ATLAS Computing and ADC Status

ATLAS GridKa Cloud Face to Face Meeting 2023

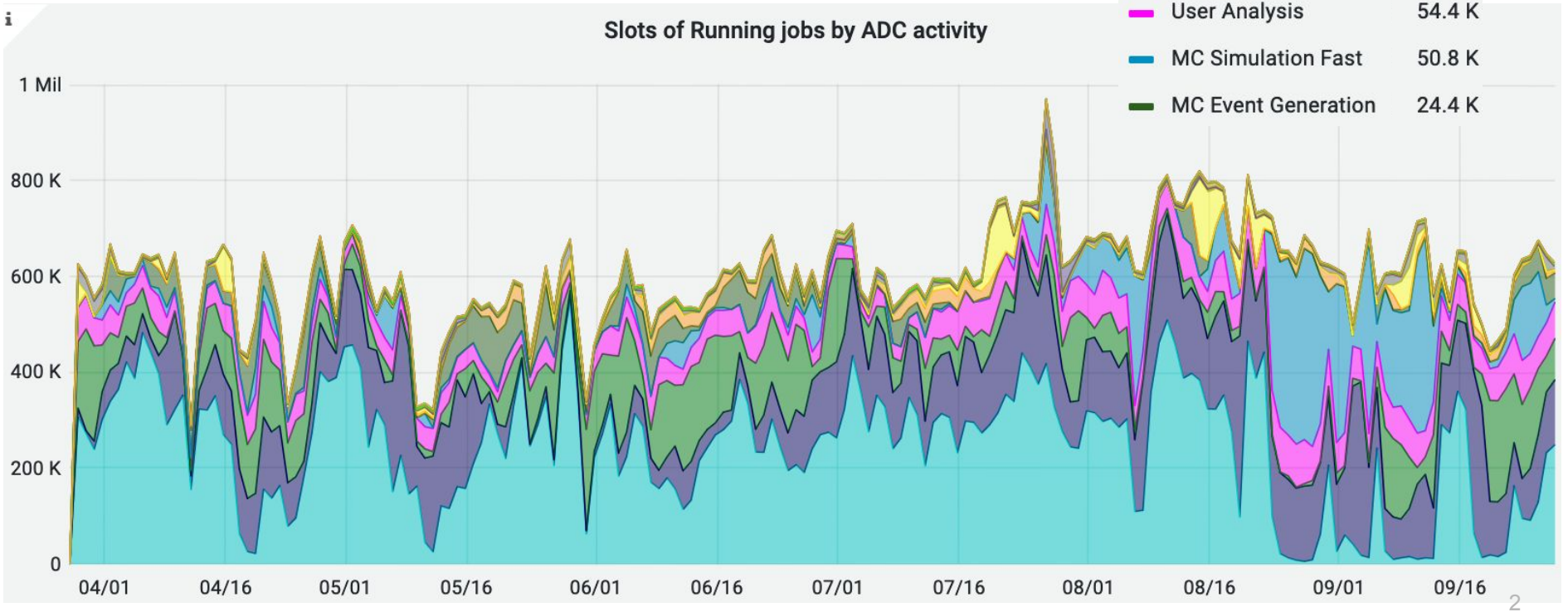
DESY, October 9th, 2023

Mario Lassnig (CERN), Zach Marshall (LBNL),
David South (DESY)

On behalf of the ATLAS S&C community
S&C Week indico *ADC@S&C Week indico*

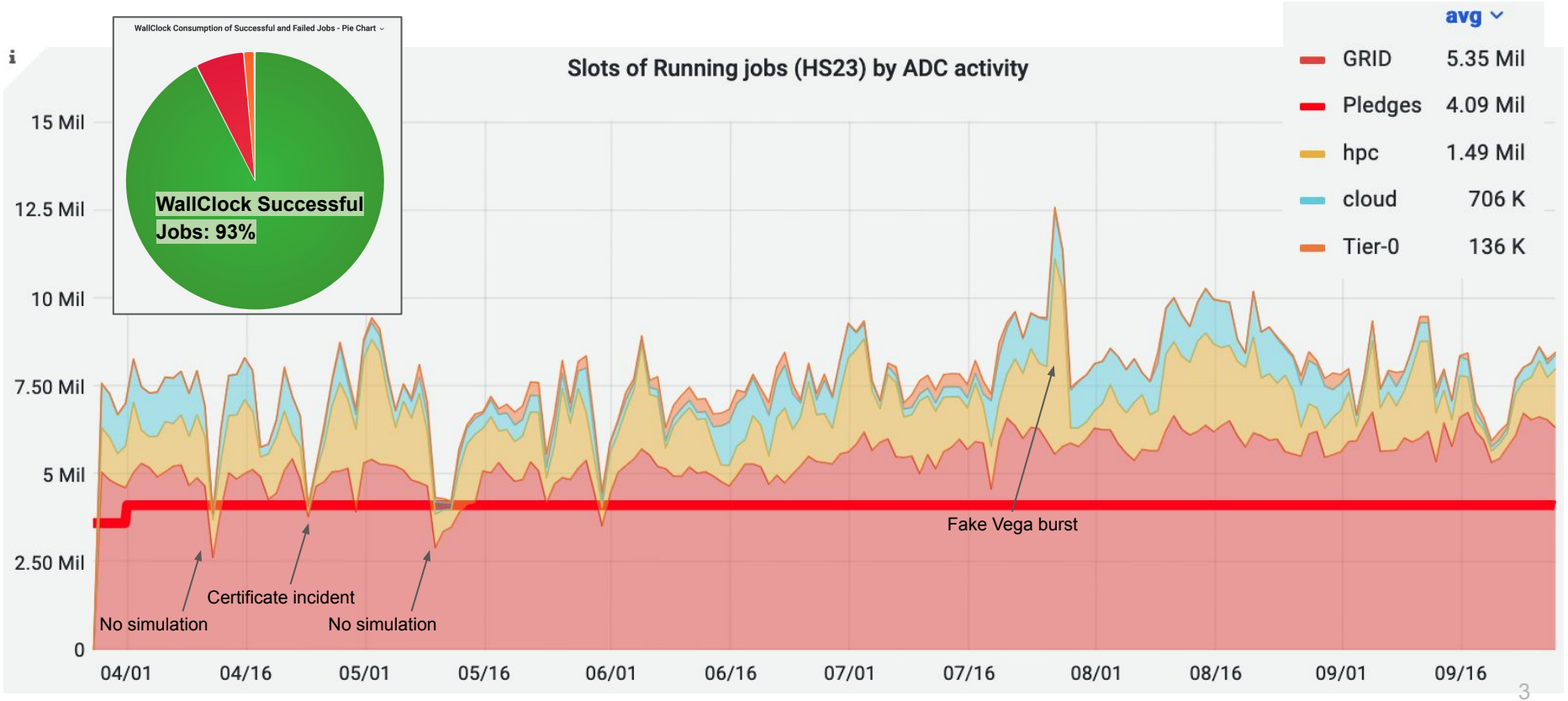
Resource Usage – Compute – Last 6 months

- Generally smooth running in the last 6 months
- MC23 ongoing, lots of fast sim recently
  - Looking forward to AF3 CP recommendations!



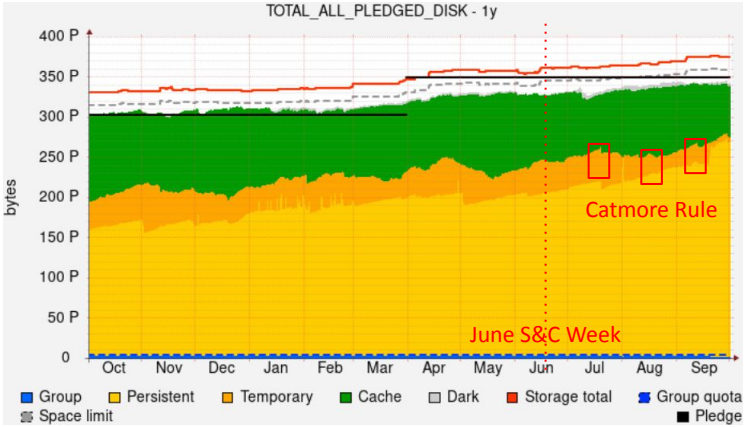
Resource Usage – Compute – Last 6 months

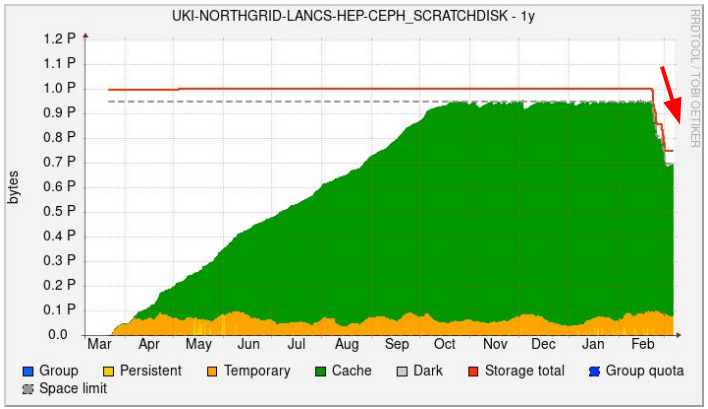
- Grid sites providing quite a lot of efficient (and over-the-pledge) CPU
- HPC and cloud (Google and Vega especially) continue to contribute

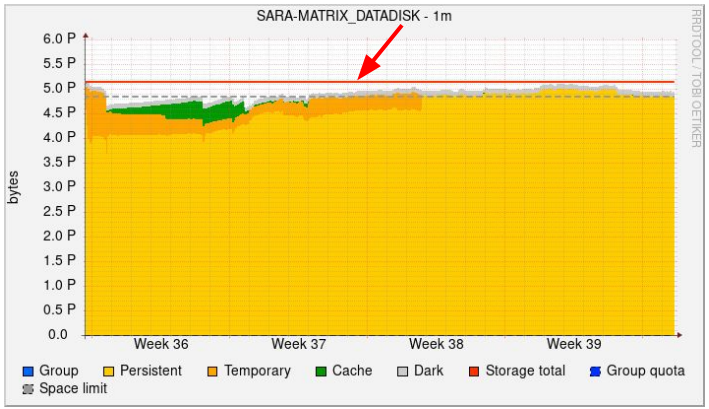


Storage situation

- Storage total well above pledge
  - Catmore rule applied often, lifetime model application soon
  - Global healthy cache/total ratio
- Hidden cache hotspots (cold spots?)
  - Require proactive data rebalancing
- SCRATCHDISK capacity
  - Capacity increases, but pinned data decreases
  - Recommendation to deploy pledge to DATADISK instead
  - More aggressive rebalancing could yield 2+ PB more
  - Long-running idea to merge completely with DATADISK - went with Cameron?

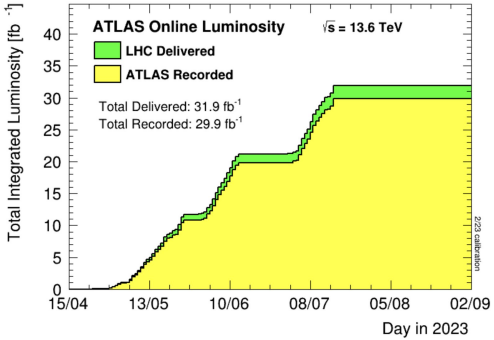




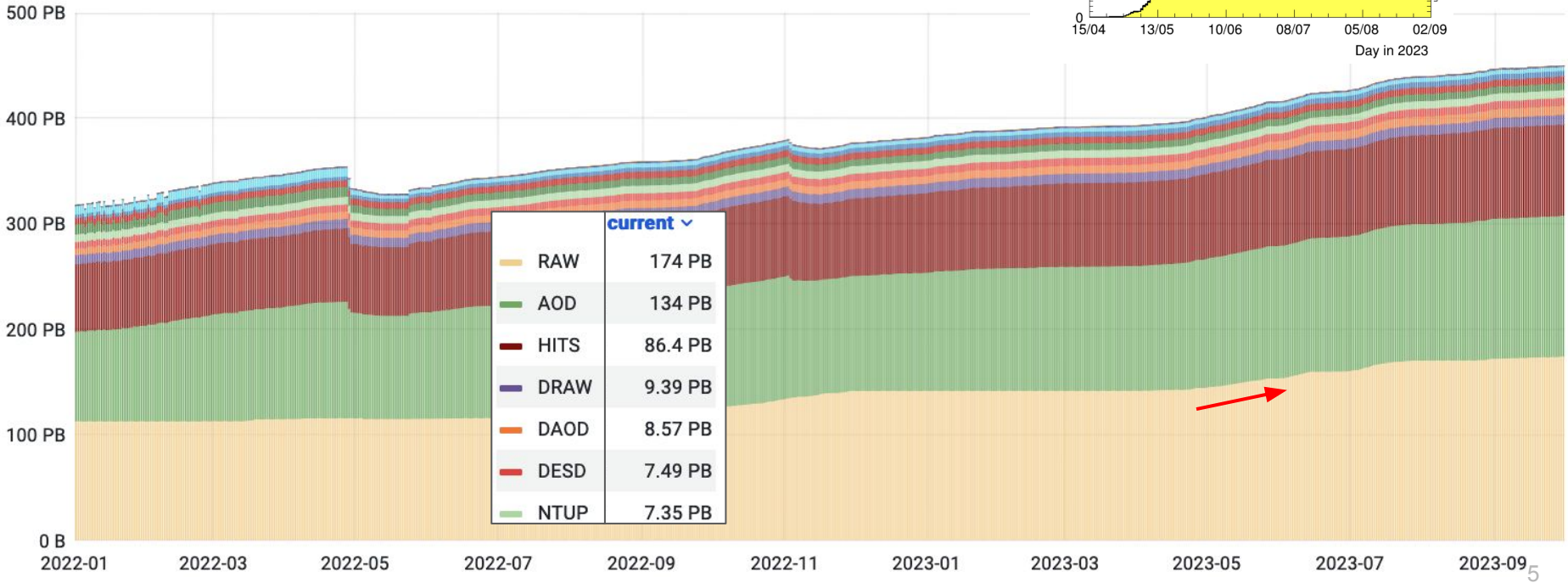


Resource Usage – Tape – Since Start of Run 3

- We are taking data:
  - Thank you to the sites, the ops teams, and all the developers that helped us get here!
  - In the last months more sites offering tape; happy to work on archival solutions with sites!

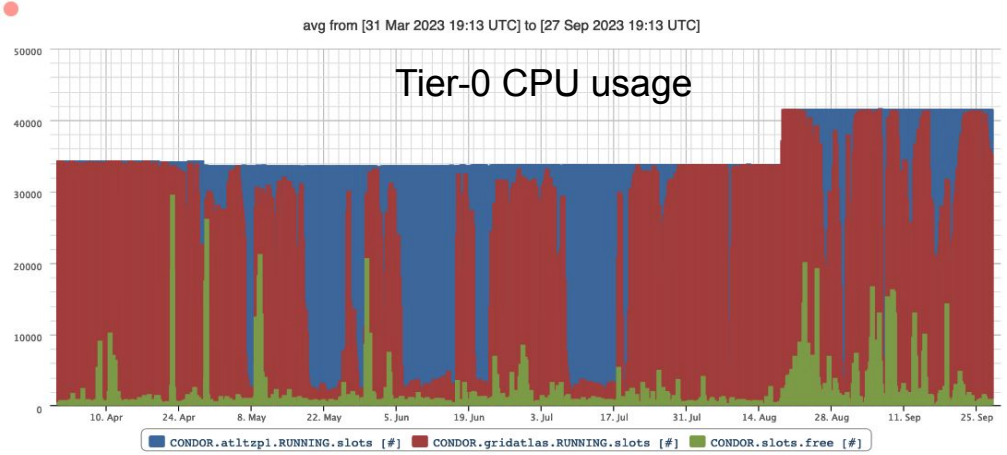


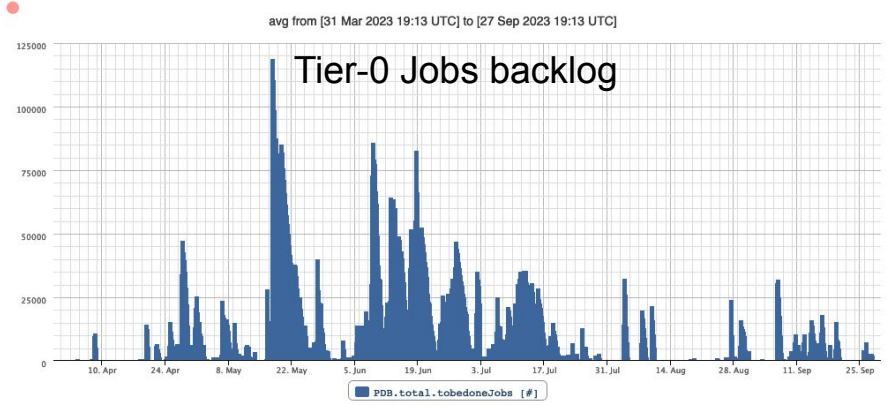
Volume per datatype_grouped

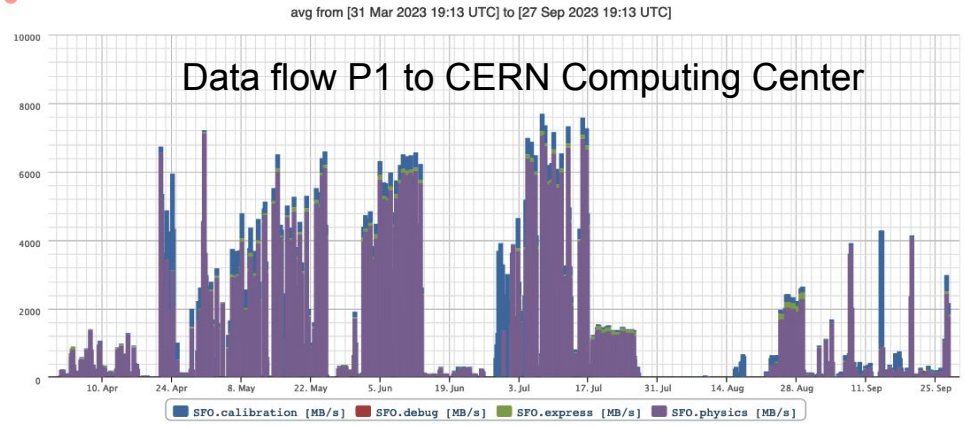


Tier 0 Operations — last 6 months

- Stable and reliable Tier0 operations
- CPU
  - Tier0 has priority, and Grid gets them when possible
  - Backlog never near to be an issue
- SFO throughput healthy, close to the tested limit
  - Let's see how the Heavy Ions do





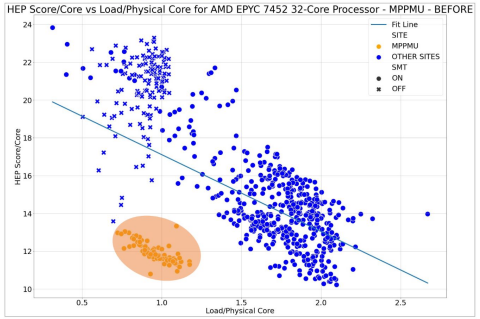


Resource and Site Evolution

- Taiwan transitioning from Tier 1 to Tier 2 this month
  - A non-trivial problem; thanks to the experts who have helped and the US for support
- Still in early stages with a few other sites
  - UAEU is trying to develop a Tier 2; thanks to the Italians for support
  - *Very* early discussions with AIMS Rwanda
- Would be nice to start to understand HPC plans and roadmaps in a few places
  - Next US HPC? TACC on ARM?
  - Fugaku 2?
  - Leonardo? How much might we capture?
  - German Cloud is the sign of things to come? Jupiter at Jülich with significant ARM?
- We have no further concrete news about the Russia / Belarus / JINR situation
  - Decision postponed to the December council meeting

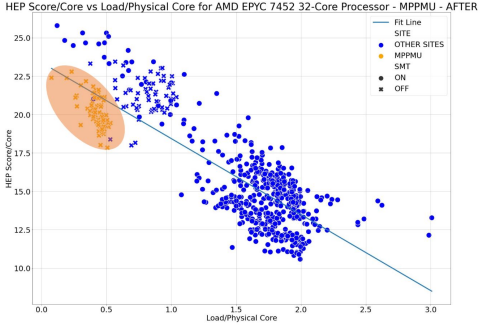
HEPScore23 to probe to ATLAS grid job slots

Load Measurements



- MPPMU performance in terms of score per core is low and comparable to other sites that have double load in the servers
- Contacted with site admin
- Wrong masking of the disabled cores (socket 2 completely disabled) while SMT was ON
- Changes have been applied

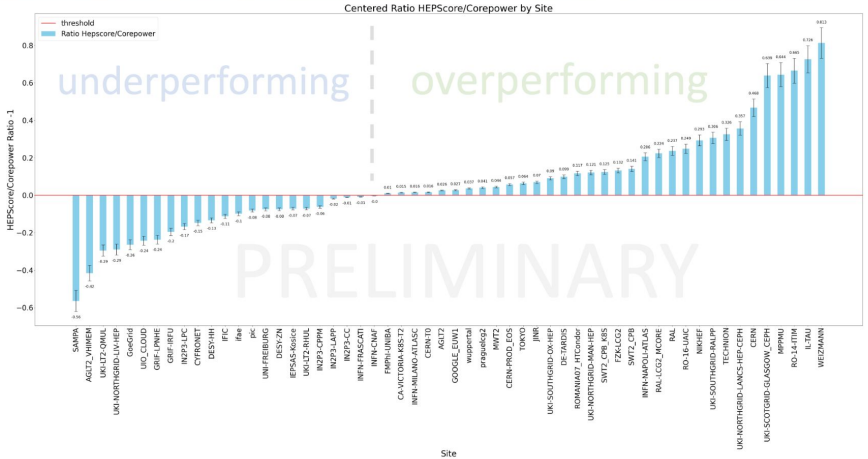
Load Measurements after changes



- Changes have been applied: SMT off fixed the problem with cores masking misconfiguration
- MPPMU performance increased, and load decreased
- That proves the applied changes had a real impact on the performance for MPPMU

- Natalia is looking into HEPScore23 measurements, and discovering things!
- Successful implementation of automated HEPScore23 submission via PanDA using HammerCloud
- Dynamic stimation of compute resource performance

HS23/Corepower Centered Ratio for sites



ARM resources

- Lots of discussion on this last week
  - Is it a game changer?
- Study ARM resources with Glasgow
  - Dave talk, Dwayne talk
  - Around 1600 cores available
  - 50M event validation sample almost finished
- Studies also at BNL - Imran's talk

Systems were evaluated during HEPscore23 runs

ARM 128 core	ARM 64 core	X86_64 Xeon
HEPscore – 2030	HEPscore – 1042	HEPscore – 1278
Max power – 370W	Max power – 252W	Max power – 601W
Idle – 125W	Idle – 107W	Idle – 179W
HEPscore / Watt – 5.49	**HS / W – 4.13**	**HS / W – 2.13**

What's the best way of going about this in the future

At some point we will want to get rid of the test CE

Option 1

- Modify exiting Queue

UKI-SCOTGRID-GLASGOW_CEPH
ce01.gla.scotgrid.ac.uk (x86)
ce02.gla.scotgrid.ac.uk (x86)
ce03.gla.scotgrid.ac.uk (x86)
ce04.gla.scotgrid.ac.uk (x86)
ce05.gla.scotgrid.ac.uk (aarch64)

- Dangerous, will impact the workflow of many VO's. Will have all the ARM traffic on one CE - not scalable in the future?

Option 2

- Add a queue pointing to the same CE's

UKI-SCOTGRID-GLASGOW_CEPH
ce01.gla.scotgrid.ac.uk (x86)
ce02.gla.scotgrid.ac.uk (x86)
ce03.gla.scotgrid.ac.uk (x86)
ce04.gla.scotgrid.ac.uk (x86)

UKI-SCOTGRID-GLASGOW_ARM
ce01.gla.scotgrid.ac.uk (aarch64)
ce02.gla.scotgrid.ac.uk (aarch64)
ce03.gla.scotgrid.ac.uk (aarch64)
ce04.gla.scotgrid.ac.uk (aarch64)

- If Job requirements can be successfully injected this seems the safest option

Option 3

- Read architecture from the jobs themselves.

UKI-SCOTGRID-GLASGOW_CEPH
ce01.gla.scotgrid.ac.uk (x86, aarch64)
ce02.gla.scotgrid.ac.uk (x86, aarch64)
ce03.gla.scotgrid.ac.uk (x86, aarch64)
ce04.gla.scotgrid.ac.uk (x86, aarch64)

- Condor_submit has architecture flags. Could try to pulling the architecture flag from condor_submit into the ARCsub, maybe modify in job description language (JDL)?
- Would potentially need to set up/inject default architecture so that "standard" x86 jobs don't get sent to ARM cores.

Option 4

- Pilots report to VO what architecture it's running on.
- That VO sends jobs of that type. Would require every VO to add this functionality to their pilots, not really default-able.
- Potentially wasteful if a site gets a pilot running on an ARM server and has no ARM work, long term solution?

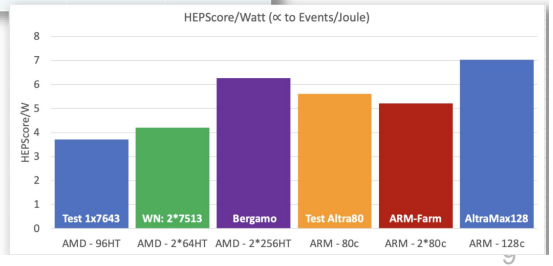
Dwayne Spiteri, University of Glasgow — Glasgow ARM Cluster: Experience and Findings — 19

GridPP UK Computing for Particle Physics

HEPScore/Watt

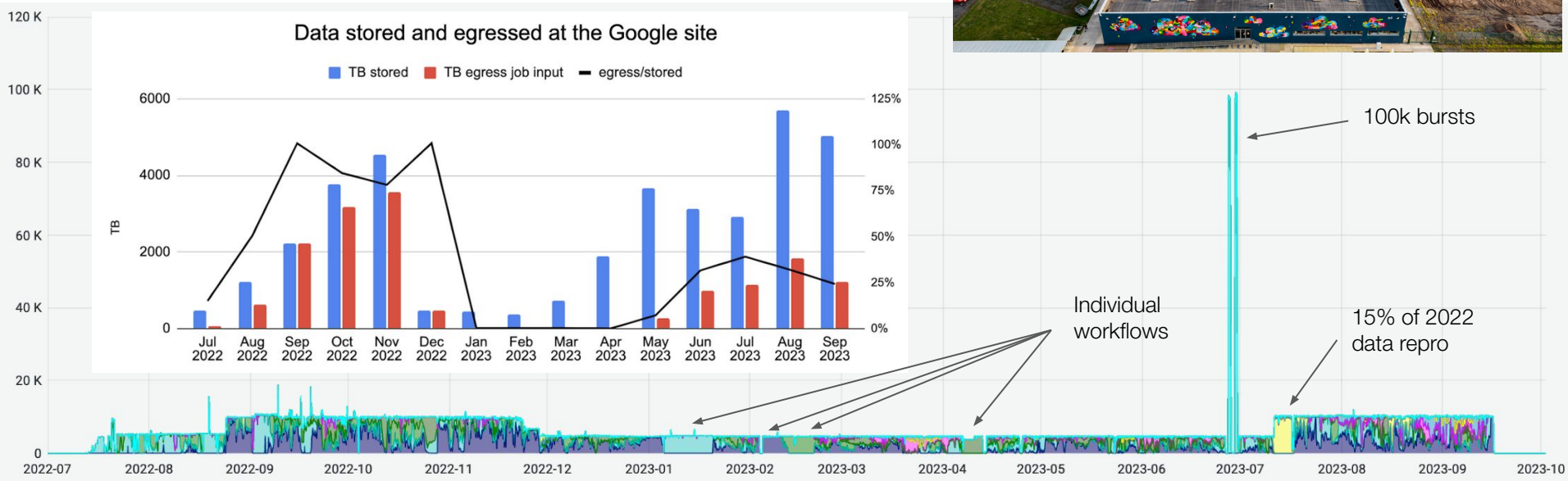
Machine	CPU	Threads	HS/Watt	
Test AMD	1x EPYC 7643 HT	96	3.7	
Std AMD WN	2x EPYC 7513 HT	128	4.2	
AMD Bergamo	2x EPYC 9754 HT	512	6.3	50% improvement from 7513 WN
Test ARM	1x ALTRA 80	80	5.6	
Farm ARM	2x ALTRA 80	160	5.2 ??	
Ampere Max	1x ALTRA Max	128	7.0	25% improvement from 80c

Note added: We use Average Watts to calculate HS/Watt. Other people also use Max Watts, so care needs to be taken comparing numbers.



Google TCO

- 15 month project utilising Google cloud resources
- Service Agreement, flat rate
- Run ATLAS workflows on site with elastic setup compute queue via k8s and Object Store RSE
- Many successful studies, talk last Friday, more to come
- TCO coming to ICB soon



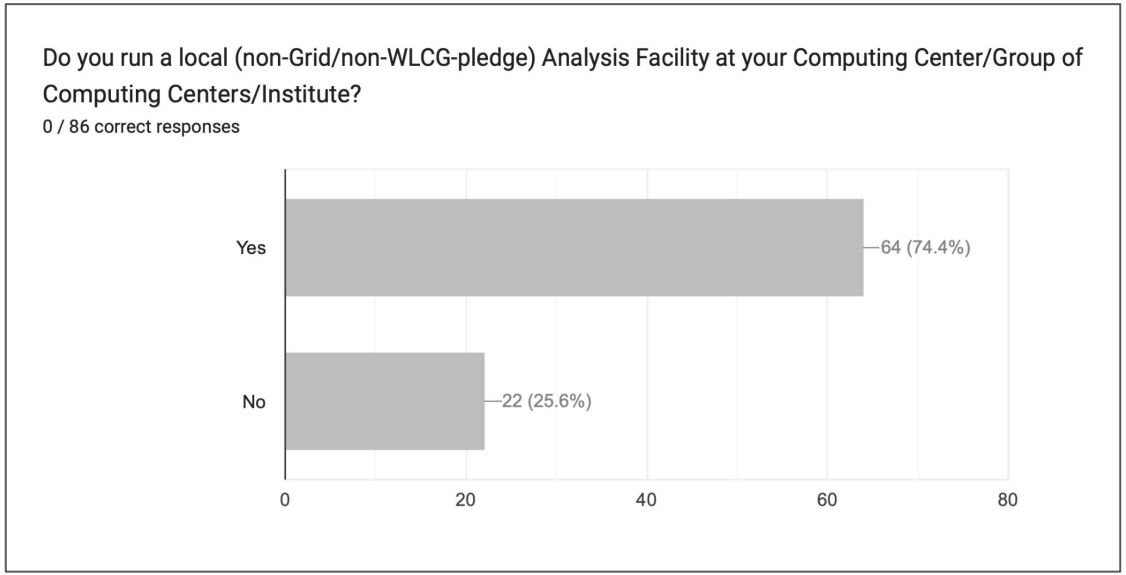
Brainstorming follow-up

- In February 2022, ATLAS management circulated a brainstorming document
  - Led to an impressive amount of open, honest discussion focused on Analysis

- We created an S&C-view follow-up document that we discussed last March, and since then we have been trying to follow up and report at each S&C week:
  - Concrete efforts to demonstrate common-base / CP Algs frameworks
  - Improved distributed analysis experience; investigated and reduced the “long tail of jobs”
  - Testing common ntupling within the prodsys for PA & CP groups to relieve analysers
  - Review (and will be reviewed more) the “remnants” (DAOD non PHYS/PHYSLITE) — some can be incorporated into PHYS via event augmentation
  - Workflows simplifications ongoing (e.g. simplified MET overlap-removal, on-the-fly systematics)
  - Target to demonstrate a realistic columnar analysis example by the end of 2024

Analysis Facilities Survey

- Analysis Facilities Survey earlier this year (it was intentionally vague about what an Analysis Facility ***is*** in the survey)
  - 85 respondents; 75% run an AF (not removing overlaps, but looks small)
  - Total ~37 PB, 800 kHS23 (almost a full copy of our CERN resources)
  - Averages: 35% lxplus; 40% local cluster; 10% Grid; 15% NAF/Other
  - More complete analysis coming soon (talk by Zach in Friday afternoon AMG)



Resource Request for 2025

- 2025 preliminary resource requests discussed with the C-RSG team (RRB)
- RAW data size set to 1.8 MB (discussion on SuperCells still ongoing):
  - Several questions from C-RSG on the RAW size topic — an item that we need to follow up

ATLAS Report to the C-RSG, September 2023

		2024 agreed @ April 2023 RRB	2025 Prel. Request @ September 2023 RRB	Balance 2025 wrt 2024
CPU	T0 (kHS23)	936	1100	17.5%
CPU	T1 (kHS23)	1516	1661	9.6%
CPU	T2 (kHS23)	1852	2030	9.6%
CPU	SUM (kHS23)	4304	4791	11.3%
Disk	T0 (PB)	49	56	14.3%
Disk	T1 (PB)	163	186	14.1%
Disk	T2 (PB)	200	227	13.6%
Disk	SUM (PB)	412	469	13.9%
Tape	T0 (PB)	207	264	27.4%
Tape	T1 (PB)	452	567	25.4%
Tape	SUM (PB)	659	831	26.1%

Table 6: Summary of the preliminary requests for computing resources in 2025.

"Old" RR for 2024 (approved)

ATLAS Report to the C-RSG, April 2023

		2023 agreed @ April 2022 RRB	2024 Request @ September 2022 RRB	2024 Request @ April 2023 RRB	Balance 2024 wrt 2023
CPU	T0 (kHS06)	740	850	936	26.5%
CPU	T1 (kHS06)	1430	1501	1516	6.0%
CPU	T2 (kHS06)	1747	1834	1852	6.0%
CPU	SUM (kHS06)	3917	4185	4304	9.9%
Disk	T0 (PB)	40	46	49	22.5%
Disk	T1 (PB)	136	162	163	20.2%
Disk	T2 (PB)	168	198	200	18.9%
Disk	SUM (PB)	344	405	412	19.8%
Tape	T0 (PB)	174	205	207	19.0%
Tape	T1 (PB)	353	448	452	28.1%
Tape	SUM (PB)	527	653	659	25.1%

Table 6: Summary of the final ATLAS requests for computing resources in 2024.

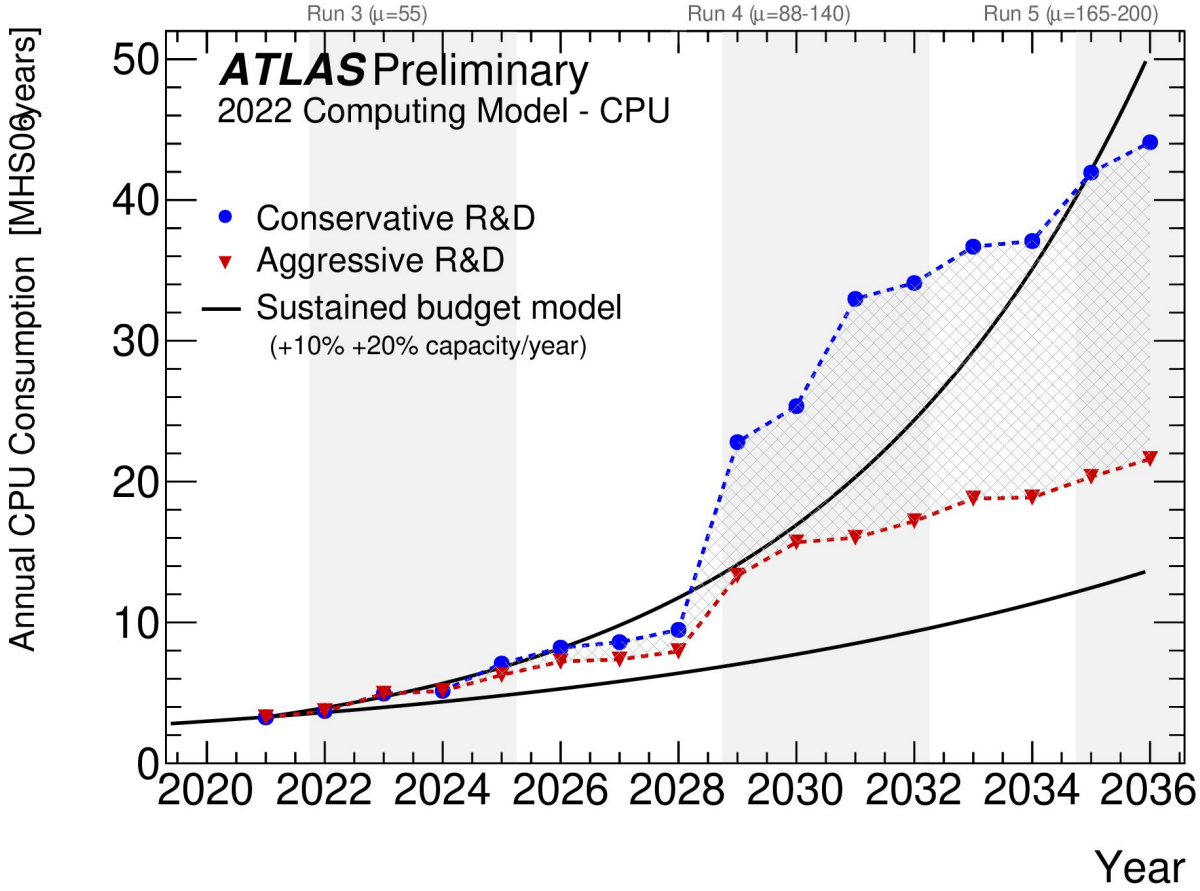
LHCC Recommendations

- LHCC met in September; report only recently available
- No news on the action to prepare benchmark analyses for HL-LHC
- New review of trigger rates and MC events for the general experiment sessions (previously they mostly glossed over these)
- Starting to think about 2026 and beyond (next slides)

- The **LHCC commends** the WLCG and the experiments on the successful data taking in 2023 and the efficient use of the computing resources.
- The **LHCC proposes** to set up a mechanism to review the evolution of the experiment-specific parameters, such as trigger rates and numbers of MC events, needed each year to determine the required resources, given the LHC parameters provided by the LPC.
- The year 2026, the first of the LS3, is expected to introduce sizable changes in the proportions each medium is needed by each experiment. The **LHCC appreciates** having a coarse projection of the computing resource needs of each experiment over the next years.
- The **LHCC notes** that broad adoption of small data formats is important and strongly depends on each experiment's policy. The **LHCC recommends** having regular updates on the status of the adoption of the small data tiers.
- The **LHCC is impressed** by the outcome of the IRIS-HEP/HSF training initiative, which proves to be unified and scalable. Sustainability is more fragile since training requires significant resources and efforts must be incentivised. The **LHCC supports** this initiative and **encourages** the participation of new instructors and mentors. The **LHCC notes** that better coordination in the organisation of general (basic) and experiment-specific courses could improve the learning of newcomers. Along the same lines, the **LHCC suggests** the experiments to make use of the HSF training centre to expose their own courses, which will increase the visibility of both training initiatives.

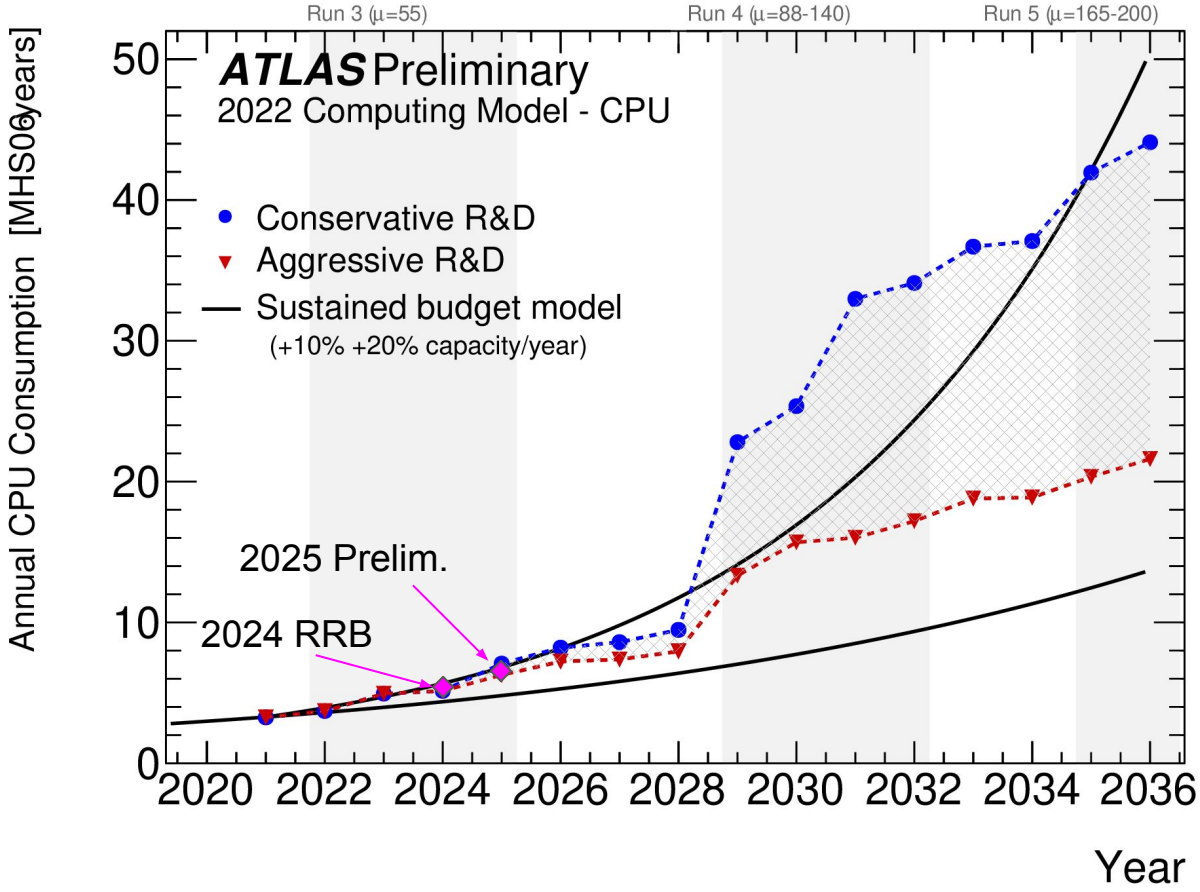
Planning for 2026 and Beyond

- You are all no doubt deeply familiar with our HL-LHC projections by now



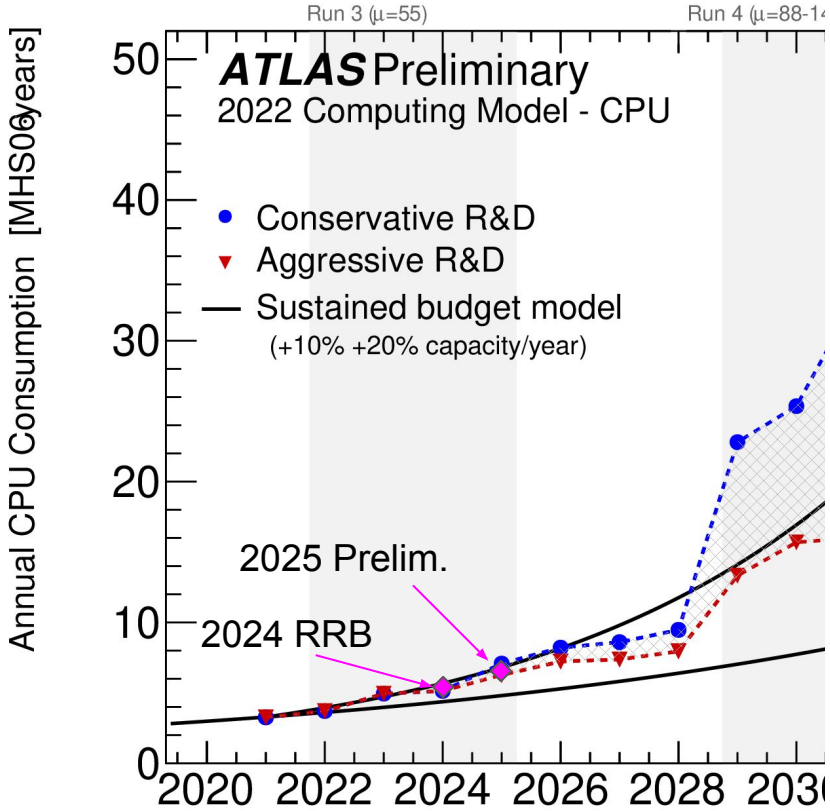
Planning for 2026 and Beyond

- You are all no doubt deeply familiar with our HL-LHC projections by now



Planning for 2026 and Beyond

- You are all no doubt deeply familiar with our HL-LHC projections by now



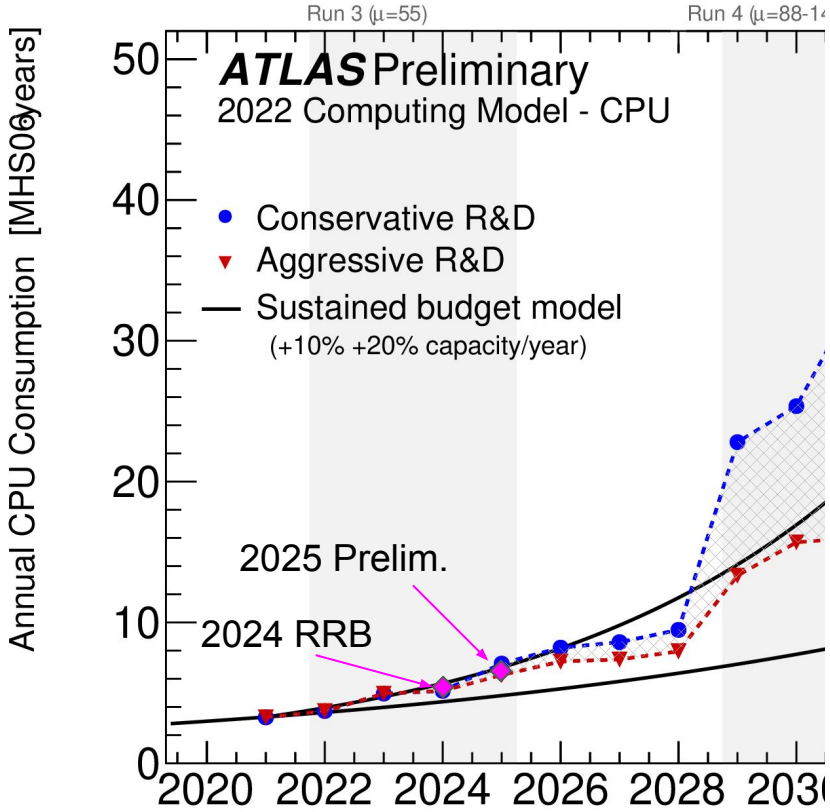
2025 Prelim.: **6672 kHS23**
HLT contrib.: 600 kHS23
HPC etc.: 1300 kHS23
WLCG sites: 3691 kHS23
(Rest is Tier0 contributions)

2026:
HLT contribution → Zero (unplugged)
→ 16% increase in WLCG site requirements assuming *no* change to MC, *no* additional data processing, etc

(The plan included increases, which is why you see the point go up in 2026)

Planning for 2026 and Beyond

- You are all no doubt deeply familiar with our HL-LHC projections by now



Current biggest change is from 2028 to 2029

2028: 7.8 – 8.4 MHS23
2029: 11.8 – 21.2 MHS23

Increase is 50–150%; roughly the same as the integrated change from 2029 to 2036

With this large single-year increase coming, **how can we formulate our requests** to best help you ensure that you are able to deliver resources in the HL-LHC period?

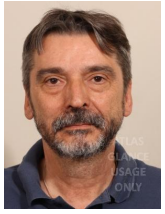
Computing Coordination Updates

S&C Coordination


Software Coordination



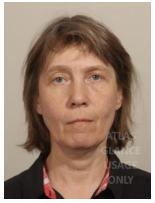
ADC Coordination




ADAM Coordination


Welcome!

- Alexei Klimentov (ADC)
- Nurcan Ozturk (ADAM)
- John Chapman (Software)
- Vangelis Kourlitis (AMG)
- Jana Schaarschmidt (Simulation)
- Serhan Mete (Core Software)
- Caterina Marcon (SPOT)
- Tatiana Ovsiannikova (SPOT)
- Marcelo Vogel (Database Operations)
- Ruggero Turra (Physics Validation)
- Yanlin Liu (MC Production)
- Emma Torro (Derivations)
- Rod Walker (Workflow Management)
- … and a few still TBA



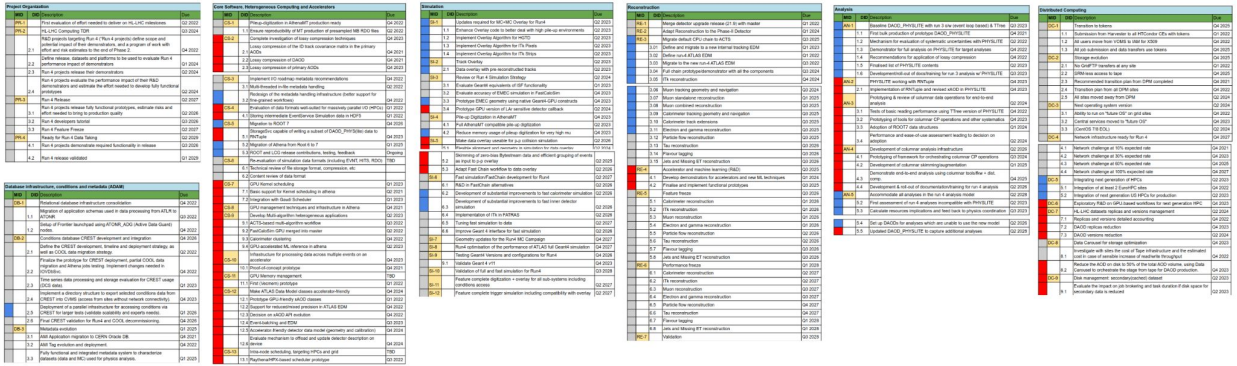
Oxana has been re-elected for another two years as ICB Chair

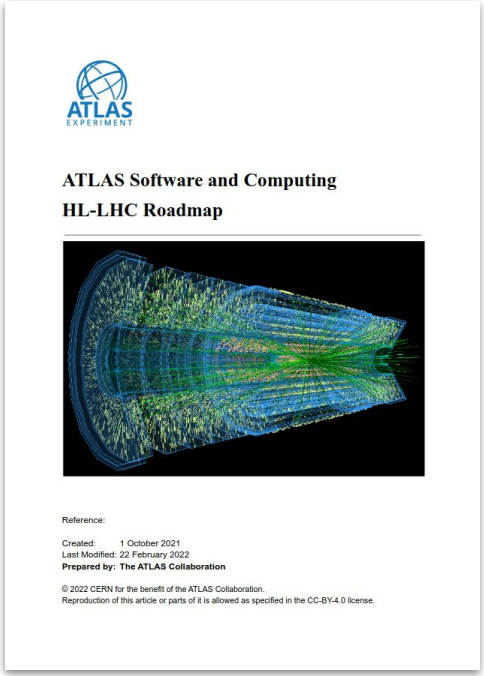
'You need to go home, take a long relaxing bath surrounded by aromatic candles and do an hour of yoga; but that's out of the question. How about a five minute smoking break?'

Looking forward: The road to HL-LHC

- LHCC performs a series of reviews of the Software and Computing plans of the LHC experiments towards HL-LHC
  - The ATLAS HL-LHC Computing Conceptual Design Report was published in May 2020

- A follow up ATLAS Software and Computing HL-LHC Roadmap was published in March 2022 with clearly defined *milestones*





ATLAS EXPERIMENT

ATLAS Software and Computing HL-LHC Roadmap

Reference:

Created: 1 October 2021
Last Modified: 22 February 2022
Prepared by: The ATLAS Collaboration

- An ATLAS HL-LHC Computing TDR is planned for 2024

ATLAS Distributed Computing milestones

Distributed Computing				
	MID	DID	Description	Due
	DC-1		Transition to tokens	Q4 2025
		1.1	Submission from Harvester to all HTCondor CEs with tokens	Q1 2022
		1.2	All users move from VOMS to IAM for X509	Q4 2022
		1.3	All job submission and data transfers use tokens	Q4 2025
	DC-2		Storage evolution	Q4 2025
		2.1	No GridFTP transfers at any site	Q1 2022
		2.2	SRM-less access to tape	Q4 2025
		2.3	Recommended transition plan from DPM completed	Q4 2021
		2.4	Transition plan from all DPM sites	Q4 2022
		2.5	All sites moved away from DPM	Q2 2024
	DC-3		Next operating system version	Q2 2024
		3.1	Ability to run on "future OS" on grid sites	Q4 2022
		3.2	Central services moved to "future OS"	Q4 2023
		3.3	(CentOS 7/8 EOL)	Q2 2024
	DC-4		Network infrastructure ready for Run 4	Q4 2027
		4.1	Network challenge at 10% expected rate	Q4 2021
		4.2	Network challenge at 30% expected rate	Q4 2023
		4.3	Network challenge at 60% expected rate	Q4 2025
		4.4	Network challenge at 100% expected rate	Q4 2027
	DC-5		Integrating next generation of HPCs	Q2 2023
		5.1	Integration of at least 2 EuroHPC sites	Q4 2022
		5.2	Integration of next generation US HPCs for production	Q2 2023
	DC-6		Exploratory R&D on GPU-based workflows for next generation HPC	Q4 2023
	DC-7		HL-LHC datasets replicas and versions management	Q2 2024
		7.1	Replicas and versions detailed accounting	Q4 2022
		7.2	DAOD replicas reduction	Q4 2023
		7.3	DAOD versions reduction	Q2 2024
	DC-8		Data Carousel for storage optimization	Q4 2023
		8.1	Investigate with sites the cost of Tape infrastructure and the estimated cost in case of sensible increase of read/write throughput	Q4 2022
		8.2	Reduce the AOD on disk to 50% of the total AOD volume, using Data Carousel to orchestrate the stage from tape for DAOD production.	Q4 2023
	DC-9		Disk management: secondary(cached) dataset	Q2 2023
		9.1	Evaluate the impact on job brokering and task duration if disk space for secondary data is reduced	Q2 2023

Maintenance & Operations | Conservative R&D | Aggressive R&D

- Several HL-LHC milestones related to distributed computing
  - Not a static list! Regularly reviewed, updated and/or expanded
  - New milestones defined since the roadmap publication
  - Essentially two types: Maintenance and Operations, R&D
- ADC Maintenance and Operations Milestones
  - These are essential changes, needed “just to get by”
    - *Tokens, evolution of storage technologies/access protocols, OS changes, network/data challenges, ..*
  - More details on these milestones in the ATLAS report at the WLCG Workshop in Lancaster last November and in other talks this week
- ADC R&D Milestones
  - Conservative R&D: New developments achievable with current effort
  - Aggressive R&D: New developments requiring extra effort
  - These are translated into “Demonstrators”, described in the following

The road to Run-4 (R2R4)

- ADC has 14 items on the R2R4 work plan

DC-1	Transition to tokens	*Job submission, queues, IAM, transfers*
DC-2	Storage evolution	*GridFTP, DPM, SRM-less tape*
DC-3	Next operating system version	*CentOS EOL, future OS*
DC-4	Network infrastructure ready for Run 4	*Data Challenges, SDNs*
DC-5	Integrating next generation of HPCs	*EuroHPC, US HPC*
DC-6	Exploratory R&D on non-x86 resources and next generation HPC	*ARM, GPU*
DC-7	HL-LHC datasets replicas and versions management	*DAOD versions/reduction/recreation*
DC-8	Data Carousel for storage optimisation	*Tape cost, AOD reduction, smart writing*
DC-9	Disk management: secondary(cached) dataset	*Impact on job brokering*
DC-10	Evaluation of commercial cloud resources	*Google, Burst scheduling, TCO*
DC-11	Optimising the user analysis experience	*Site evaluation, tails, waiting times*
DC-12	Sustainability in ATLAS Computing - session	*Failed wall times, potential adaptations*

Details: Mario's talk from last week

Sustainable ADC

- ADC is delivering great results!
  - Thanks to a lot of very dedicated individuals
- Some observations
  - Overall available personpower is slowly but surely shrinking
    - Temporary placements (6m-1y) to fill the most critical gaps
    - Adds significant overhead to find funding, replacements, work reshuffling and prioritisation
  - Senior technical people are increasingly spending time in coordination
    - Replenishment by new & young people stagnated
    - Important projects and tasks have become dormant
  - Many areas are dependent on single individuals
  - But people also fragmented over many different areas and tasks: Lots of 5% - 10% OTP work classifications
- Can we do better?
  - Gain more time for development and operations
  - Clearer communication flow
  - Set up a more sustainable coordination framework
  - Relax the historical boundaries within ADC
  - Make it easier for site contributions - we're thinking of CRC here
  - Attract new people (and excite the oldtimers!)

New ADC Organigram

ATLAS DISTRIBUTED COMPUTING
OCTOBER 2023

ADC COORDINATION

Mario Lassnig, Alexei Klimentov

PHYSICS

Production Coordination
M. Borodin

Analysis Coordination
A. Forti

Centralised Production
- Monte Carlo Production
- Group Production
- Data Reprocessing
- Physics Validation
- HLT Reprocessing

Physics Analysis
- User Analysis Tools
- Analysis Model Group
- DAST

FABRICS

Computing Run Coordination
Rotating appointment

Infrastructure
- Tier-0
- Grid
- HPC
- Cloud
- BOINC
- Analysis Facilities

Operations
- DA Operations
- DPA Operations
- Central Services
- CRIC
- HammerCloud
- Monitoring
- ADCoS

DATA MANAGEMENT

Coordination
S. McKee, P. Vokac

Development
- Rucio

Operations
- Rucio Deployment
- DDM Central Operations
- Monitoring

R&D
- Networks
- Caches
- Storage
- Cloud

WORKFLOW MANAGEMENT

Coordination
A. Pacheco, R. Walker

Development
- Workflow Definition
- Workload Management
- Workload Execution

Operations
- System Deployment
- BigPanDA Monitoring

R&D
- Data Analytics
- Analysis Facilities
- Cloud
- HPC

Organisation and communication

- ADC Coordination Board
  - Consists of the area coordinators and meets weekly
  - The CRC has the option to join if they want to have a bigger say in things!
- Rewritten mandates
  - Previously they were very detailed, almost like job adverts ("do task X", "look after service Y", …)
  - Completely rewritten to focus on the expected outcomes of the activity areas
  - The coordinators are your point of communication, no need to funnel through ADC Coordination
  - Give the coordinators the freedom to act without micromanagement from the top
  - Give the engineers the freedom to act in a more grassroots fashion
- Recognition of Task Forces
  - No more static “boxology”, but a more flexible way to deal with limited duration, focused topics
  - (And it's the way we do things anyway…)
- ADC Weekly
  - Very positive response — people seem to enjoy that it's short and to the point!
  - Possibilities to improve the meeting? Let's discuss!

S&C events in **2023** and **2024**

- 16–20 October: ATLAS Week, CERN
- 16–20 Oct: Rucio & Dirac Workshop, KEK, Japan
- 30 Oct–3 Nov: Induction Day and in-person Tutorial (remote 13-17 Nov)
- 31 Oct–1 or 2 Nov: A GPU Tutorial will be held (indico, info)
  - Day 3 may focus on pythonic GPU usage, including analysis. Register soon!

- S&C Weeks:
  - Feb 5–9, 2024
  - June 10–14, 2024 (TBC…)
  - Oct 7–11, 2024
    - CERN rooms are booked for all; proposals welcome from hosts for an external week

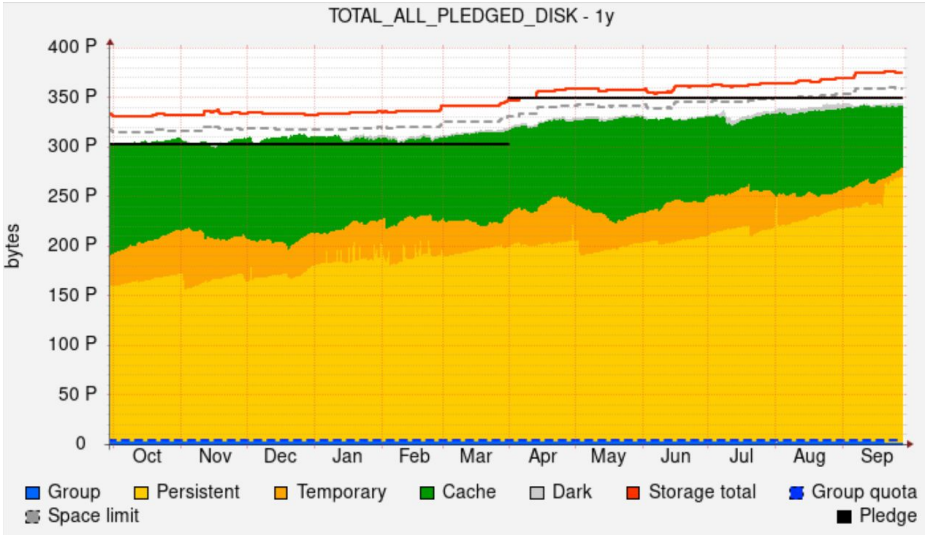
- 10–15 March 2024: ACAT in Stony Brook, NY, USA
  - Start preparing your abstracts soon!!
- 19–25 Oct 2024: CHEP, Krakow, Poland

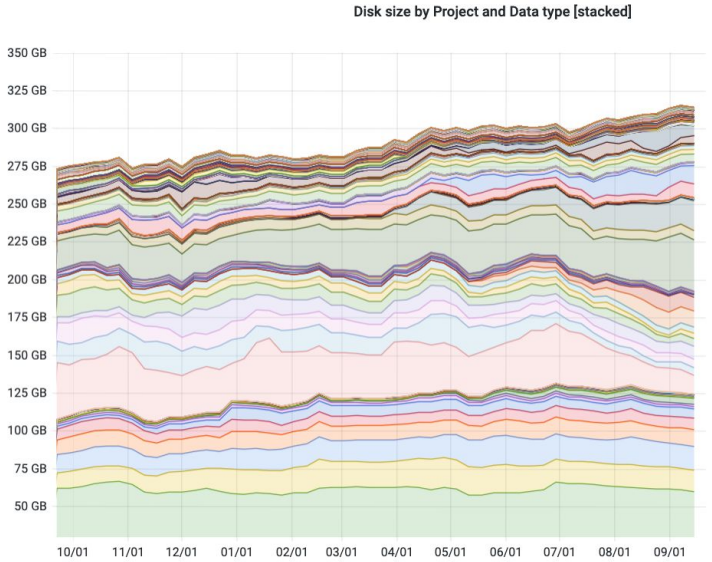
If you know of other conferences, please make sure they're on the list

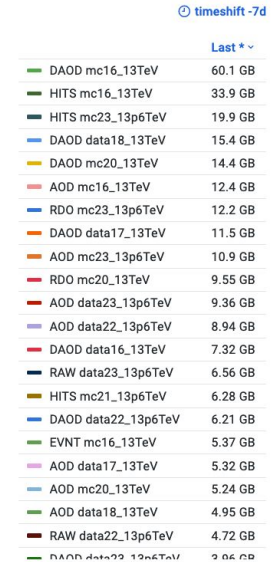
Extras

Resource Usage – Disk – Last 12 months

- Business as usual, no major issues at the moment.
- Space still dominated by R21 MC16, but getting closer

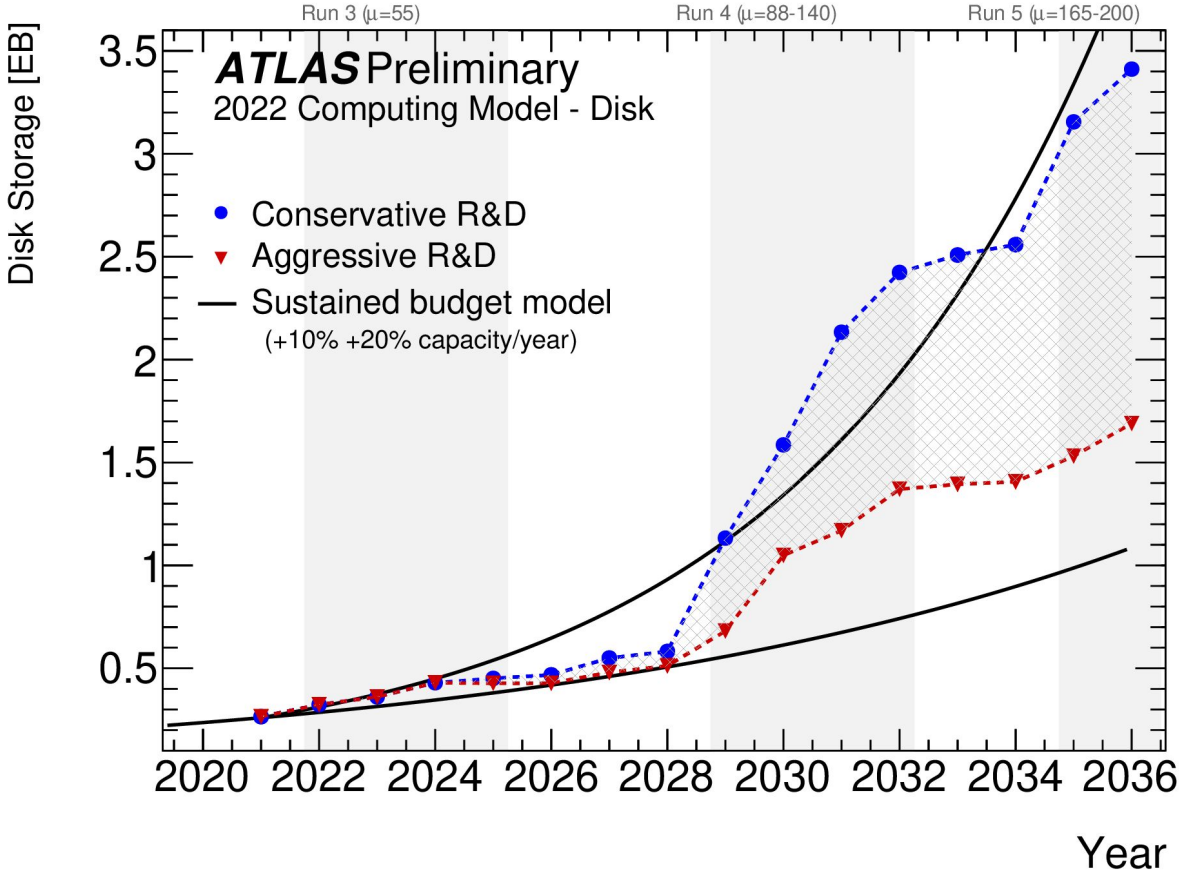






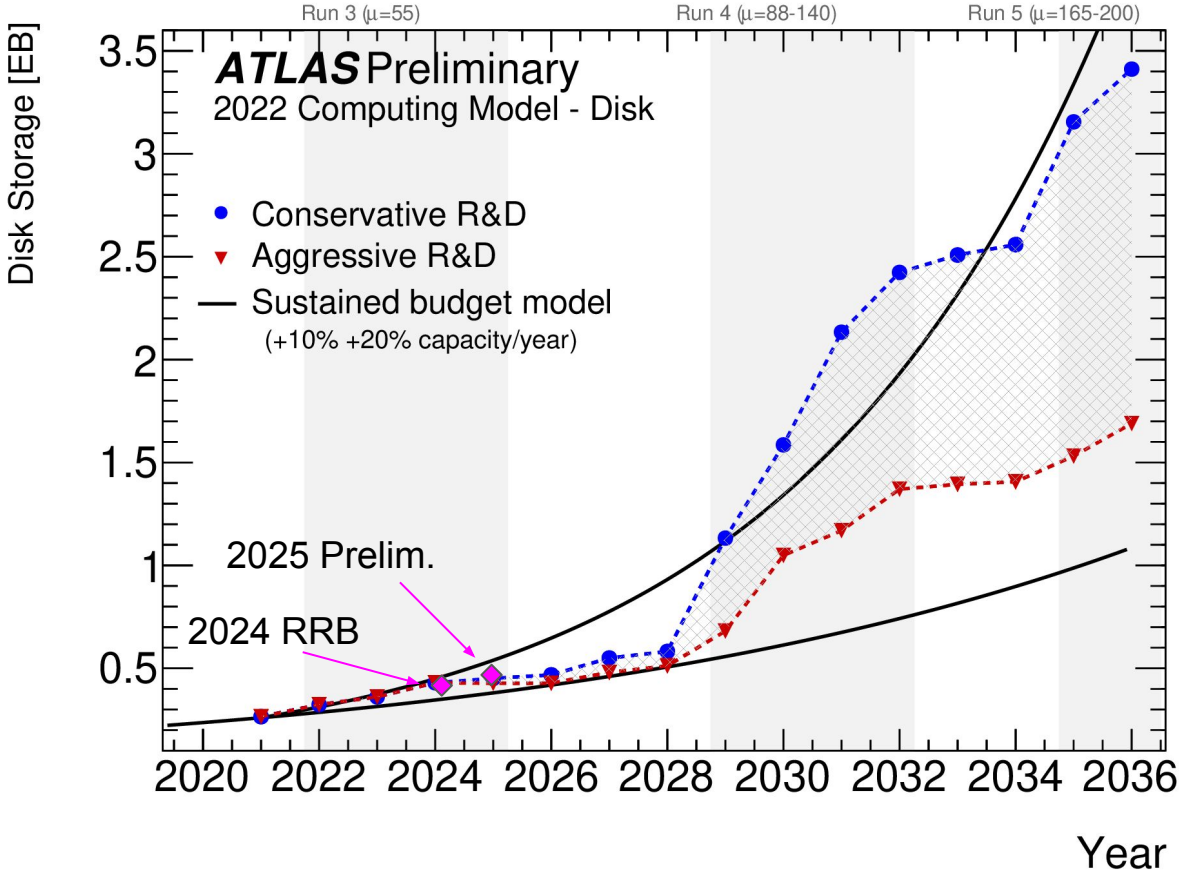
Planning for 2026 and Beyond

- You are all no doubt deeply familiar with our HL-LHC projections by now



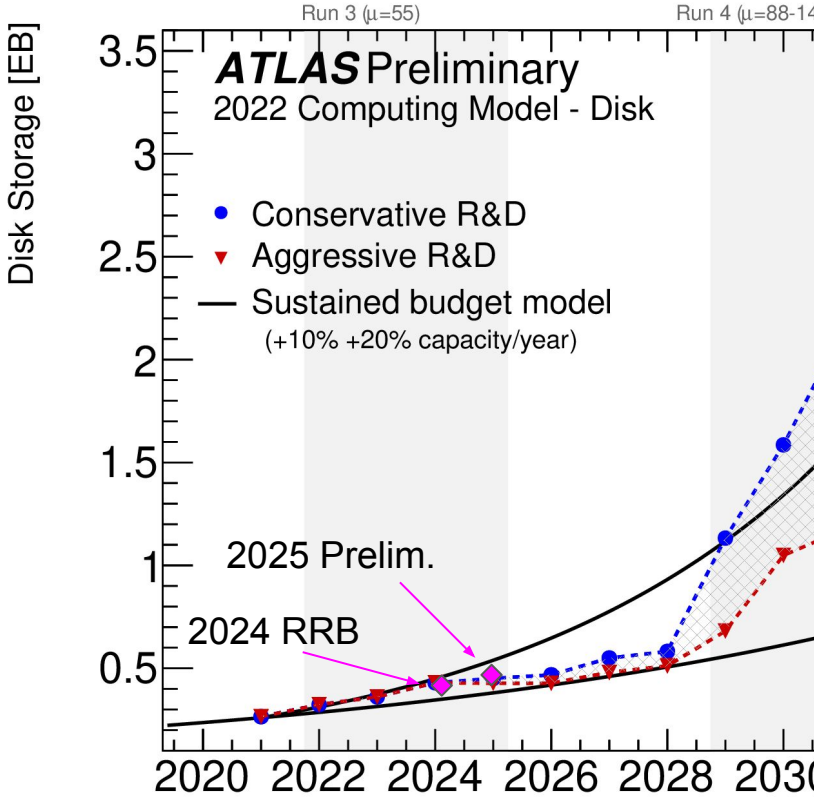
Planning for 2026 and Beyond

- You are all no doubt deeply familiar with our HL-LHC projections by now



Planning for 2026 and Beyond

- You are all no doubt deeply familiar with our HL-LHC projections by now



2025 Prelim.: **469 PB**
Tier0 contrib.: 56 PB
WLCG contrib.: 410 PB

2028: 550–610 PB
2029: 680–1220 PB
2030: 1050–1630 PB

23–100% increase in 2029
55–34% increase in 2030