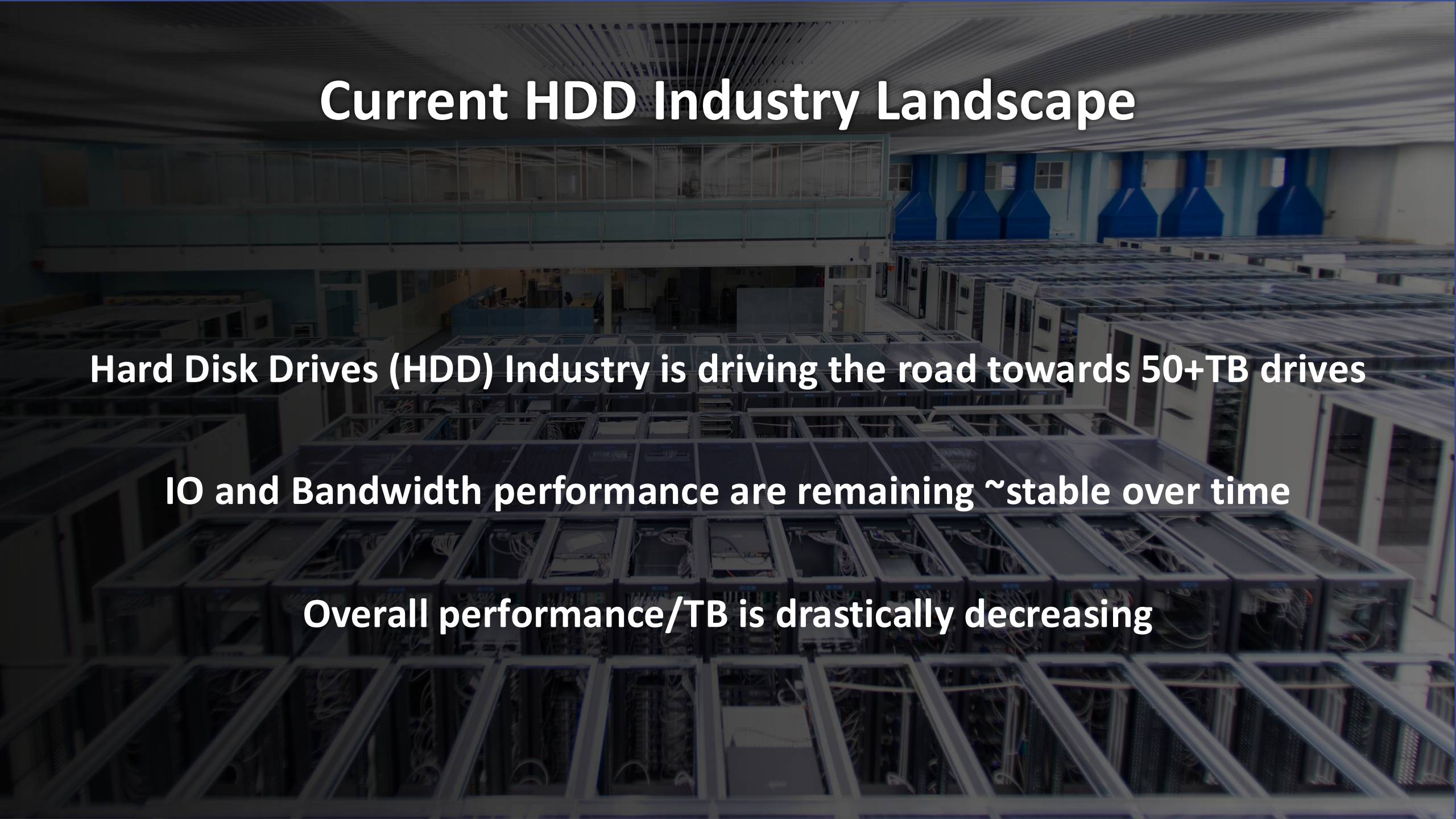
Next-Generation Exascale Flash Storage

CERN openlab PURESTORAGE®

CERN openlab collaboration

Luca Mascetti
CERN openlab Storage CTO
Storage and Data Management

Current HDD Industry Landscape



Hard Disk Drives (HDD) Industry is driving the road towards 50+TB drives

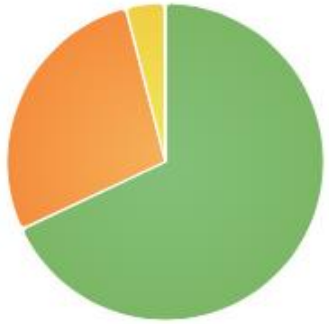
IO and Bandwidth performance are remaining ~stable over time

Overall performance/TB is drastically decreasing

CERN Physics Storage – EOS – 2024 in numbers

Total amount of files read	Total amount of bytes read	Total amount of files writte	Total amount of bytes written
41.9 Bil	8.23 EB	7.54 Bil	1.20 EB

Export: Amount of files read

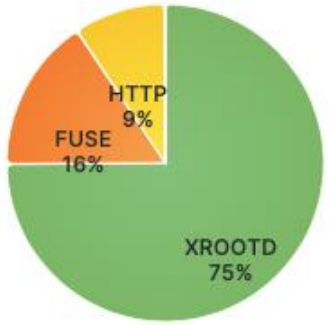


	Value	Percent
XROOTD	28.5 Bil	68%
FUSE	11.7 Bil	28%
HTTP	1.62 Bil	4%
GRIDFTP	1.20 Mil	0%

Ingestion: Amount of files written

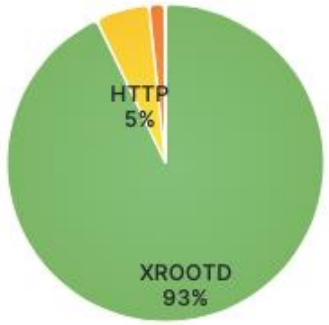
	Value	Percent
FUSE	5.01 Bil	66%
XROOTD	1.79 Bil	24%
HTTP	742 Mil	10%
GRIDFTP	1.02 Mil	0%

Total data read per protocol and instance: All



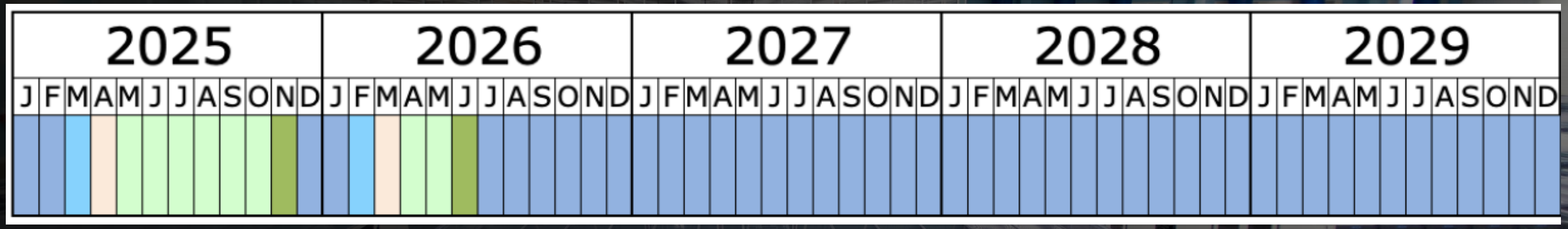
	Value	Percent
XROOTD	6.18 EB	75%
FUSE	1.29 EB	16%
HTTP	763 PB	9%
GRIDFTP	39.2 TB	0%

Total data write per protocol and instance: All



	Value	Percent
XROOTD	1.11 EB	93%
HTTP	65.0 PB	5%
FUSE	18.1 PB	2%
GRIDFTP	20.8 TB	0%

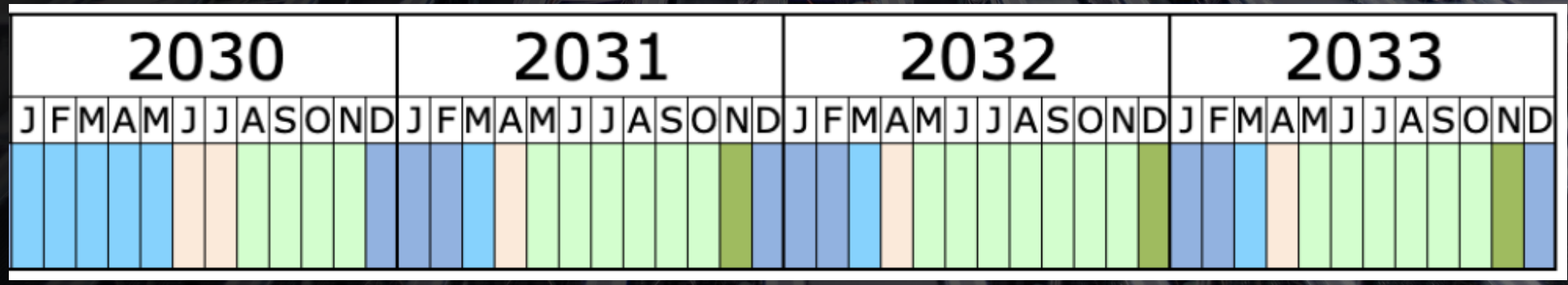
Hi-Luminosity LHC Schedule



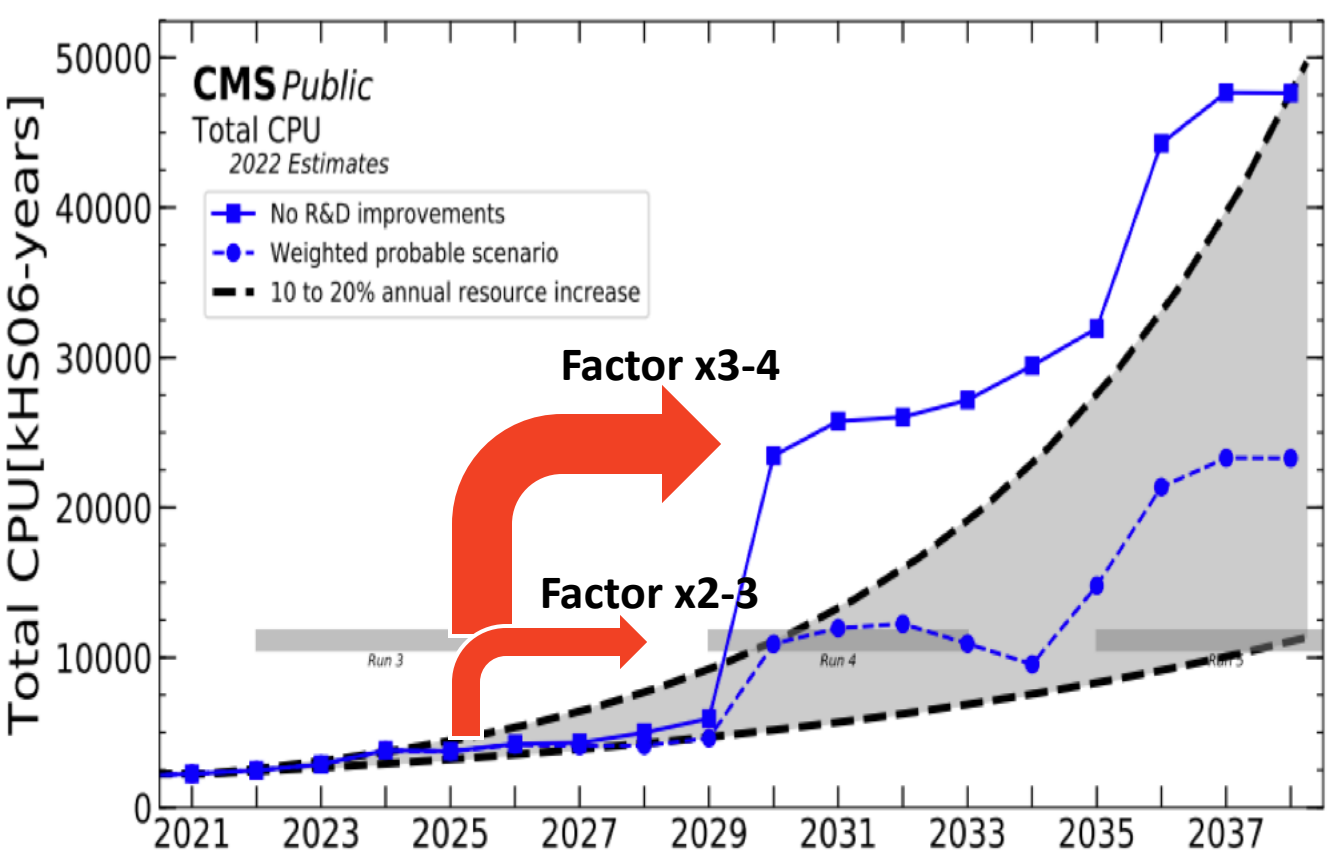
Extended LHC Run3
Operations

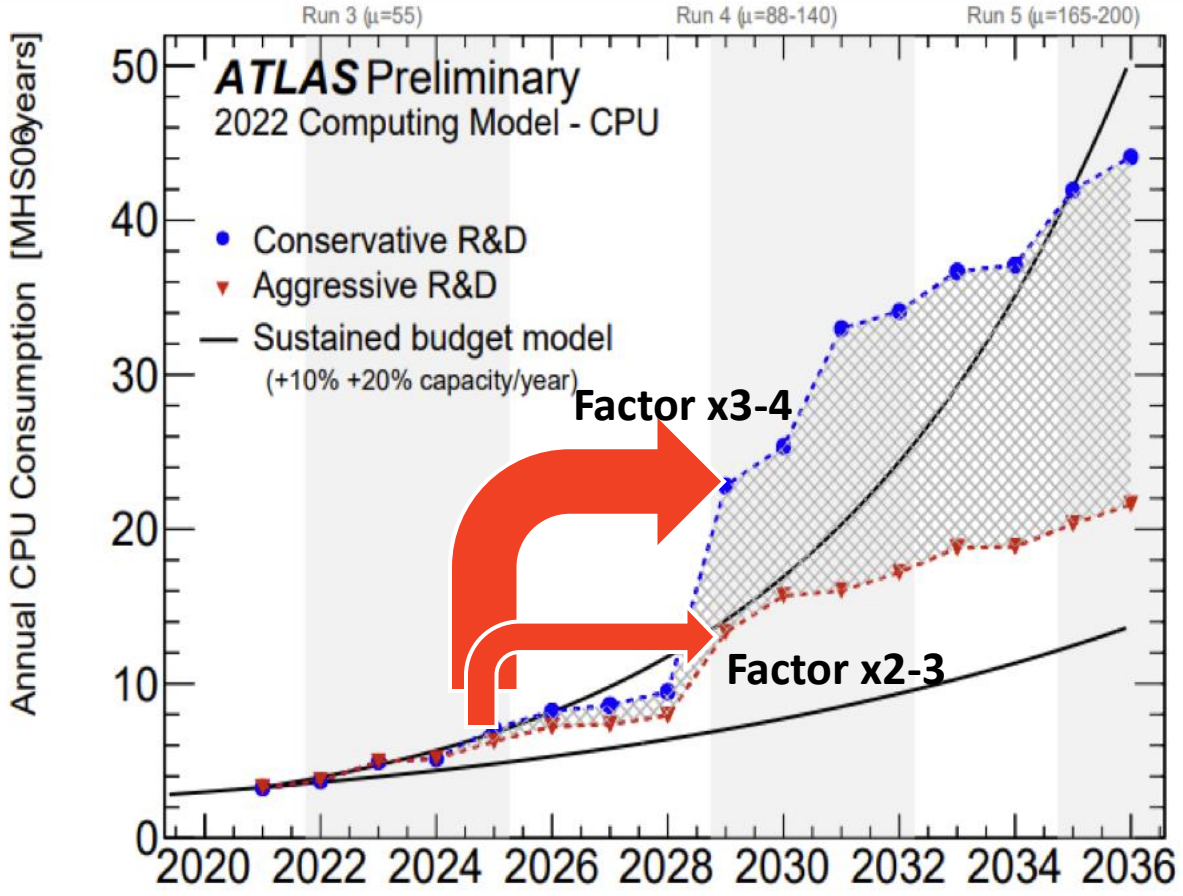
Long Shutdown 3
Accelerator Complex and Experiments Upgrades

Hi-Luminosity LHC Operations
Run4



Experiments' Computing Model



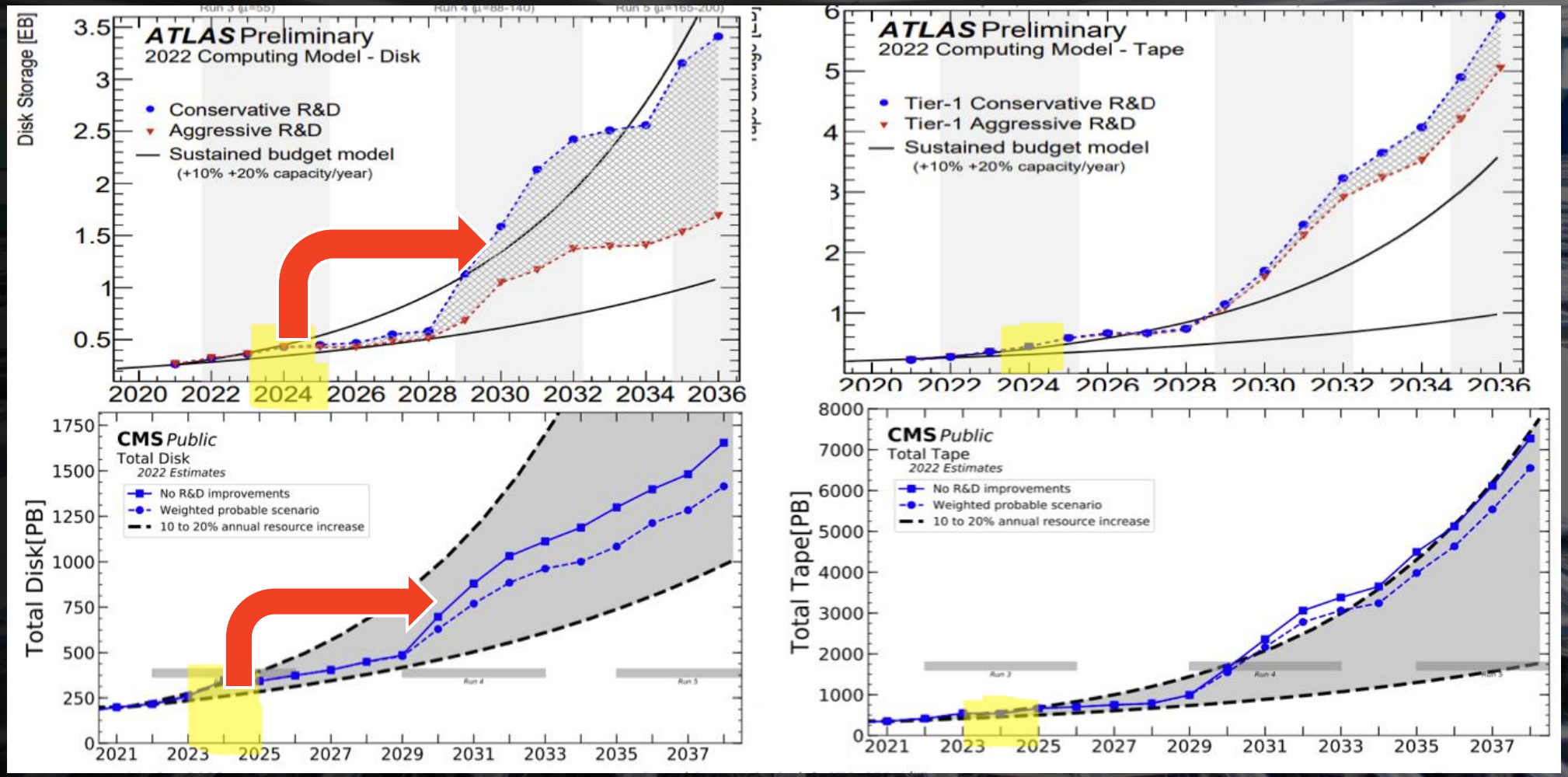


With ~7M HS06 (CMS+ATLAS) in WLCG we see 220k parallel streams on CERN Physics storage system (EOS)

- How many parallel stream should we expect with 50 MHS06?

- In Q4-2024 CMS demonstrated remote reconstruction from WLCG Tier-1s reading directly from CERN
  - Will this be a general trend for the future?

Experiments' Storage Model



In ~2030 CERN IT will need to provide ~3-4x more of the current storage capacity to LHC Experiments

Note: capacity does not automatically translate into required performance!

What is the role of SDDs?

More and more use-cases (e.g. AI, ML, etc..) have unusual IO patterns

We expect to see in the future more random access (both read and write) to our storage. SSD will be fundamental to address these new requirements.

Can we “transparently” integrate these to our distributed systems?

Is the initial cost (drop) promise of SSD still valid?

In the future manufacturers will providing JBOD-like and internally build redundancy to their storage

- SMR, HAMR, NANDs (Triple-Level-Cells, Quad-Level-Cells, Penta-Level-Cells)
- Built-in trade-offs between performance, reliability, endurance, price and capacity

Next-Generation Exascale Flash Storage

Exploring next-generation large-scale, high-density and low-cost storage methods using high performance storage resources based on NAND flash memory technologies.

Investigate the latest flash technologies available in industry to build an high performance and environmentally-friendly scalable solution up to the exabyte level, harnessing the cost-efficiency rooted in low level NAND flash memory technologies.

Motivation, Interests and Benefits

- Ensuring exploration of future NAND technologies
  - Assess and evaluate large-scale impact of these technologies for the coming years
    - Scalability, Performance
    - Space density, environmental friendliness
    - Cost challenges
- Exploit low-level “bit-storing” technologies (PLC, QLC, etc…)
- Need to maintain wider expertise in multiple storage technologies
- Support Storage and Data Management R&D to look forward to future directions

Project Roadmap

Phase 1

- PoC to validate efficiency, environmental impact and cost assessment of a DirectFlash-based solution compared to the current deployment.
- Assess PoC scalability, space density, power efficiency, environmental friendliness and cost challenges.
- Publish scientific material on these results.

Project Roadmap

Phase 2

- Integration of DirectFlash technology into CERN's storage system (EOS) to further increase the scalability and its efficiency.
- Demonstrate DirectFlash and HDDs EOS auto-tiering model
- Demonstrate viability and value in HEP environment, evaluate possible solutions that might replace long-term HDDs.

Project Roadmap

Phase 3

- Exploration of additional benefits in the use of DirectFlash technology, opportunity for new workflows, applications and system design.

- Evaluation of additional area in the storage infrastructure that could be simplified and improved.

Outlook

- Storage flexibility is key!
- Exploration and assessment of new hardware technologies
  - NAND-based storage will become the backbone of storage solutions
- Dedicated software development to help in reducing TCO costs
  - e.g. Auto-Tiering, Auto-Caching, Conversion Policies…
- Lots of interesting work ahead…

Thanks for the attention!

Accélérateur de science