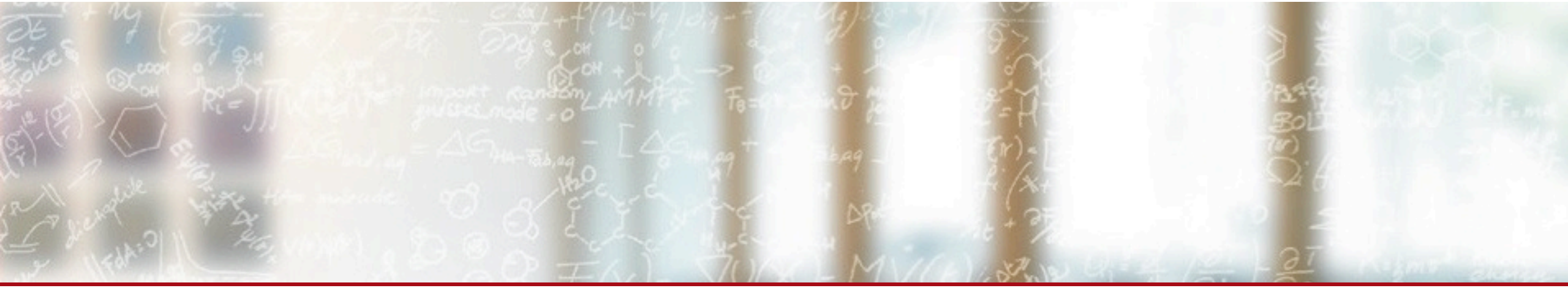
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



Storage & Data Strategy

CSCS User Lab Day 2023

Miguel Gila, CSCS

September 04, 2023

Evolution of our storage & data strategy

Evolution

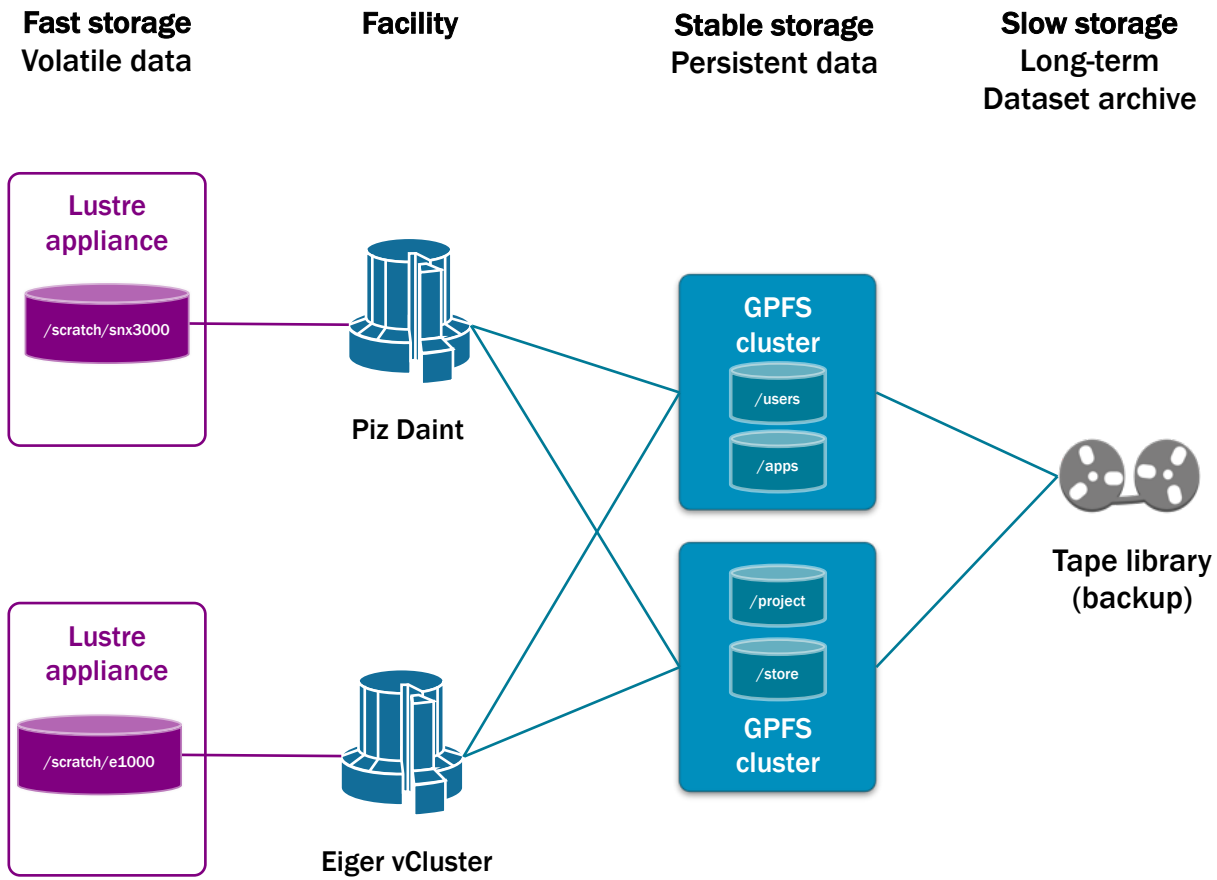
- As with the case of our computing infrastructure, our storage infrastructure and platforms also evolve over time

- This is due to technology improvements, cost changes, or an update in data strategy and policies

- Focusing, for comparison, on two periods of time:
  - The Piz Daint era, from 2012 to today
  - The Alps era, from 2020 to…

Storage in the Piz Daint era

- Different storage backends depending on
  - Use-case
  - Performance envelope
  - Technology constraints
  - Historical reasons
- Persistent data and tape access stored on IBM Spectrum Scale (GPFS) clusters and a myriad of backends (SAS or FC disks, SAN, JBODs, etc.)
- Volatile data with fast access patterns on Lustre appliances



Storage quotas and performance profiles

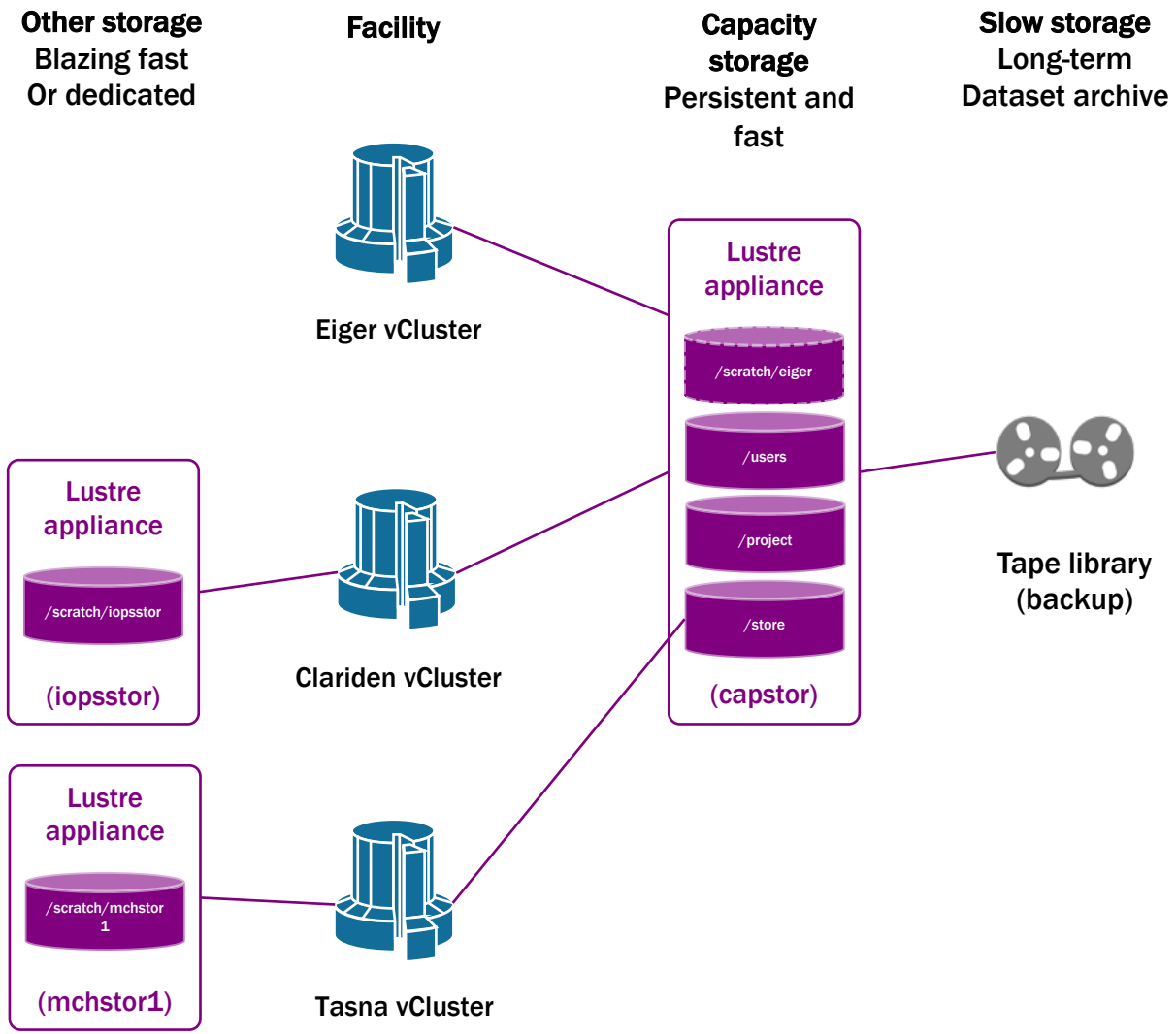
- https://user.cscs.ch/storage/file_systems

Eiger

	/scratch (Piz Daint)	/scratch (Alps)	/scratch (Clusters)	/users	/project	/store
Type	Lustre	Lustre	GPFS	GPFS	GPFS	GPFS
Quota	Soft (1M files)	Soft (1M files)	None	50GB/user and 500k files	Maximum 50k files/TB	Maximum 50k files/TB
Expiration	30 days	30 days	30 days	Account closure	End of the project	End of the contract
Data Backup	None	None	None	90 days	90 days	90 days
Access Speed	Fast	Fast	Fast	Slow	Medium	Slow
Capacity	8.8 PB	9.1 PB	1.9 PB	160 TB	6.0 PB	7.6 PB

Storage in the Alps era

- Consolidate backends and technologies
- Most storage areas will be based on Lustre, which is a much more mature product now
- Filesystems/spaces available:
  - Capstor
  - Iopsstor
  - Purpose, dedicated filesystem
- Metadata ops for the different areas hitting different servers



Storage quotas and performance profiles

- Details to be determined, but the bulk of the policies for most users should remain similar

- Introducing the concept of soft quotas and hard quotas, allowing certain use-cases to temporarily exceed usage

- Some use-cases with specific requirements (lots of iops, more strict SLAs/SLOs, etc.) to use other storage components, in some cases dedicated exclusively

Daint era storage specs and pictures

- Sonexion 3000 (Daint's scratch)
  - 8.8 PiB Cray Sonexion 3000
  - Spinning disks
  - Raw performance:
    - 112 GB/s write | 125 GB/s read
  - 1640 HDDs (8TB each)
  - 2 metadata servers (20x 800GB SSD each)
  - 3 full racks
  - Infiniband FDR

Comparable performance

- Alpstor
  - 10 PiB HPE ClusterStor E1000
  - Spinning disks
  - Raw performance:
    - 120 GB/s
  - 240 HDDs (16TB each)
  - 2 metadata servers
  - ~2 half racks
  - Slingshot 10

Alps era storage specs and pictures

In the process of being accepted

- Capstor
  - 100 PiB HPE ClusterStor E1000D
  - Spinning disks
  - Raw performance:
    - **~1TB/s**
    - 300k Write iops | 1.5M Read iops
  - 8480 spinning disks (16 TB each)
  - 6 metadata servers
  - 11 full racks
  - Slingshot 11

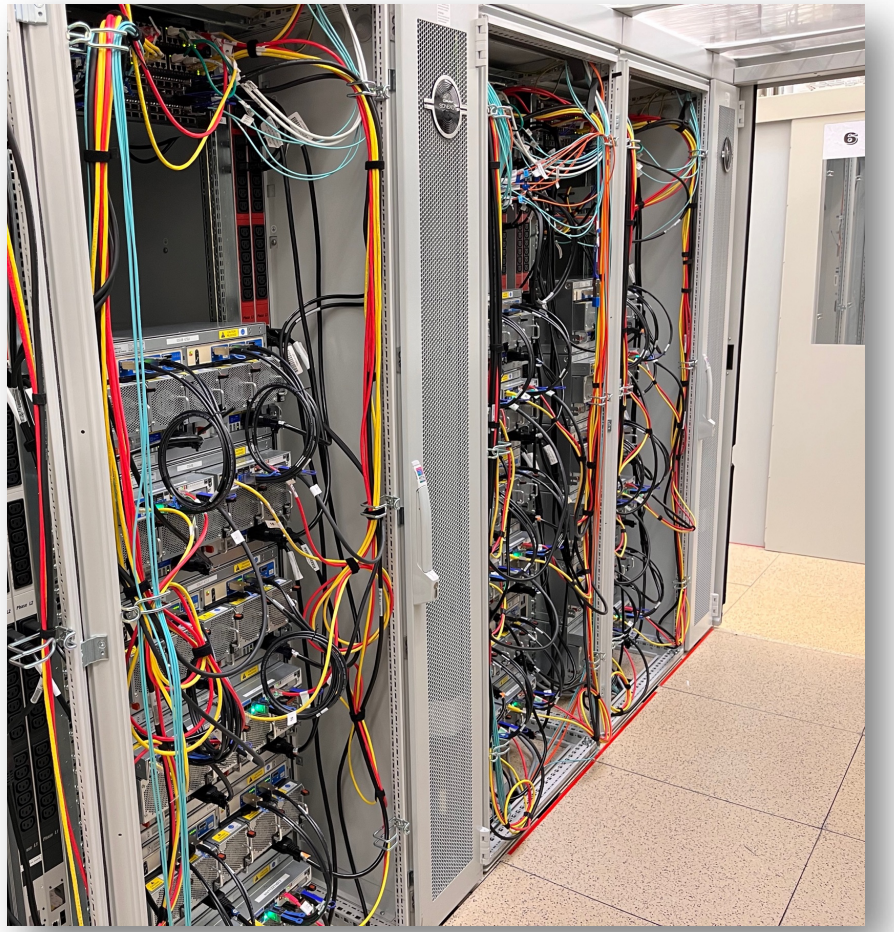
Accepted, rolling to users soon

- iopsstor
  - 3.2 PiB HPE ClusterStor E1000F
  - All flash
  - Raw performance:
    - 240GB/s Write | 600 GB/s Read
    - **13.5M Write iops | 18.4M read iops**
  - 240 NVMe devices (30TB each)
  - 2 metadata servers
  - 1 rack
  - Slingshot 11

Some more pictures

Capstor: 126 disks per enclosure!

Back of iopsstor



Back of Daint's Sonexion 3000

Tape library

Lots of tapes!

What changes for you?

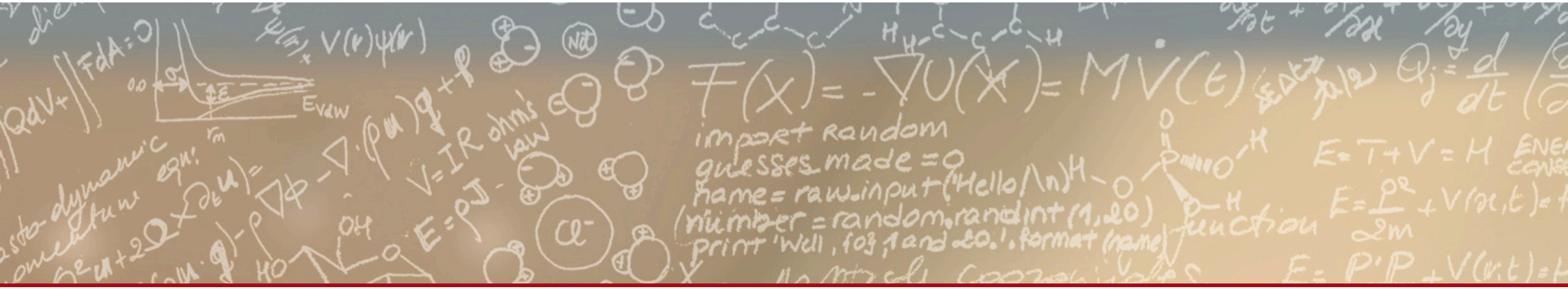
Short-medium term

- If you're only a user of Piz Daint, then nothing changes
- If you're a user of Eiger, in the next months capstor will be rolled into production, so at some point there will be a new storage component for $HOME, $SCRATCH, $PROJECT and $STORE, and data will be migrated
- If you're user of any of our other vClusters… you probably have been converted to the new environment
- Overall raw performance is much higher with new filesystems
  - DVS no longer used
  - Distribute metadata ops for the different storage areas to different servers
- Snapshots won't work anymore (~/.snapshot/….)

Long-term

- Data movement from vCluster to vCluster
  - Datamover service is being re-engineered
  - Invite users to utilize the datamover service to sync data as needed

- This is only the beginning of some long-term changes
  - We’re constantly moving forward, evaluating other storage alternatives with different capabilities



Q&A

Thank you for your attention