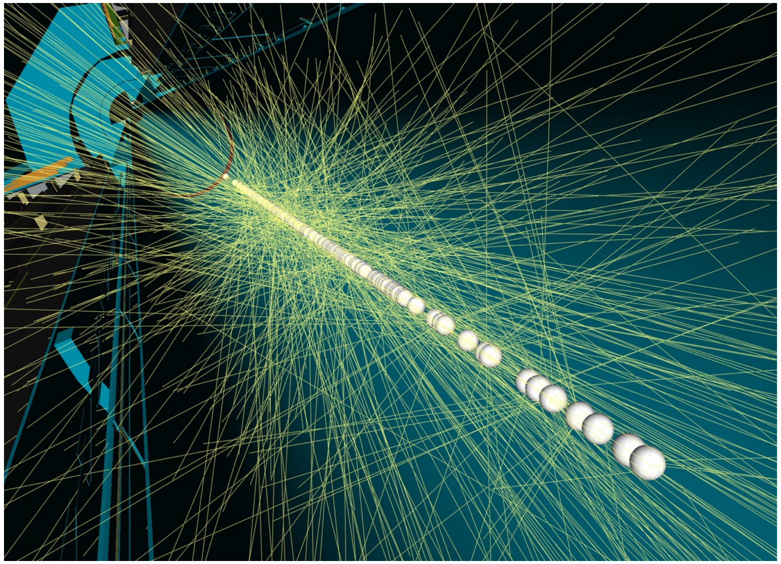
Models and Requirements for LHC Computing 2020-2030

- Introduction
- WLCG evolution and Hardware Trends
- Plans Alice & LHCb – challenge @ Run3
- Plans ATLAS & CMS – challenge @ Run4
- Conclusions

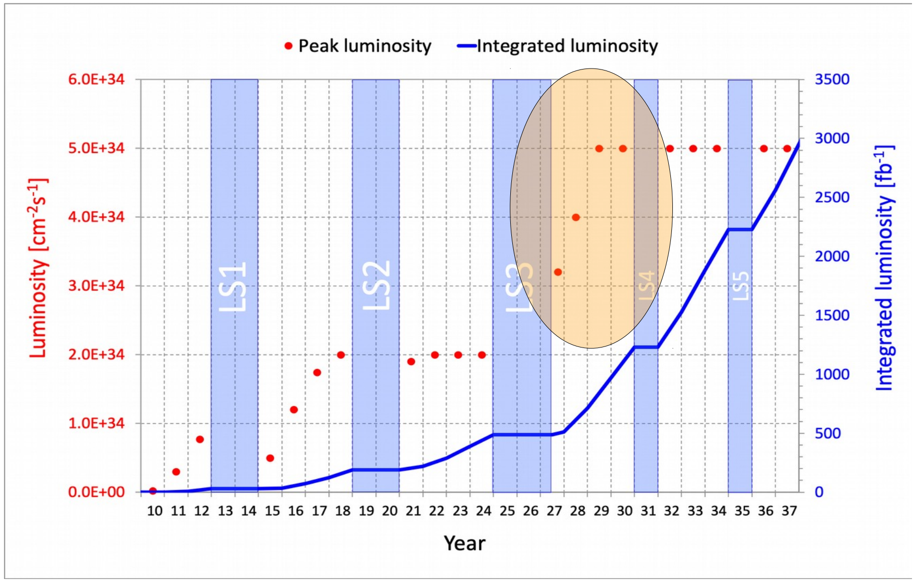


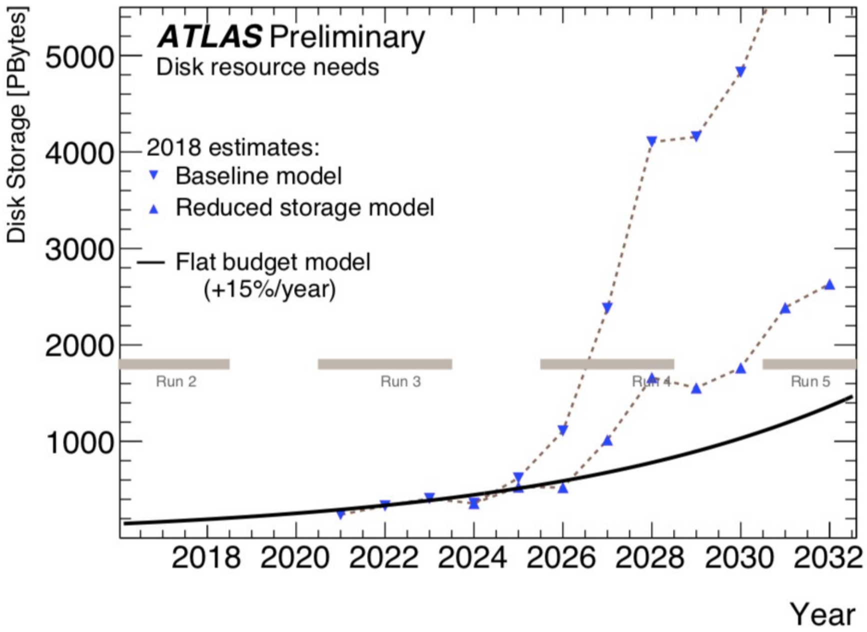
Input and material form J.Marks, K.Schwarz, C.Wissing, M.Elsing and many others

Please note: 3 pages different/excluded in indico version
due to not-yet public status of material

Introduction

- Well-known massive increase of luminosity and detector upgrades for Run-4
  - affecting ATLAS&CMS after 2027
- 1st resource projections looked scary
  - ~factor 10 beyond flat budget
- HL-LHC computing review in progress right now
  - detailed CDRs close to final for ATLAS, CMS, WLCG & Doma
- For ALICE & LHCb major changes already w/ Run-3 → separate issue



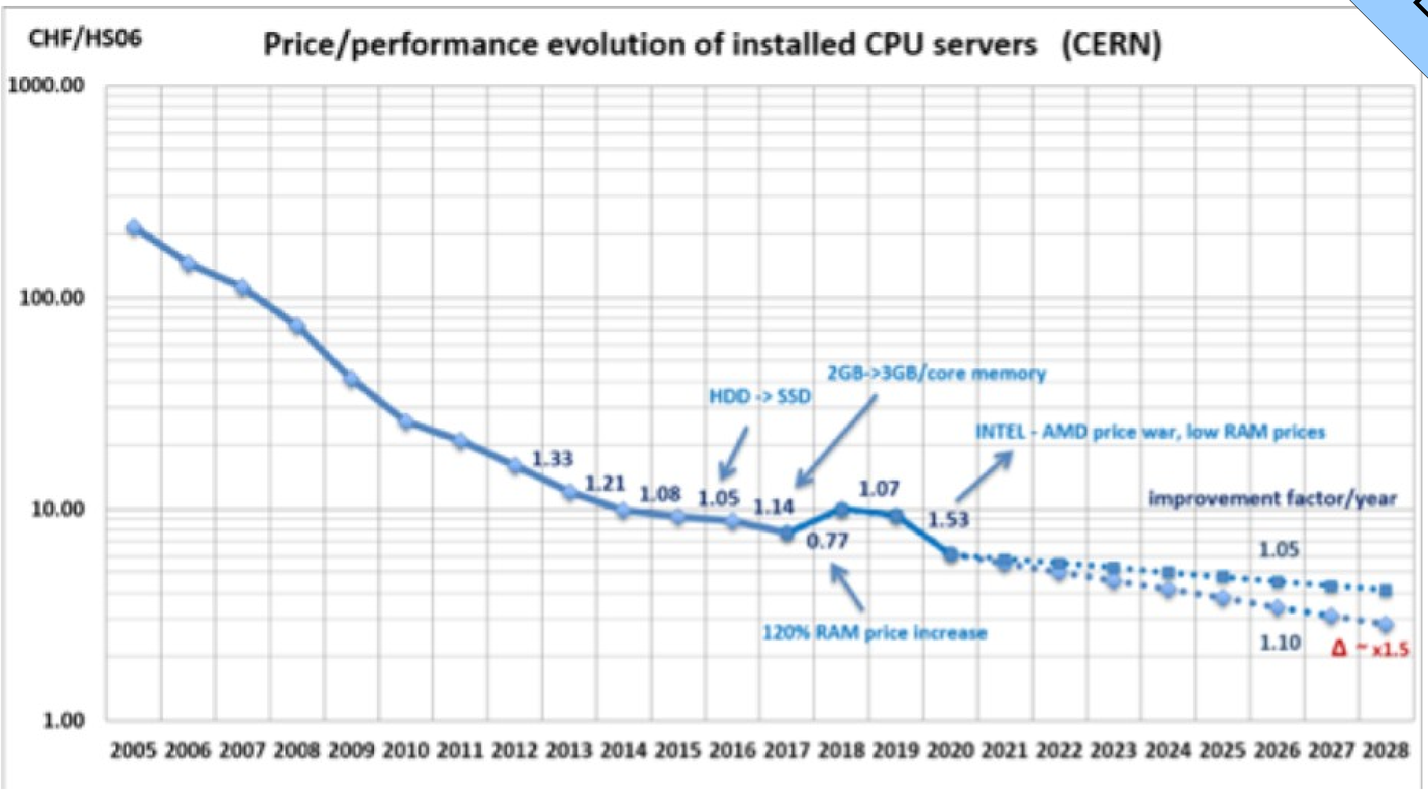


WLCG/Resources perspective

- WLCG will enter ExaByte data processing regime with HL-LHC
- Limited budget
  - assume flat budget for LHC computing
- Large uncertainties on hardware evolution
  - traditional assumptions on efficiency increase ~15-20%/year not on solid grounds
  - use of accelerators, impact of ML, future of tapes, ...
- Enhance common use of software tools and services
  - data management, event generation, detector simulation, reconstruction, ...
- Other scientific domains approaching similar resource requirements
  - enhance cooperation

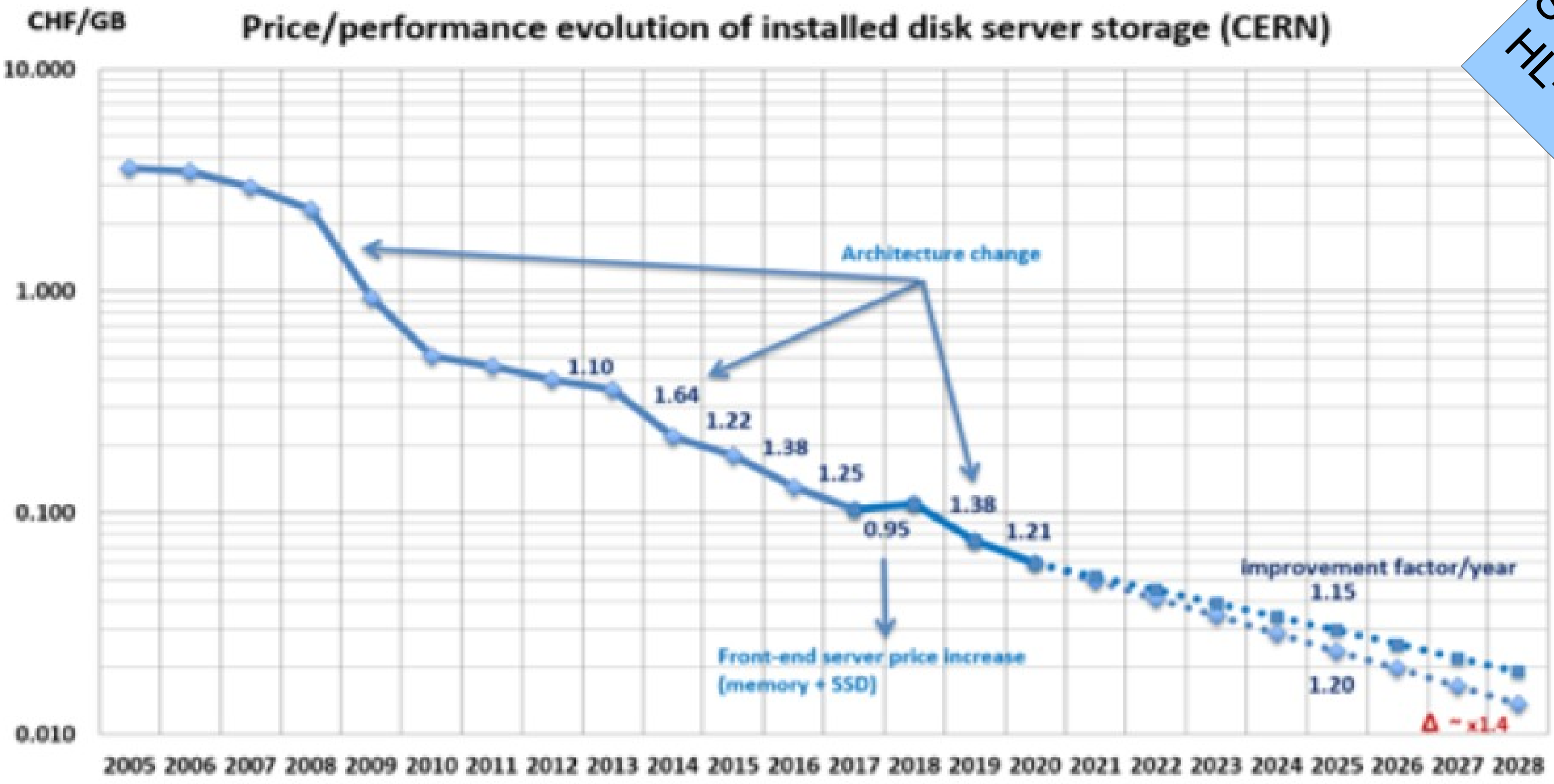
Technology evolution – CPU

Source: WLCG HL-LHC report



- Clear slow-down of efficiency increase for CPU in recent years
  - 10%/y seems already optimistic ...
- Similar trends for GPU accelerators

Technology evolution – Disk



Source: WLCG HL-LHC report

- Rather irregular performance evolution for disks
  - new recording techniques on the horizon (HAMR, MAMR)
  - clearly better trend than for CPU:
    ~15%/y might be realistic

Technology evolution – Tape & Network

- Tapes:
  - still factor ~5 cheaper in cost per TB
    - optimized data placement: move data Disk → Tape
    - important factor in potential storage cost-savings
  - technology wise stable progress and good prospects
  - but shrinking market and very few manufacturers
  - longer-term availability unclear:
    - huge cost increase in case of Tape → Disk
- Networking
  - evolution still fast, expected growth ~35%/year
  - presumably least critical component for HL-LHC

WLCG Computing Model evolution

- Addressed in DOMA (Data Organisation, Management and Access) project
  - from traditional Tier-1/2/3 model
    - with strictly defined roles and responsibilities
  - to a more flexible model with federations and data-lakes, CPU-only sites with cache-storage, opportunistic HPC and Clouds
    - less storage endpoints, lower operation effort and cost
  - DE case:
    - new/extended role for Helmholtz centers to act as data-lake center
- Tools and Technologies discussed and evaluated
  - QoS (Quality of Service) model
    - more fine-grained storage classification (not just disk & tape)
    - Data Carousel – just-in-time staging of needed data to disk
  - Caching services, e.g. XCache
  - non-SRM Third-Party-Copy (xrootd, http protocols)

Plans and Requirements for Alice and LHCb

- Huge changes in operation and requirements already needed for upcomping Run-3
- Most effort focused on preparing for this phase
- So far assume ~flat budget evolution sufficient from Run-3 → Run-4
- No dedicated Computing CDRs at this stage from Alice and LHCb for HL-LHC

ALICE upgrade for Run 3/4

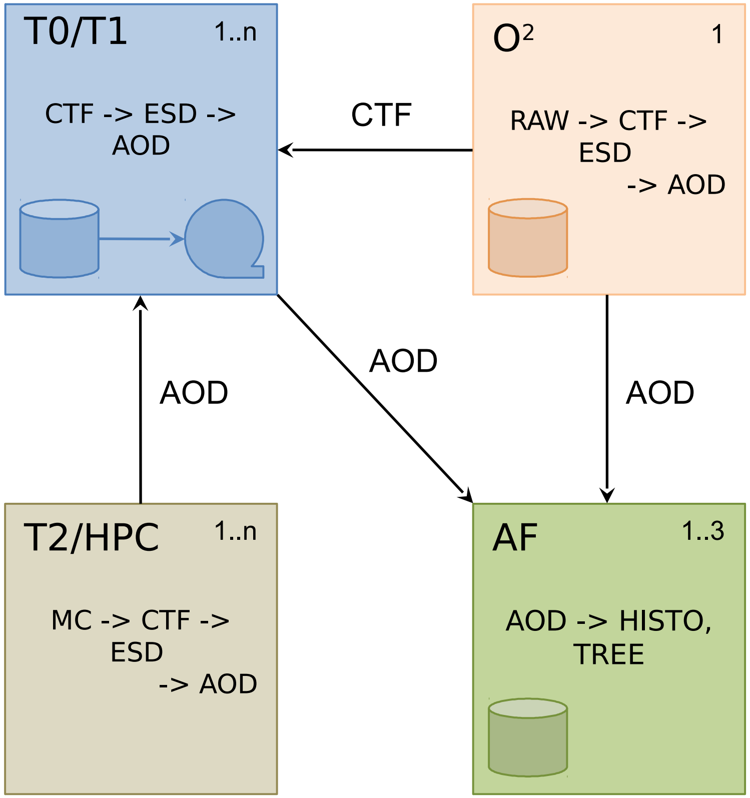
- ALICE is in the critical phase of the Run 3/4 upgrade preparation
- Aim of the hardware upgrade in Run3/4 is to boost amount of collected collisions by factor ~100
- Operating ALICE will require cardinal change in approach to data processing
  ⇒ **O2 framework**
- Main challenges for handling the data to come:
  - The volume and the rate of data to store:

  ⇒ requires combination of data reduction and compression by factor >35
  - **(~3.5 T/s → <100 GB/s)** ⇒~50 PB/y from Pb-Pb and reference pp data
  - Computing power to process these data:

  ⇒ requires drastic improvements in reconstruction code (factor ~10 already achieved wrt Run2) and adapting heaviest operations to GPU processing

Recap of ALICE Run 3/4 Computing Model

Reconstruction
Calibration
Archiving
Analysis

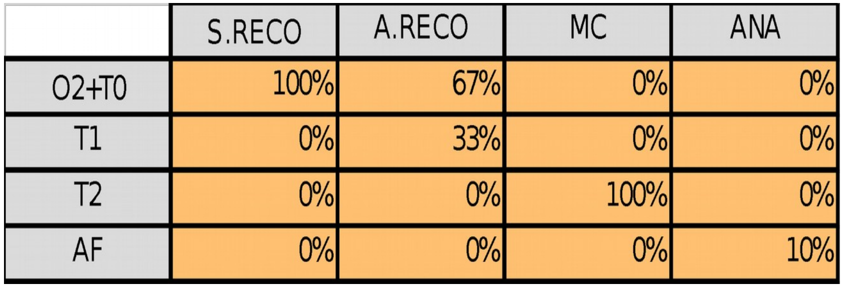


Calibration
Reconstruction
Compression

Simulation

Analysis

O2 TDR, Ch. 4



	S.RECO	A.RECO	MC	ANA
O2+T0	100%	67%	0%	0%
T1	0%	33%	0%	0%
T2	0%	0%	100%	0%
AF	0%	0%	0%	10%

- Subject to fine tuning
- MC can be run as a backfill

2/3s of CTFs processed by O² + T0 and archived at T0;
1/3 of CTFs exported, archived and processed on T1s;

One calibration (sync.) and two reconstruction passes (async.) over raw data each year;

CTFs removed from disk before a new data taking period starts; Only AODs are kept on T0/T1 disk and archived to tape;

10% of AODs sampled and sent to the Analysis Facility for quick analysis and cut tuning;
Analysis of full data sample across T0/T1s only upon Physics Board approval.

Resource Requirements for 2020 and for 2021

ALICE		2018		2019		2020		2021		
		Req.	C-RSG	Req.	C-RSG	Req.	2020/2019 CRSG	Req.	2021/2020 Req	**Annual Growth 2018 -> 2021 (*)**
CPU [kHS06]	Tier-0	350	350	430	430	350	-18.60%	471	34.60%	**10.40%**
	Tier-1	307	307	365	365	365	0.00%	498	36.40%	**17.50%**
	Tier-2	313	313	376	376	376	0.00%	515	37.00%	**18.06%**
	Total	**970**	**970**	**1171**	**1171**	**1091**	**-6.80%**	**1484**	**36.00%**	**15.23%**
Disk [PB]	Tier-0	26.2	26.2	34.3	34.3	31.2	-9.00%	45.5	45.80%	**20.20%**
	Tier-1	30.5	30.4	37.9	37.9	44	16.10%	53.3	21.10%	**20.45%**
	Tier-2	29.6	29.7	33.9	33.9	39	15.00%	44.8	14.90%	**14.81%**
	Total	**86.3**	**86.3**	**106.1**	**106.1**	**114.2**	**7.60%**	**143.6**	**25.70%**	**18.50%**
Tape [PB]	Tier-0	49.1	49.1	44.2	44.2	44.2	0.00%	86	94.60%	**20.54%**
	Tier-1	40.9	42.2	37.7	37.7	37.7	0.00%	57	51.20%	**11.70%**
	Total	**90**	**91.3**	**81.9**	**81.9**	**81.9**	**0.00%**	**143**	**74.60%**	**16.69%**

(*) annually compounded rate

Long-term estimates for Run 3/4

- Pre-covid-19 estimates: step of requests in 2021, but CPU, disk and tape will likely stay compatible with standard resources growth (15-20%) for Run3, Run4 and LS3
- The ALICE computing requests are subject to a scrutiny process, which is by construction limited to two years in the future
- Software algorithms and updated computing model allow to fit into the standard Grid resource growth for Run 3/4 for CPU/disk
- Tape requirement will depend on the data taking: archive on T1 tape 1/3 of CTF's and AOD's, i.e. 22 PB per HI period (about 5.5 PB for GridKa)
- Year-by-year adjustments will be made following the regular C-RSG/RRB process.

LHCb Run 3 / 2021

- LHCb
  - Neues Computing Model:

- Das LHCb upgrade data processing (Reconstruction, Kalibration und Alignment) wird Online durchgeführt und liefert 3 Output Ströme

  - FULL stream (RDST, optionally RAW)
  - Turbo stream (TurboRaw)
  - TurCal (RDST+ RAW)

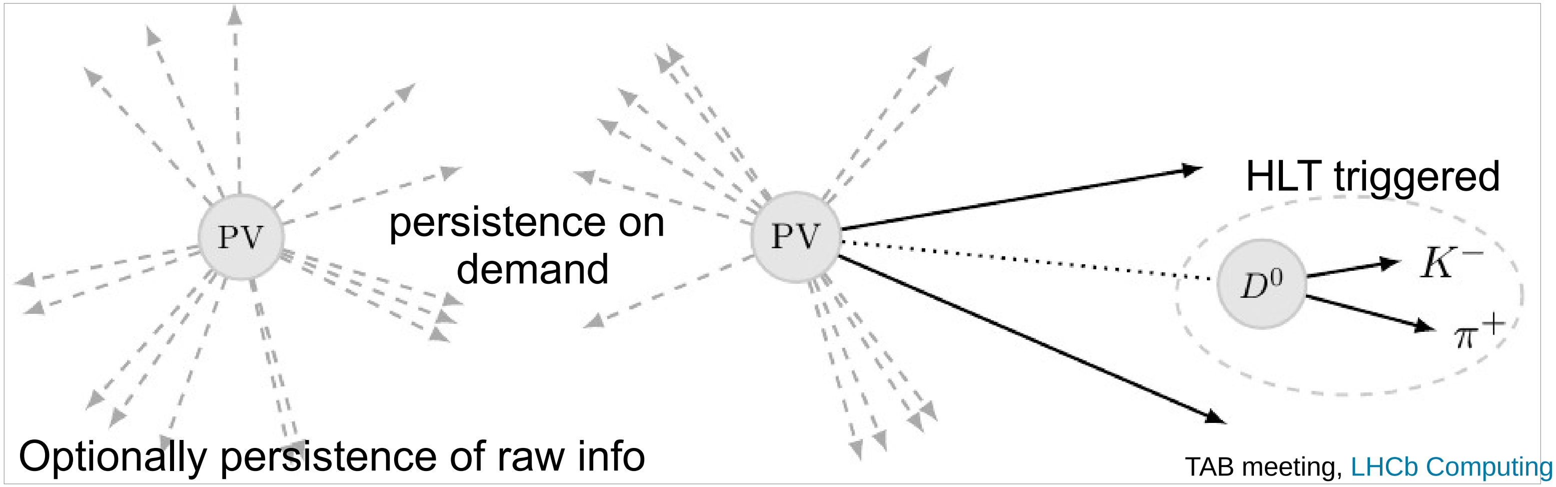
data filtering → processing

DST, MDST
TurboRaw, MDST

tape (Tier0, Tier1 (Full and TurCal))

disk(1-250 kB)

- Dynamische Speicherung der Daten des Turbo stream

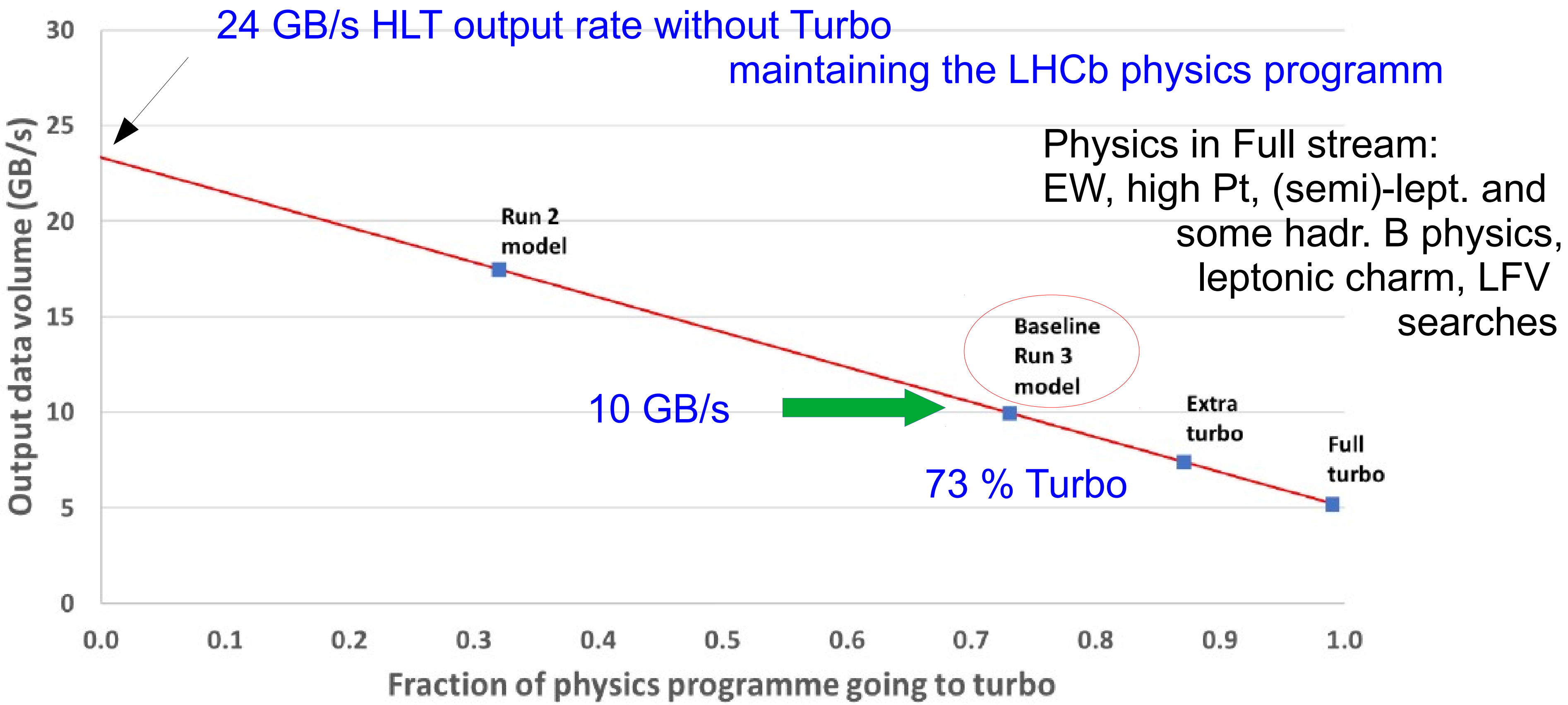


TAB meeting, LHCb Computing Upgrade Model, J. Marks

Turbo stream wurde in Run2 getestet: Rohdaten verwerfen, Speichern auf Tape, nur MDST auf Disk → Einsparung von Disk Resourcen

LHCb Run 3 / 2021

- LHCb
  - Neues Computing Model:
    - Die Daten Output Rate hängt vom gewählten Physikprogramm und dem Turbo Anteil ab



24 GB/s HLT output rate without Turbo maintaining the LHCb physics programm

Physics in Full stream:
EW, high Pt, (semi)-lept. and some hadr. B physics, leptonic charm, LFV searches

10 GB/s

73 % Turbo

LHCb Run 3 / 2021

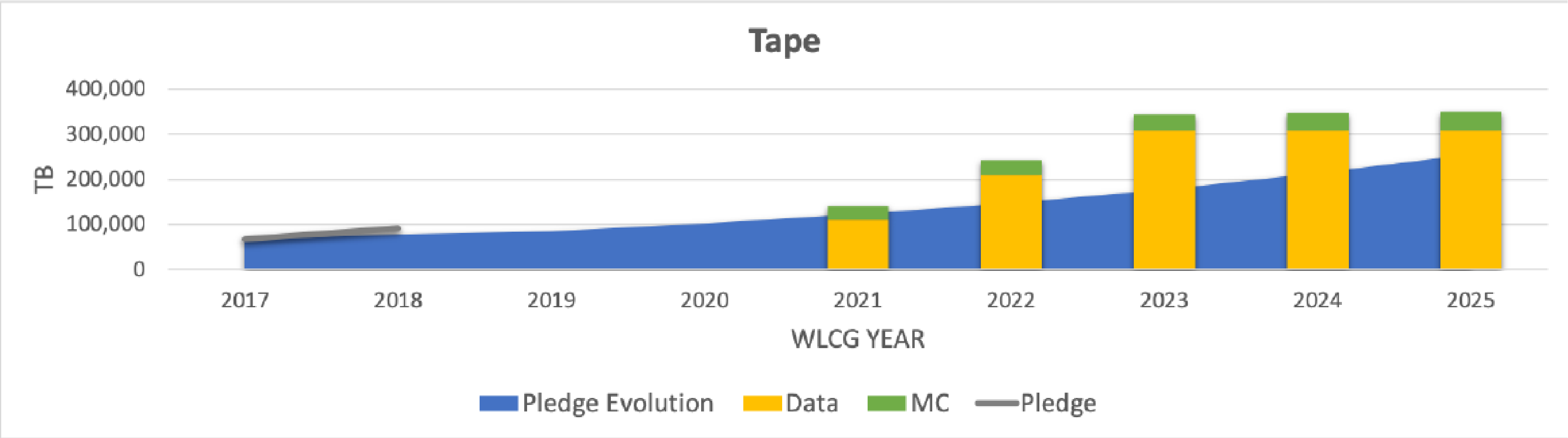
- LHCb
  - Neues Computing Model
  - Baseline Modell
    → Speichern von 70% der Daten als Turbo Strom (keine Rohdaten)
    → 30% (Full und TurCal Strom) werden auf Tape gespeichert (7.5 GB/s), rekonstruiert, gefiltert, selektiert und als MDST auf Disk gespeichert (1 GB/s)
  - Reprozessieren des Full und TurCal Stroms erfolgt im LHC winter-shutdown vom Tape (2nd copy im Tier 1)
    → 2021 (2022) bei 2 Monaten staging time Bandbreiten von 3.6 GB/s (7.2 GB/s) erforderlich
  - 90% der Rechenzeit für Simulationsrechnungen (40% detailliert)
  - Resourcenbedarf für 2021-2023 steigt um Faktor 1.4-1.8

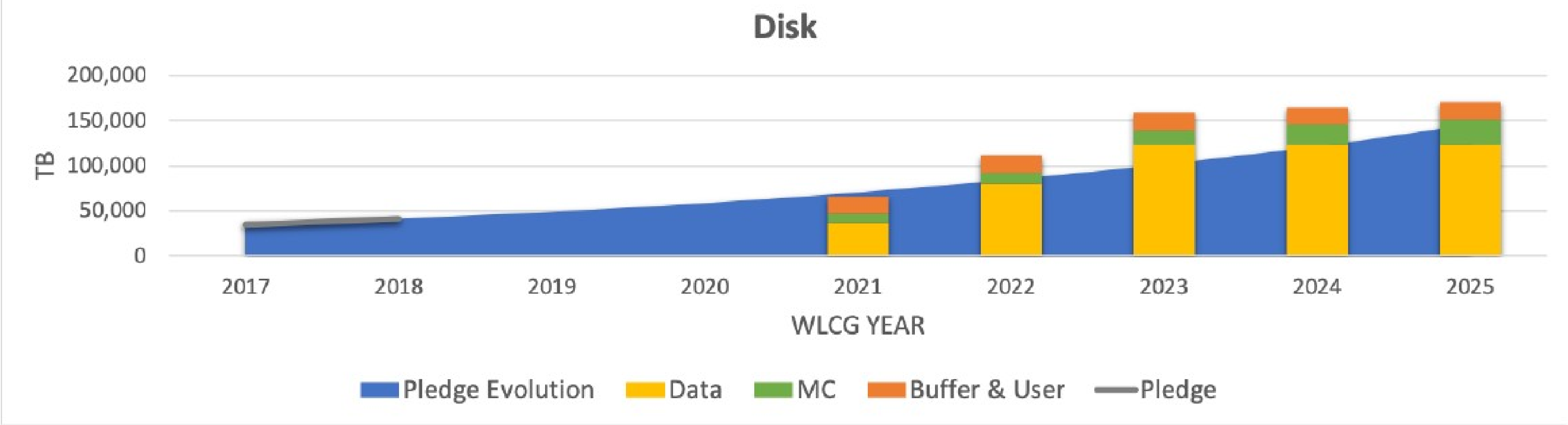
	[all / DE-KIT]	[all Tiers]
CPU [kHEPspec]	367 / 124.5	734
Disk [PB]	41.4 / 7.0	70
Tape[PB]	96 / 16.3	152

LHCb-PUB-2019-012, October 2019

Resource Requirements

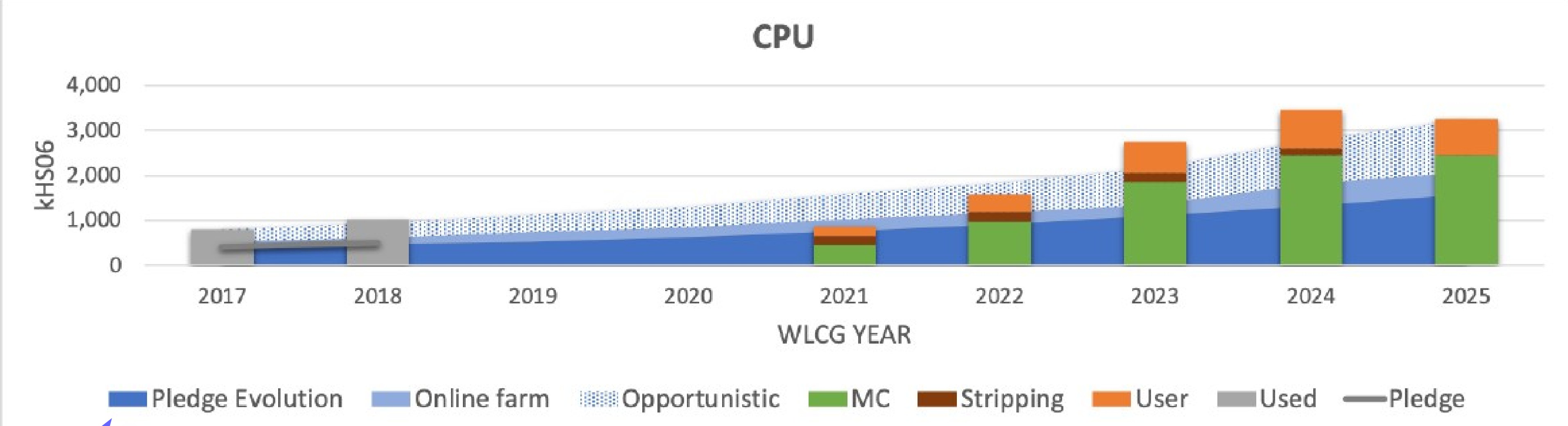
- LHCb baseline scenario: assuming an HLT output of 10 GB/sec



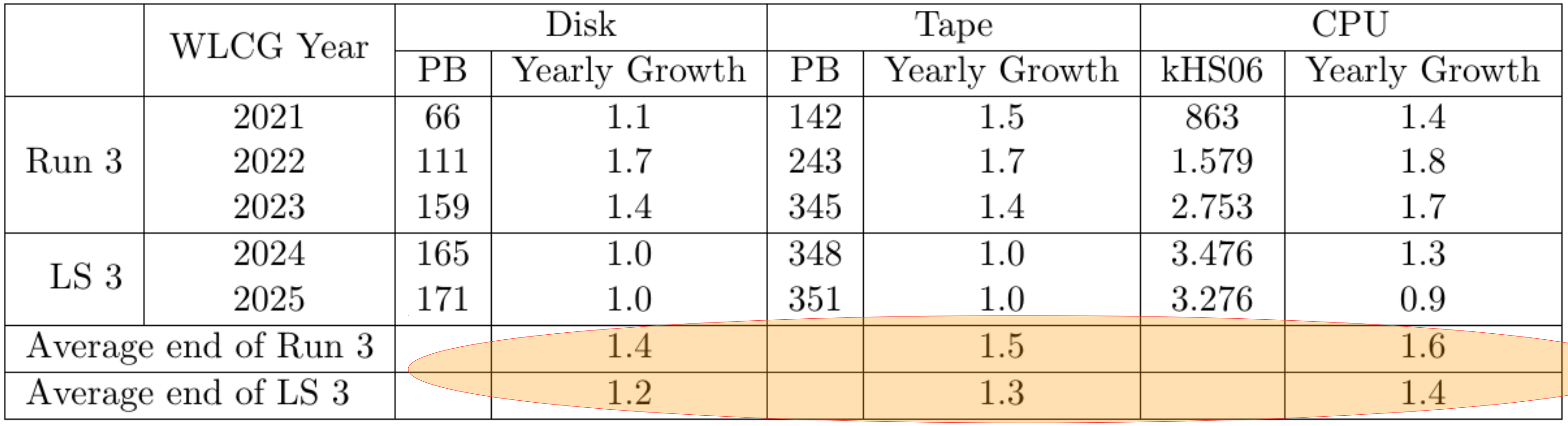


Resource Requirements

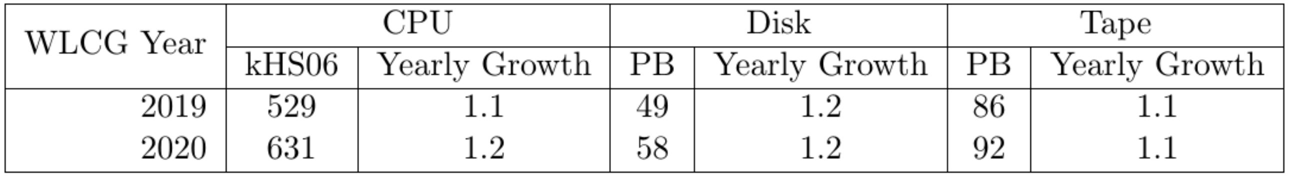
- LHCb baseline scenario: assuming an HLT output of 10 GB/sec



+20%/y



	WLCG Year	Disk		Tape		CPU	
		PB	Yearly Growth	PB	Yearly Growth	kHS06	Yearly Growth
Run 3	2021	66	1.1	142	1.5	863	1.4
	2022	111	1.7	243	1.7	1.579	1.8
	2023	159	1.4	345	1.4	2.753	1.7
LS 3	2024	165	1.0	348	1.0	3.476	1.3
	2025	171	1.0	351	1.0	3.276	0.9
Average end of Run 3			1.4		1.5		1.6
Average end of LS 3			1.2		1.3		1.4



WLCG Year	CPU		Disk		Tape	
	kHS06	Yearly Growth	PB	Yearly Growth	PB	Yearly Growth
2019	529	1.1	49	1.2	86	1.1
2020	631	1.2	58	1.2	92	1.1

Plans and Requirements for ATLAS and CMS

- Modest increase of requirements during Run-3
  - in-line with flat-budget projections
- Big change with HL-LHC/Run-4
  - Higher luminosity 7.5 10^{34} cm^{-2} s^{-1} with <PU> up to 200
  - upgraded detector (more channels)
  - up to 10 kHz trigger rate
- Dedicated Computing CDRs in preparation (~final now) by ATLAS and CMS, in coordination with WLCG
  - common LHCC scenario: **10 kHz HLT, <PU>=200, 7.5 10^{34} cm^{-2} s^{-1}**
  - review and discussion now starting in LHCC
  - plan to evolve into Computing TDR by 2023
- Material shown here provisional
  - under review by LHCC

ATLAS – where to optimize (1)

- 60% of CPU used for simulation (generation, detector interaction, reconstruction)
  - ev-gen NLO/NNLO becoming CPU consuming
  - Geant4 R&D
  - fast simulation improvements
    - crucial factor: ratio Fast:Full

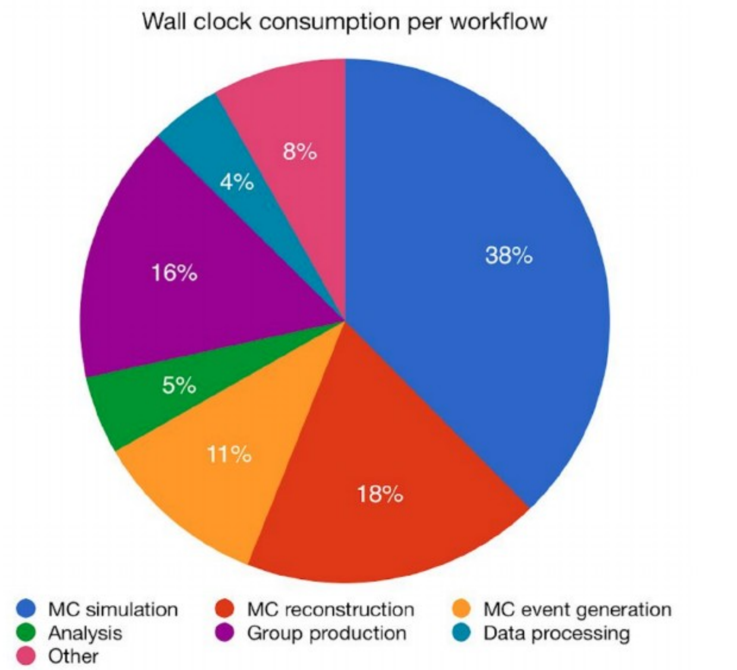
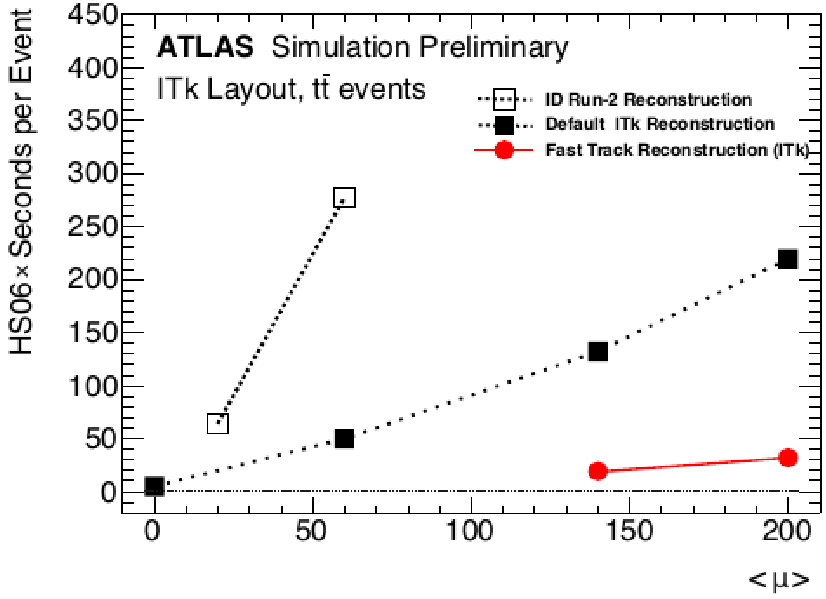


Figure 1: ATLAS CPU hours used by various activities in 2018



- Reconstruction
  - current ID tracking very sensitive to pile-up
  - Huge improvement with new ITk Reco
  - further optimization prospects

ATLAS – where to optimize (2)

- Analysis:
  - Run-2: AODs & DAODs (many variants) fill ~90% of disk storage
  - Run3(4): Unified DAODPhys & DAODPhysLite for most analyses
    AODs partially on tape → data carousel
  - ROOT evolution (RDataFrame, …)
  - Python/Jupyter/DataScience Ecosystem
- GPU – how to use efficiently ?
  - accelarator-based systems increasingly popular (e.g. ML & HPC)
    - strong push for usage by (some) funding agencies
  - rather straightforward for ML training (in ML environment)
  - non-trivial to port HEP reco & simul code
    - big effort & expert manpower

ATLAS: Three Scenarios (1)

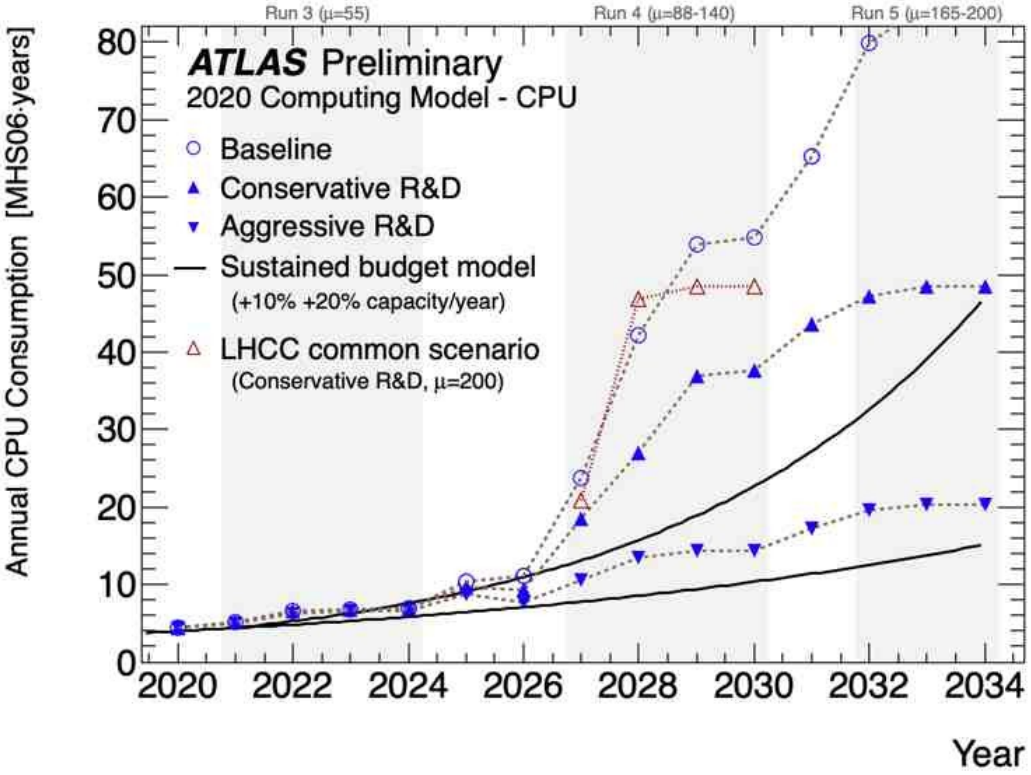
	Baseline - Run 3	Conservative R&D	Aggressive R&D
Event generation	Time/event same as Run 2, compromises made in physics quality	Better physics performance for same CPU	CPU/event halved via a combination of software improvements, event sharing and physics choices
Simulation	Mixture of G4/fast sim MC+MC overlay digitisation used, but high memory queues needed to produce the premixed RDOs	Fast sim becomes default, Static compilation of Athena code with G4, dependencies implemented Digitisation uses AthenaMT to reduce memory for RDO production	Substantial G4 speedup ACTS-based FATRAS used to simulate ITk reco in fast simulation rather than G4 EVNT→AOD in single step Trigger like algorithms used to discard events prior to reco
Reconstruction (<PU> = 200)	default ITk reco tracking/total: 214/584	optimized ITk reco tracking/total: 50/295	

ATLAS: Three Scenarios (2)

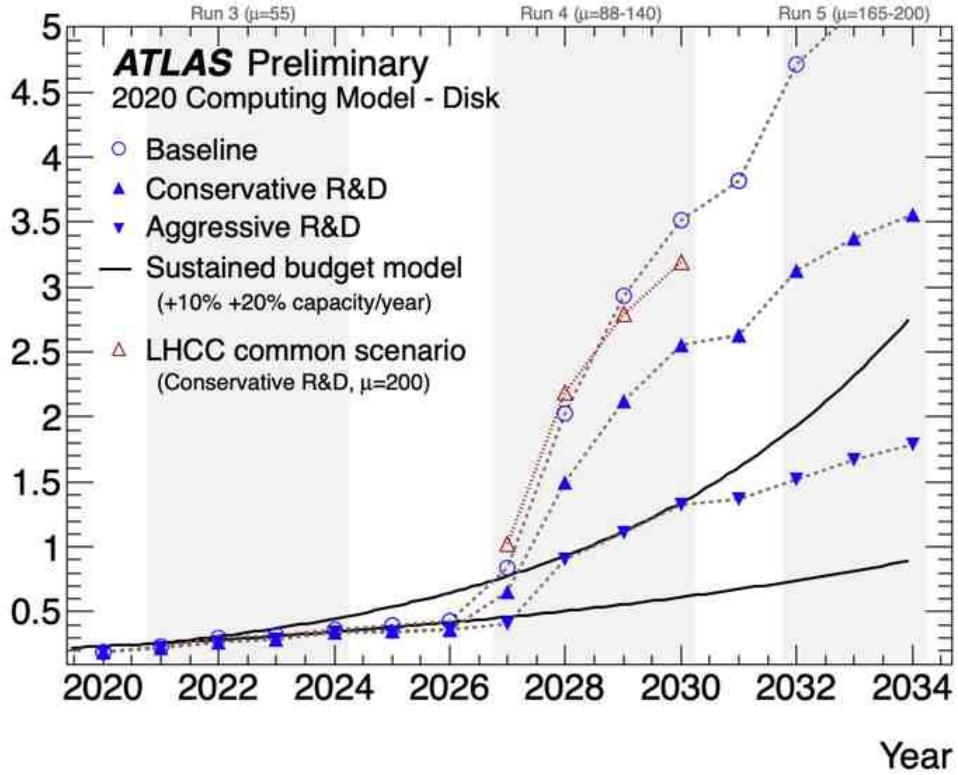
	Baseline - Run 3	Conservative R&D	Aggressive R&D
Analysis model	4 DAOD processings/y Usage: 50% PHYS, 10% LITE, 20% other DAOD, 20% DRAW	3 DAOD processings/y Usage: 60% LITE, 20% PHYS, 20% DRAW	2 DAOD processings/y Usage: 80% LITE, 20% DRAW
MC statistics	As 2018 projections 3 x data events per year	2.5 x data events per year, Re-reconstruct past three years each year, Full reproduction of MC needed for ongoing analysis every 6 years (including new evgen)	As conservative but with 2 x data events
Fast:Full simulation ratio	Run 3: 1:1 Run 4: Start at 1:2 rising to 2:1	Run 3: 1:1 at start rising to 4:1 by the end Run 4: 2:1 at start rising to 8:1 by the end	

ATLAS Resource Projections Run-4

CPU



Disk

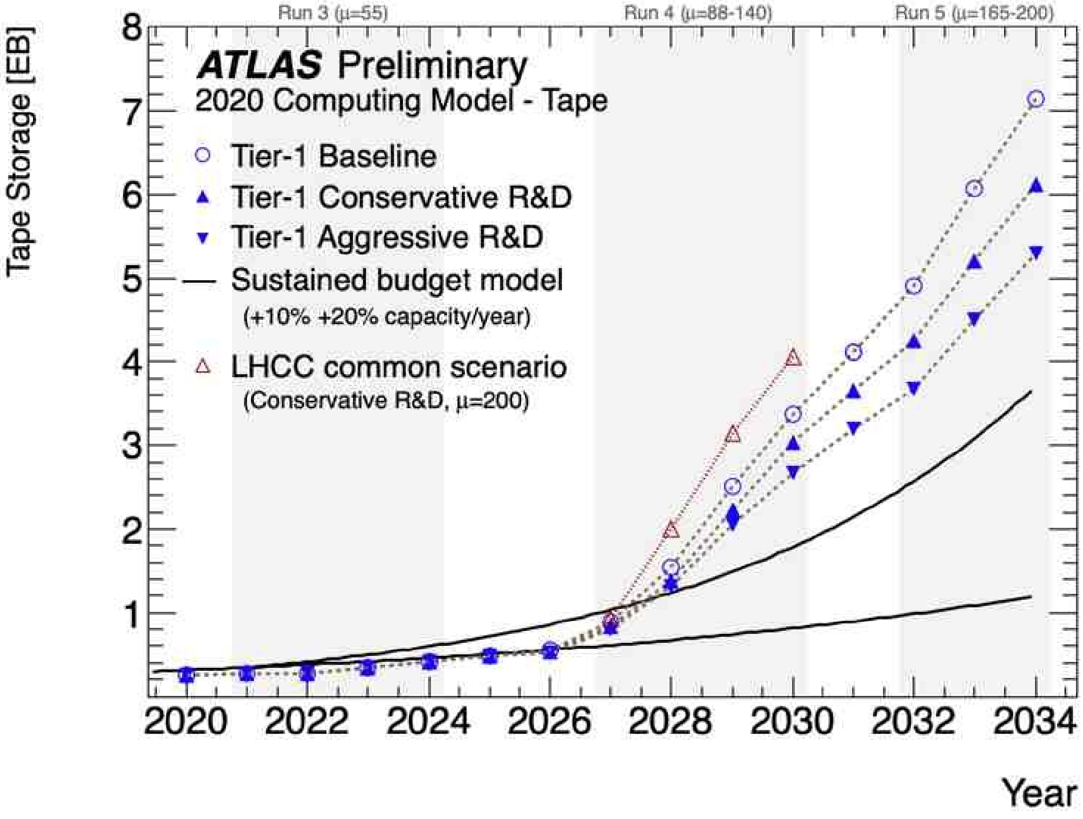


Aggressive R&D within flat-budget band (10-20%/y)
→ relies on substantial developments in many areas, large risks!

ATLAS Resource Projections Run-4

- Less difference for tape between 3 scenarios
  - mostly driven by RAW data needs

Tape



CMS plans & reqts - 1

- Reconstruction (data & MC) is dominating CPU usage: ~80% @ Run-4
  - Big effort already undertaken to use GPUs for Reco
    - integrated in CMSSW framework
      - one codebase supports multiple backends
    - plan to have GPU based HLT farm
  - Further development/optimization ongoing
- Simulation less CPU-critical for CMS
  - investigating in fast-simul, optimizing NLO generators
- Analysis data formats
  - push MiniAOD and NanoAOD
- 'Declarative programming' for analysis

CMS plans & reqts - 2

- Evaluate/move to common tools for computing operations
  - Rucio data management system
    - originated at ATLAS/Cern but now used by many other experiments
  - CRIC (Computing Resource Information Catalogue)
    - information system for resources, also ATLAS and others will use it
  - DD4HEP
    - detector description tool
    - used by ILC/CLIC and evaluated by LHCb
  - ...

CMS Resource Projections Run-4

(CMS plots of resource projections not yet public)

Extrapolation of CPU & Disk requirements based on current (LS-2) status
→ gap of ~factor 10 for Run-4 wrt flat-budget

Assuming Reco optimizations (10%/y) , reduced # AOD replicas, usage of MiniAOD/NanoAOD for most analyses
→ gap of factor 3-4 in CPU and 2-3 in Disk for Run-4 wrt flat-budget

Conclusions - 1

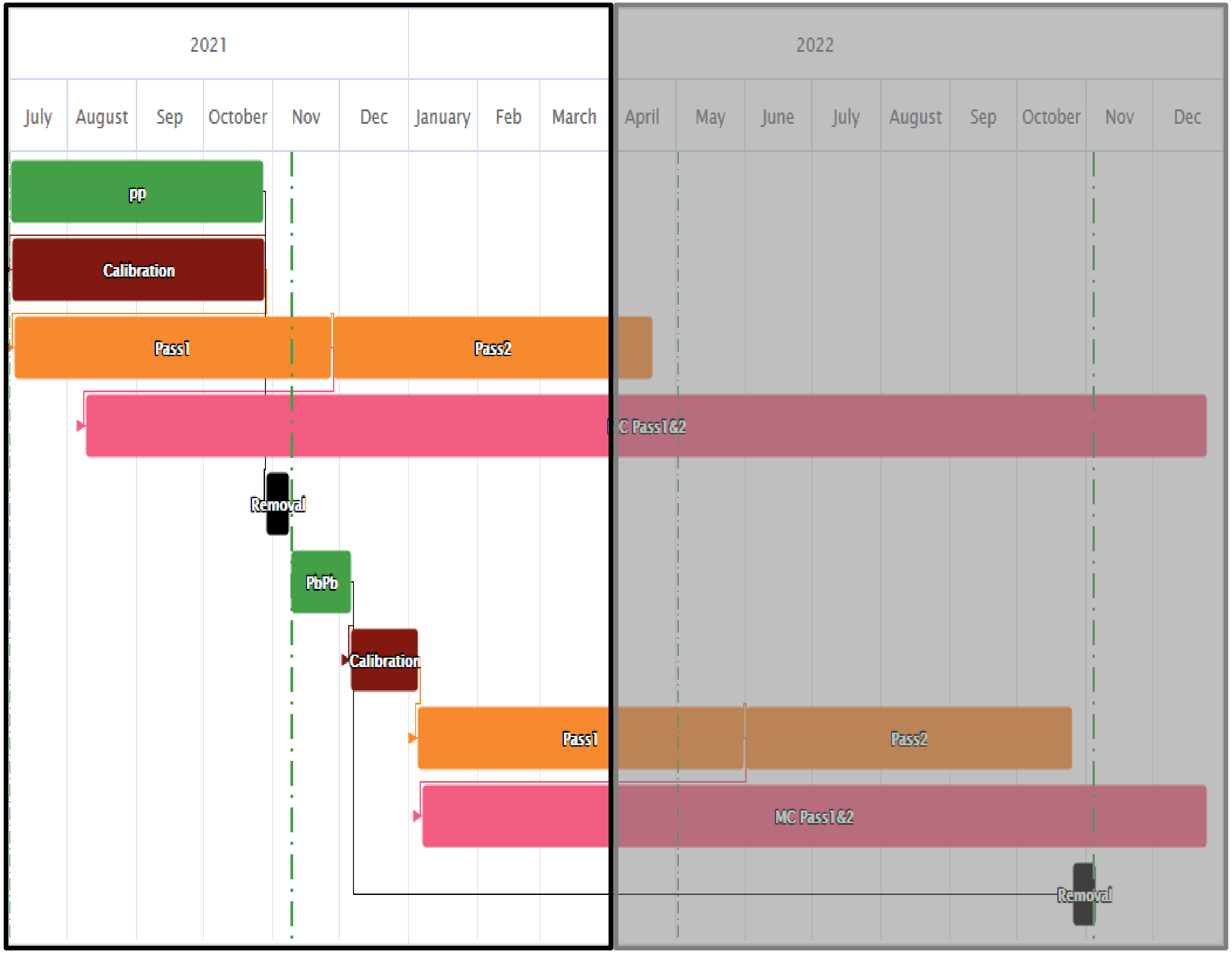
- Massive increase of demands due to different operation parameters & conditions
  - Coming with Run-3/2021 for Alice & LHCb
  - Coming with Run-4/2026 for ATLAS & CMS
- Large uncertainties in extrapolation of hardware efficiency gains for CPU, Disk and Tape
  - 'flat budget evolution' of 10%/year might be optimistic
    - 5% vs 15% yearly gain translates into factor 2 difference by 2028
- ALICE
  - Run-3: factor ~100 increase in recording rate
  - new O2 model: requirements at upper limit flat budget: ~+20%/y
- LHCb
  - Run-3: factor 30 increase in data volume → Turbo concept to mitigate
  - signficant increase of requirements : +30-40%/y

Conclusions - 2

- **Run-3**: ATLAS & CMS well within 'flat-budget'
- **HL-LHC/Run-4**: Massive increase of demands: factor 3-5 above 'flat-budget' with direct extrapolation
  - Simulation most critical for ATLAS, Reconstruction for CMS
- Mitigation strategies under discussion
  - 'aggressive R&D' scenario in ATLAS brings requirements back to 'flat-budget'
  - Requires massive R&D effort in all areas:
    - Fast Simul, optimized reco, evgen/NLO optimization&share, minimalistic analysis data model, data carousel, …
- Crucial to join efforts in common areas
  - data management, event generation, detector simulation, operation tools, ...

Backup

Pre-COVID-19 resource requirements for 2021



pp period (~3 months):

- Commission of detectors and validation of physics results (7 PB to be archived to tape)

HI period (28 days):

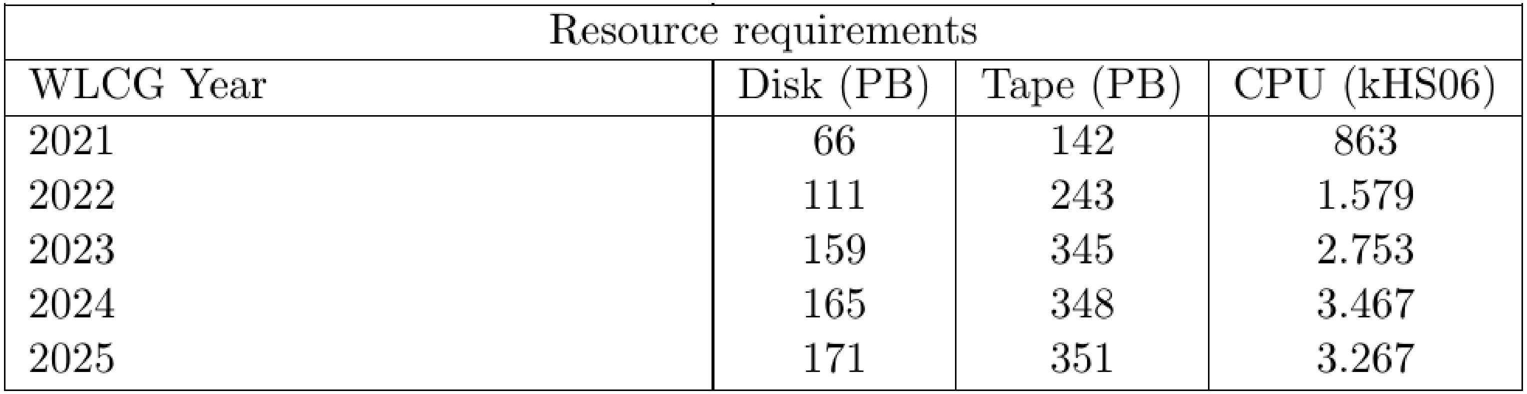
- Data volume: ~54PB CTFs, ~20 PB AODs on disk
- Both CTF and AOD will be archived on tape
- Computing power: ~430 kHS06 for 10 months for **two asynchronous passes** executed in the 10 months

MC:

- 1 Pb-Pb pass: ~500 kHS06
- 1 MC pp pass: ~270 kHS06

Summary

- LHCb upgrade experiment will record a factor 30 in volume due to higher luminosity, trigger efficiency and pileup
- While the computing upgrade for ALTAS and CMS starts for Run 4 LHCb and ALICE need an optimized scheme and increased computing resources already for Run 3
- The expanded use of the Turbo concept represents a major step forward in handling the large data volumes of the LHCb upgrade
- The baseline trigger output bandwidth of 10 GB/sec is recorded to tape with a disk usage reduced by a factor of 2 due to selection/reduction of full and calibration data streams
- The CPU power is determined by the production of simulation data needed



Resource requirements			
WLCG Year	Disk (PB)	Tape (PB)	CPU (kHS06)
2021	66	142	863
2022	111	243	1.579
2023	159	345	2.753
2024	165	348	3.467
2025	171	351	3.267