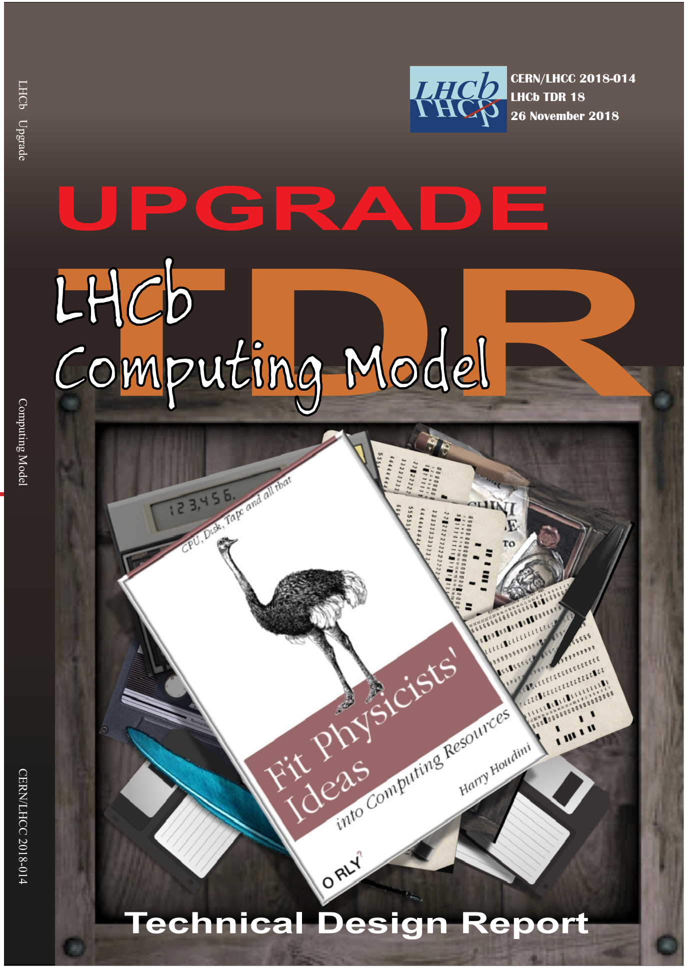
Upgrade Computing and resources

Jahrestreffen der deutschen LHCb-Gruppen

October 1st 2019

Concezio Bozzi

Istituto Nazionale di Fisica Nucleare



https://cds.cern.ch/record/2319756

Overview

- Building the LHCb Upgrade Computing Model
  - From Run2 to Run3
  - Storage: dealing with the deluge of data from the pit
  - CPU: understanding the need for simulated events
- Offline computing requirements for Run3 and LS3
- Mitigation strategies
- Risk analysis
- Outlook

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

Run 2 event sizes

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%

Event size:
Turbo/FULL ~0.5

N.B Turbo event size is an average. It ranges from a few kB (minimal persistence) to full event size

Extrapolation of Run2 rates to Run3 conditions

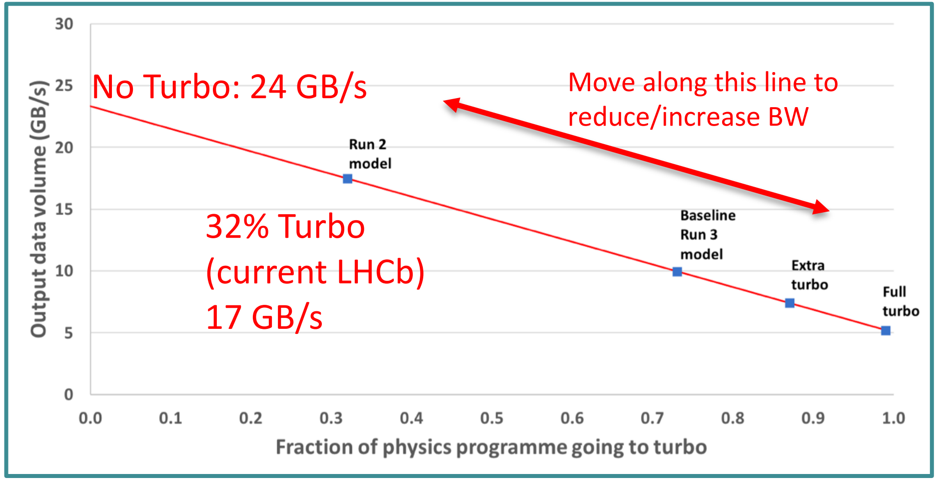
- With the upgrade conditions several factors need to be applied
  - Luminosity $4*10^{32}$ $cm^{-2}s^{-1}$ to $2x10^{33}$ $cm^{-2}s^{-1}$
  - HLT efficiency increase because of removal of L0 hardware trigger
  - Raw event size increase due to pileup, according to simulation
- Without any changes the HLT output rate would increase in Run 3 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

Event size:
Turbo/FULL ~0.167

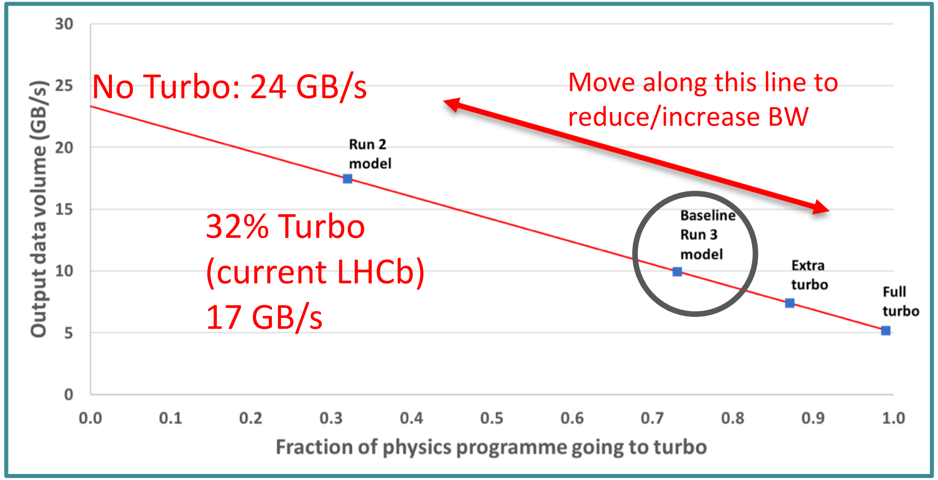
Evolution of physics programme

- Moving a larger fraction of the physics programme to Turbo decreases the output bandwidth
- Turbo events are considerably smaller (16 % of Full size)
- Some selections need to stay in Full
  - Keep some flexibility, recover from eventual errors, develop new analysis ideas



Evolution of physics programme

- For the baseline model we assume 60% of the physics selections currently on FULL stream migrating to Turbo
- Massive migration, not trivial!
- Baseline model assumes 73% of the physics selections on Turbo
- Corresponds to a BW of 10 GB/s

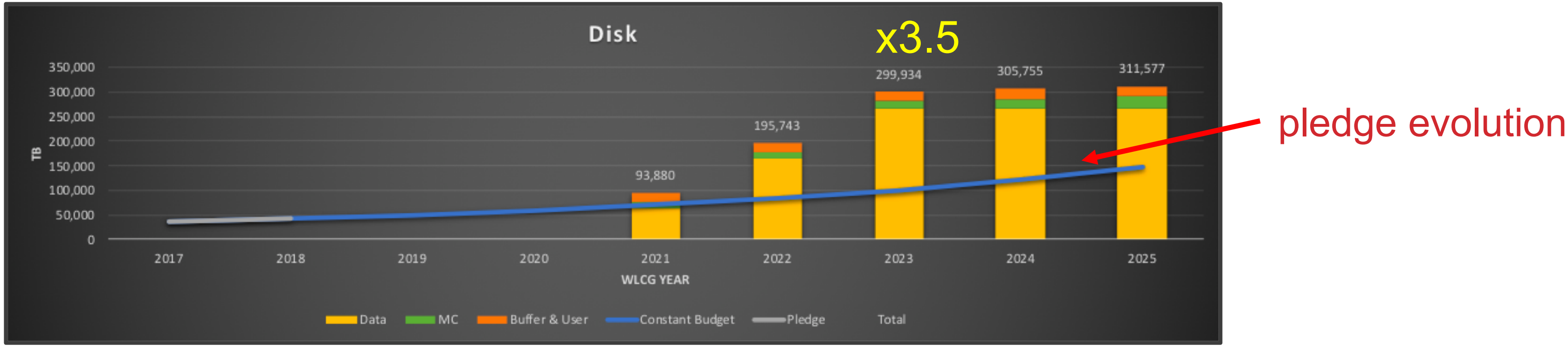


Turbo	FULL
Exclusive charm	Inclusive beauty
Exclusive light quarks	Multi-leptonic (LFV)
	Electroweak
	High pT
Exclusive beauty	

Potential trigger selection repartition among streams (excluding calibration)

Baseline bandwidth: evolution of the model

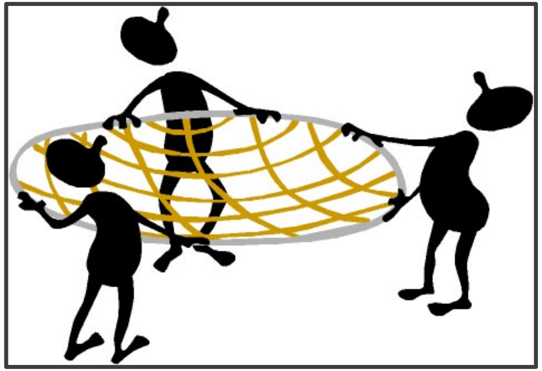
- Can we fit 10 GB/s in a reasonable amount of storage resources ?
- First attempt, presented in summer to LHCC and WLCG resulted in an amount of disk **3.5 times larger** than what expected in a “constant budget” evolution model !
- mitigation strategies clearly needed



First attempt to fit upgrade data on disk (summer 2018)

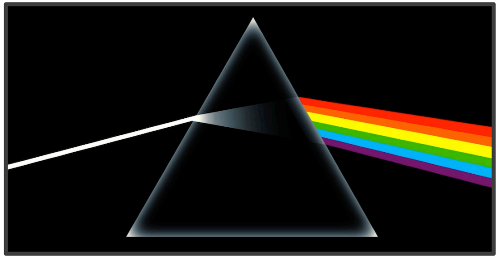
Baseline bandwidth: evolution of the model

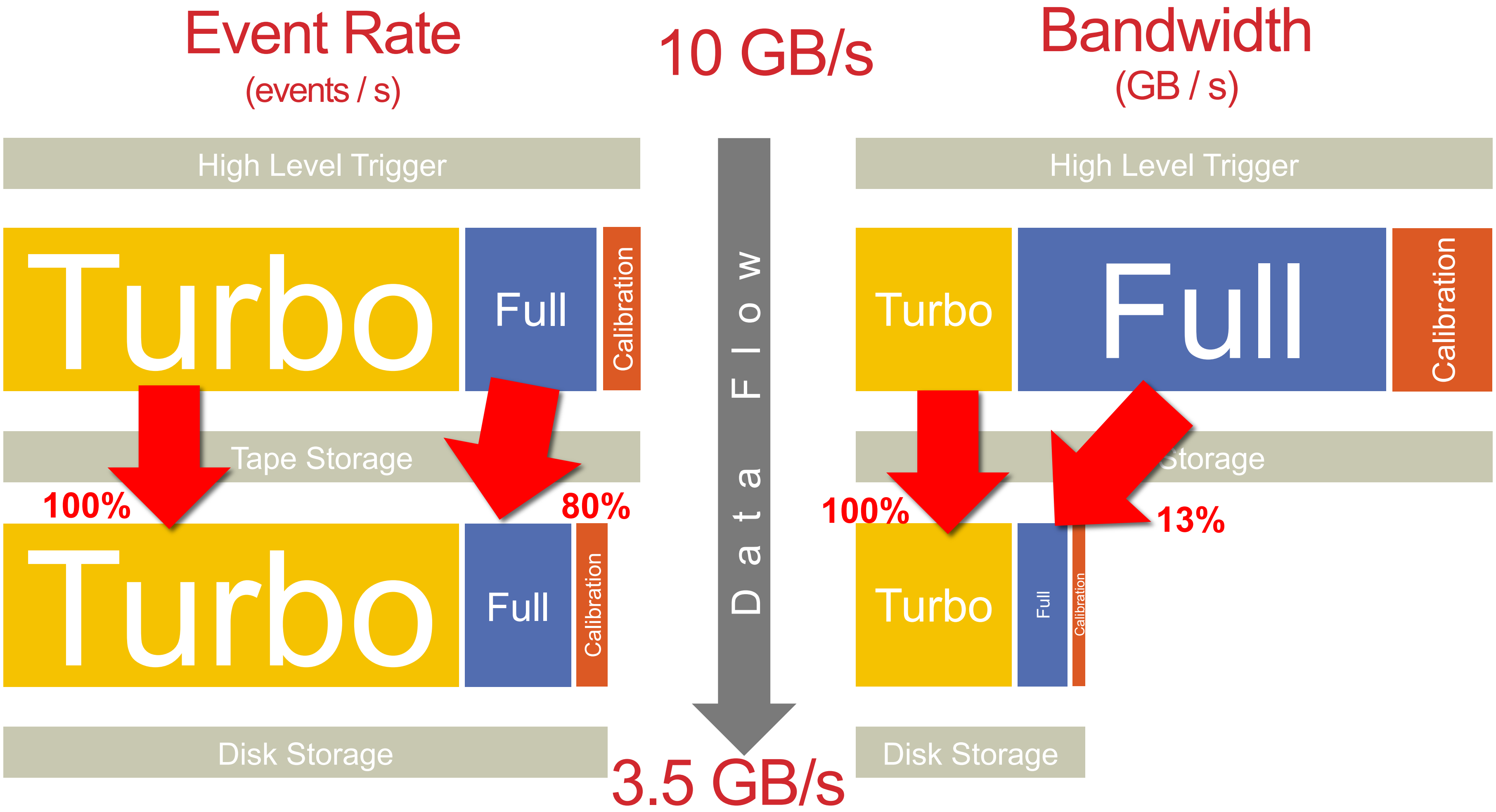
- Idea! Use cheap storage as a **safety net** :
  - save the desired BW on tape
  - Profit of *stripping* to reduce data volume to disk.
  - …but with the possibility of reprocessing



- Operationally more challenging
- Much safer from the physics point of view

- Stripping == offline processing of data with a large set ($O(10^3)$) of specialised selections analysis oriented
  - Similar to Turbo trigger selections
  - High event retention (~80%)
  - Use selective persistence to substantially reduce data volume
  - Output format is MDST



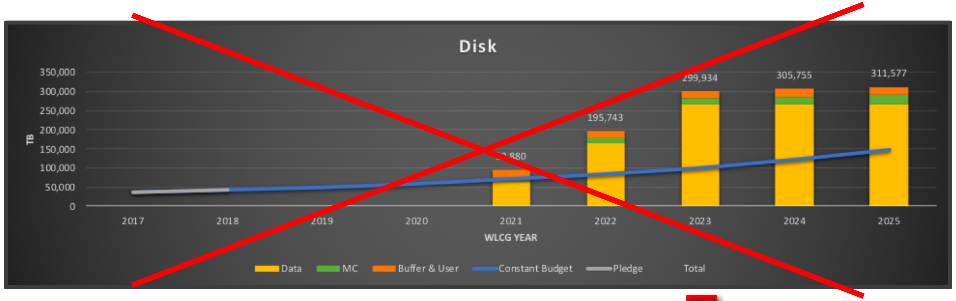

Event Rate
(events / s)
10 GB/s
Bandwidth
(GB / s)
High Level Trigger
Turbo
Full
Calibration
Tape Storage
100%
80%
Turbo
Full
Calibration
Disk Storage
Data Flow
3.5 GB/s
High Level Trigger
Turbo
Full
Calibration
Storage
100%
13%
Turbo
Full
Calibration
Disk Storage

Baseline bandwidth: evolution of the model

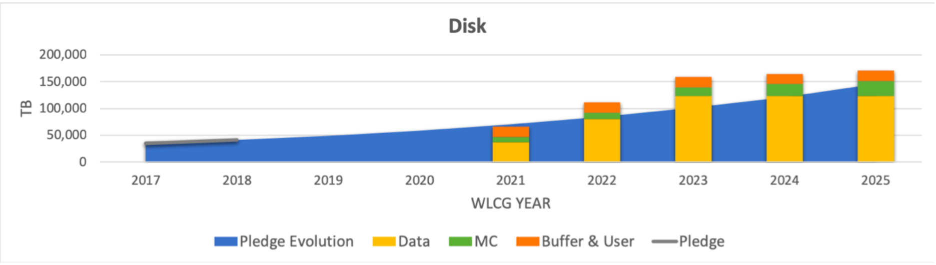
- Can we fit 10 GB/s in a reasonable amount of storage resources ?
- New model:
  - 10 GB/s to tape
  - Reduce by ~1/6 FULL and Calibration data volume with stripping
- Save **3.5 GB/s** to disk!

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%



Old version (summer)



TDR model

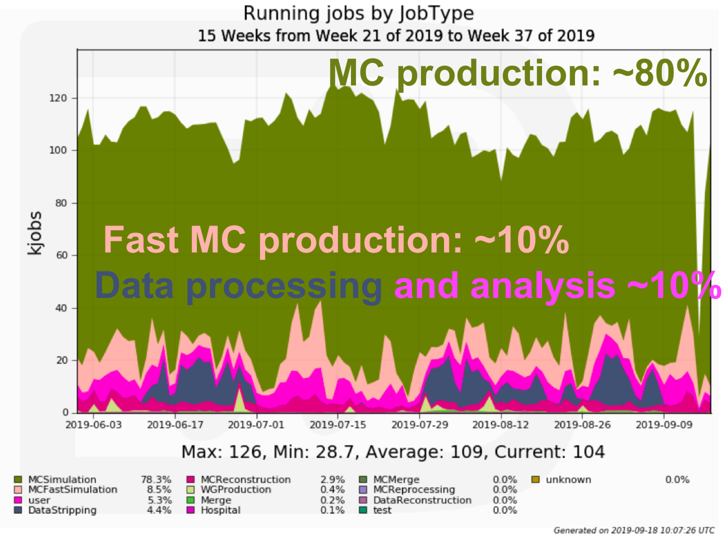
The model: what about CPU ?

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

- Scale current MC production to estimate the CPU needs

- Number of needed MC events scale with luminosity
- Seen “experimentally” in Run 2
- Well justified by physics
  - Events signal-dominated
  - Generally pure selections
  - $L_{int} \times \varepsilon_{trig}$ is a good proxy for yield

- Assume the same scaling for Upgrade



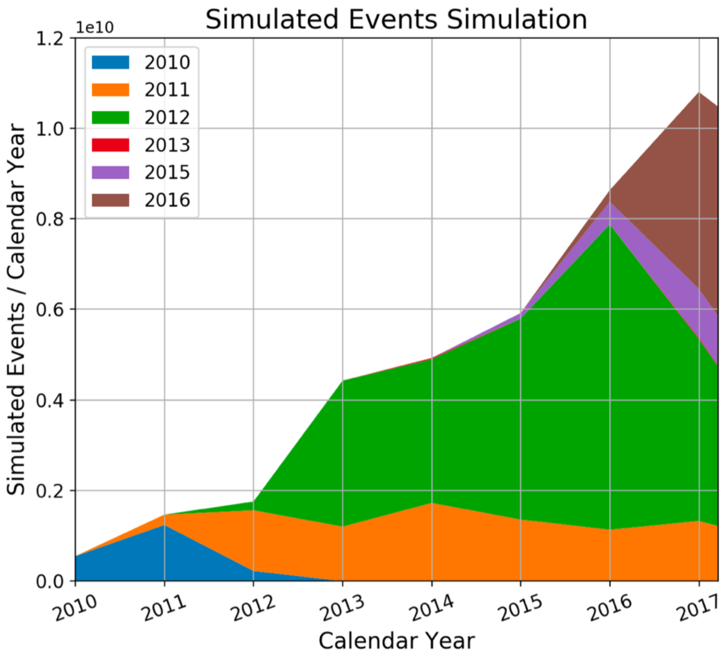
	2015	2016
simulated events/year ($\times 10^9$)	0.7	3.7
Recorded luminosity (fb^{-1})	0.3	1.6
Simulated events/ fb^{-1}/year ($\times 10^9$)	2.3	2.3

The model: main assumptions

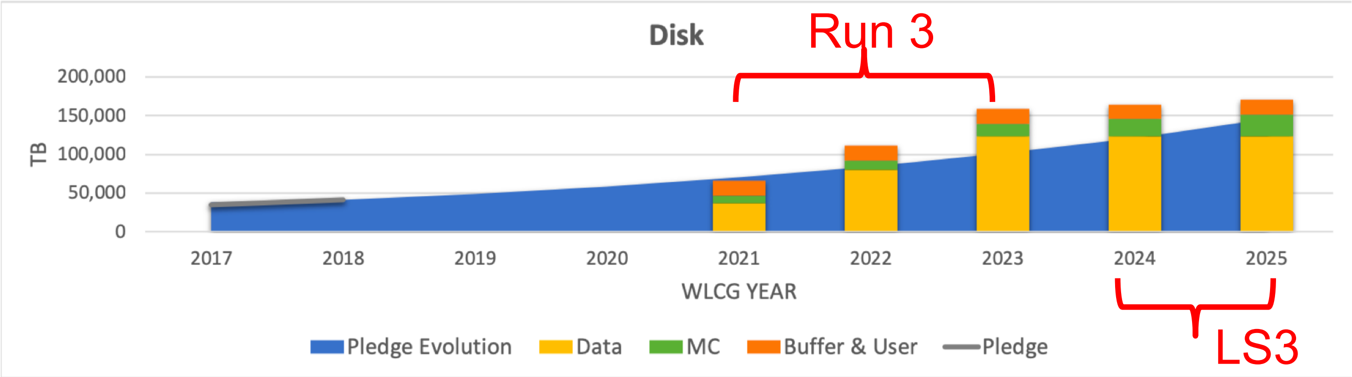
- Assumptions on simulated event volume
  - N. of MC events scales with L_{int}
  - MC production for a data taking years extends over the following 6 years
  - MC events saved in MDST format (x40 size reduction!)
- Assumptions on replicas

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



The model: storage requirements - disk



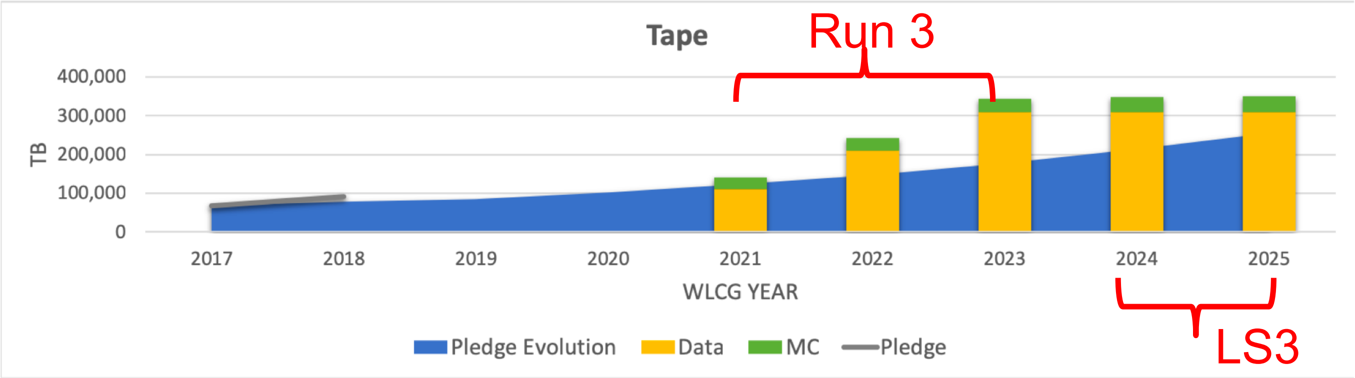
	WLCG Year	Disk	
		PB	Yearly Growth
Run 3	2021(*)	66	1.1
	2022	111	1.7
	2023	159	1.4
LS 3	2024	165	1.0
	2025	171	1.0
Average end of Run 3			1.4
Average end of LS 3			1.2

- Pledge evolution assumes a "constant budget" model (+20% more every year)
- Given as a gauging term

- Max deviation from this model: x1.6
- In line with the model by the end of LS3

(*) 2021 requests slighly updated to reflect current knowledge of LHC running conditions

The model: storage requirements - tape



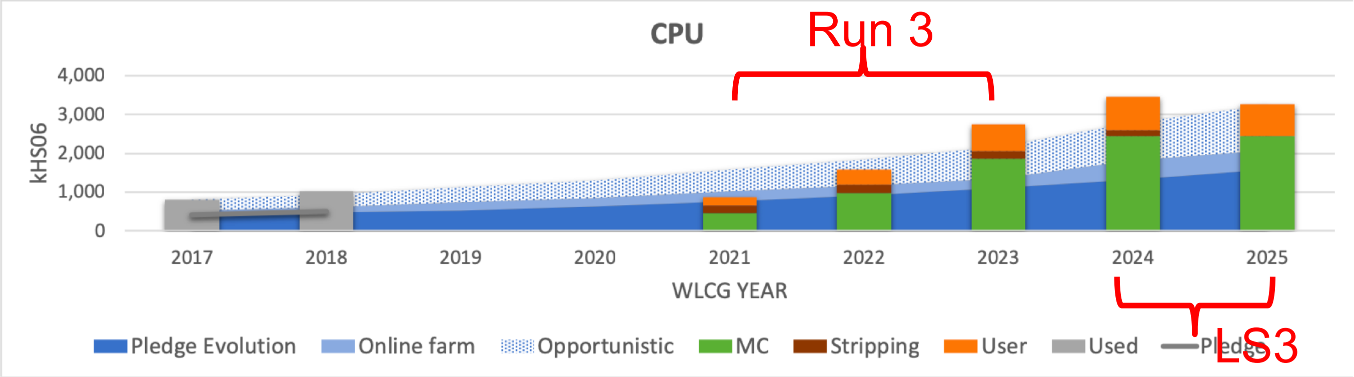
	WLCG Year	Tape	
		PB	Yearly Growth
Run 3	2021(*)	142	1.5
	2022	243	1.7
	2023	345	1.4
LS 3	2024	348	1.0
	2025	351	1.0
Average end of Run 3			1.5
Average end of LS 3			1.3

- Pledge evolution assumes a “constant budget” model (+20% more every year)
- Given as a gauging term

- Max deviation from this model: x1.9
- ~ in line with the model by the end of LS3

(*) 2021 requests slighly updated to reflect current knowledge of LHC running conditions

The model: CPU requirements



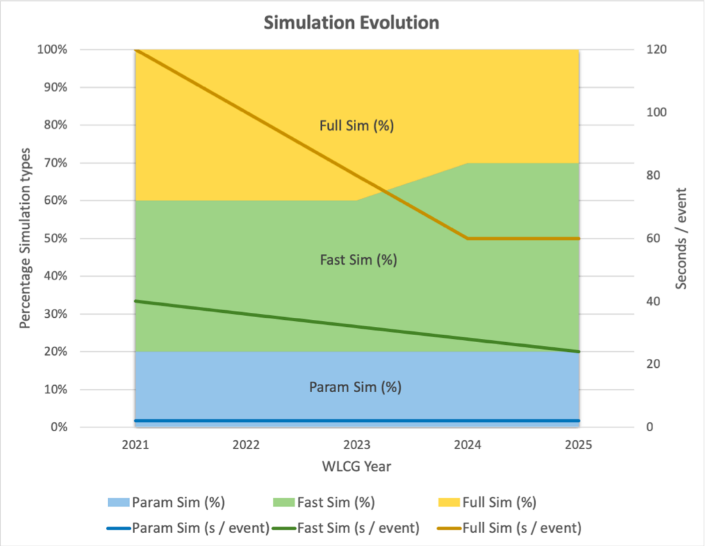
	WLCG Year	CPU	
		kHS06	Yearly Growth
Run 3	2021(*)	863	1.4
	2022	1.579	1.8
	2023	2.753	1.7
LS 3	2024	3.476	1.3
	2025	3.276	0.9
Average end of Run 3			1.6
Average end of LS 3			1.4

- Pledge evolution assumes a "constant budget" model (+20% more every year)
- Given as a gauging term
- Max deviation from this model: x2.5
- Plan to use opportunistic resources, which are however not granted
- Online farm used opportunistically when idle

(*) 2021 requests slighly updated to reflect current knowledge of LHC running conditions

Alternative: Aggressive simulation development

- 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



	WLCG Year	Percentage full/fast/parametric	Timing (s)	CPU kHS06	CPU Yearly Growth
Run 3	2021	40/40/20	120/40/2	863	1.4
	2022	40/40/20	100/36/2	1.423	1.6
	2023	40/40/20	80/32/2	2.051	1.4
LS 3	2024	30/50/20	60/28/2	1.844	0.9
	2025	30/50/20	60/24/2	1.542	0.8

The model: alternative options

	WLCG Year	Disk		Tape	
		PB	Yearly Growth	PB	Yearly Growth
Run 3	2021	58	1.0	142	1.5
	2022	95	1.6	243	1.7
	2023	134	1.4	345	1.4
LS 3	2024	140	1.0	348	1.0
	2025	146	1.0	351	1.0
Average end of Run 3			1.3		1.5
Average end of LS 3			1.2		1.3

Data parking

- Reduce disk need
- No effect on tape
- No effect on CPU
- Operationally challenging

	WLCG Year	Disk		Tape	
		PB	Yearly Growth	PB	Yearly Growth
Run 3	2021	67	1.1	129	1.4
	2022	114	1.7	205	1.6
	2023	164	1.4	282	1.4
LS 3	2024	170	1.0	285	1.0
	2025	176	1.0	288	1.0
Average end of Run 3			1.4		1.5
Average end of LS 3			1.2		1.3

Reduced HLT output bandwidth

- Reduces tape need
- Sub-optimal use of Turbo + Stripping: may result in slightly larger disk need!
- No effect on CPU
- Effect on physics

The model: risk analysis

- The largest storage requirements concern tape which is relatively cheap
- Can reduce the tape need reducing HLT output BW
  - Impact on physics reach
  - Requires very aggressive use of Turbo
  - It comes with no gain on expensive resources (disk, CPU)

- Mitigation of disk resources can be achieved with data parking
  - Operationally challenging: need high tape throughput or very long processing time (driven by tape staging time)
  - Impact on experiment's competitiveness, long waiting times to access data sets

- CPU needs can be reduced with aggressive use of faster simulation
  - Needs a lot of development
  - No guarantee yet that we can achieve the assumed time/event

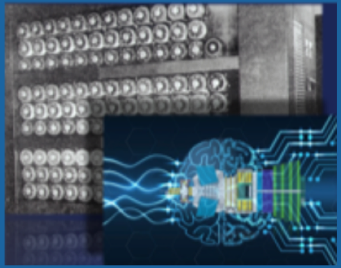
Conclusions

- The LHCb Upgrade experiment will collect ~x10 signal yields than the current LHCb
- An extrapolation of the current LHCb data rates would yield x30 in data volume
- 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 computing resources
  - Maintain the full breadth of the physics programme
  - Flexible: can incorporate future technology advancements

Outlook

- Several changes ahead towards Run3 physics analysis
- Many stripping lines will have to be converted to HLT2 lines
  - …and optimised for speed
- A single selection framework is being built for both HLT2 and the successor of “stripping”
  - Join the upcoming hackathon and start testing!

- The workflow for user analysis will be overhauled
  - Centralised production instead of «chaotic» user submissions
  - Building on the experience gained with «working group productions» in the past few years



12th LHCb Computing Workshop

https://indico.cern.ch/event/831054/

18-22 November 2019
CERN
Europe/Zurich timezone

Search...

- Overview
- Timetable
- Registration
- Participant List

The 12th LHCb computing workshop will be held at CERN, starting in the afternoon of Monday, November 18th and ending at lunchtime on Friday November 22nd.

The Programm will consist of plenary sessions only, in the domains of

- simulation
- computing infrastructure, monitoring and documentation
- core software
- RTA
- distributed computing
- offline analysis

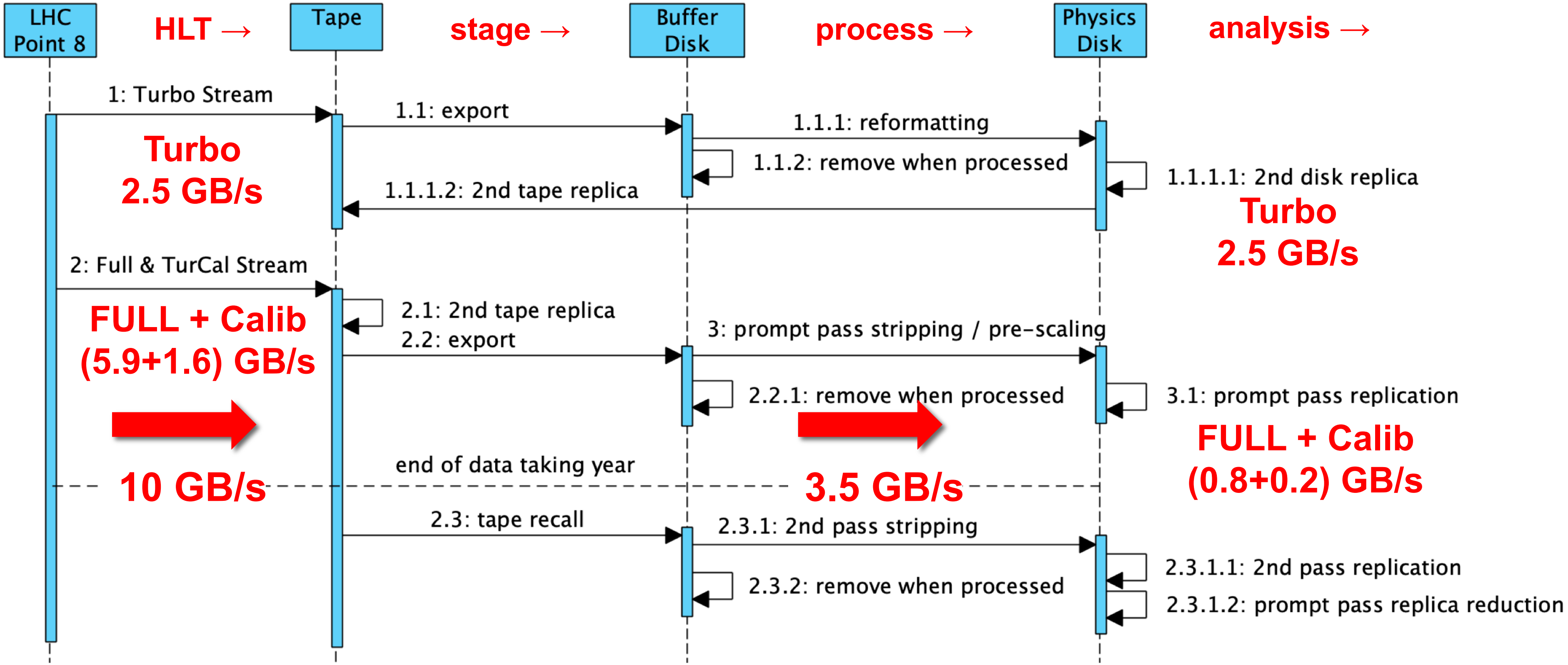
Registration is open
Please register, even if the event will be held at CERN

Starts 18 Nov 2019, 14:00
Ends 22 Nov 2019, 13:00

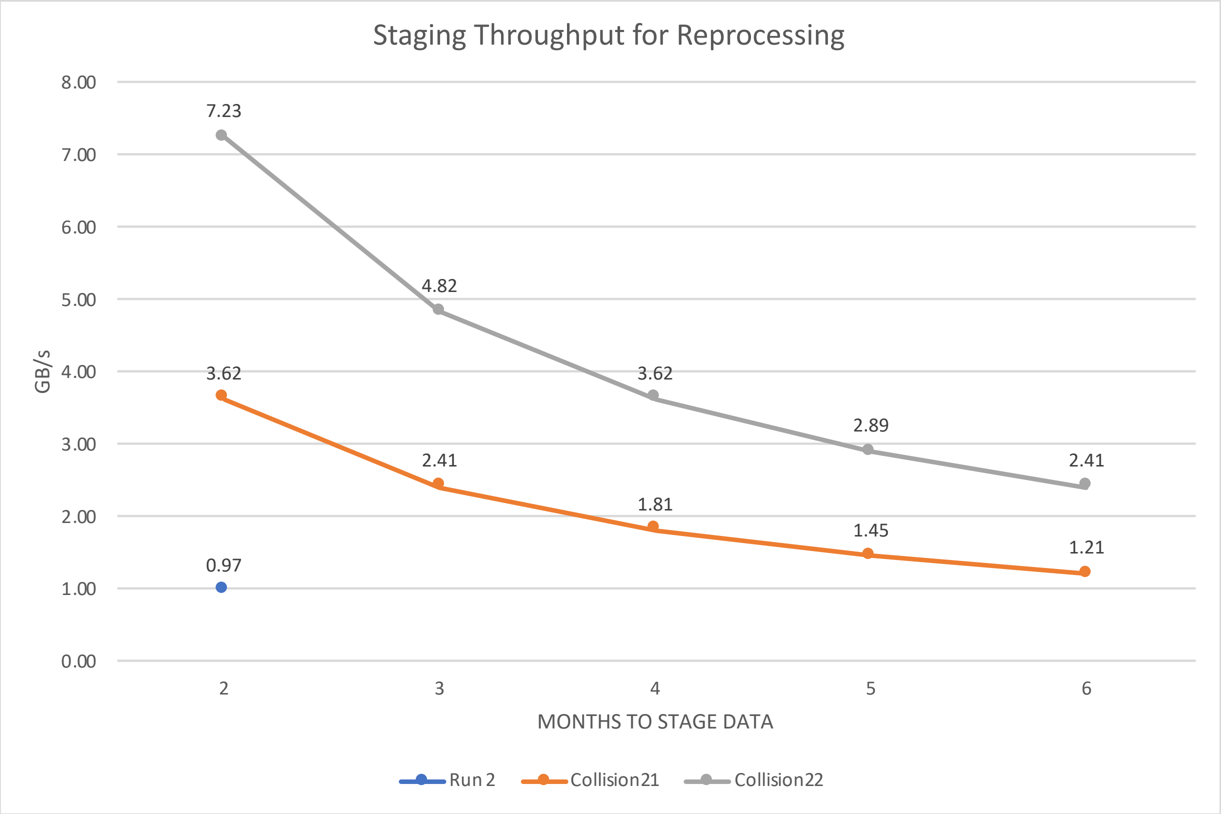
CERN
222/R-001

Backup

Data Processing Workflow per Data Taking Year



Tape Reading Throughput for Reprocessing



Offline computing requests for 2021

- Preliminary requests have been sent to the C-RSG
- Same model as in LHCb Upgrade Computing Model TDR
  - Minor adjustments following latest prescriptions on instantaneous ($1x10^{33}$) and integrated ($3fb^{-1}$ baseline, $7fb^{-1}$ contingency) luminosities
  - Contingency used for tape requests only
    - Large increase
- N.B.: 2020 pledges due by September 30th

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**