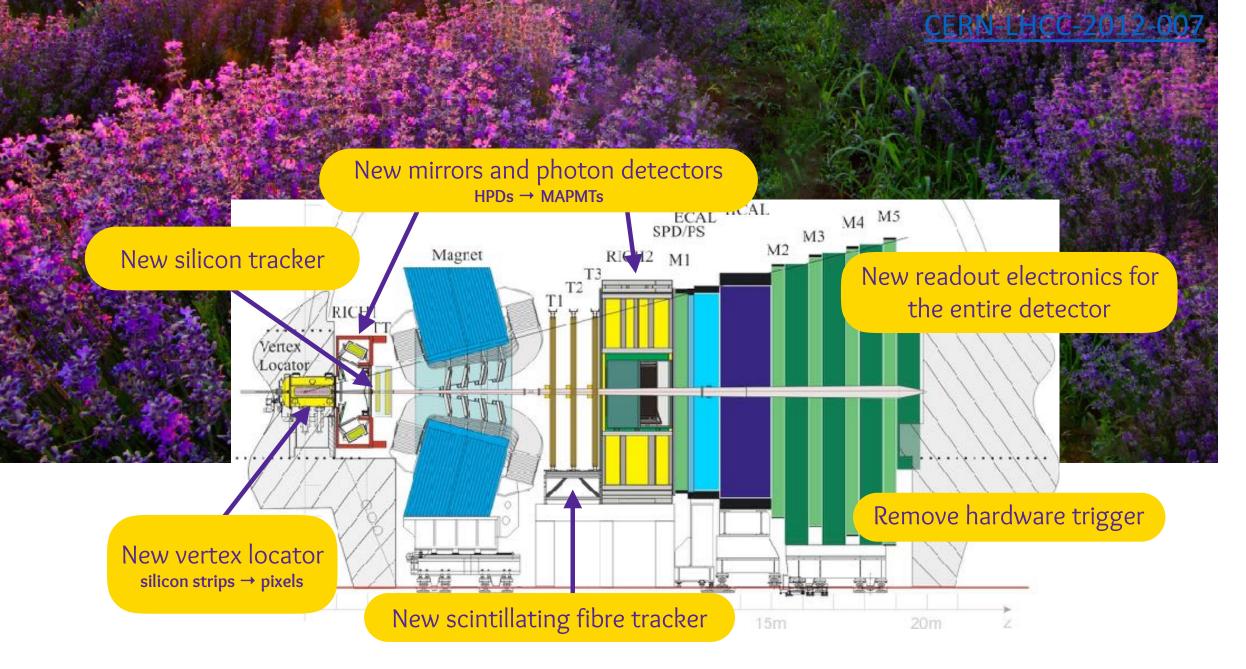
LHCb update

Ben Couturier, CERN Concezio Bozzi, INFN Ferrara

WLCG workshop Lancaster, November 7th 2022

Overview

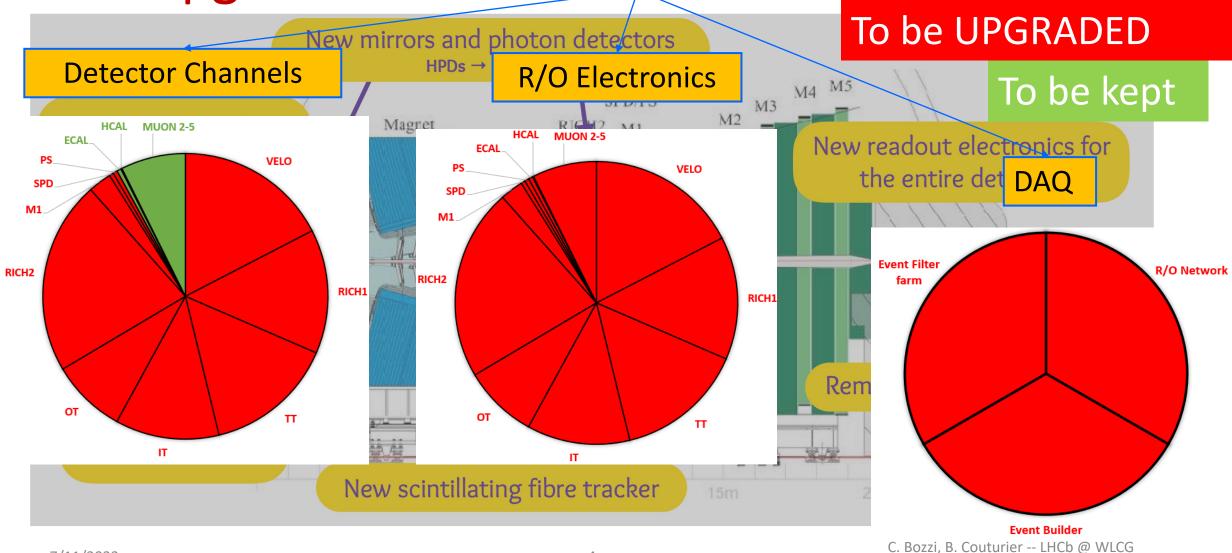
- Run3 + Run4 computing model
- Resource usage
- HPC status
- Analysis facilities



Chr/1 B/2022 LHCb full-detector real-time alignment and calibration: Latest devælopments and perspective • CHEP 2018, Sofiazi, B. Couturier -- LHCb @ WLCG workshop

workshop

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
 - 4x10³² to 2x10³³ cm⁻²s⁻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)



Fit Physicists Ideas

CPU, Disk, Tape And All That

Into Computing Resources

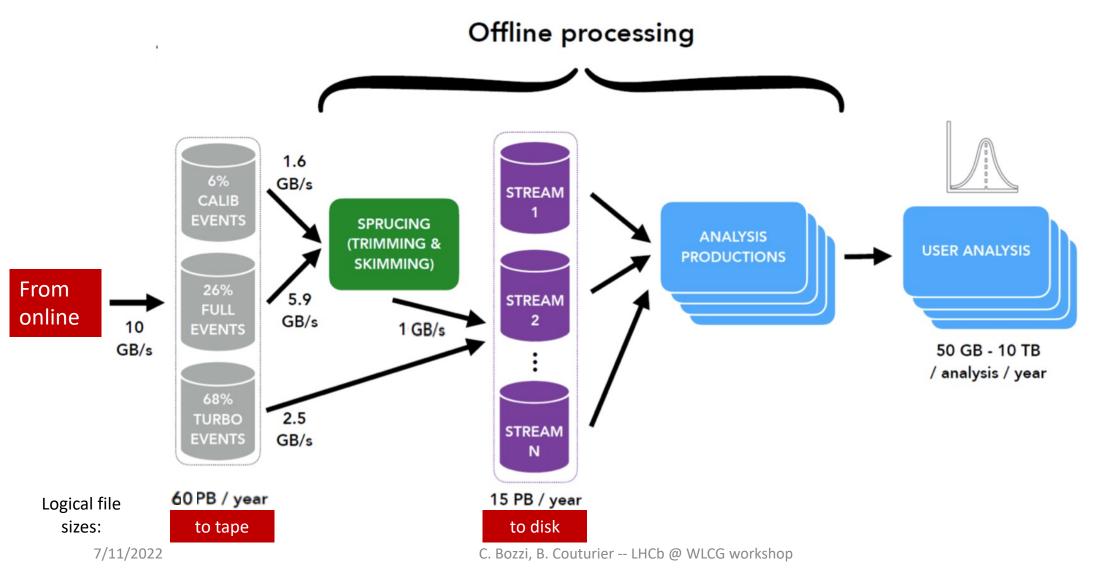
O RLY[?]

Harry Houdini

Data streams and dataflow

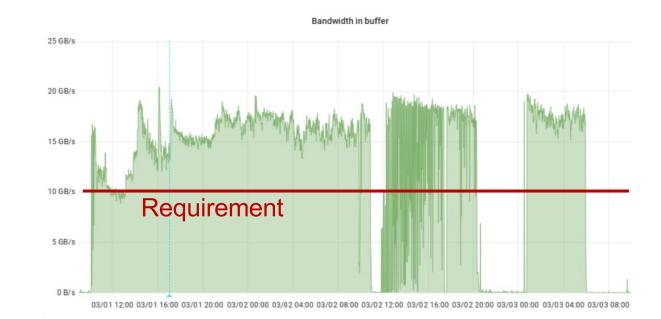
- Data from the LHCb detector organised in 3 streams; in all cases; events are reconstructed online at the HLT farm
 - FULL: «classic» stream, where information from the entire event is persisted in DST format and input to offline «sprucing» i.e. «slimming and skimming» for subsequent physics analysis
 - 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

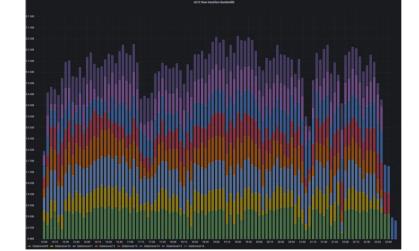


Data challenges

- 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

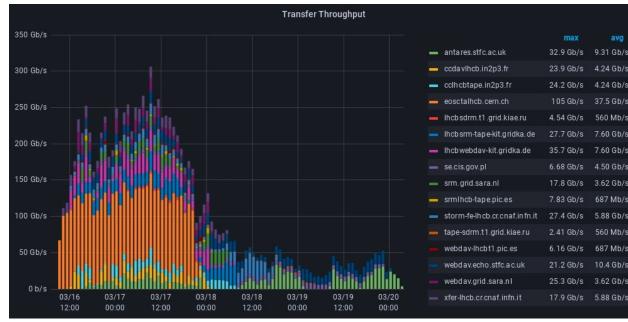


Data transfers from the 8 movers



Corresponding data links traffic

Tape challenges

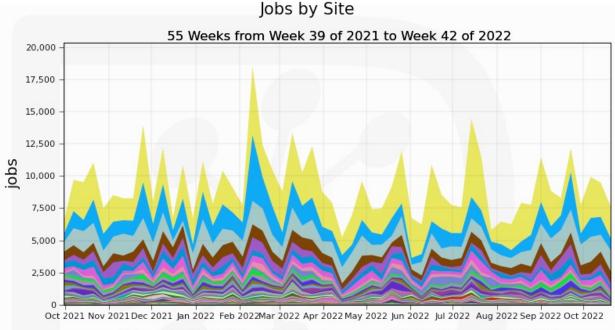


$\begin{array}{c} \text{Write} \\ \text{CERN disk} \rightarrow \text{T}^{2} \end{array}$		Read tests T1 tape \rightarrow T1 disk		
Site	expected Speed (GB/s)	Site	expected Speed (GB/s)	
CERN	11	CERN	1.90	
CNAF	1.72	CNAF	1.35	
GRIDKA	2.23	GRIDKA	1.36	
IN2P3	1.25	IN2P3	0.98	
NCBJ	1.32	NCBJ	0.91	
PIC	0.2	PIC	0.17	
RAL	2.96	RAL	1.93	
RRCKI	0.25	RRCKI	0.21	
SARA	1.07	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)



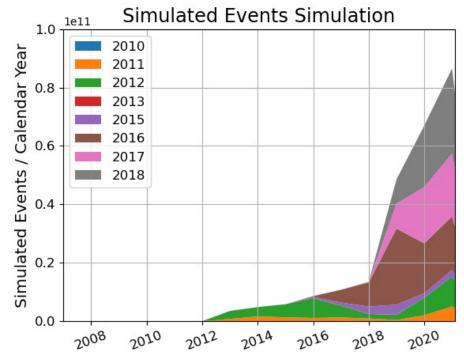
Max: 18,487, Min: 5,242, Average: 9,245, Current: 7,657

LCG.CERN.cern	27.8%	LCG.PIC.es	1.8%	LCG.UKI-LT2-IC-HEP.uk	0.8%
LCG.CNAF.it	14.8%	LCG.NCBJ.pl	1.7%	LCG.Beijing.cn	0.8%
LCG.RAL.uk	13.4%	LCG.SARA.nl	1.6%	LCG.Liverpool.uk	0.5%
LCG.GRIDKA.de	7.6%	LCG.Manchester.uk	1.4%	LCG.Oxford.uk	0.5%
LCG.IN2P3.fr	4.8%	LCG.UKI-LT2-QMUL.uk	1.3%	LCG.UKI-LT2-Brunel.uk	0.5%
LCG.RRCKI.ru	4.1%	LCG.RAL-HEP.uk	1.2%	LCG.NIPNE-07.ro	0.5%
LCG.NIKHEF.nl	3.9%	LCG.NCBJ-CIS.pl	1.1%	LCG.LAL.fr	0.4%
LCG.CSCS.ch	2.2%	LCG.CPPM.fr	1.0%	LCG.Durham.uk	0.3%
LCG.Lancaster.uk	2.0%	LCG.MIT.us	0.9%	plus 45 more	

Generated on 2022-10-24 07:15:46 UTC

Monte Carlo simulation

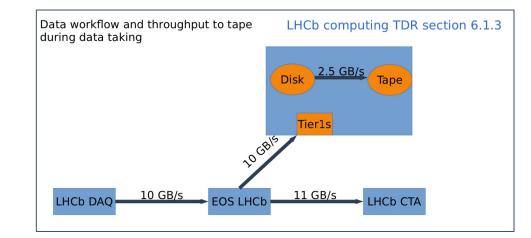
- No input data required. Starting from random seed!
 - Pile-up significantly smaller than GPDs
- Simulation dominates (95%) CPU work, runs everywhere
 - Improvements in simulation and introduction of fast simulation significantly decrease CPU work per event
- Simulation reconstruction is heavily filtered
 - 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

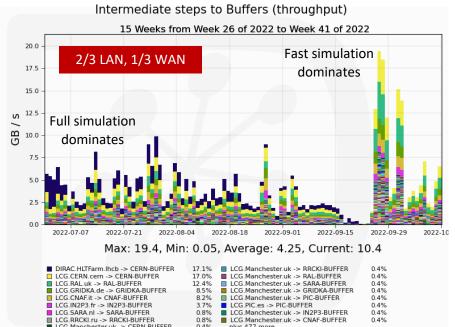


Year	Simulated events (10 ⁹)	Stored events (10 ⁹)	Ratio	CPU work kHS06.y	CPU per event kHS06.s	LFS TB
2017	10.3	4.2	40.3%	817	2.50	640
2018	12.0	3.0	25.3%	1009	2.65	550
2019	45.0	6.9	15.2%	1290	0.90	1110
2020	67.0	16.8	31.7%	1357	0.81	2010
2021	80.0	11.1	13.9%	1815	0.72	2030

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.

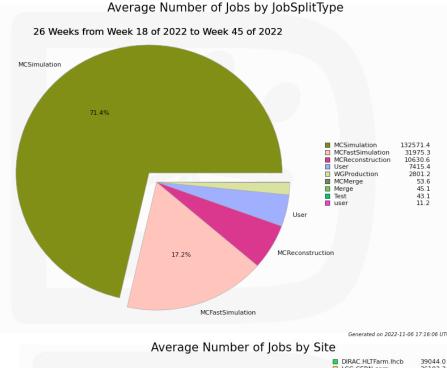


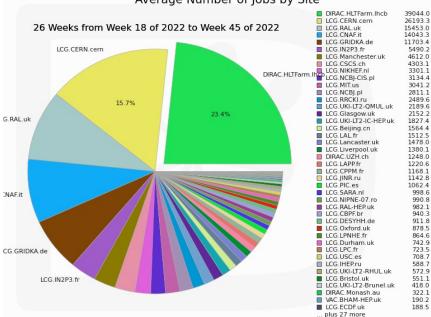


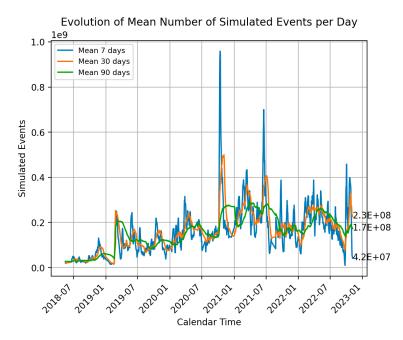
Generated on 2022-10-20 19:49:57 UTC

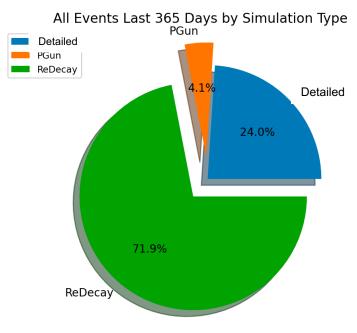
Distributed computing operations

- 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





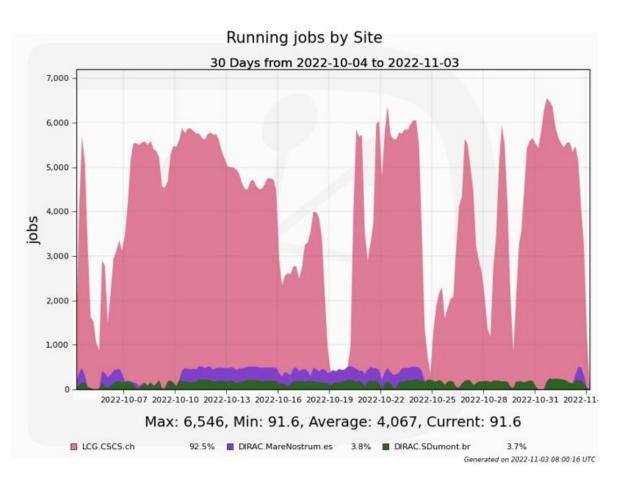




Generated on 2022-11-06 17:15:27 UTC

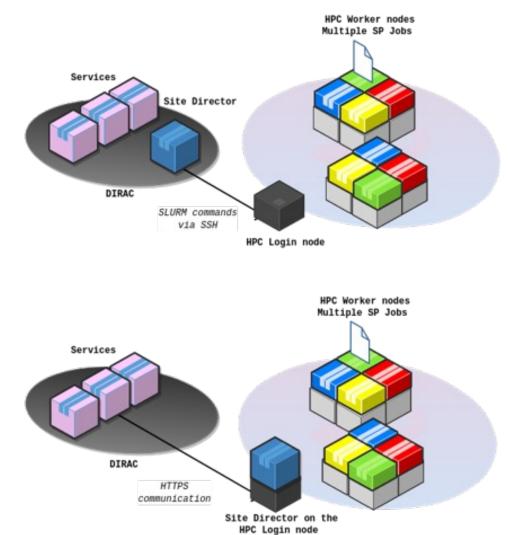
Progress on HPC

- 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

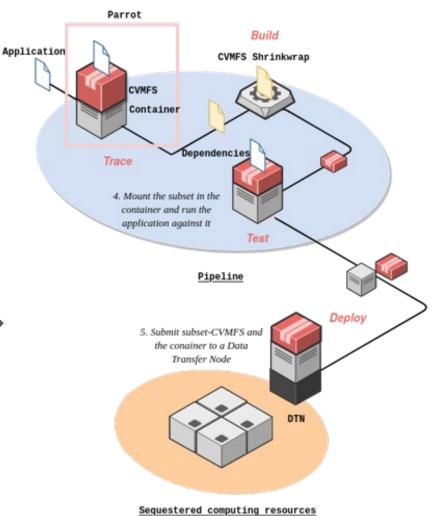
- HPCs with external connectivity:
 - Support AREX (ARC) services
 - Keep leveraging CSCS computing resources
 - Support multi-node allocations
 - Useful when a limited number of large allocations is available
 - Only work with resources orchestrated via SLURM
 - Test a Site Director installation on 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 <u>CI pipeline</u> 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

1. A new 2. Execute and monitor the 3. Get the dependencies and application comes application with CVMFS and a create a subset of CVMFS from it in container image



DIRAC news



- Rolling out major DIRAC release (v8.0)
 - First DIRAC release dropping py2 support \rightarrow 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

r or or or o	nple_tupling_f	-	1					
WG	Application	Data Type (set elsewhere)	Input Type (not set)	CondDB tag		DDDB tag (set elsewhere)	Desired Priority	
_		(set elsewhere)	(90K Sel)	(set elsewh	ere)	(set elsewhere)	16	
Inputs /	Outputs					Size (this job)	Size (entire samp	(m)
Input		06789_1.test_stream_	A.dst			222.4 MB	27.59 68	
Output		00789 2.dvntuple.roo				996.43 kB	~ 123.59 %8	
	were processed in 00:01:58	l on a 30.5x machine.						
_	reproduce 3768358 "FEST	" "example_tupling_f	ull_linei" 0					
Checks								
State	Check		Trees	Messages				
PASS	example_tupling_full_1 at_least_100_entries	line1/	K50ToPipPim/DecayTre	e Found 13815 3	n KS0ToPipPim	/DecayTree (190 required)		
_								
PASS	example_tupling_full_1	line1/Ks0_branches	K50ToP1pP1m/DecayTre	e All required	branches were	found in Tree KS0ToPipPim	/DecayTree	
PASS	example_tupling_full_	line1/Ks0_mass_hist	KS0ToPipPim/DecayTre	e Histogram of	KS0_M success!	fully filled from TTree KS	ToPipPim/DecayTree (conti	ins
PASS	example_tupling_full_1	line1/Ks0_mom_hist	KS0ToPipPim/DecayTre	e Histogram of KS0ToPipPim/D	(KS0_PX**2 +) ecayTree (con/	CS0_PY**2 * KS0_PZ**2)**0. tains 13426.0 events)	5 successfully filled from	11
PASS	example_tupling_full_1 Ks0_mom_xy_hist	line1/	KS0ToP1pP1m/DecayTre	e Histogram of events)	KS0_PX, KS0_P	f successfully filled from	TTree KS0ToPipPim/DecayTr	ee
	example tupling full line1/KSI	0ToPipPim DecayTree Ks0 n	sass hist	+	,ease	ple tupling full line1/KS0ToPipPin	DecayTree Ks0 mom hist	
3000				128	3500		Binde	
2500				69	3000		Sinde	
							Deed	
1 2000		+ +		0	Ispani	+	8	
Entries / 1.6 [article]					10000/18	•	-B B	
Ene					1500 E		Bio	
1000					_	· .	Binde	
					1000	•	Inded	
500					500		B	
	60 470 480	490 500 KSQ_M [unit		530 540	0.0008+0	2.000e-4 4.000e-4 (RS0_PX**2	6.000+4 8.000+4 + KS0_PY**2 + KS0_PZ**2)**0.5 (unit	ŋ
300	example_tupling_tuil_line1/K3	S0ToPipPim DecayTree Ks0_	mom_xy_hist					
				800 + JO				
200				ÚP				
100				600 E				
				0				
						11		Г
14.00				400		09	se JS	h
- 100								Ì
-100								
						te	st pr	1

Analysis Productions

h Home

Pipelines
 Submissions
 Mattermost
 Documentation

fest / spruce_exclusive_feb_2022

State	ACTIVE
Version	v0r0p3657063
Size	 (NaN% ready on disk)
Ownership	christopher.burr@cern.ch
Merge Request	https://gitlab.cern.ch/lhcb-datapkg/AnalysisProductions/-/merge_requests/215
🖵 JIRA Task	https://its.cern.ch/jira/browse/WGP-288
Tags	
config	fest
eventtype	90000001

DIRAC Production Request 96783

is assigned sample ID 6929 and comprises the following transformations:

comprises 1 ste	ep - output is not kept
Step ID	154271
Application	Moore/v53r4
Options	<pre>\$ANALYSIS_PRODUCTIONS_BASE/FEST/sprucing_excl.py \$APPCONFIGOPTS/Persistency/Compression-ZLIB-1.py</pre>
Extra Data Packages	AnalysisProductions.v0r0p3657063 ProdConf

Transformation 157186 comprises 1 step - output is kept

Step ID 154272

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

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



- 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)	$5 \times 10^6 (2.5 \times 10^6 \text{ in } 2021)$							
Integrated luminosity	10 fb ⁻¹ (5 fb ⁻¹ in 2021)							
Trigger rate fraction (%)	action (%) 26 / 68 / 6 Full/Turbo/Tu							
Logical bandwidth to tape (GB/s)	10	(5.9/2	2.5/1.6	.5/1.6 Full/Turbo/TurCal)				
Logical bandwidth to disk (GB/s)	3.5	5 (0.8 / 2	2.5/0.2	EFull/T	'urbo/Tur	Cal)		
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 (Tu	urbo); 3	(Full,	TurCal)			
MC replicas on tape			1 (N	(IDST)				
MC replicas on disk			T, 30%	of the	total data	aset)		
Resource r	equire	ements						
WLCG Year	Disl	k (PB)	Tape	(PB)	CPU (kHS06)			
2021 2022 2023 2024 2025	66 111 159 165 171	1.1 1.7 1.4 1.0 1.0	142 243 345 348 351	1.5 1.7 1.4 1.0 1.0	863 1579 2753 3467 3267	1.4 1.8 1.7 1.3 0.9		

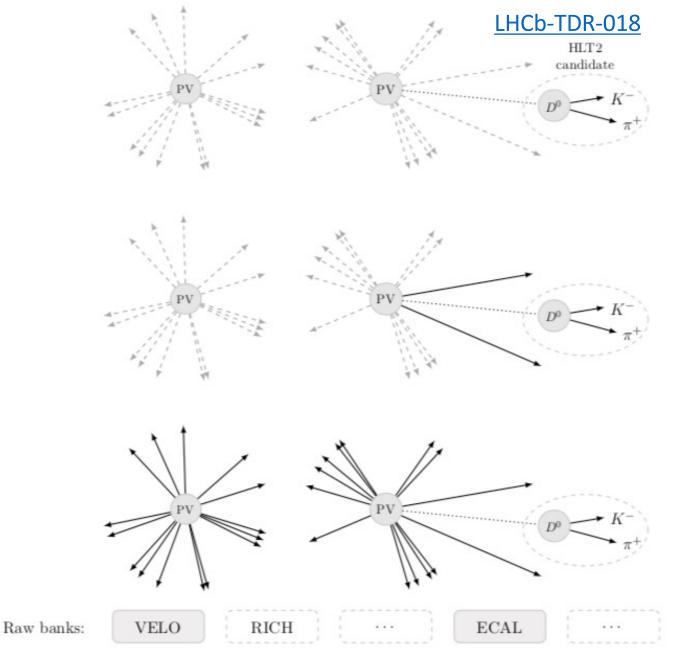
^a corresponding to 120, 40, 2s on a 10HS06 computing core

^b simulation of year N starts in year N+1

Data persistency

• Different levels of persistency:

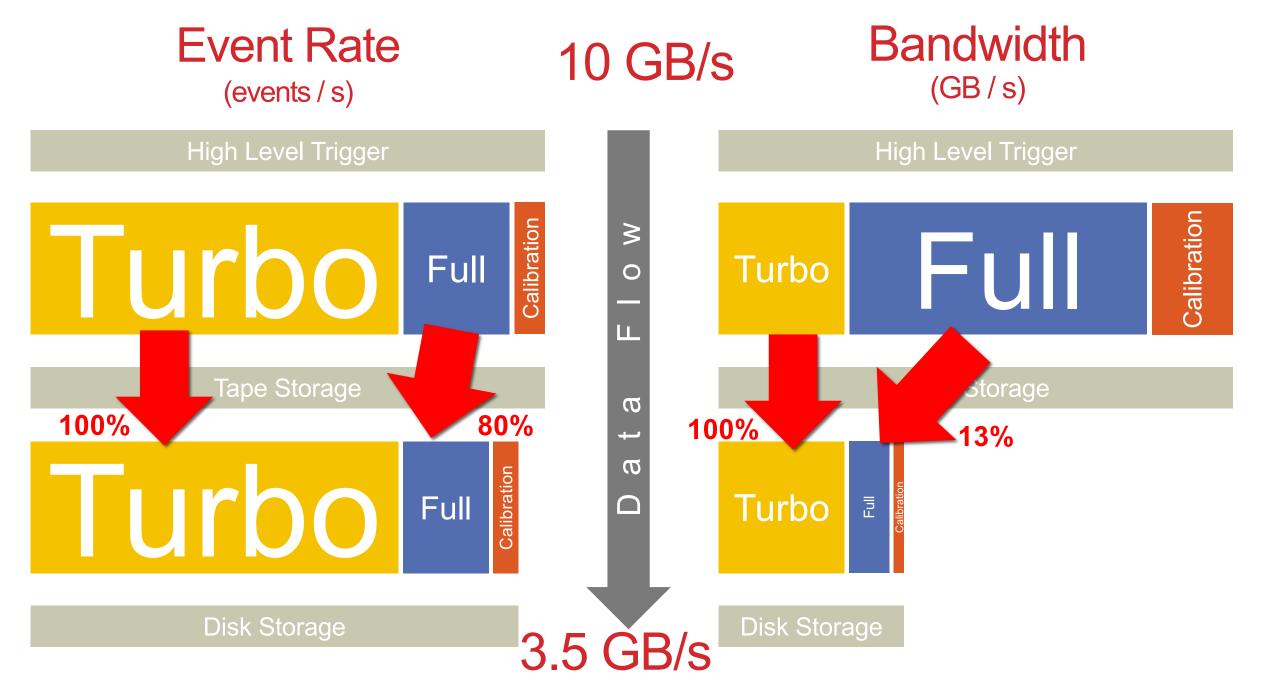
- FULL and TURCAL: the full event is persisted
- TURBO: selective persistency, ranging from candidate firing the trigger to the entire event, optionally including some RAW subdetector data banks



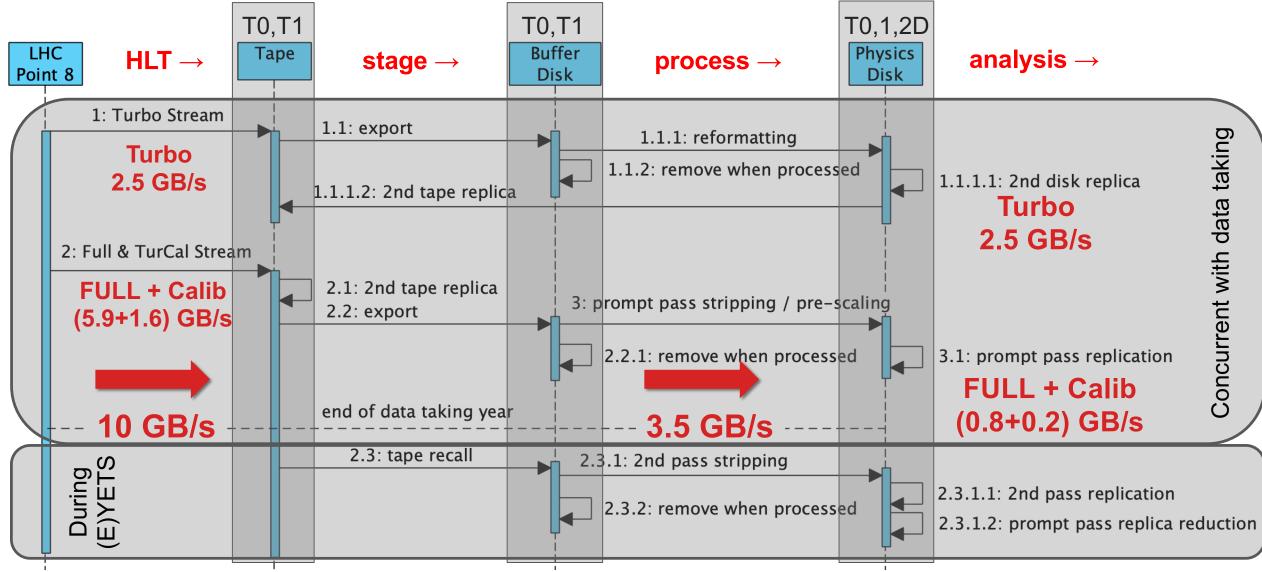
HLT output bandwidth

- 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 Throug	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%	



Data Processing Workflow per Data Taking Year



C. Bozzi, B. Couturier -- LHCb @ WLCG workshop