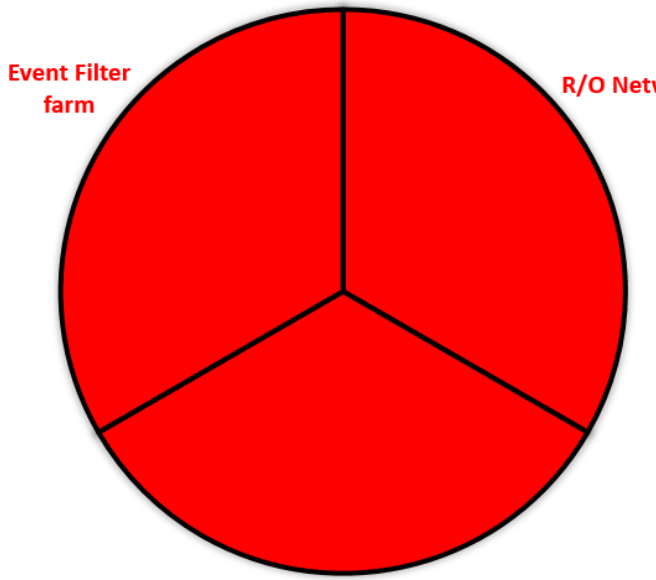
Detector Channels

R/O Electronics

DAQ







A big challenge in data handling

- Major expansion of LHCb physics programme through:
  - 5-fold increase in instantaneous luminosity
    - $4x10^{32}$ to $2x10^{33}$ $cm^{-2}s^{-1}$
  - 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 to mitigate throughput from online to offline
  - Nevertheless, from ~0.65GB/s (Run2) to 10GB/s (Run3-4)

Four activities of LHCb data management

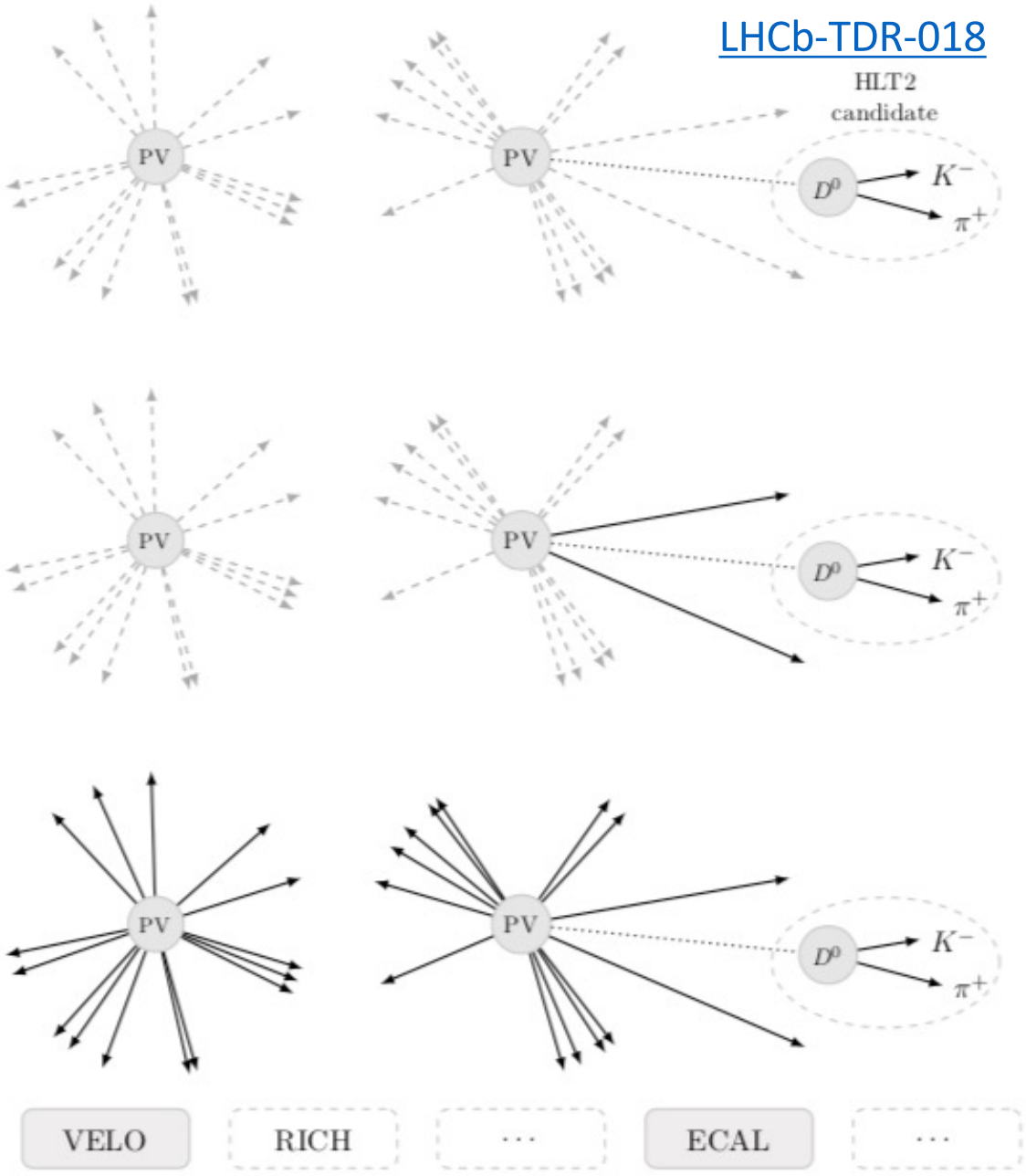
- Distribution (with FTS) of data originating from the LHCb experiment for custodial storage
- Usage of buffer disks for the intermediate processing steps of
  - Data: LAN only
  - Monte-Carlo: LAN and WAN
- Consolidation of processing outputs in fewer, larger files, aka "merging": mostly LAN
- Replication (with FTS) of data and Monte-Carlo samples for physics analysis

Data streams from the LHCb detector

- Data from the LHCb detector organised in 3 streams; in all cases, events are reconstructed on 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

Data persistency

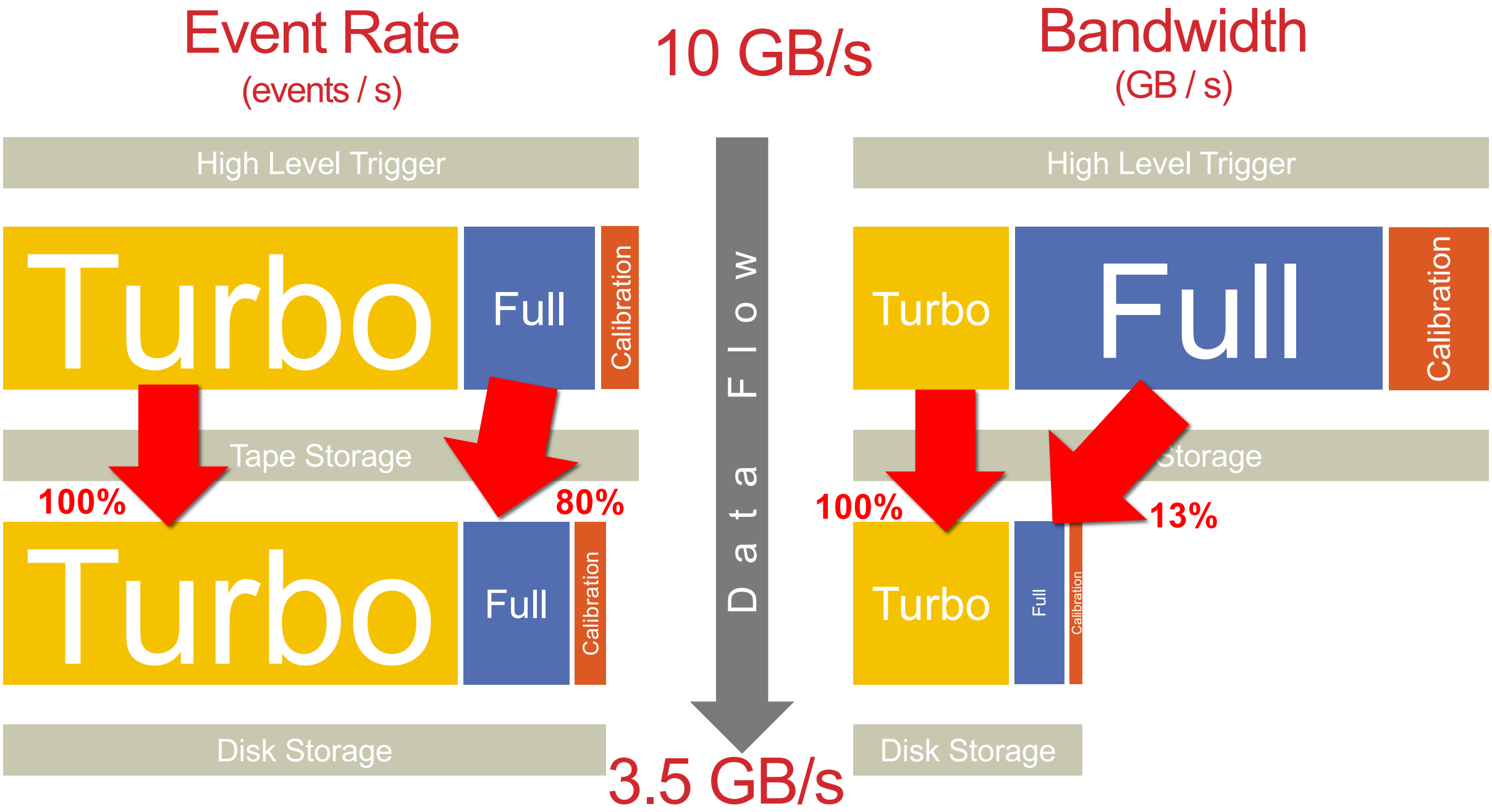
- Different levels of persistency:
  - FULL and TURCAL: the full event is persisted
  - TURBO: selective persistency, ranging from candidate firing the trigger to the entire event, optionally including some RAW subdetector data banks



HLT output bandwidth

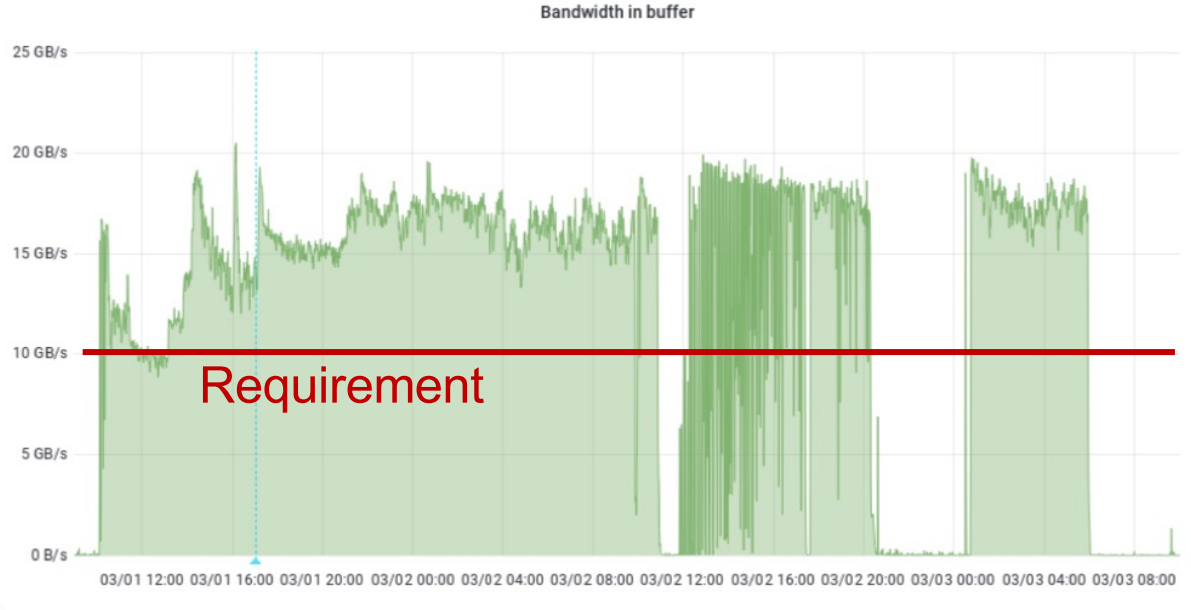
- Due to selective persistency, emphasis has shifted from trigger rate (Hz) to bandwidth (bytes/s)
  - save less information and give more rate for a given bandwidth!
- About 60% of the physics selections on FULL in Run2 are migrating to TURBO in Run3
  - Massive migration, not trivial!
- Logical bandwidth 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%

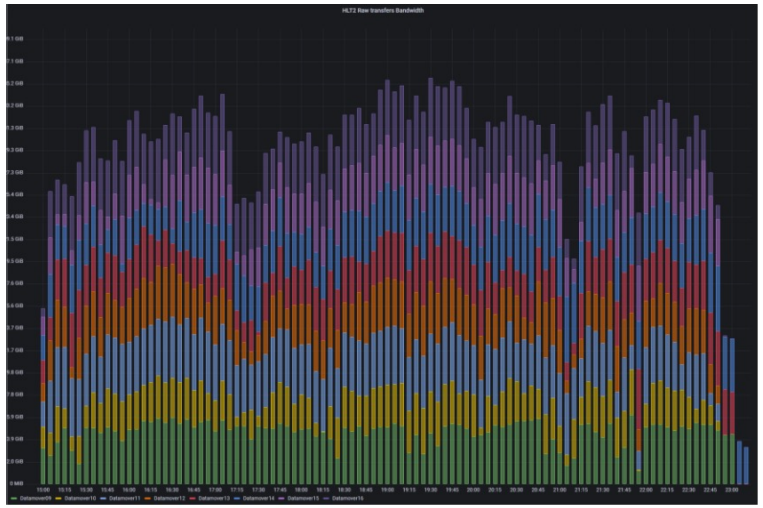

Event Rate
(events / s)
10 GB/s
Bandwidth
(GB / s)
High Level Trigger
Turbo
Full
Calibration
Tape Storage
100%
80%
Turbo
Full
Calibration
Disk Storage
Data Flow
3.5 GB/s
High Level Trigger
Turbo
Full
Calibration
Storage
100%
13%
Turbo
Full
Calibration
Disk Storage

Data challenges

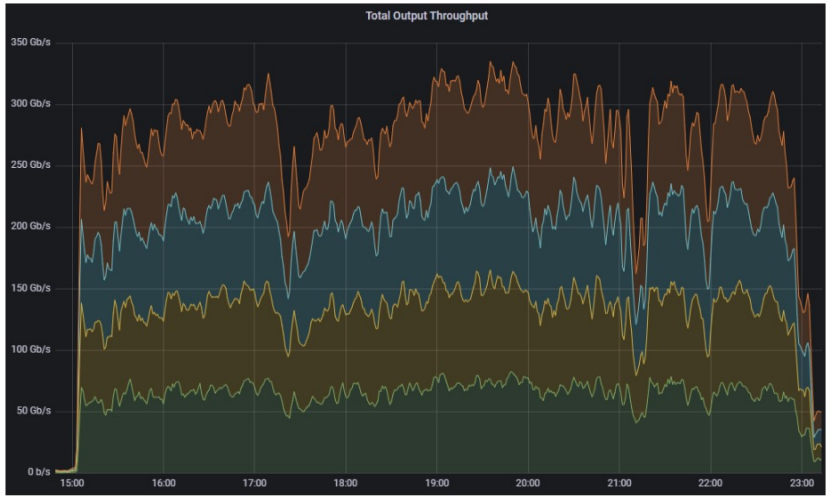
- Large-scale tests for data export from P8 to EOS/CTA performed
  - February 2022: Throughput exceeding target (16 GB/s > 10GB/s)



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

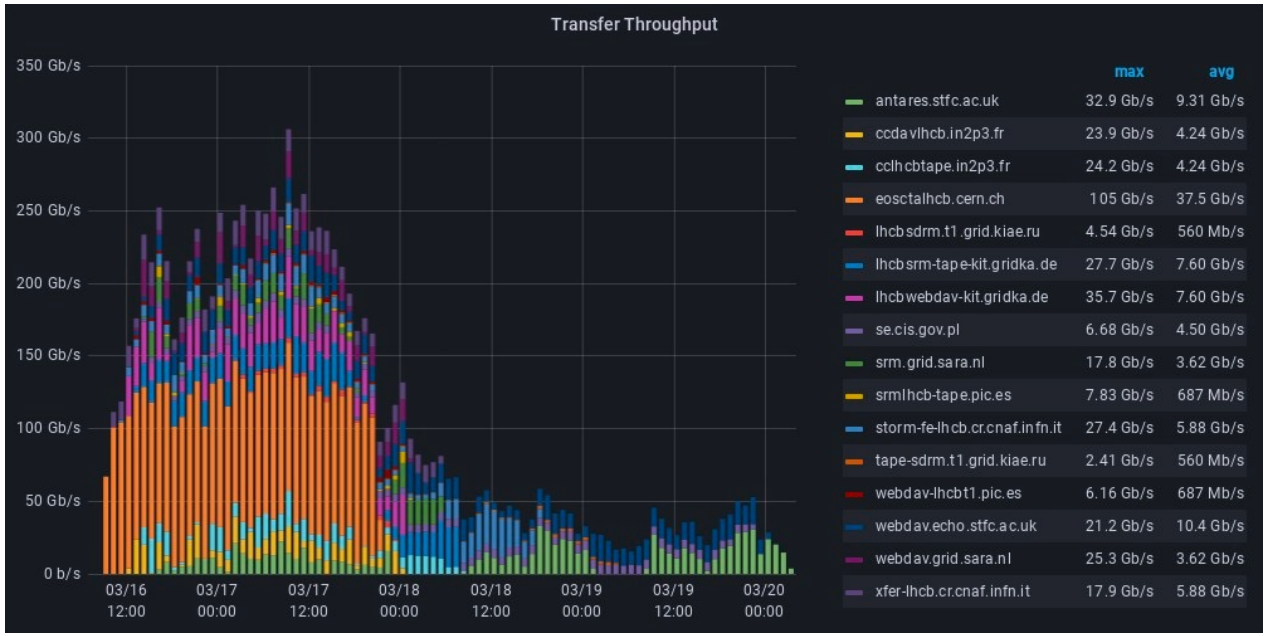


Data transfers from the 8 movers



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

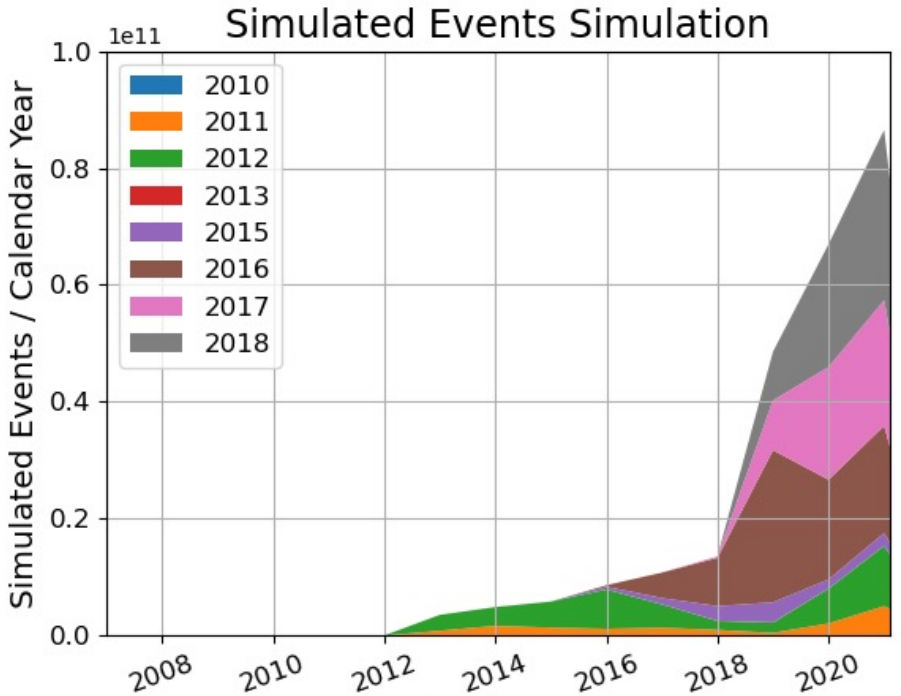
- Both write and read tests OK
  - Requirements exceeded in most sites
- A couple of sites needed following up
- No ”stress test” with real data so far
  - 2022 is a commissioning year for LHCb new detectors and software trigger

Four activities of LHCb data management

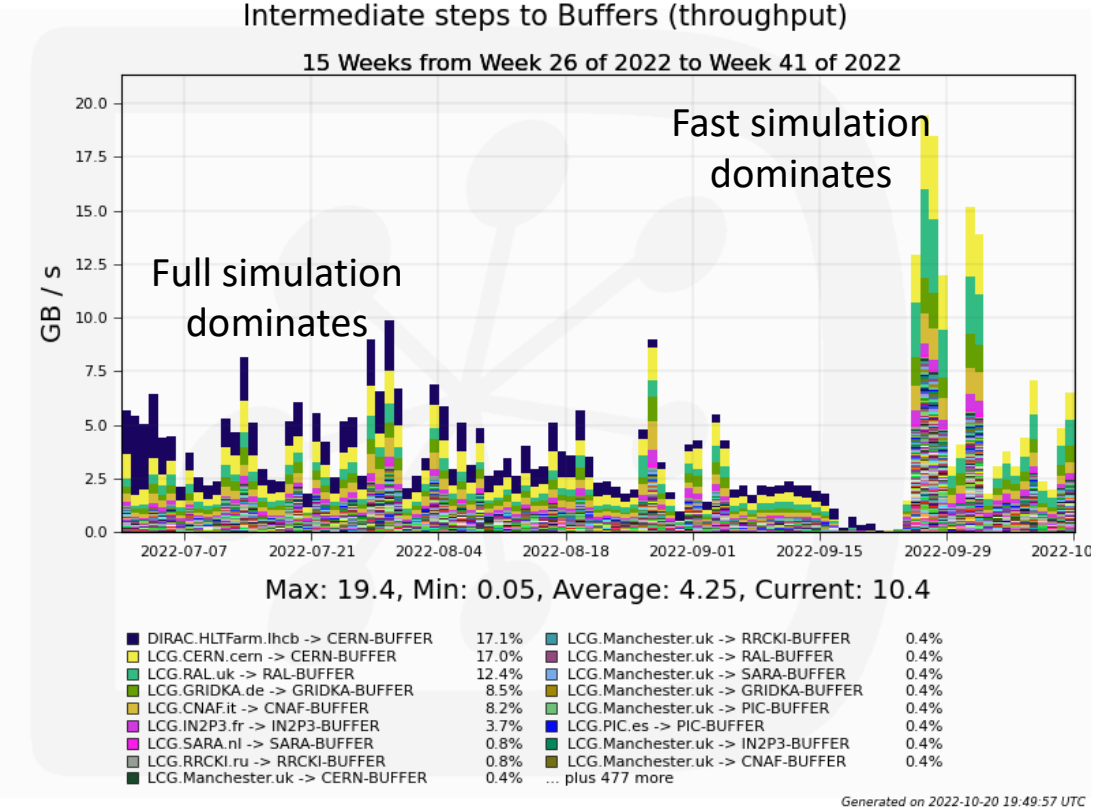
- Distribution (with FTS) of data originating from the LHCb experiment for custodial storage
- Usage of buffer disks for the intermediate processing steps of
  - Data: LAN only
  - Monte-Carlo: LAN and WAN
- Consolidation of processing outputs in fewer, larger files, aka “merging”: mostly LAN
- Replication (with FTS) of data and Monte-Carlo samples for physics analysis

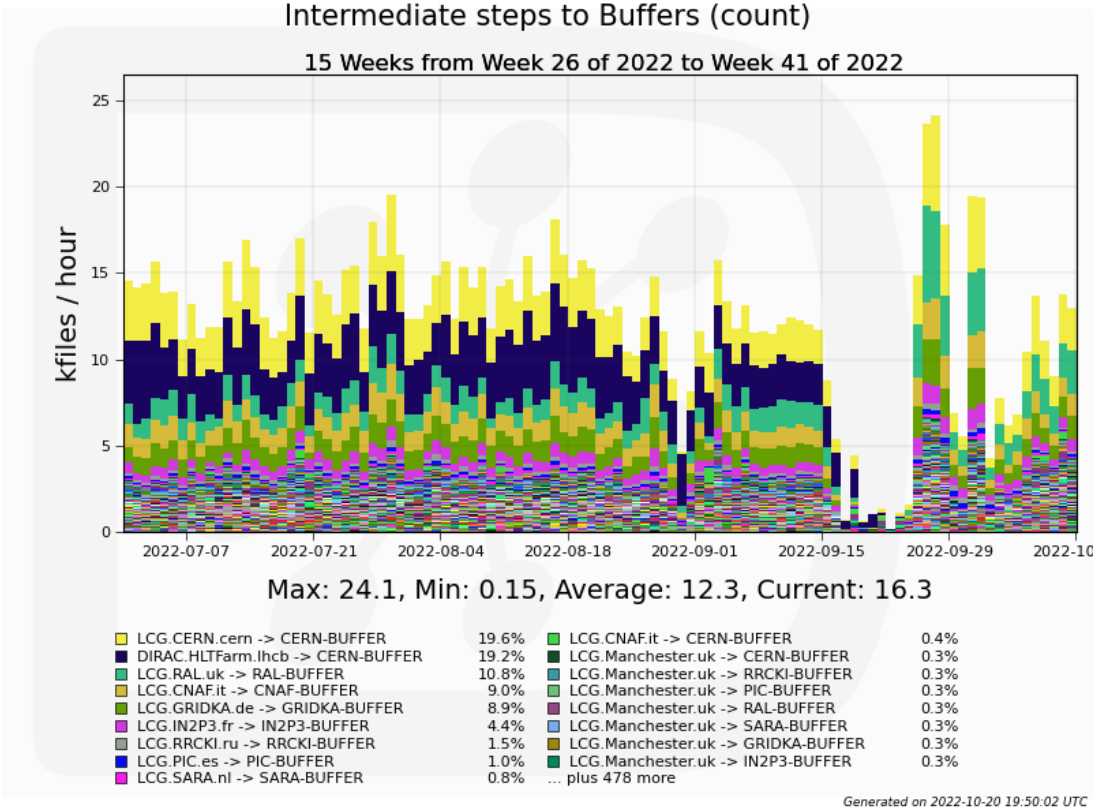
Monte Carlo simulation

- No input data required. Starting from random seed!
  - Pile-up significantly smaller than GPDs
- Simulation dominates (95%) CPU work, runs everywhere
  - Tier0, Tier1: WAN used only for replication
  - Tier2: WAN usage depends on several factors:
    - MC reconstruction / filtering run locally or somewhere else
    - amount of fast simulation
- Simulation reconstruction is heavily filtered
  - E.g. 80B events simulated in 2021 but only 11B stored, corresponding to 2PB logical volume added



Monte Carlo simulation: intermediate steps



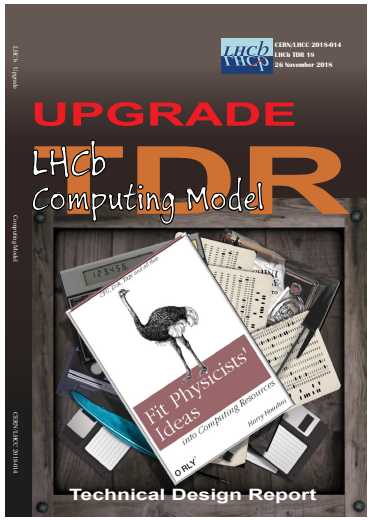


- Network usage: 2/3 Tier0/Tier1 (LAN only), 1/3 other sites (WAN)
- Fast simulation requires 3x bandwidth

- Total measured WAN throughput: O(1.5-5GB/s) depending on fast simulation
- Factor ~2 expected in Run3-4

Real data: intermediate steps

LHCb-TDR-018



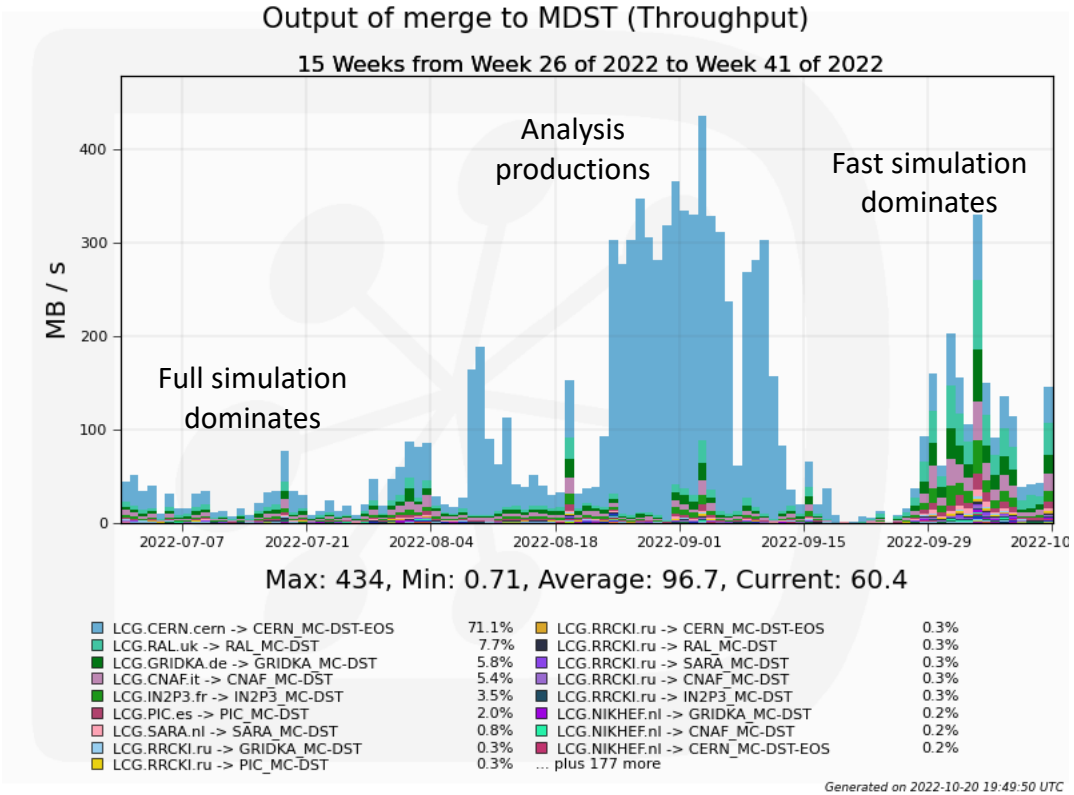
Run3 computing model, **aggregated** throughputs

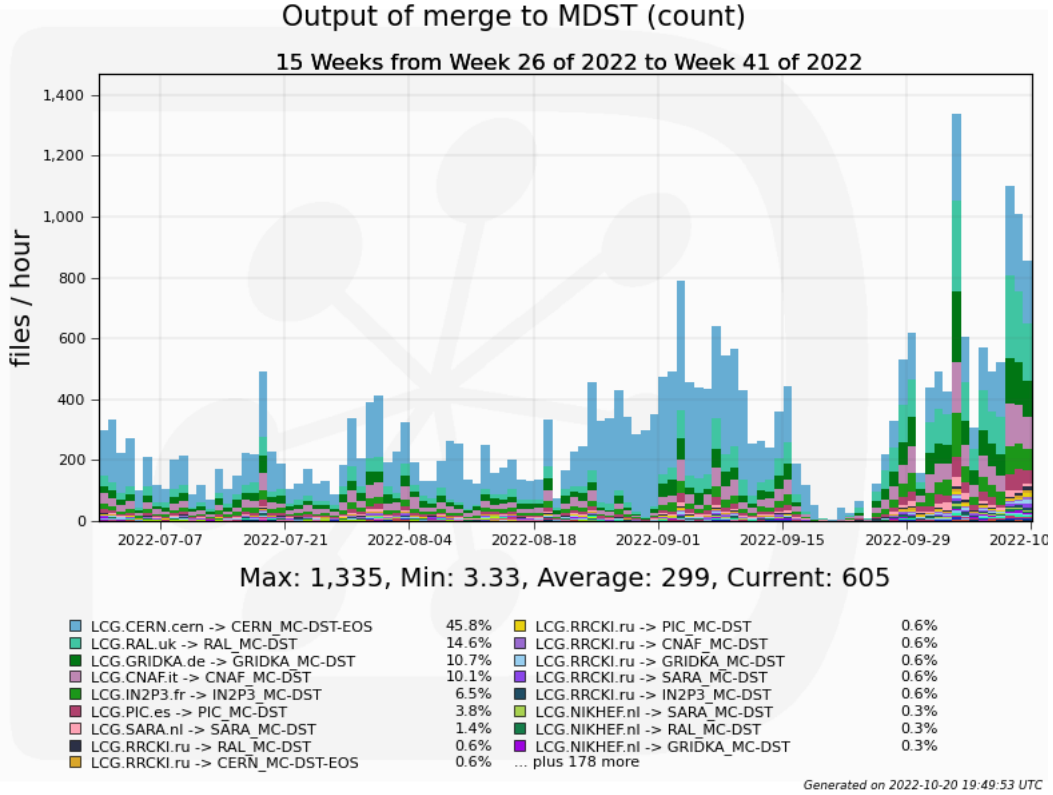
- Processing concurrent with data taking:
  - FULL+TURCAL sprucing output → 1GB/s LAN (0.2 @ T0, 0.8 @ T1)
- End-of-year reprocessing over two months:
  - FULL+TURCAL re-sprucing output → 1GB/s LAN
- These data are input to «merging» (see next step)

Four activities of LHCb data management

- Distribution (with FTS) of data originating from the LHCb experiment for custodial storage
- Usage of buffer disks for the intermediate processing steps of
  - Data: LAN only
  - Monte-Carlo: LAN and WAN
- Consolidation of processing outputs in fewer, larger files, aka “merging”: mostly LAN
- Replication (with FTS) of data and Monte-Carlo samples for physics analysis

Merging: mostly LAN



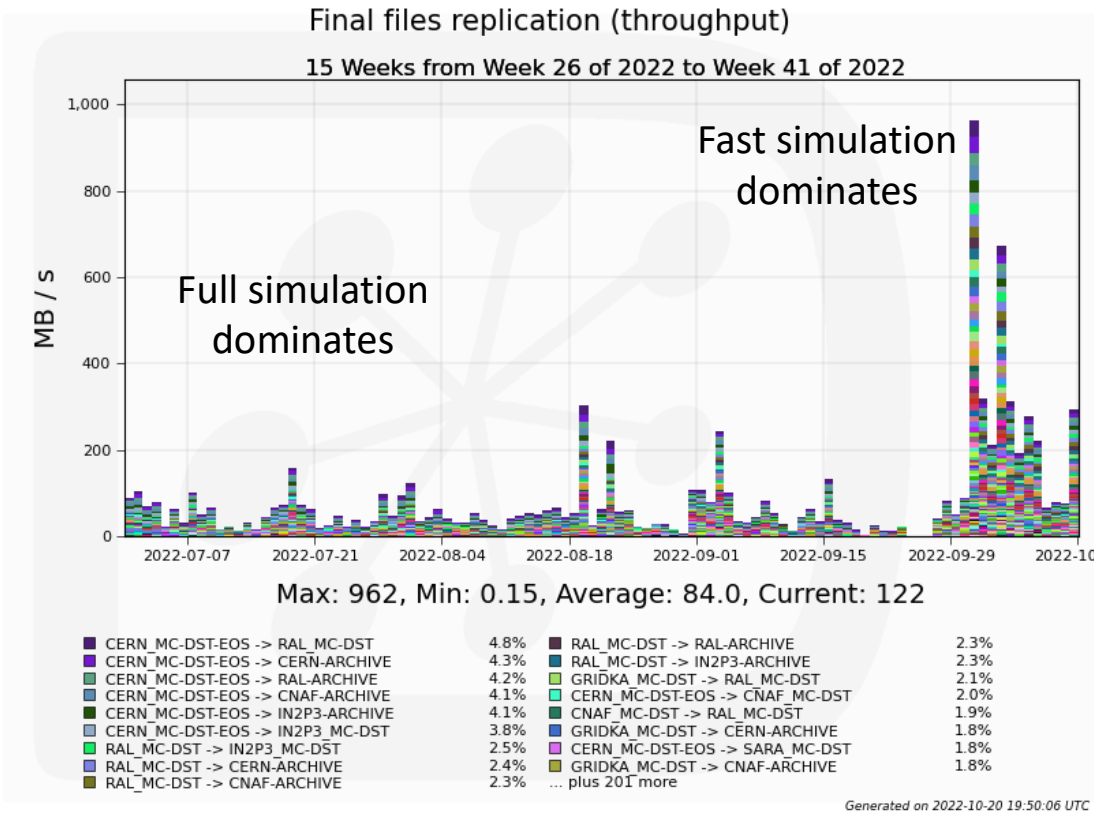


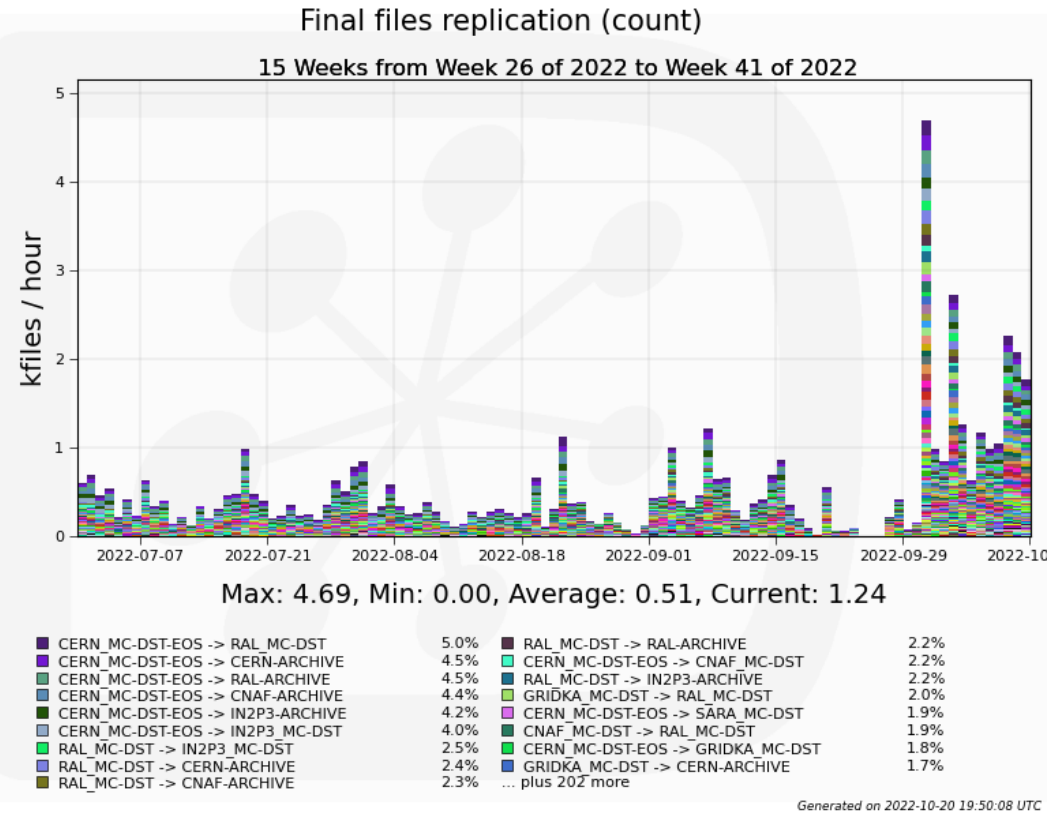
- In the above plots: simulation + artefact at CERN due to “analysis productions”
  - Extrapolation to Run3 simulation: O(100 MB/s) depending on mixture
  - Small WAN utilization to cope with specific situations (e.g. Russian T1)
- Merging of stripping output of Run3 data: 1GB/s from computing model

Four activities of LHCb data management

- Distribution (with FTS) of data originating from the LHCb experiment for custodial storage
- Usage of buffer disks for the intermediate processing steps of
  - Data: LAN only
  - Monte-Carlo: LAN and WAN
- Consolidation of processing outputs in fewer, larger files, aka "merging": mostly LAN
- Replication (with FTS) of data and Monte-Carlo samples for physics analysis

Monte Carlo simulation: dataset replications

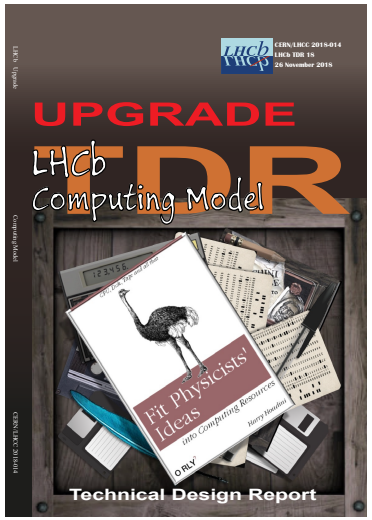




- Fast simulation requires ~3x bandwidth
- Total measured WAN throughput: O(0.1-0.4 GB/s) depending on fast simulation
- Factor ~2 expected in Run3

Real data: dataset replications

LHCb-TDR-018

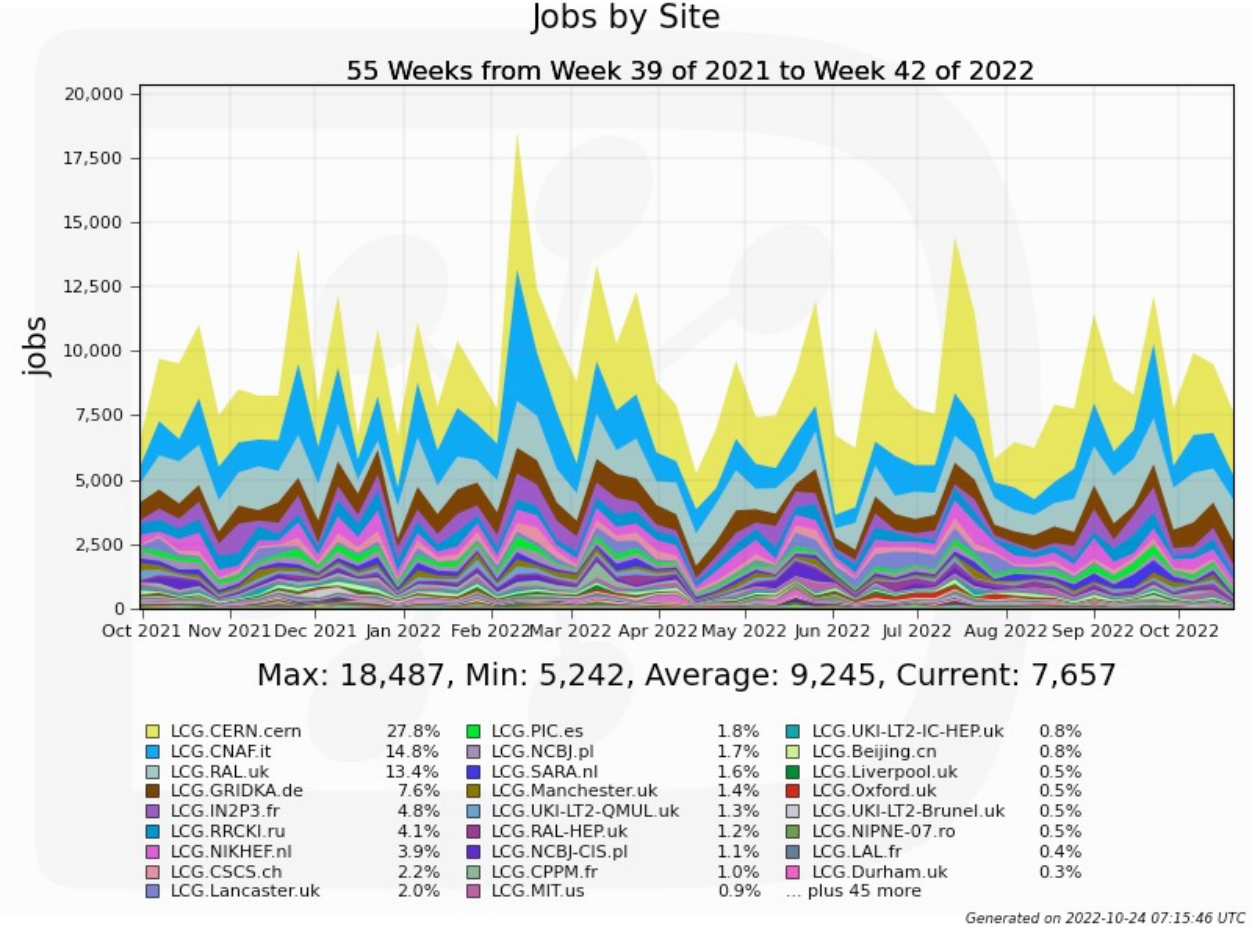


Run3 computing model, **aggregated** throughputs

- Processing concurrent with data taking:
  - 2 copies of TURBO → 2.5GB/s LAN + 2.5GB/s WAN
  - 2 copies of FULL+TURCAL sprucing output → 1GB/s LAN + 1GB/s WAN
- End-of-year reprocessing over two months:
  - 2 copies of FULL+TURCAL re-sprucing → 1GB/s LAN + 1GB/s WAN
  - N.B. we delete one copy of previous sprucing cycle

Data placement 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
- No impact on WAN other than dataset replication (2 copies)



Summary

(*) depending on processing model and mixture of full / fast simulation
(**) small, specific utilization if needed

Activity	Data type	Channel	During data taking		Winter reprocessing		Continuous	
			LAN (GB/s)	WAN (GB/s)	LAN (GB/s)	WAN (GB/s)	LAN (GB/s)	WAN (GB/s)
Custodial storage	Real data	P8 → T0 T0 → T1	10 10	10 10				
Intermediate processing steps	Real data	T0→T0 T1→T1	0.2 0.8	0 0	1.9 7.6			
	Simulation	Any → T0,T1					7-25(*)	3-10(*)
Merging	Real data	T0→T0, T1→T1	1	0	1	0		
	Simulation	T0→T0, T1→T1					0.2-0.8(*)	0(**)
Dataset replication	Real data	T0,T1→T0,T1,T2D	3.5	3.5	1	1		
	Simulation	T0,T1→T0,T1,T2D	0.1-0.4(*)	0.1-0.4(*)			0.1-0.4(*)	0.1-0.4(*)
		TOTAL	**25.6 25.9(*)**	**23.6 23.9(*)**	**11.5**	**1**	**7.3 26.2(*)**	**3.1 10.4(*)**

Final remarks

- LHCb will increase network usage by an order of magnitude in Run3 and beyond
  - Dominated by real data coming from the detector
  - A couple of knobs to turn for Monte Carlo simulation
- 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.

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

Data Processing Workflow per Data Taking Year