Multilayer Automated Storage (MÁS)

Eric Lançon on behalf of Hironori Ito, Yinzi Wu, David Cameron, Vincent Garonne, Mario Lassnig, Martin Barisits, Cedric Serfon, Doug Benjamin, Xin Zhao, and many more attended MAS meetings.

WLCG/QoS workshop, Feb. 7, 2020

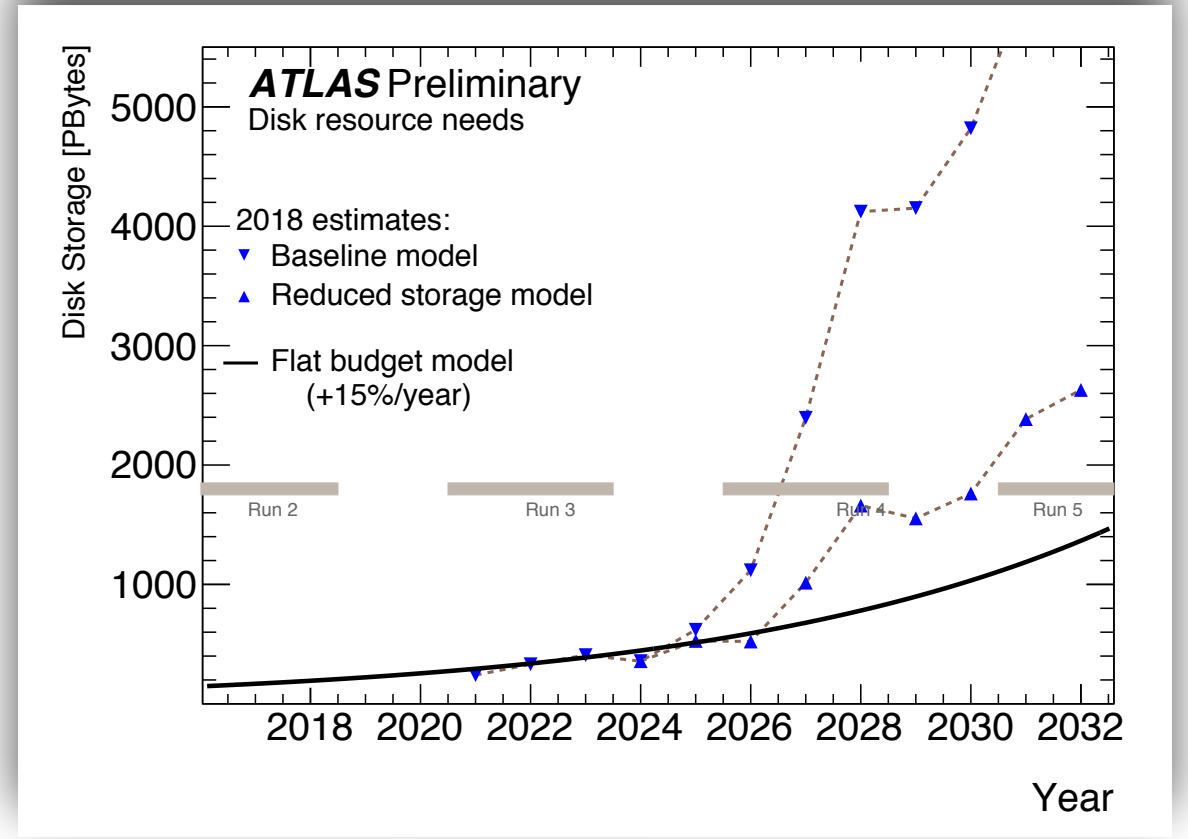
BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

Genesis: THE HL-LHC computing Question: How to reduce storage cost ?

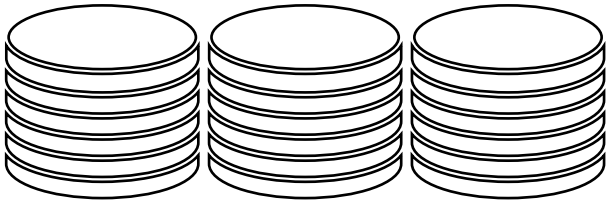
- Many answers
  - A1/ Store less data
    - Event size
    - Data format
    - Versioning
    - …
  - A2/ Reduce data replication
  - A3/ Trade disk for tape
  - A4/ …
  - Ax/ A mixture of the above



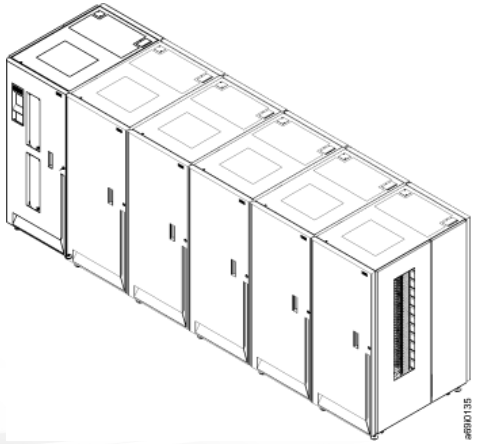
A3/ Trading disk for tape … some of the options

- **Store more data on tape**
  - Need efficient access of tape-resident data
  - Organized access via **Data Carousel**
- **Store differently on disk** and store more data on tape
  - Disk storage is not only about volume…
  - It is about which kind of disk storage? with which local replication factor (typical range: 1 -3)?
  - At which cost?
- With today's implementation, all data have the same 'value' (cost), regardless of their usage and their type
  - Cost of one TB of never used data = cost of one TB of heavily used
  - One TB of log files = one TB of real data
- **MÁS** : R&D investigating intermediate class of storage with dynamic usage of low-cost disk & tape

Today’s storage hardware & software



- Two classes of permanent storage:
  - ‘High performance’ disk
  - Tape
- Both with large disparities in performance and reliability between sites
- With very different costs:
  - Disk / Tape (LTO) cost ~ 5 - 10
  - Depending of disk storage implementation (number of copies, RAID, Erasure formulae,…)
- While non permanent disk storage (lower redundancy and reliability) can have a much lower cost
  - Disk / Tape (LTO) cost ~ 3 - 4



Today's storage usage

- Two types of storage with widely different capabilities and reliabilities between sites: Tape & Disk
- Tape systems are largely underutilized
  - Designed for peak usage (reprocessing)
  - Usage is expected to increase with Data Carrousel
- Disks filled with large fraction of infrequently (never) used data
  - Most data are not used over long periods (several months)
  - From economical perspective they better be stored on tape

QoS in storage

- Today it is black or white
- Clients are expecting data to be on-line (disk) or off-line (tape)
- And they don’t know how fast the data will be delivered
  - Non-consistent delivery performance between sites and configurations
- Because of this binary classification, storage providers implement ‘high-quality' disk storage
  - Even if data is not used
- Quality of the service allows users/clients to specify the performance of the data delivery
  - Clients set the expected performance for the data in need
  - Providers implement storage according to needs

QoS and cost

- Many opened questions
- How to relate QoS to global cost?
- What are the metrics?
  - What is ‘High-Quality’ storage?
  - What is ‘Fast’ storage?
- What is the overall cost optimization function?
- How QoS is guaranteed (SLA?) and monitored ?
- How to pledge resources?
- All these need to be kept in mind…

MÁS what it is?

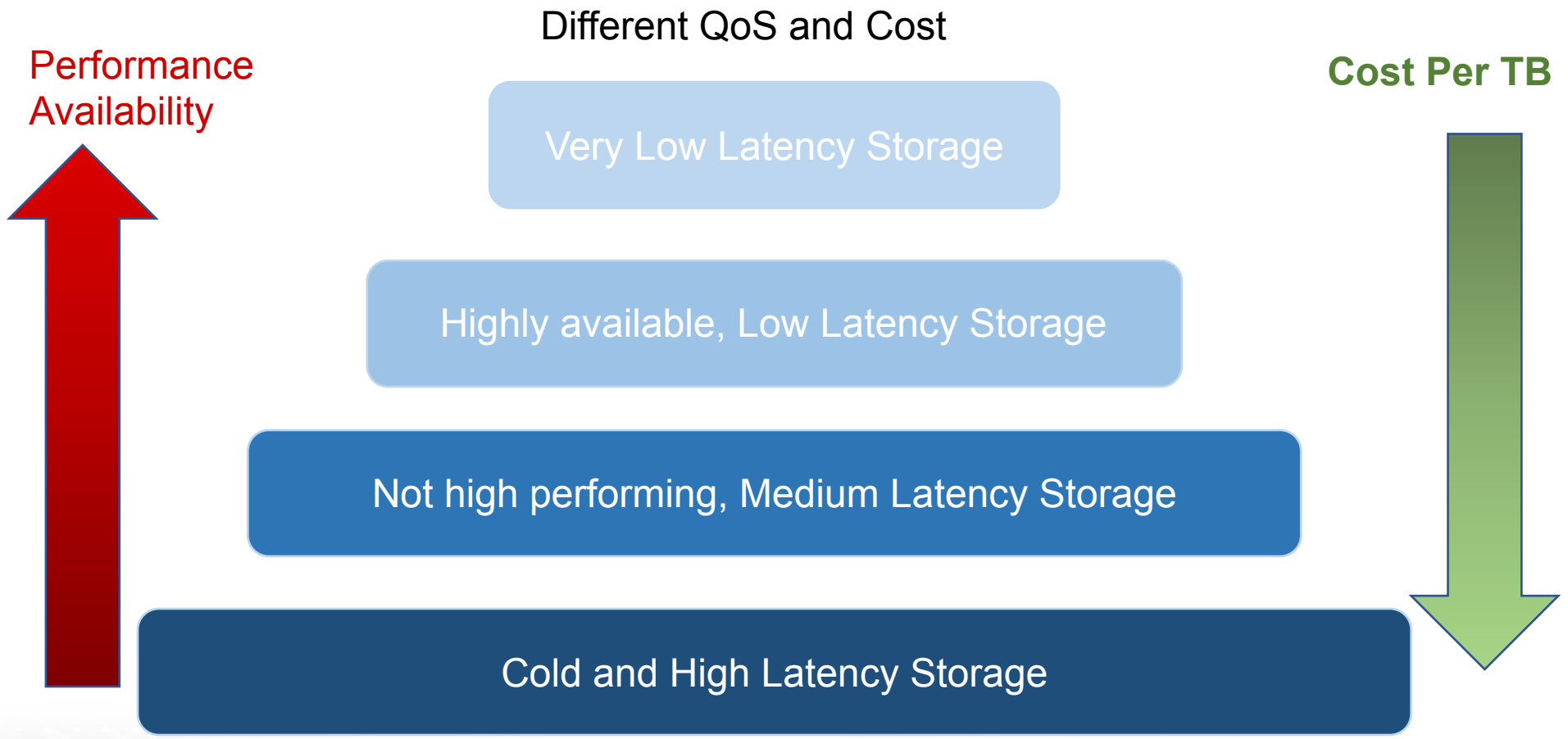
A path away from binary solution

Redefining disk storage as high availability/reliability

Introducing an intermediate storage class

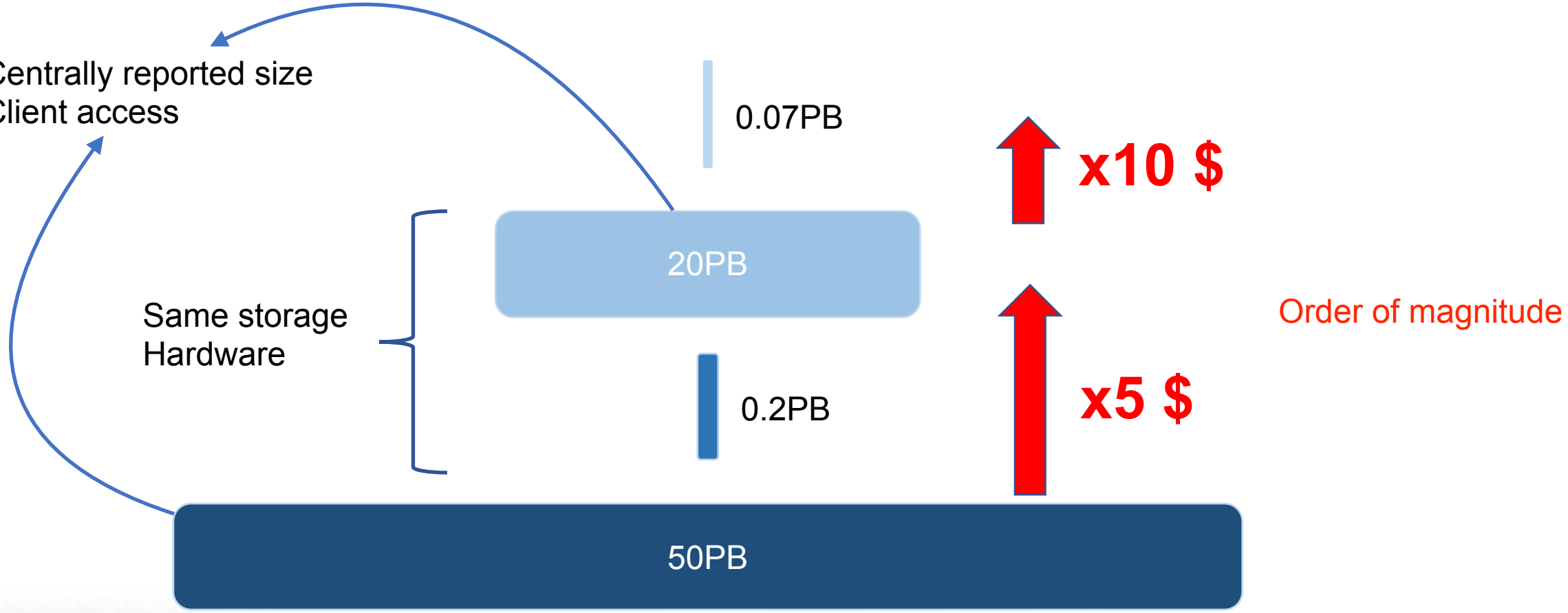
Managing QoS transitions

Multi Layer Local Storage: Cost vs Performance



Size distribution of existing storage (BNL case)

- Centrally reported size
- Client access

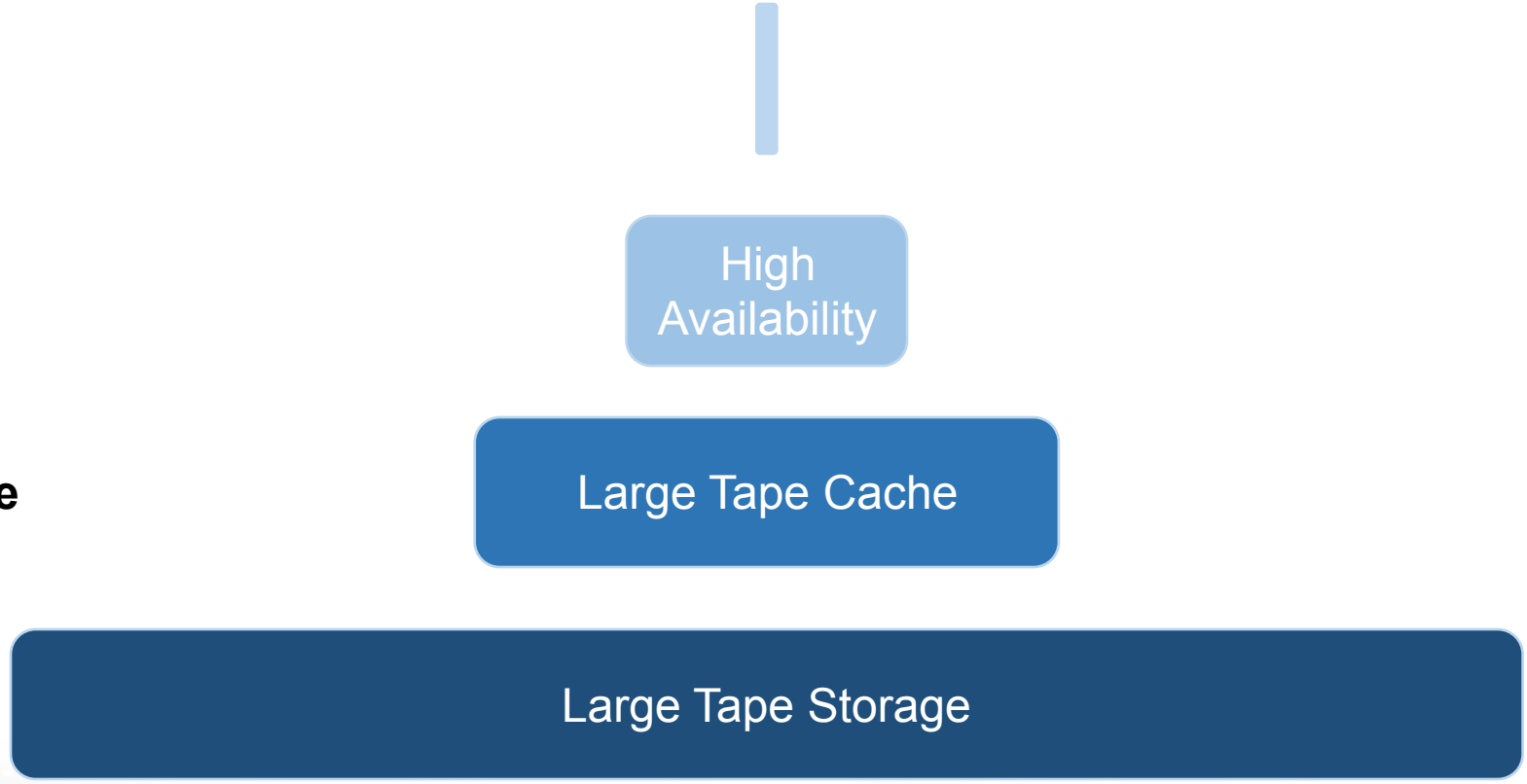


Desired distribution of different storage types

Highly Used data

More Frequently Used Data

Different Hardware



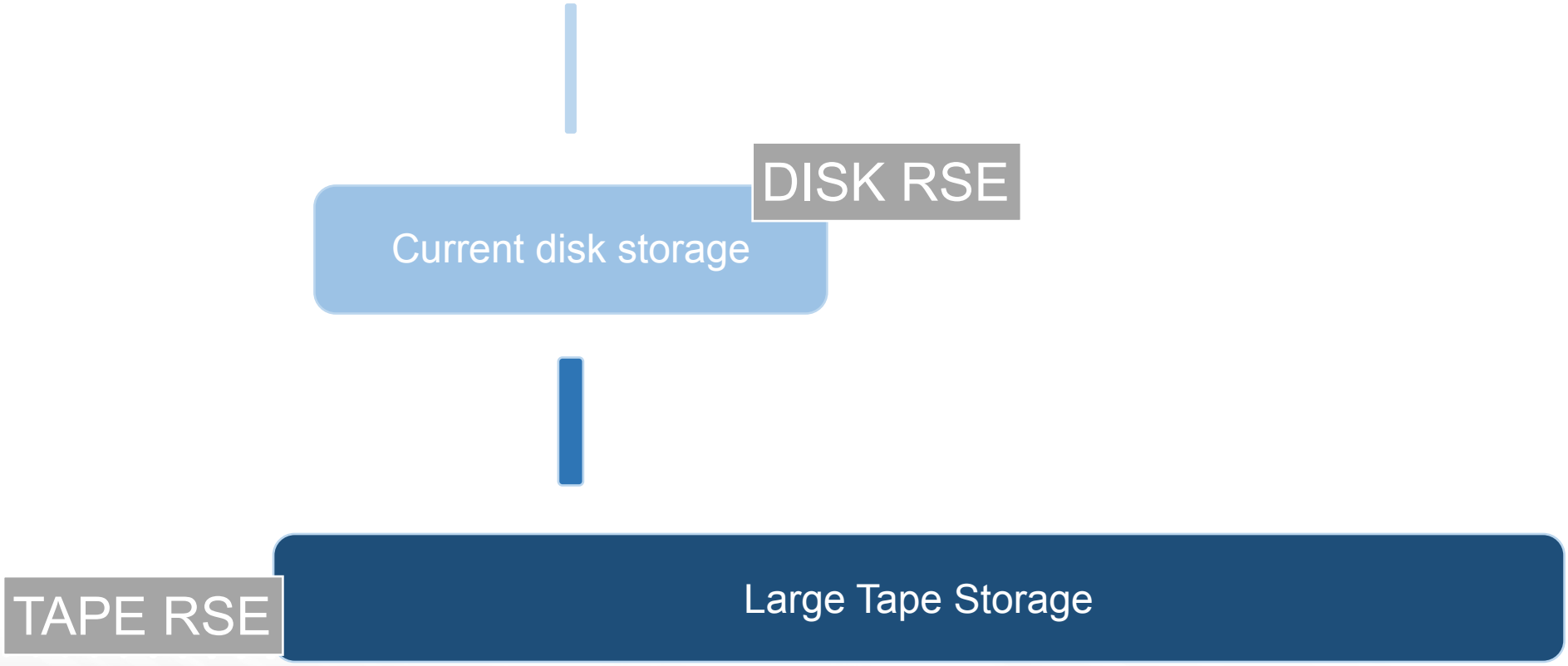
MÁS R&D path - I

- **Trade high performance disk storage for tape & low cost disk storage**
  - To reduce overall cost of storage
- '*High-cost*' disk (HCD) storage reserved for frequently used data only.
  - This is today's disk end-point (N copies on storage)
- Implement large '*Low-cost*' disk (LCD) buffer storage
  - Simple JBOD (no data replication, end-of-life equipment)
- Active data migration between elements:
  1. Data on HCD, untouched for a long period are moved to LCD **and** tape
  2. Data on LCD are removed from LCD when LCD space gets full and kept on tape
  3. Data migrated from tape to LCD when needed
- Data Management System (Rucio) is informed of migration from HCD to LCD (step 1)

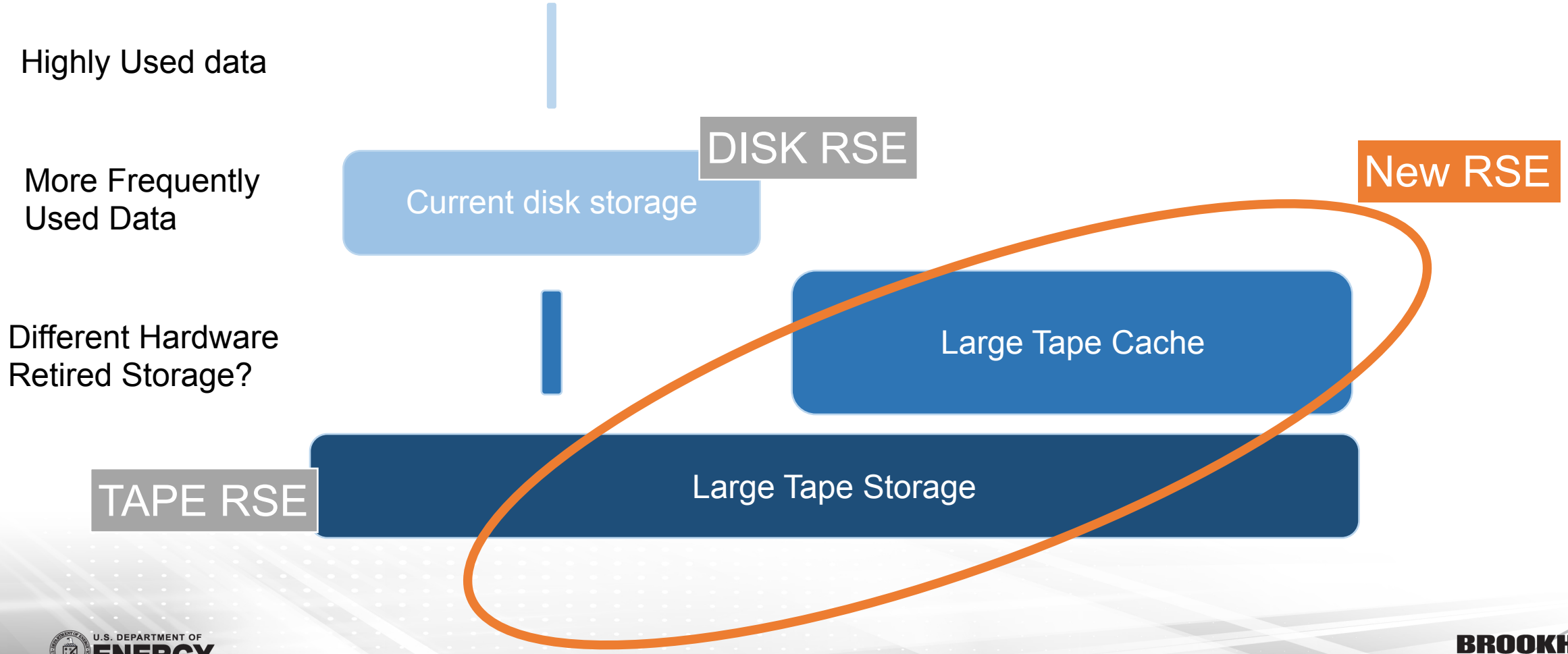
MÁS R&D path - II

- **Implicitly 3 classes (QoS) of storage AVAILABILITY**
  1. High Availability: Disk (High cost)
  2. Medium Availability: Tape and/or Disk (Low cost)
  3. Low Availability: Tape (Very low cost)
- There will still be large differences in performance (CAPABILITY) between sites implementing that model
  - Disk storage configuration
  - Network
  - Tape system
  - They should not be forgotten

Current implementation



MÁS implementation



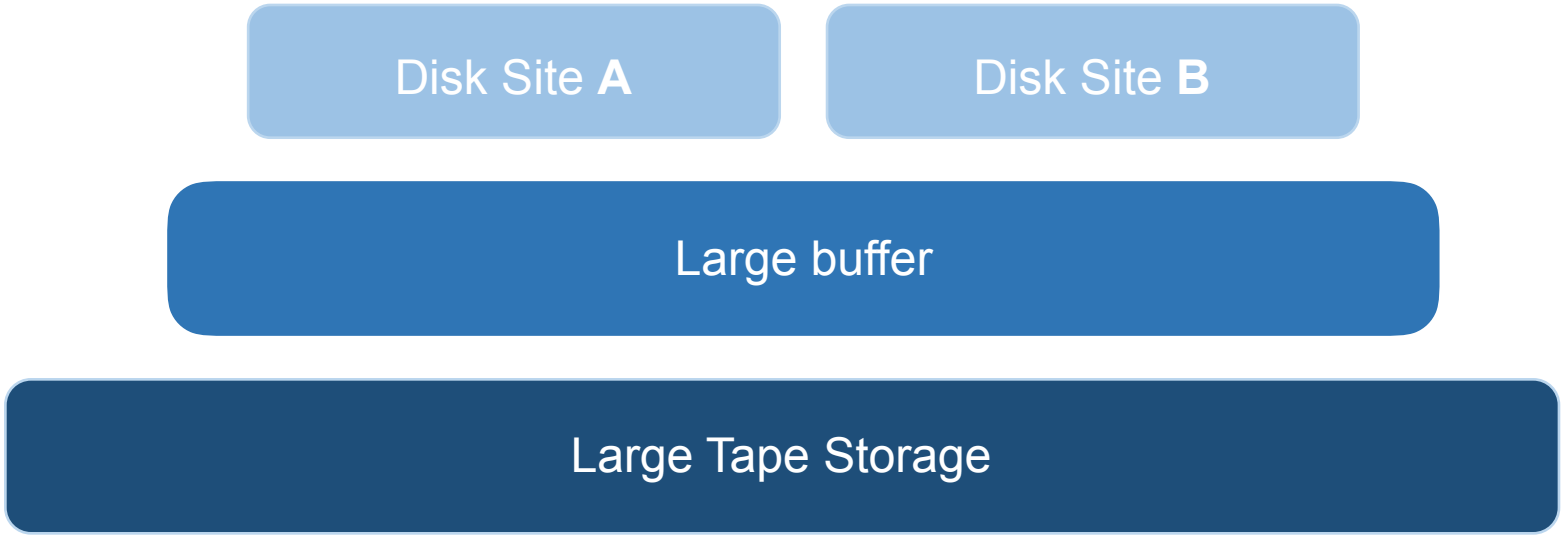
Current status

- New RSE implemented (3PB), Rucio knows it is a low latency end-point by the name
- Using ATLAS Rucio service, unused data identified by previous analyses, are being copied to MÁS end-point from regular disk area.
- Migration status:
  - 1PB of untouched data have been copied so far
  - Data are being removed from regular disk (2 steps procedure for now)
  - All migrations are known to Rucio
- Dedicated PanDA queue has been setup to monitor performance

Steps forward

- Migration to dCache 5.2 (next month) will allow tape+disk data status (but not tape-or-disk)
- Optimisation MAS storage data management:
  - Size,
  - Which data to keep when buffer gets full? FIFO, LRU, TLRU, MRU, … algorithms?
  - Gain experience
- Rucio missing functionalities?
  - Currently: QoS implicit by RSE name, should become an attribute of RSE
  - Multi-QoS RSE? currently one RSE per QoS per site
- For now MÁS touched superficially only one aspect of QoS: availability, other metrics related to capabilities should also be addressed (to reflect the diversity of storage capabilities)
- Richer QoS should also be defined
  - Bandwidth/latency to storage,
  - Delivery time of data
  - Guaranteed QoS lifetime

MÁS MÁS



And …
if we had the money …

- We would buy this instead…
- Integrated multi-storage with data management

DDN®
AI · BIG DATA · HPC

Data Sheet

EXA5 Product Family

The Next Generation Parallel Filesystem Solution

Global Data Platform
- Easy to deploy
- Integrated and optimized
- Comprehensive monitoring
- Data security and governance
- Resiliency and data protection
- Multi-tenancy, containers
- On premise and multicloud

Scales Effortlessly
- Start with a single appliance
- Scale seamlessly to more than 20PB per appliance and hundreds of PBs per filesystem

Fastest
- 1TB/s per rack

Densest
- 12PB per rack

Flexible Architecture
- All NVMe, SSD, HDD
- Hybrid configurations
- Standard networking

Most Experience
- 15+ years of Lustre development and support
- Get support directly from the Lustre maintainers

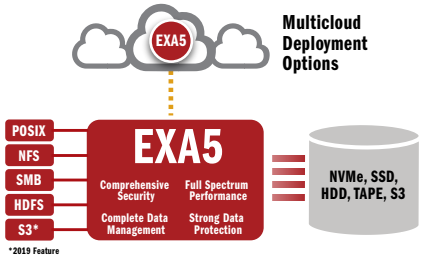
Exascale-Ready
- Already deployed at scale, with 100s of PBs in a single filesystem

Artificial Intelligence (AI), Analytics and High-Performance Computing (HPC) enable organizations everywhere to gain insights from their data with unprecedented velocity and accuracy. Data-intensive workflows put storage infrastructure squarely in the critical path—driving requirements for extreme, dependable performance that scales easily to meet evolving workload requirements on premise and in the cloud.

EXA5: The World's Most Advanced Lustre Solution

Developed and optimized using the latest advances in filesystem software technology, EXA5 delivers extreme performance, scalability, capability, reliability and simplicity. Augmented with feature-rich enhancements, EXA5 delivers a true global data platform capable of enabling and accelerating a wide-range of data-intensive workflows, at any scale. Fully-integrated DDN EXAScaler appliances combine the world's fastest hyperconverged data storage platform with a truly parallel filesystem software in a package that's easy to deploy, managed and backed by the leaders in data at scale.

EXA5 is fully-optimized to deliver data with high-throughput, low-latency and massive concurrency using a shared parallel architecture that can scale flexibly for most technical and economic benefits. Building on over a decade of experience deploying parallel filesystem solutions in the most demanding environments around the world, DDN delivers unparalleled performance, capability and flexibility for users looking to manage and gain insights from massive amounts of data.



Data Management: Automated, Optimized, and Secure

EXA5 introduces several new data management and integrity filesystem features developed by DDN and only available in its appliances. Stratagem is a powerful data orchestration engine that gives users comprehensive data residency controls using policy-based placement. Hot Pools intelligently moves data between high-performance flash and large capacity disk and ensures efficient use of storage. A native T10DIF implementation ensures that data is handled with full integrity from application to disk. Several dozen other new features in EXA5 deliver unique value for users looking to deploy the most demanding workloads on premise, in the cloud.

EXA5 enables organizations everywhere to create, analyze and keep more data than ever before, and accelerates time-to-insight with unequalled speed, precision and simplicity.

+1.800.837.2298 · sales@ddn.com

ddn.com/products/lustre-file-system-exascaler

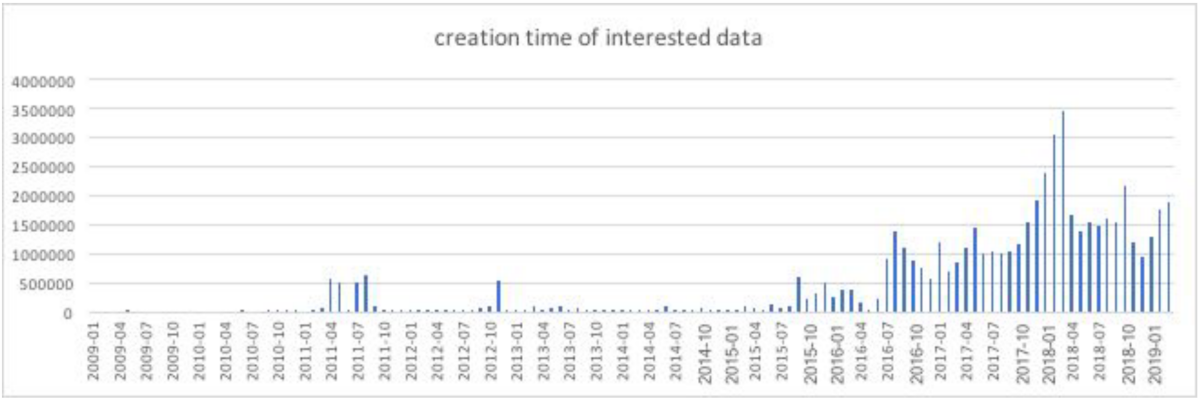
U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

Backup/additional slides

Analysis of non-access data

- We have analyzed the file metadata and storage logs to identify the data access pattern of the disk-resident data.
  - Analysis is done a the end of Oct.
  - BNL had 16PB with 76M files on disks.
  - Among 76M files, there were 55M (3.7PB) files that are created before Apr 1st and never accessed within the 6 months period between Apr 1st to Aug 31st.



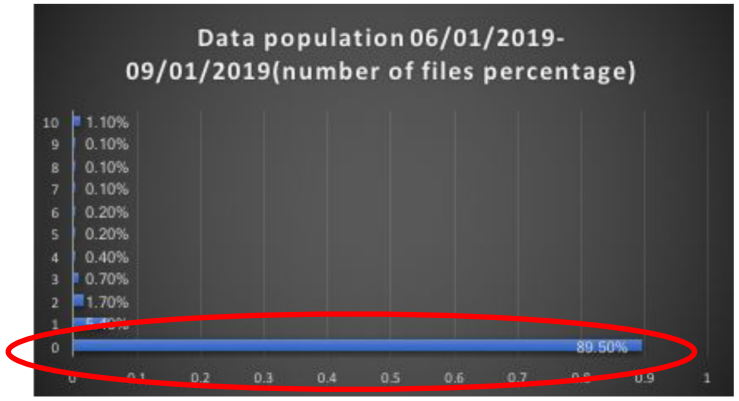
Creation date of Datadisk files created before 04/01/2019 but never accessed during 04/01-08/31/2019

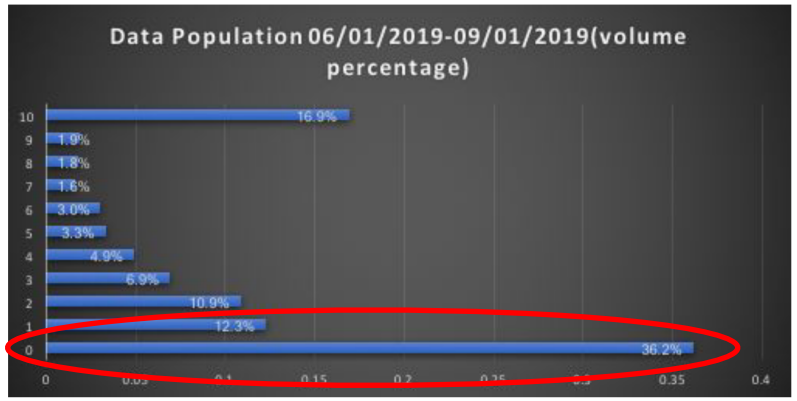
X-axis: date
Y-axis: number of files

Analysis of access frequencies

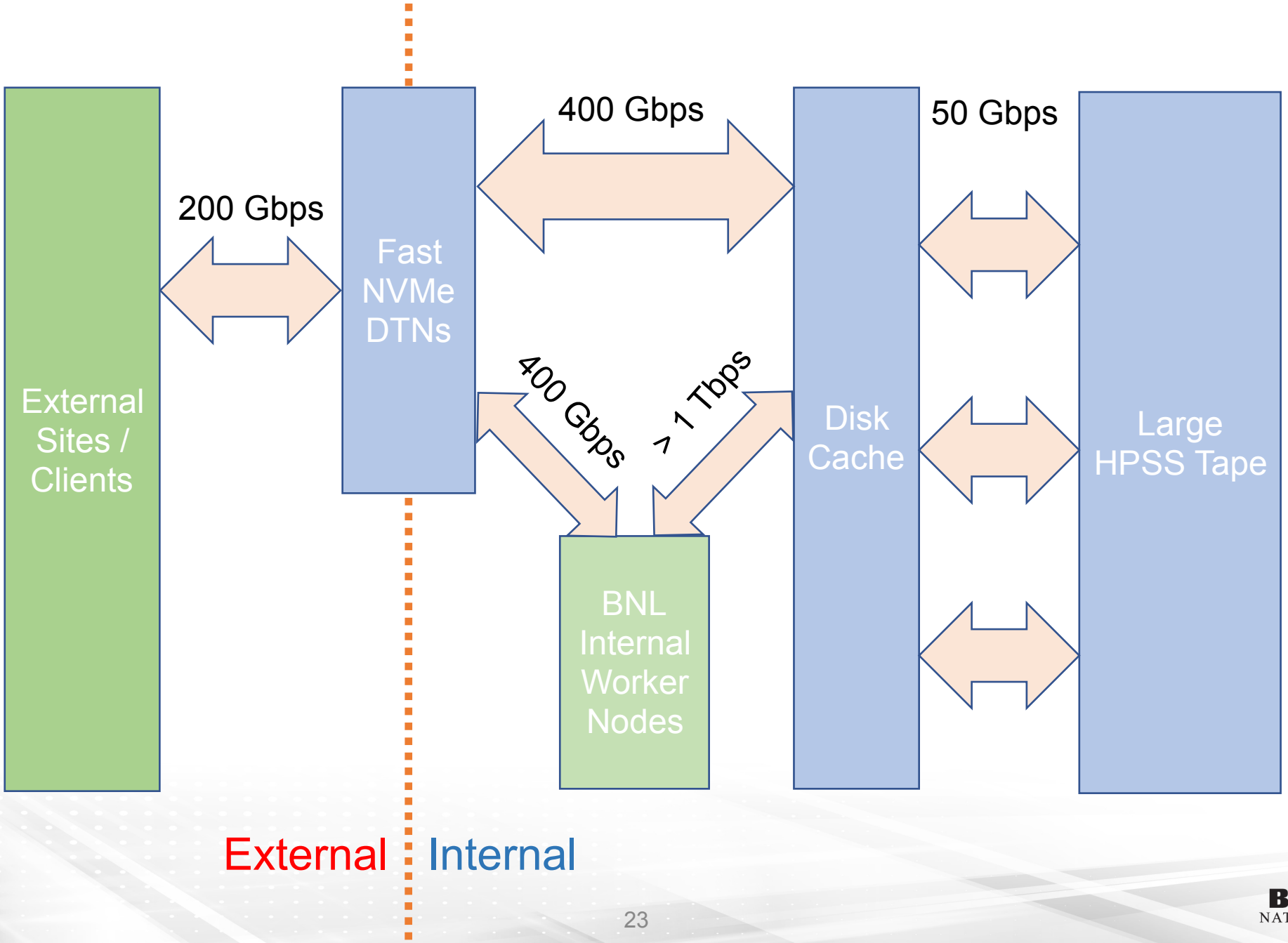
- We have analyzed the frequencies of the data access between June 1st and Sept 1st .
  - Large fractions of disk resident data are not accessed

Access pattern of Datadisk files during 06/01/2019-09/01/2019.





QoS


External Sites / Clients
200 Gbps
Fast NVMe DTNs
400 Gbps
50 Gbps
400 Gbps
> 1 Tbps
Disk Cache
Large HPSS Tape
BNL Internal Worker Nodes
External
Internal


MÁS