Considerations on storage in the HL-LHC area

Eric Lancon, Doug Benjamin, Vincent Garonne, Chris Hollowell, Qiulan Huang, Tejas Rao, Ofer Rind, Alex Zaytsev

US ATLAS Facilities workshop, SLAC, Nov. 2022

https://indico.cern.ch/event/1201515/

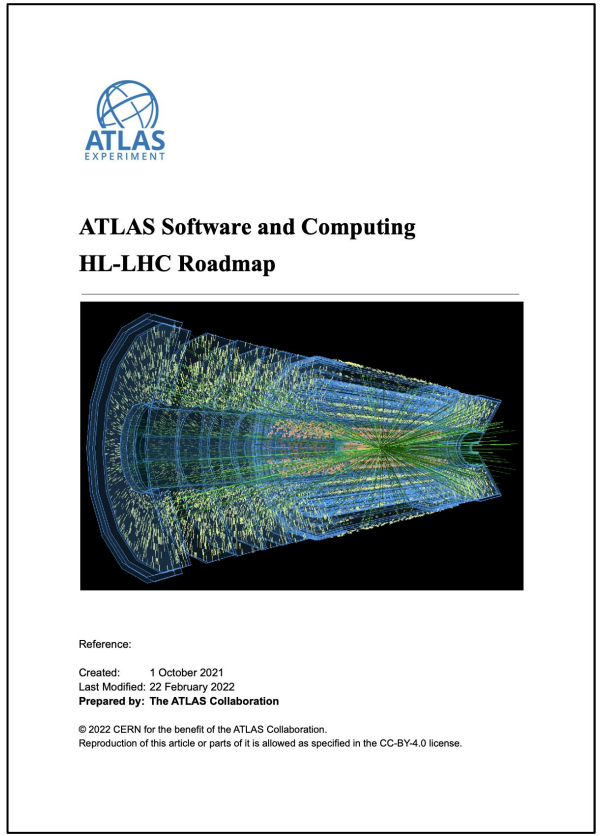
Challenges for Efficient Facility Operation into HL-LHC Era

- Managing anticipated hardware volume for HL-LHC is going to be challenging for facilities, in particular (disk) storage
- Additionally:
  - HEP solutions fall behind current trends and may come with additional costs in a multi-program environment
  - Requirements for Federated Identity and compliance with cyber regulations may be challenging

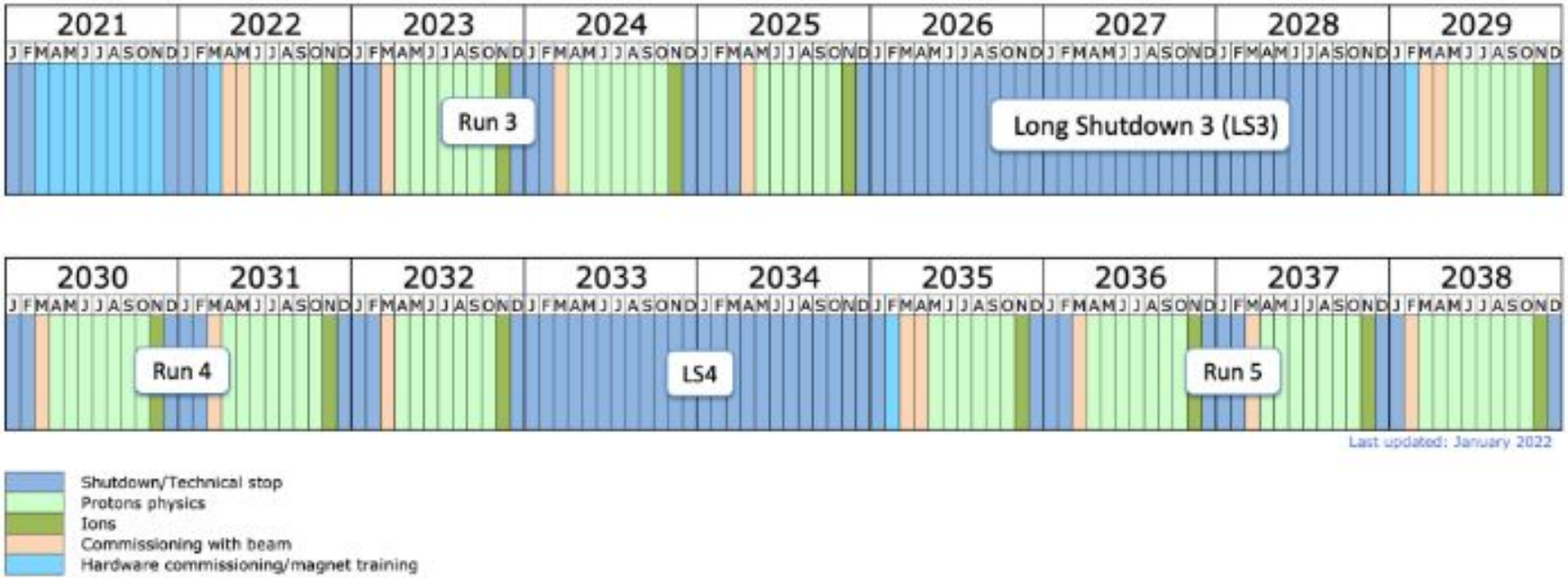
E. Lancon et al., US ATLAS Facilities workshop (SLAC Nov. 2022)

Hardware volume and budget

- Budget exercise for US ATLAS Tier-1 into the HL-LHC era
  - Internal BNL costing model applied to ATLAS hardware forecast (inflation not taken into account)
  - Costing model provides qualitative budgetary assessments into Run4 (2029-2032), derived from hardware requirements
  - Not-surprisingly, costs at Tier-1 facility driven by storage

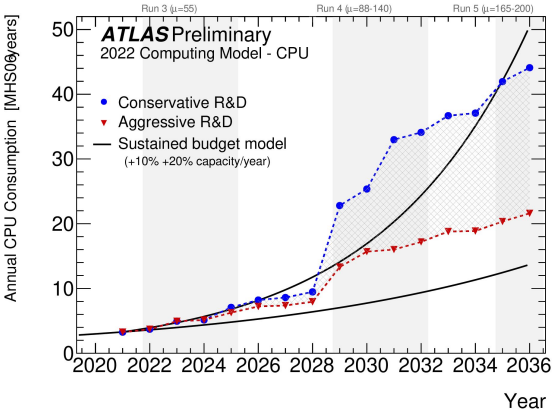


LHC Run 4: 2029 - 2032



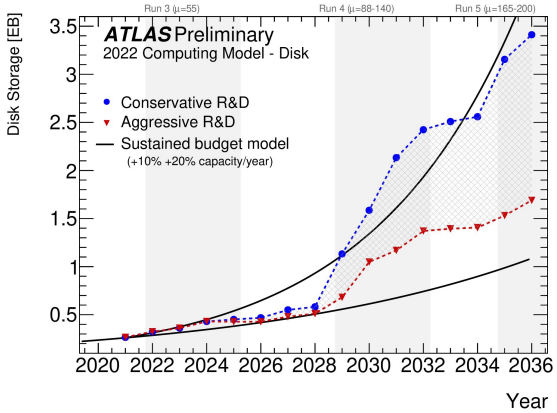
Hardware volume profile into HL-LHC era

CPU



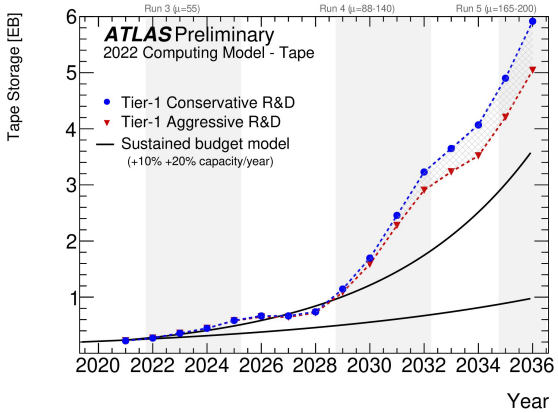
2030: **3** x 2023

Disk



2030: **3** x 2023

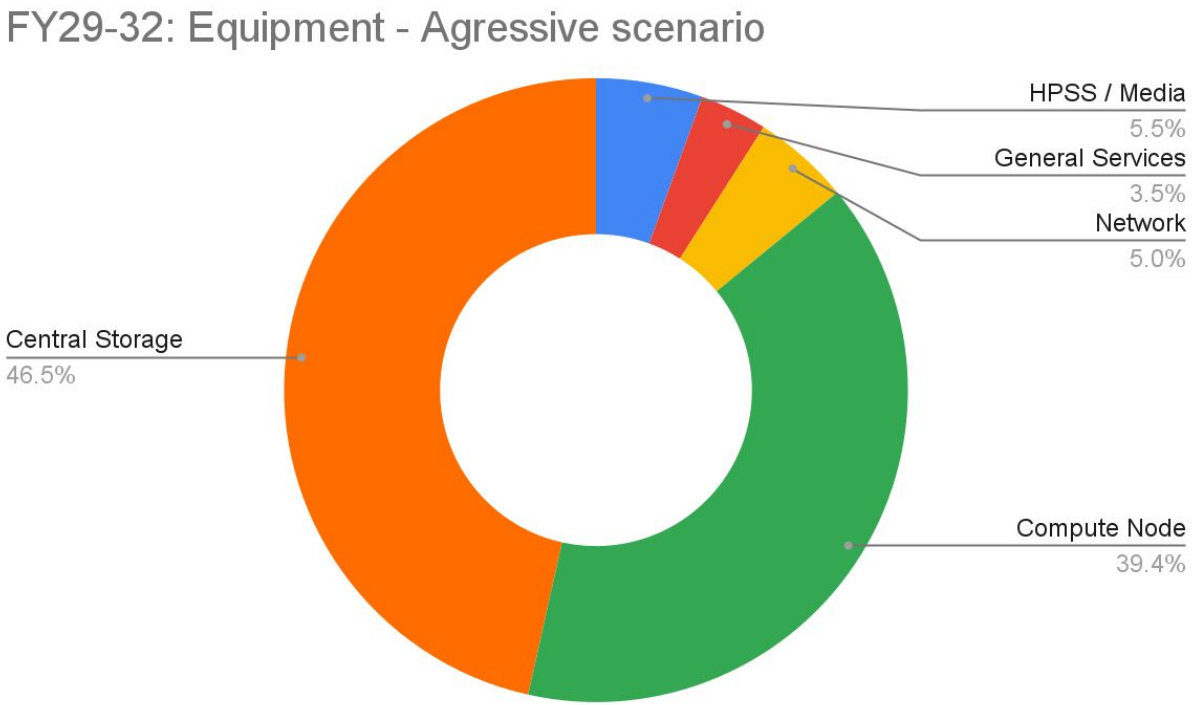
Tape



2030: **4** x 2023

Analysis not included

Storage is the most costly resource

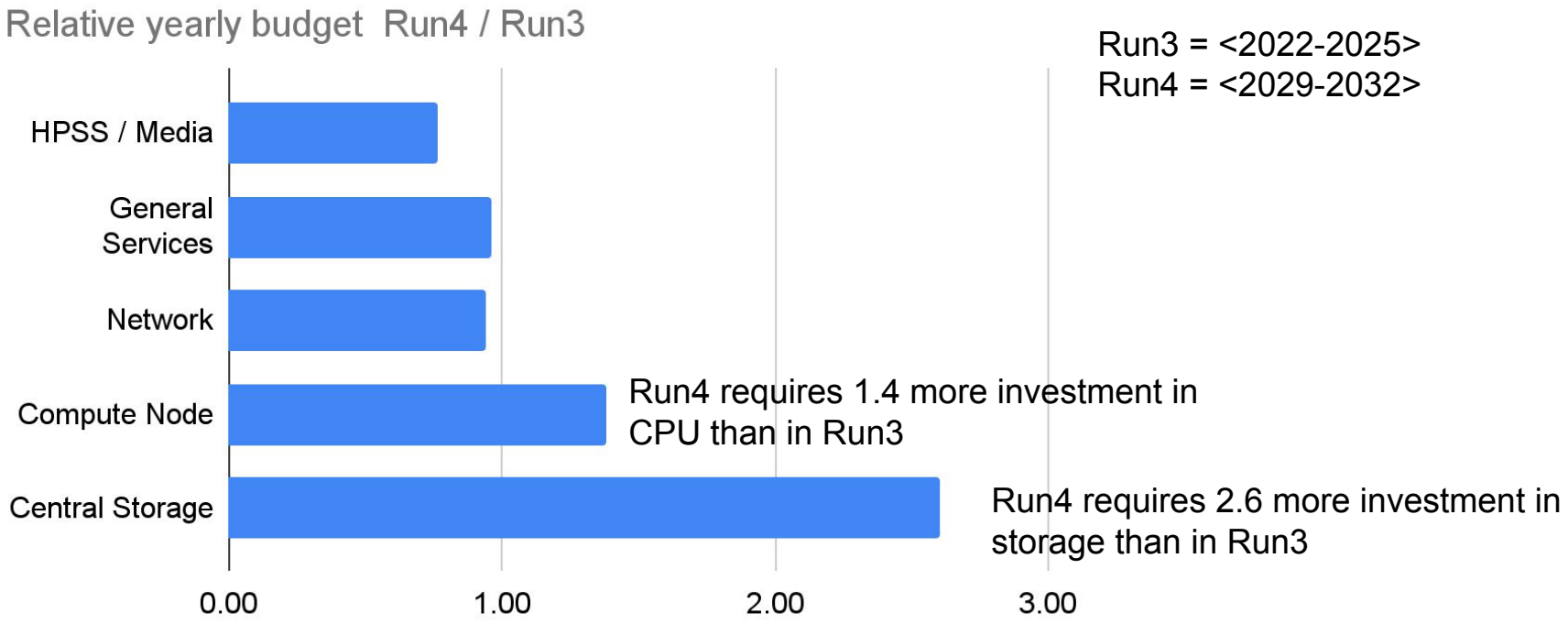


Brookhaven National Laboratory

E. Lancon et al., US ATLAS Facilities workshop (SLAC Nov. 2022)

requires

Relative equipment yearly budget Run4 / Run3



Storage is the most costly resource

How to reduce budget requirement for (disk) storage?

1. **Store less** (requirement is 3x RAW data volume)
   - Address event size (content and improved compression)
   - Versioning,
   - Replication policies.

2. **Store differently**
   - Use of different storage technologies tailored for each usage,
   - Currently one class of storage for all types of data and usages

Storage is the most costly resource

How to reduce budget requirement for (disk) storage?

1. **Store less** (requirement is 3x RAW data volume)
   - Address event size (content and improved compression)
   - Versioning,
   - Replication policies.

Implementation: ATLAS

2. **Store differently**
   - Use of different storage technologies tailored for each usage,
   - Currently one class of storage for all types of data and usages

Implementation: Facilities

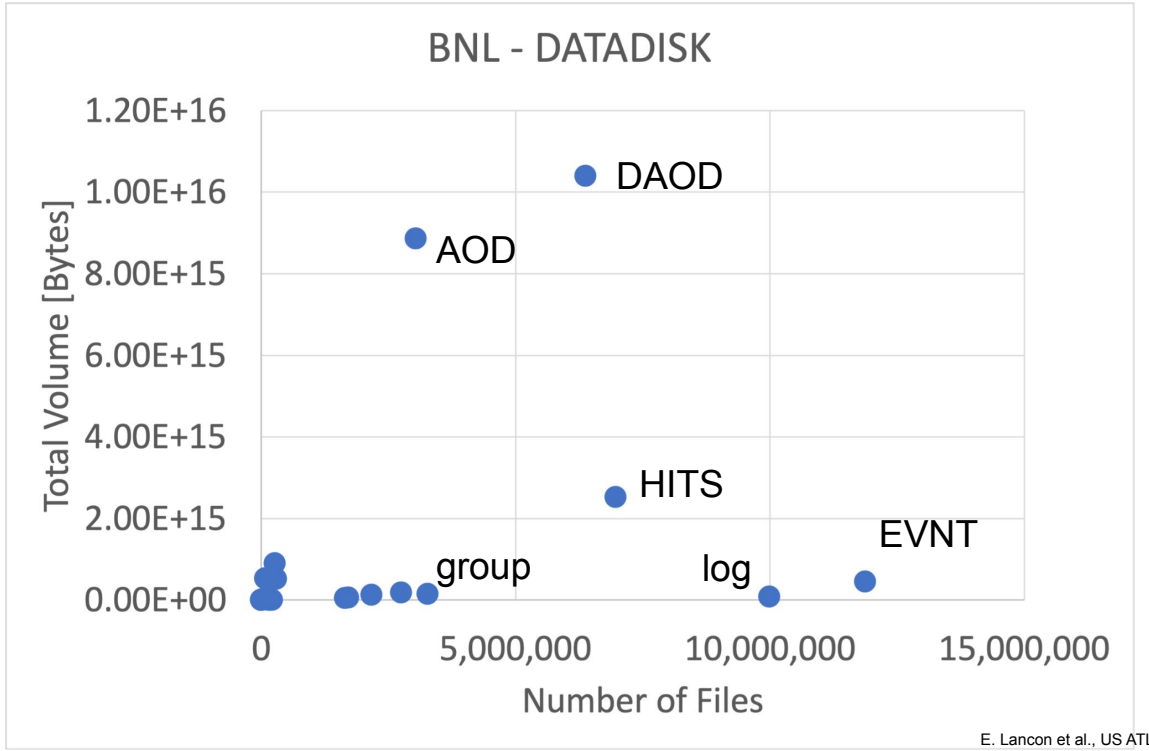
Store differently

- Issues with current disk storage:
  - Filled with warm/cold data
  - All data types are treated the same, even if they have very different values (DAOD have much higher value than logs, Experimental Data has more value than Simulation, …)
  - All data types are expected to be available immediately everywhere
  - Designed for IO while most applications are not IO limited or critical
  - Not even optimized for IO intensive applications like interactive analysis
- More optimal foundation for supporting HL-LHC activities would be:
  - Bulk storage : Object store (better scaling, operational benefits, globally accessible, …)
  - IO intensive: dedicated POSIX storage - high IOPS design
  - Archive/Cold storage: backup/frozen data
  - And a tiered storage solution to effectively leverage storage “classes”

Storage matching workflows

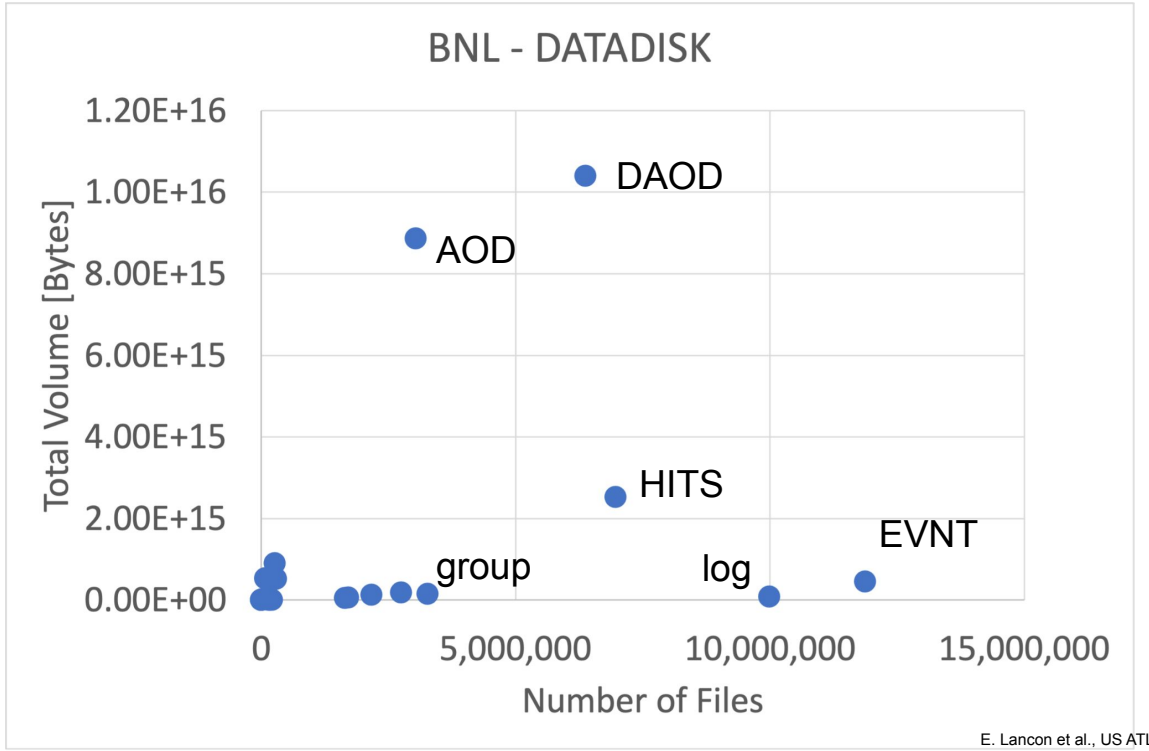
- Different workflows have different storage requirements
  - Production workflows typically spend more time on processing than IO operations
    - Capacity is a more important criteria than IOPS
    - Entire events are read into memory and processed. The IO access pattern is different from user analysis workflows
  - User analysis workflows tend to require more IOPS
    - The IO access pattern is different from reconstruction or simulation. Users use only part of the event record and more random access pattern.
    - IOPS instead of Bulk capacity is the most important optimization criteria.
- Columnar Analysis workflows should benefit from High IOPS flash storage (SSD/nvme)
- New storage architectures <-> new access methods

DATADISK today at BNL:
Total size vs number files per data type



DATADISK today at BNL

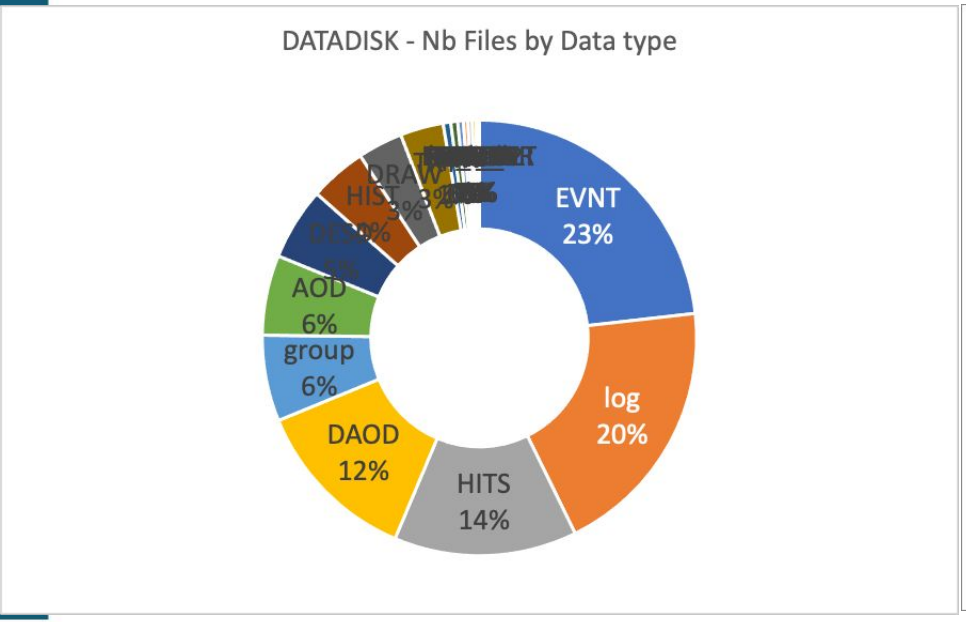
Today:
- Millions of files mostly small files
- Do not require high storage availability

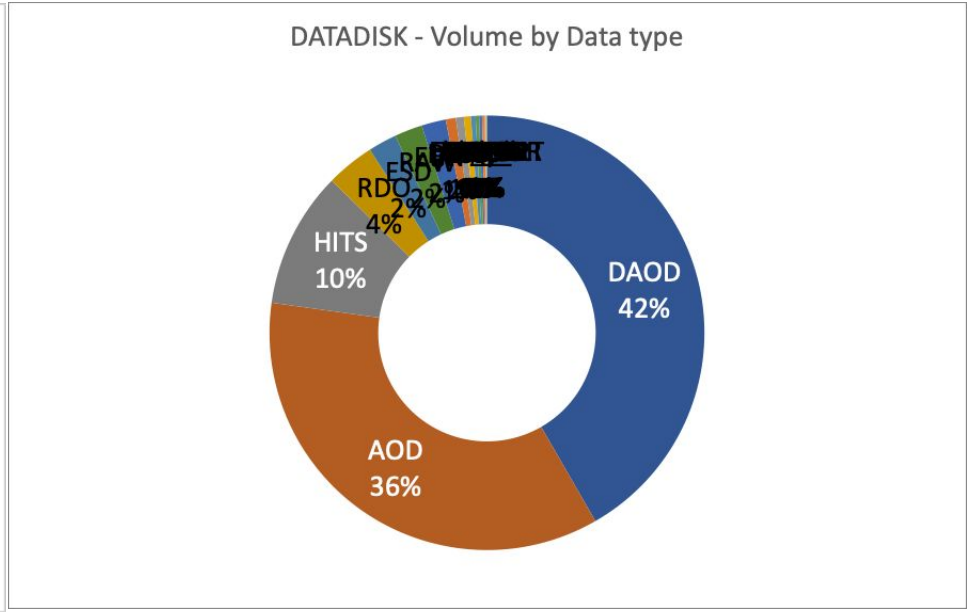


DATADISK polluted by small and low IO requirement files

E. Lancon et al., US ATLAS Facilities workshop (SLAC Nov. 2022)

~50% of namespace for 3% of volume





Let's move small files and low access frequency to a different class of storage

A new class of storage for the tens of millions small files

Let's start with small file case by using Object Store type technology

Advantages

- Object Store scale well for 100s million of files
- Can be deployed on dedicated low capability hardware
- First stage of a multi-tiered storage, next stage would be for high frequency access files on IO performant storage

Implementation

- Needs to be transparent to ATLAS
- Special dCache pools is one possibility
- Storage can possibly be used by all of US ATLAS facilities

Takeaway

- One type of storage for all is not optimal and likely will not scale into the HL-LHC era (3 x today's disk space)
- Need to implement different disk storage solutions for different use cases (workflows)
- Start an R&D implementing differentiation of disk storage with small and unused files