Offline computing requirements for Run4, Run5, Run6

5th Workshop on LHCb Upgrade 2

April 1st 2020

Concezio Bozzi

INFN

Istituto Nazionale di Fisica Nucleare

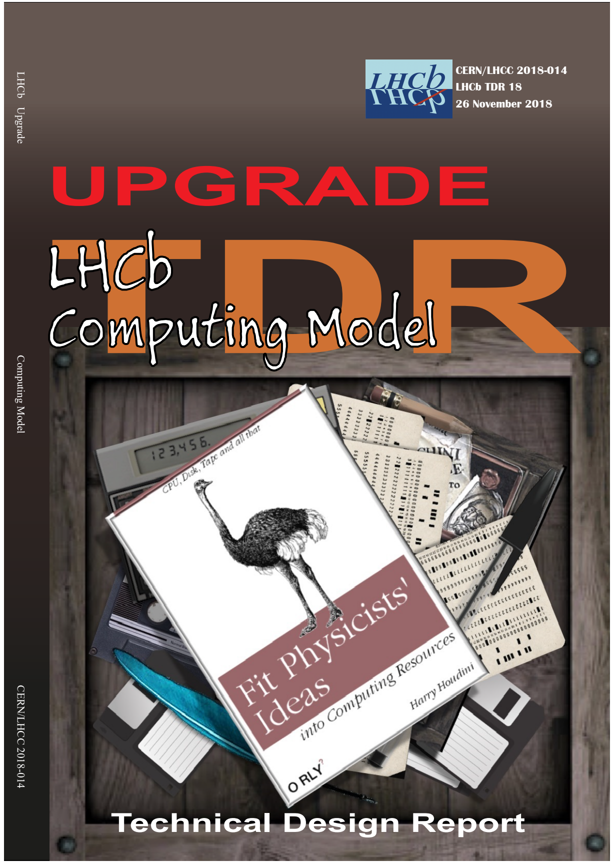
LHCb

Disclaimer

- Looking 10+ years ahead in software and computing is challenging
- Many emerging technologies, in both computing, see e.g. Conor’s talk, and storage
- Paradigm shifts ahead, e.g.
  - in-memory and neuromorphic computing
  - Non-volatile memory systems
  - Advanced tape technologies
  - Quantum technologies
- Not clear if and how all these will fit in the offline computing environment of HEP experiments
- In the following, an attempt is made to extrapolate the U1 computing model to U1b and U2, with the only purpose of qualitatively establishing (or not) its viability

Overview

- Upgrade 1 computing model
- Extrapolation assumptions
- Resource requirements
- Outlook



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

Upgrade 1: storage scales with bandwidth

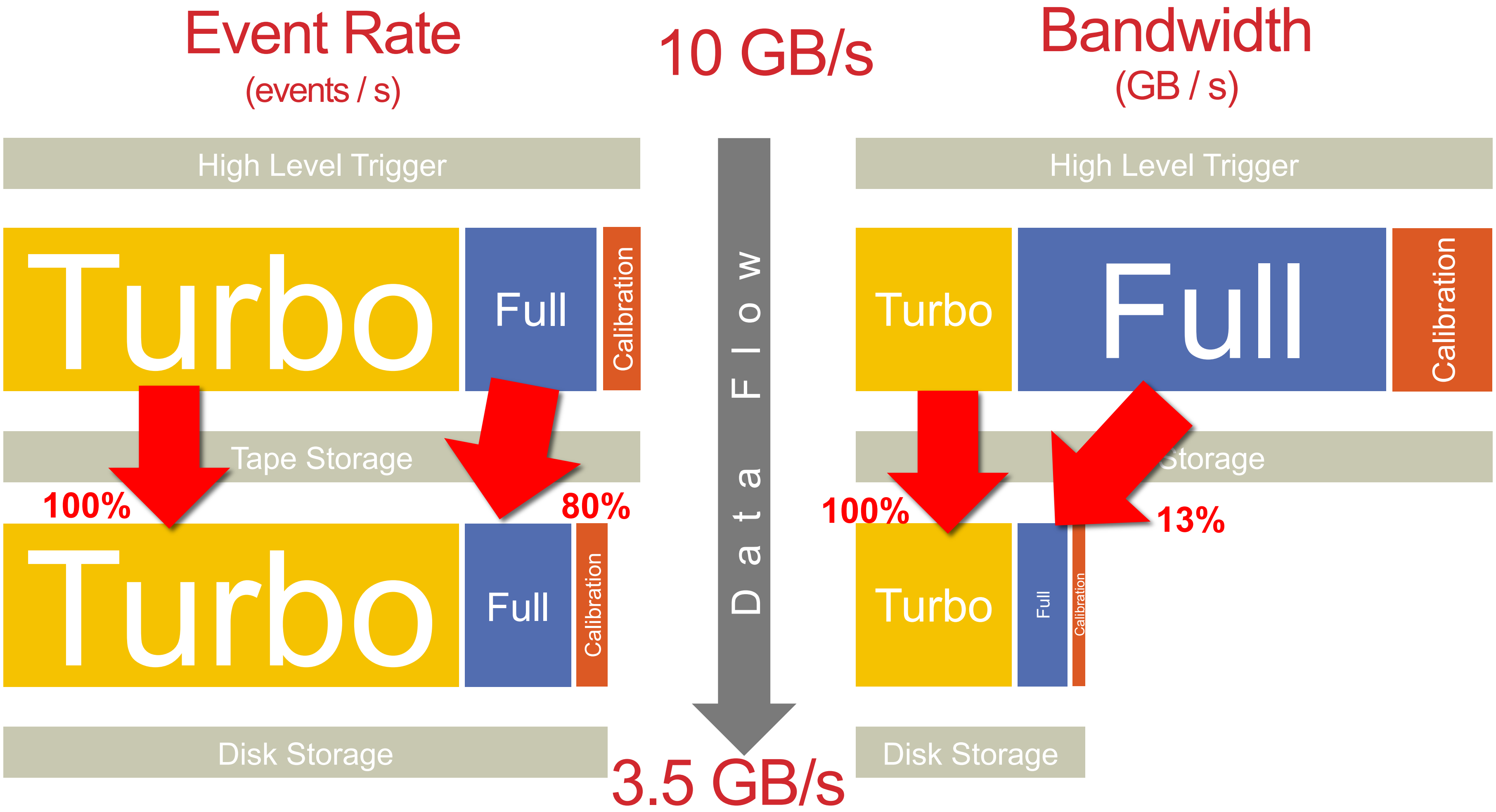
- Bandwidth from online to offline:10 GB per live LHC second
  - Saved to to tape
- Reduce by ~1/6 FULL and Calibration data volume with “sprucing”
  - **3.5 GB/s saved** to disk

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%

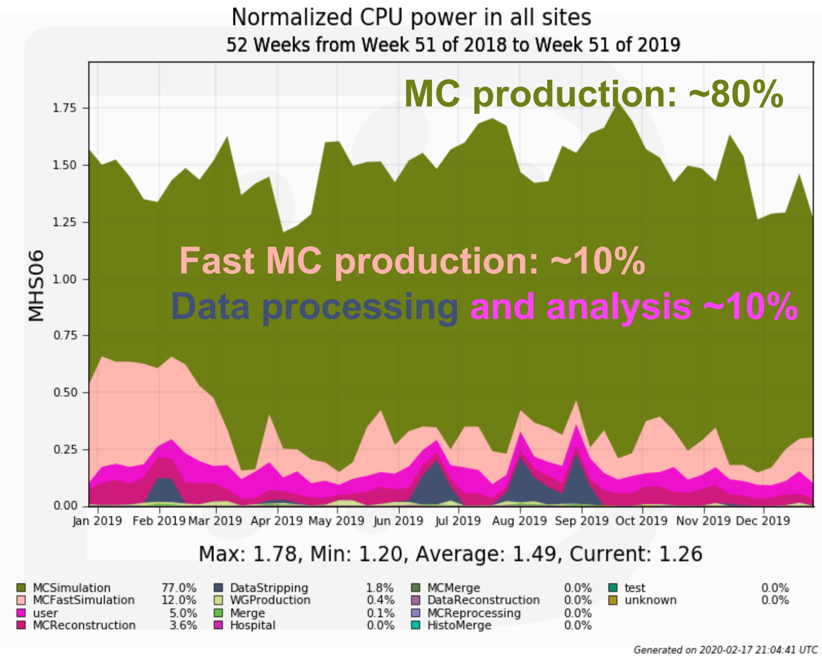
Throughput to disk

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


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

Upgrade 1: CPU dominated by MC production

- MC simulation is the main consumer and it will stay so in the future
- Current MC production is scaled to estimate the CPU needs in Upgrade 1
  - Simulation of a given data taking year continues during the following 6 years, starting slowly and ending gracefully
- 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



LHCb-PUB-2020-001

Upgrade 1 Computing Model parameters

- Storage accommodates trigger output BW of 10 GB/s
  - Fully saved on tape
  - Reduced to 3.5GB/s on disk after sprucing FULL and TURCAL streams
- CPU dominated by MC production
  - Massive use of fast(er) simulation techniques

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]	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)

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

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

Resource requirements: Upgrade 1

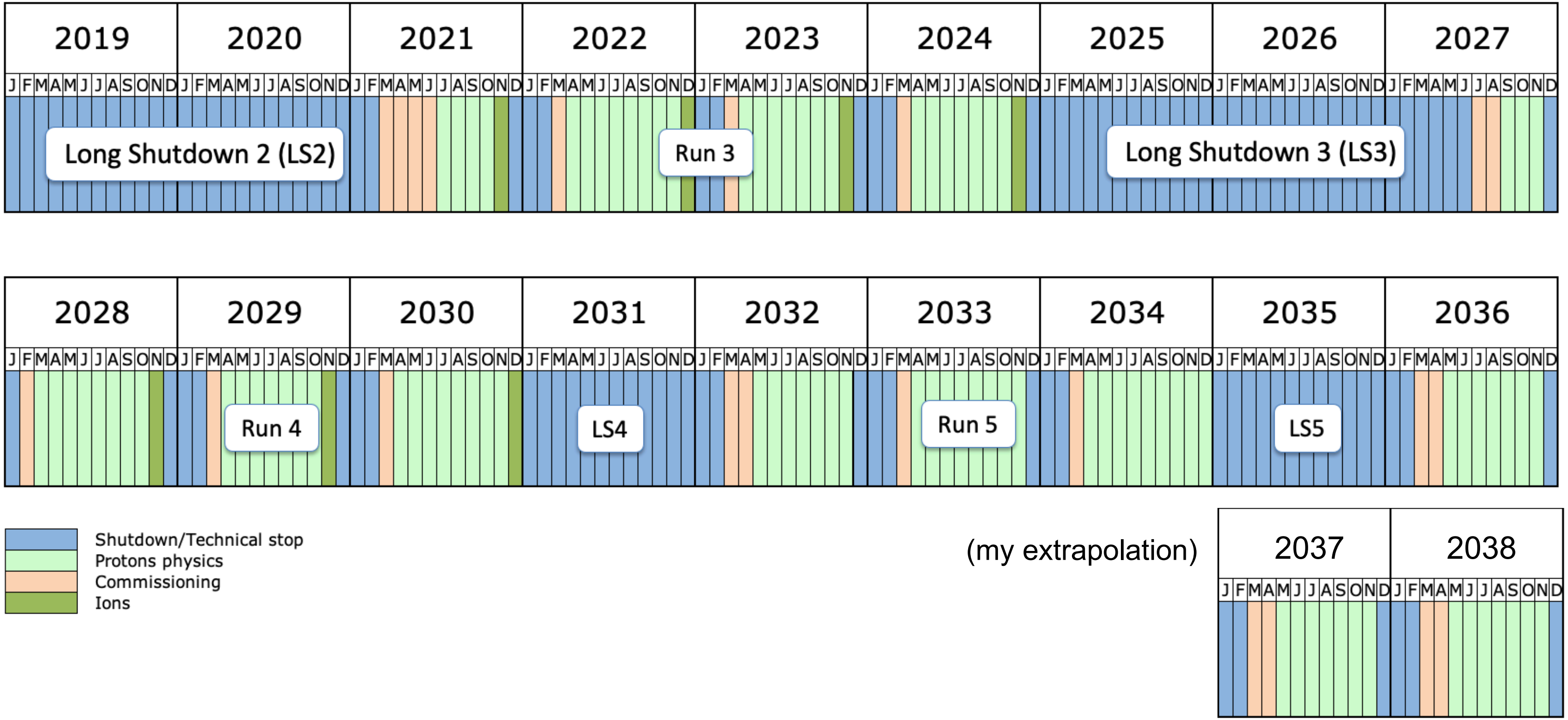
- Taking into account the new LHC schedule: 2024 is a running year
- Pledge evolution assumes a "constant budget" model of +10% more every year
  - Given as a gauging term
  - This used to be +20%
- As a consequence, no longer on flat budget at the end of LS3

Evolving to Upgrade 1b / Upgrade 2

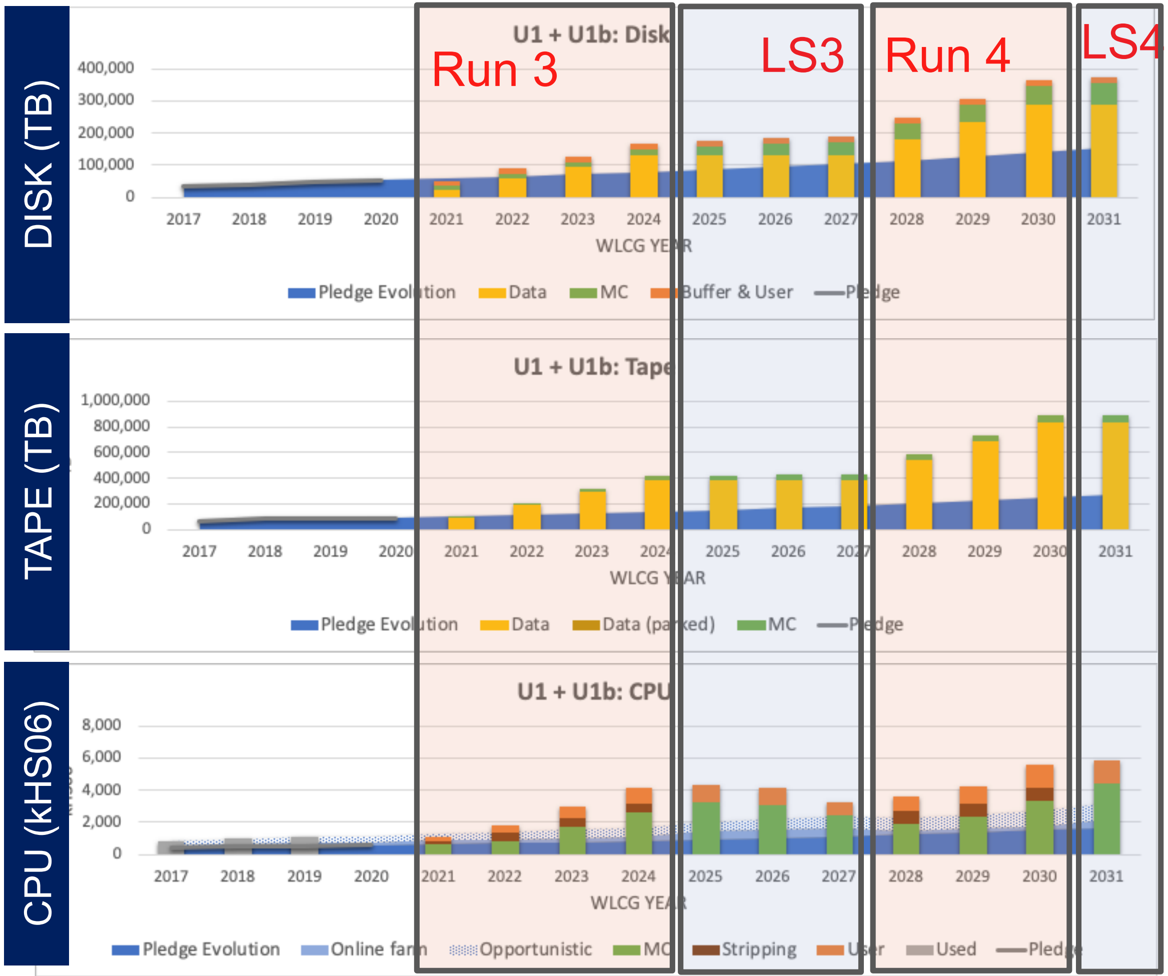
- Scale bandwidth to offline with luminosity: x1.5 (U1b), x10 (U2)
- Apply same factors to scale the number of events to simulate
- Other parameters of the Upgrade 1 computing model are kept unchanged, e.g.
  - Number of disk replicas
  - MC simulation model
  - Bandwidth reduction factor from sprucing
  - Bandwidth division between TURBO:FULL:TURCAL
  - CPU work for MC, stripping and analysis
    - Assume that increase due to increased event complexity is balanced by code speed-up

- Keep it simple, give a ballpark estimate

LHC schedule

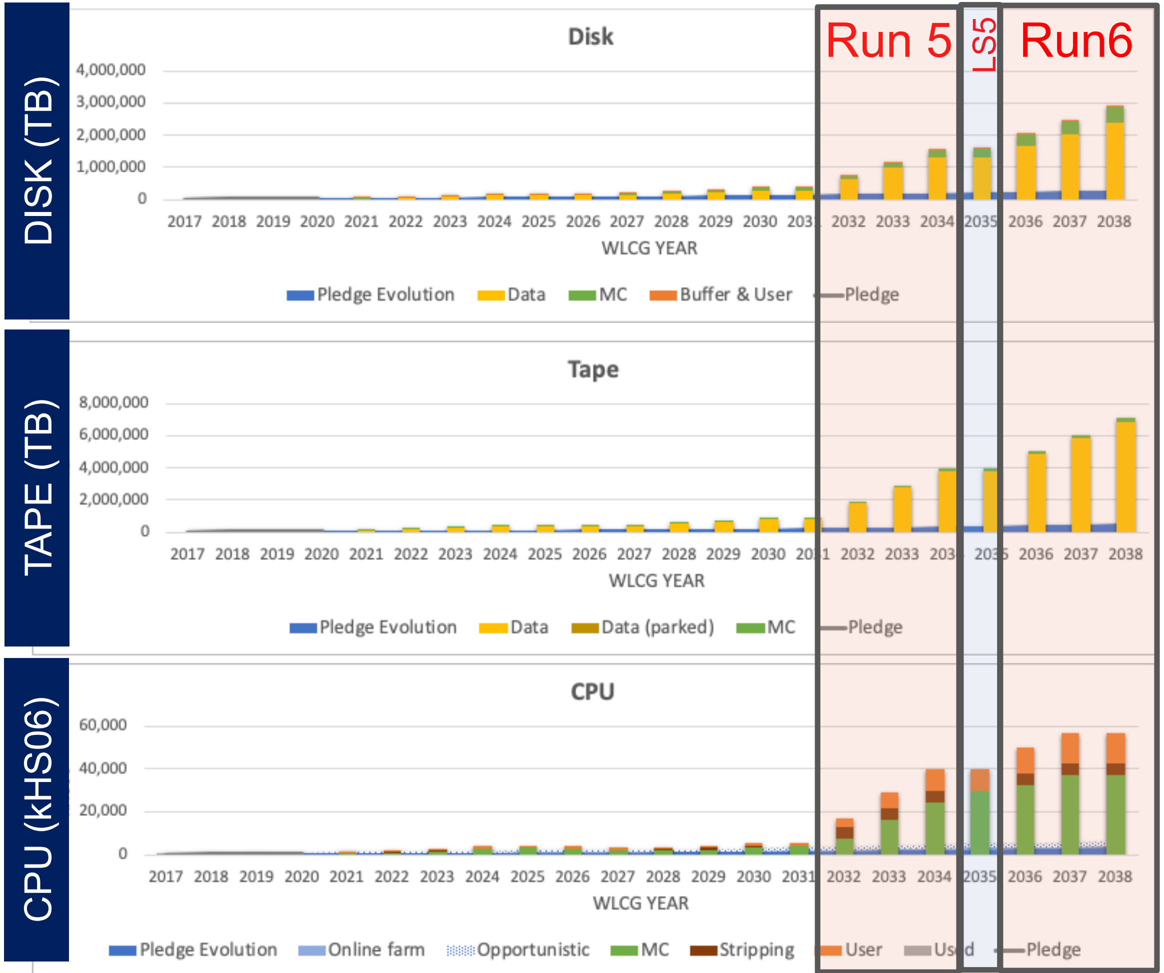


Resource requirements: U1 + U1b



- Overshoot wrt “constant budget” increases (~2x)
- At LS4:
  - Disk: 375PB
  - Tape: 900PB
  - CPU: 5MHS06
    - ~500k cores
- Compare with e.g. ATLAS+CMS 2020 pledges:
  - Disk: 400PB
  - Tape: 600PB
  - CPU: 5MHS06

Resource requirements: U1 + U1b + U2



- Entering a different regime
  - Storage: a few exabytes
    (x10 wrt constant budget)
  - Compute: tens of MHS06
    (x20 wrt constant budget)
- End of Run5 (Run6):
  - Disk: 1.5 (2.9) EB
  - Tape: 4 (7) EB
  - CPU: 40 (57) MHS06
    - ~4M (6M) cores
- Compare with e.g. ATLAS(*) end of Run4 (Run5):
  - Storage: 4 (8+) EB
  - CPU: 60 (80) MHS06

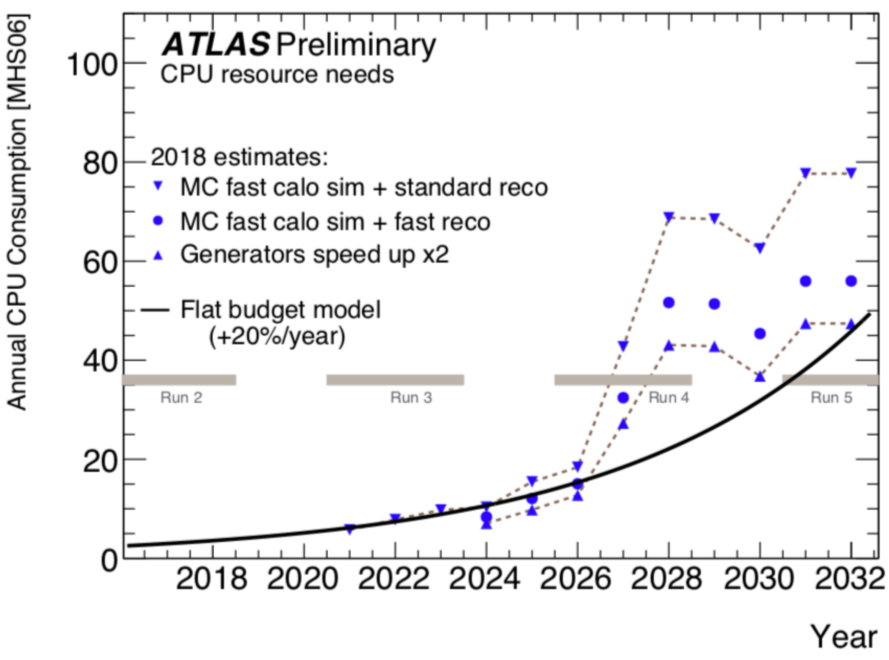
(*) see backup: baseline model, before mitigations

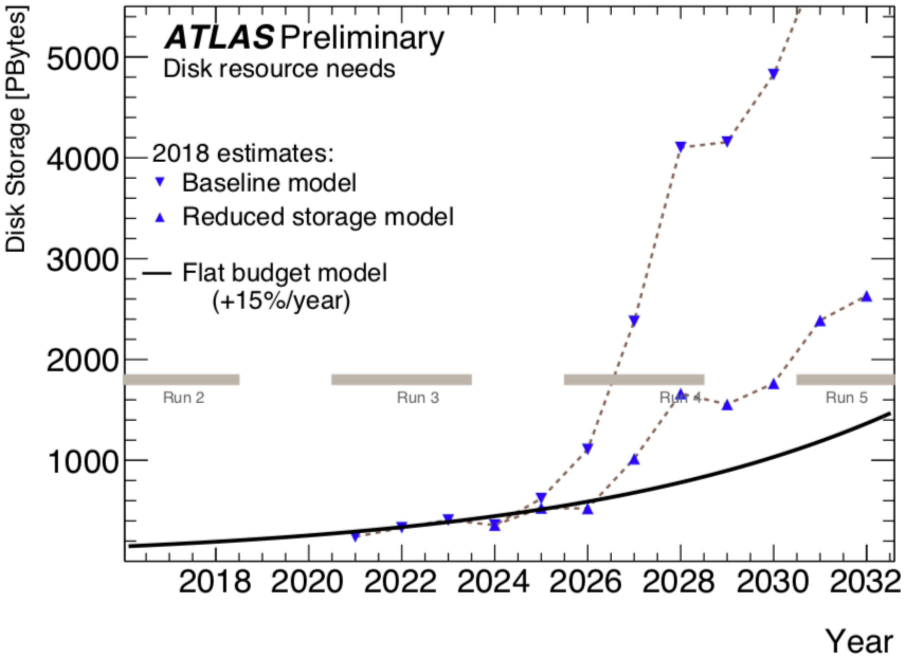
Outlook

- The U1 computing model might still be sustainable for U1b, with some readjustments
- It is definitely not sustainable for U2
  - Turning the usual handles (e.g. replicas, filtering) will not buy the required mitigation factors
  - Will a significant reduction of the bandwidth to offline be possible without compromising our physics programme?
  - Is there such a need for simulation? Can we reduce the number of simulated events?
  - Will there be one (or more) technological white knight(s) coming to rescue?
- We need to invest in R&D and re-think our strategy
  - ATL+CMS are already doing this for Run4

Backup

ATLAS HL_LHC


ATLAS Preliminary
CPU resource needs
2018 estimates:
MC fast calo sim + standard reco
MC fast calo sim + fast reco
Generators speed up x2
Flat budget model
(+20%/year)
Run 2
Run 3
Run 4
Run 5
Annual CPU Consumption [MHS06]
0
20
40
60
80
100
2018
2020
2022
2024
2026
2028
2030
2032
Year


ATLAS Preliminary
Disk resource needs
2018 estimates:
Baseline model
Reduced storage model
Flat budget model
(+15%/year)
Run 2
Run 3
Run 4
Run 5
Disk Storage [PBytes]
1000
2000
3000
4000
5000
2018
2020
2022
2024
2026
2028
2030
2032
Year

TOP 500 The List.

PRESENTED BY ICL Innovative Computing Laboratory, Berkeley Lab Lawrence Berkeley National Laboratory, ISC Group Moving Forward.

FIND OUT MORE AT top500.org

	SYSTEM	SPECS	SITE	COUNTRY	CORES	RMAX PFLOP/S	POWER MW
1	**Summit**	IBM POWER9 (22C, 3.07GHz), NVIDIA Volta GV100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/SC/ORNL	USA	2,414,592	148.6	11.4
2	**Sierra**	IBM POWER9 (22C, 3.1GHz), NVIDIA Tesla V100 (80C), Dual-Rail Mellanox EDR Infiniband	DOE/NNSA/LLNL	USA	1,572,480	94.6	7.44
3	**Sunway TaihuLight**	Shenwei SW26010 (260C, 1.45 GHz) Custom Interconnect	NSCC in Wuxi	China	10,649,600	93.0	15.4
4	**Tianhe-2A (Milkyway-2A)**	Intel Ivy Bridge (12C, 2.2 GHz) & TH Express-2, Matrix-2000	NSCC Guangzhou	China	4,981,760	61.4	18.5
5	**Frontera**	Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR	TACC/U of Texas	USA	448,448	23.5	-

Performance Development