Tape development @ DESY

Tigran Mkrtchyan on behalf of the DESY data management team

pre-GDB, 09.02.2021

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

What is DESY *(as storage)*

Experiments/Community	Service	Role
EuXFEL, Petra-III, ILC, Accelerator R&D, ...	Source of the data. Primary data site. Provides online, near-line and archival storage.	Tier-0
Belle-II, ...	Provides online and near-line storage.	Tier-1
Atlas, CMS, LHCb	Online only.	Tier-2
H1, Hermes, Hera-B, Zeus , ...	Provides online and archival storage.	DP

Multiple Faces of Tape

At data source

- High data ingest rate
- Multiple parallel streams
- High durability, multiple copies on different media
- Long-term near-line access
- Small file handling

At analysis facility

- Automatic data accessibility migration
- Bulk recall on periodic basis
- Long-term near-line access
- Recall prioritization

Data Archive

- Manual data accessibility migration
- Long-term preservation
- Automatic technology migration
- Self-healing

Technologies in Place

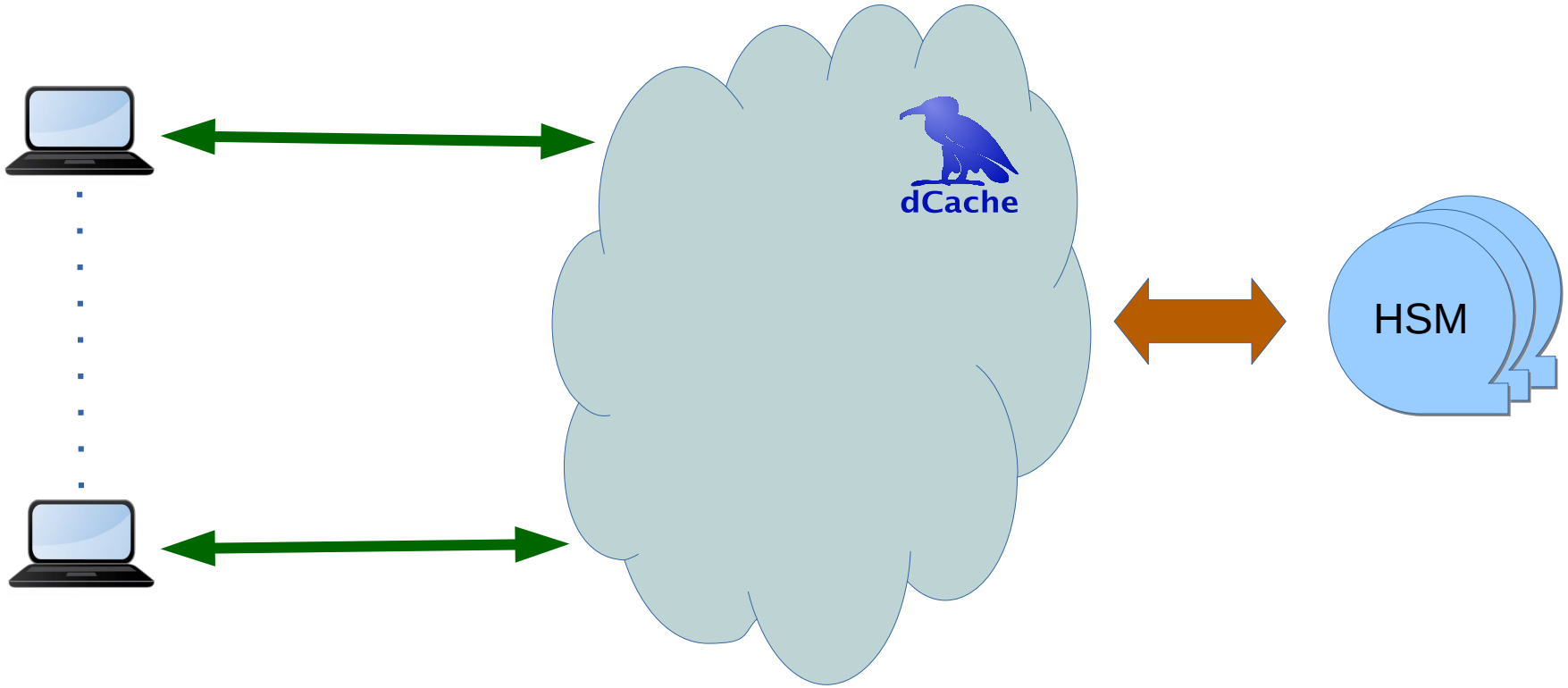
Hardware

- 2x Oracle SL8500
  - 26x LTO-8 drives
- 2x IBM TS4500 – in progress
  - 12x Jaguar
  - 8x LTO-9
  - Different buildings (500m)

Software

- TSM (IBM Spectrum Protect) – classic backup
- dCache – interface to HSM system
  - Scientific Data
  - AFS/Mail backup
- OSM (Open Storage Manager)
  - Since 1994, multiple local modifications

dCache+HSM Tandem



All access to scientific data on tape goes exclusively through dCache!

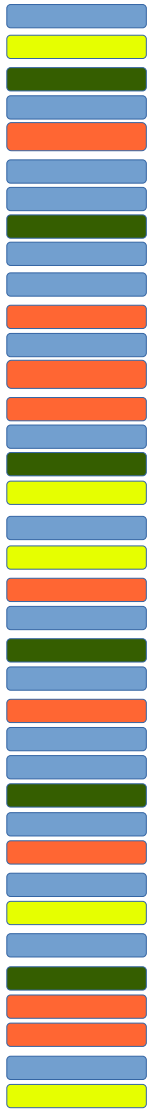
dCache Tape Connectivity

- Write-back / Read-through cache behavior
- Transparent for the users
- Available via all protocols (if not restricted)
- Multiple HSM on a single instance
- Tape location as an opaque HSM specific data

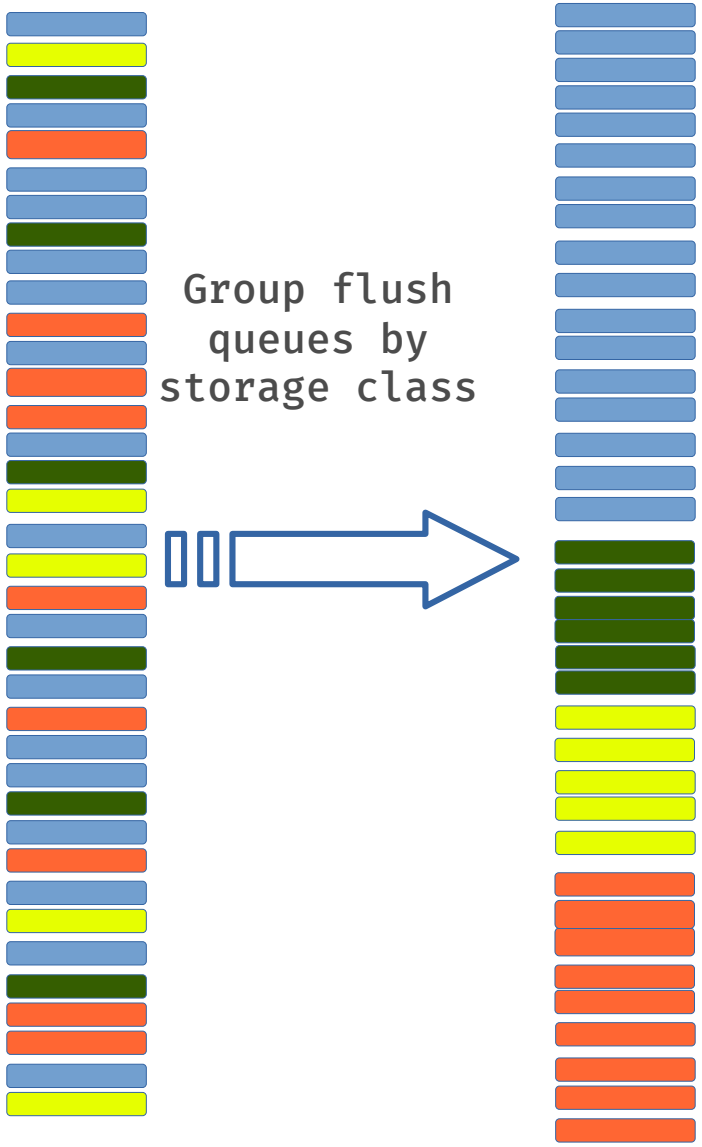
Flush to Tape `rules`

- Flush requests grouped by storage group
  - Storage group (typically) associated with a set of tapes
  - Multiple storage group can be flushed in parallel
  - Re-use tape mount
  - This is per pool decision!
- Flush triggered by:
  - Max time on disk
  - Number of files
  - Number of bytes

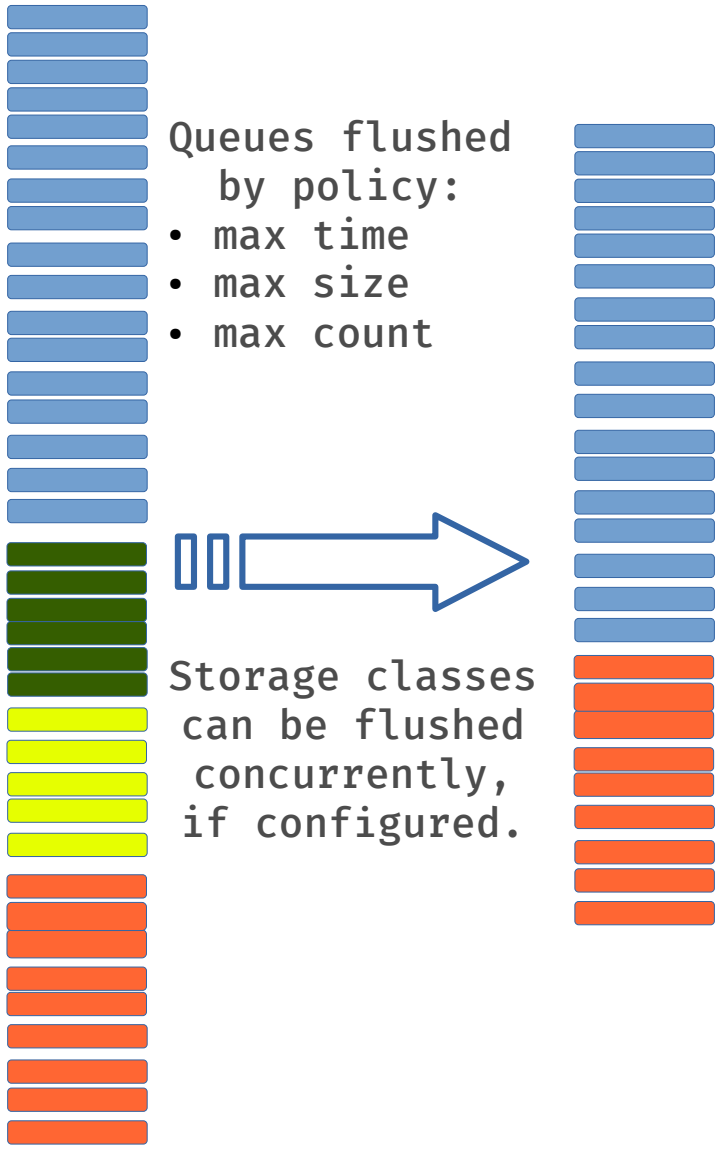
Flush Queue



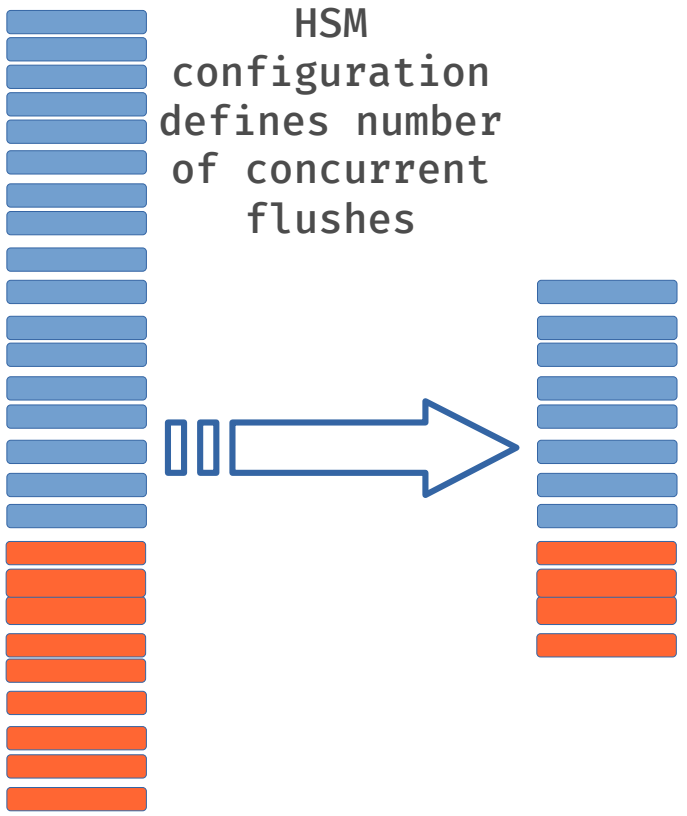
Flush Queue



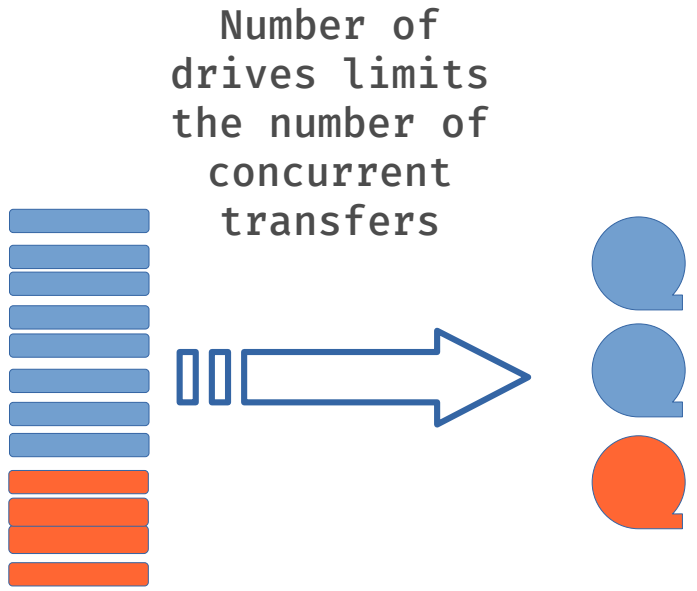
Flush Queue


Queues flushed
by policy:
• max time
• max size
• max count
Storage classes
can be flushed
concurrently,
if configured.

Flush Queue



Flush Queue

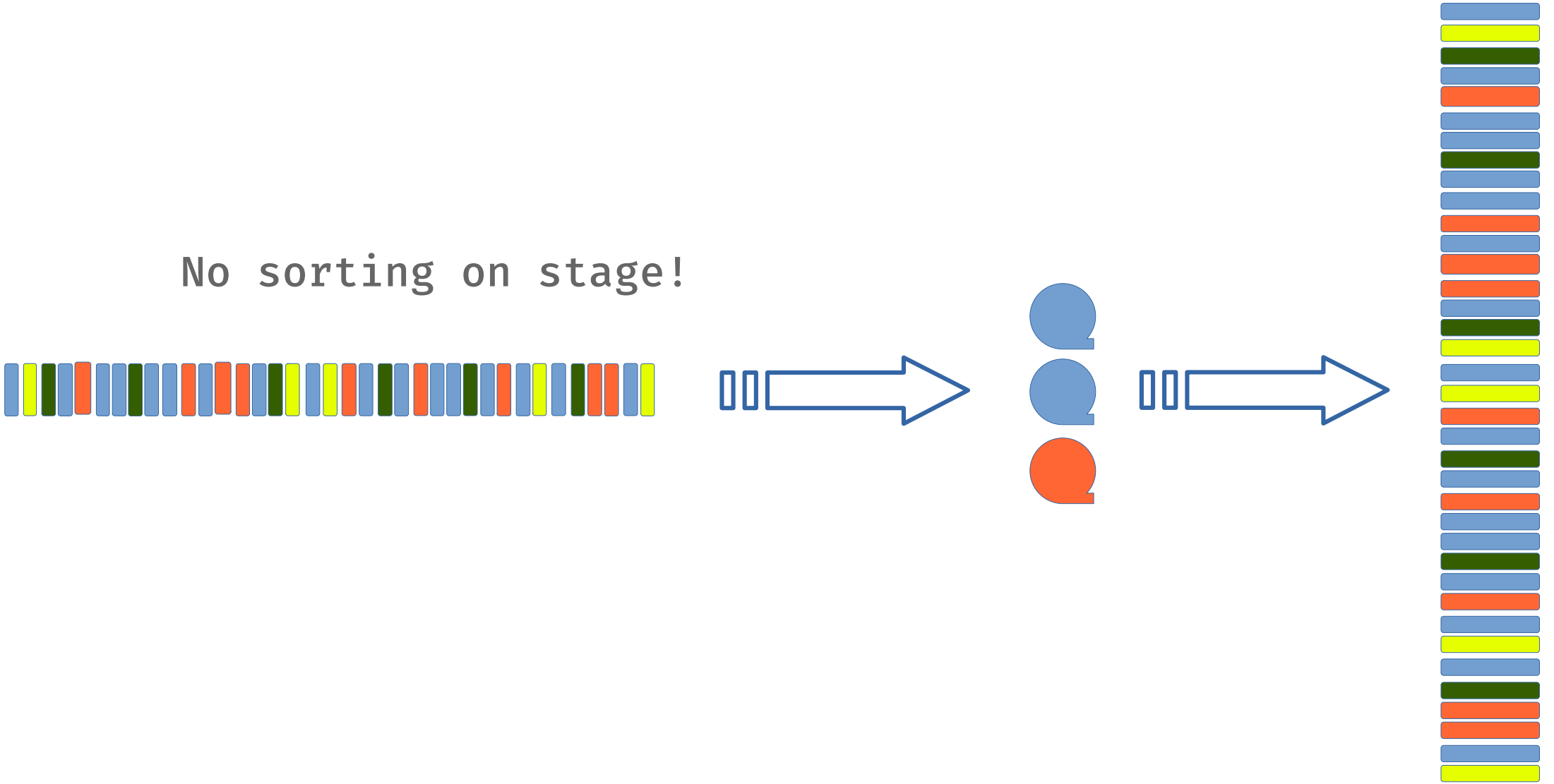

Number of drives limits the number of concurrent transfers

Restore from Tape

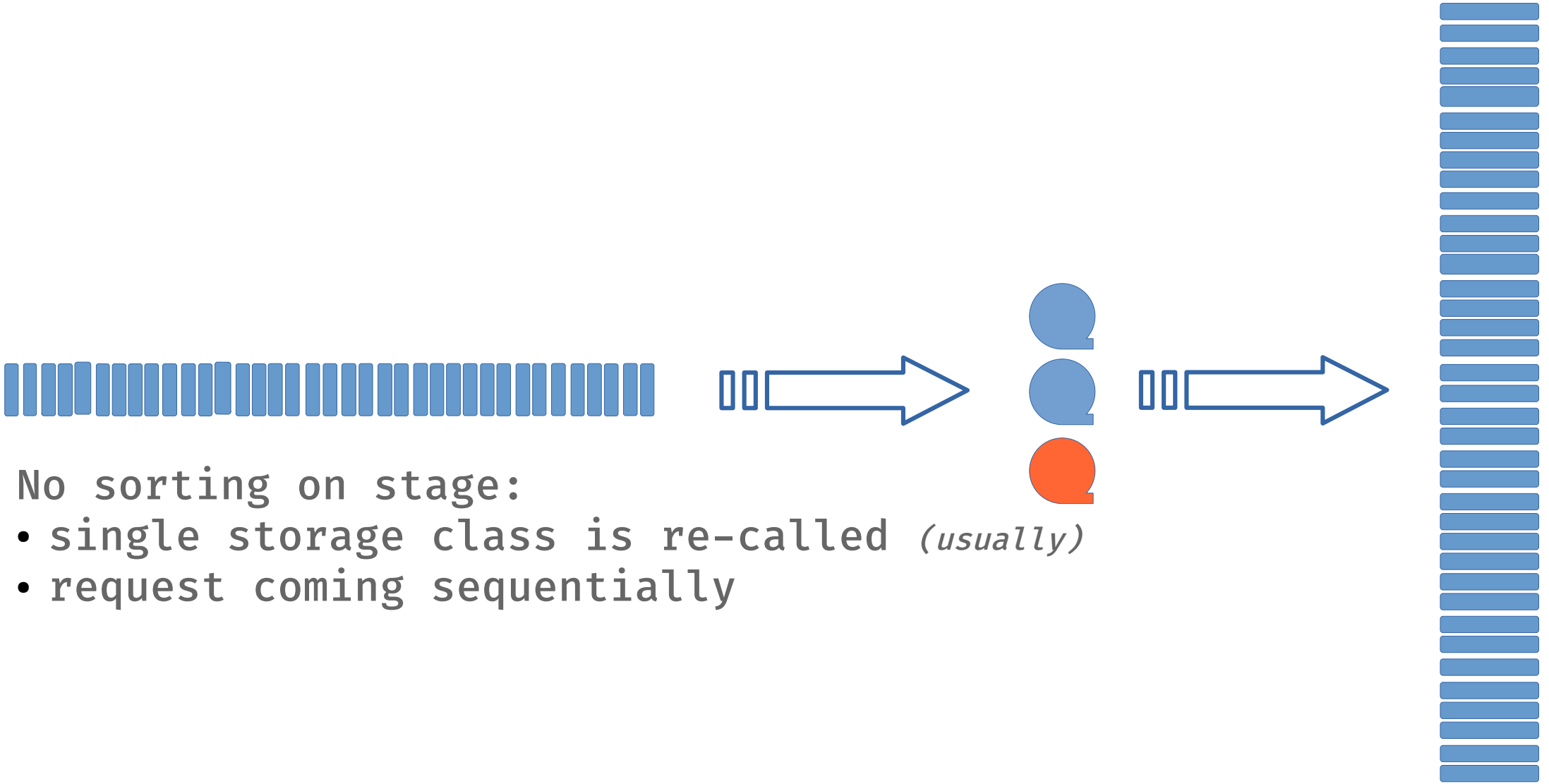
1. Appropriate stage pool is selected
   - HSM type, load, space
2. Space allocated on disk
   - Never block on space allocation when tape is mounted!
3. Requests sent to tape

The Restore Queue

No sorting on stage!



The Restore Queue



No sorting on stage:

- single storage class is re-called *(usually)*
- request coming sequentially

Work in Progress

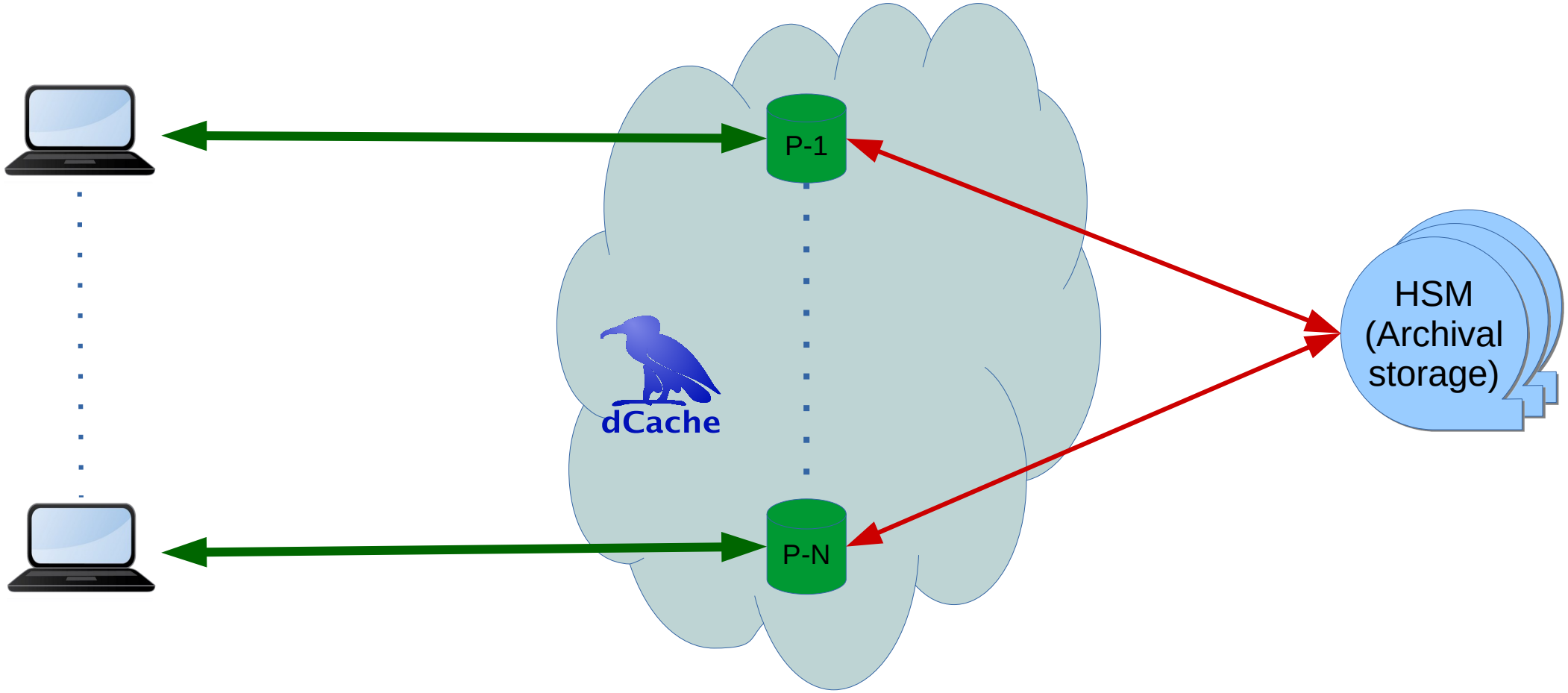
Grouping Requests by Tape

- Proof-of-concept implementation
  - Implemented for SRM
  - General re-call will follow
- Groups requests by tapes
  - Tape inventory is required
- Re-calls triggered based on policy
  - Number of files
  - last request seen
- Test deployment expected in 4 weeks

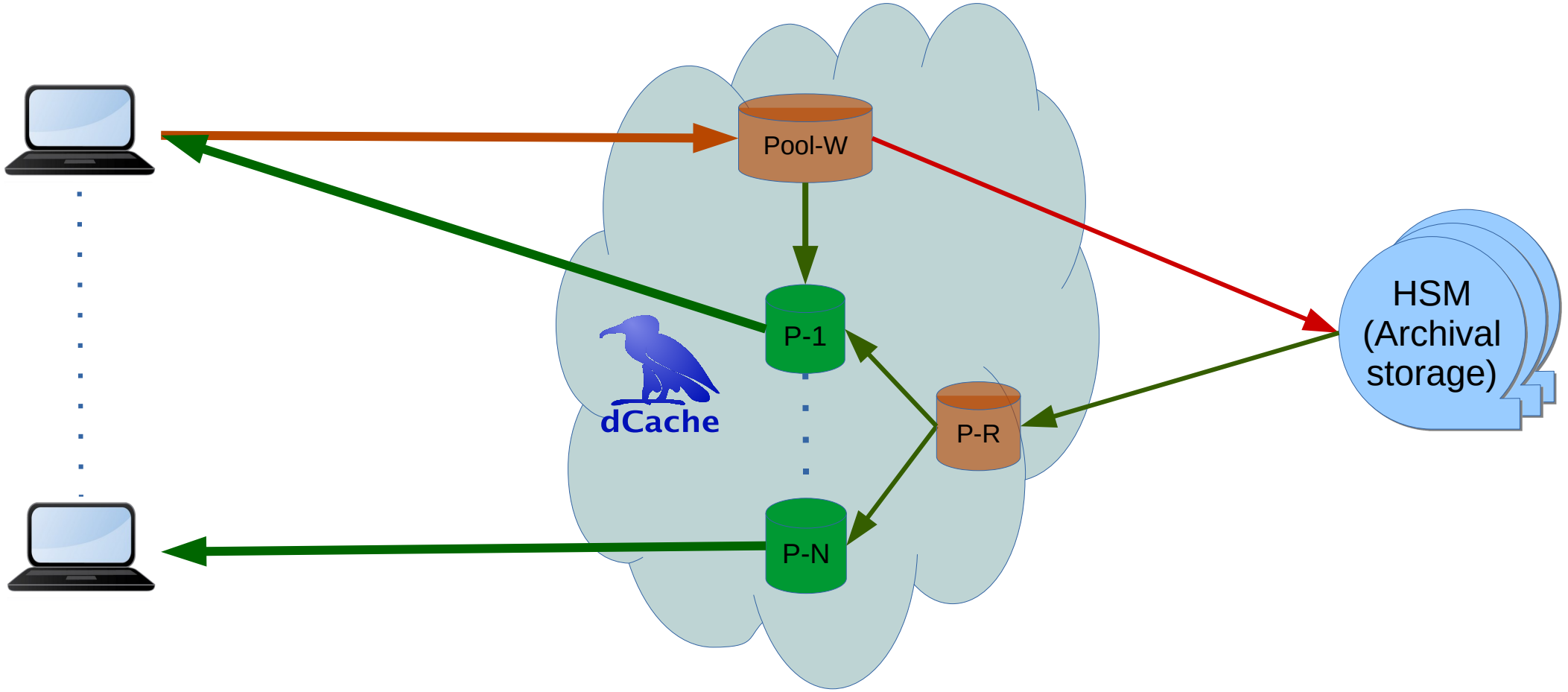
Layered/Tiered Model

- Maximize IO throughput for a given activity
- Clear role for data servers
  - User facing servers
  - Tape fronting servers
- Achieved today by sophisticated dCache setup
  - FermiLab
  - FZK

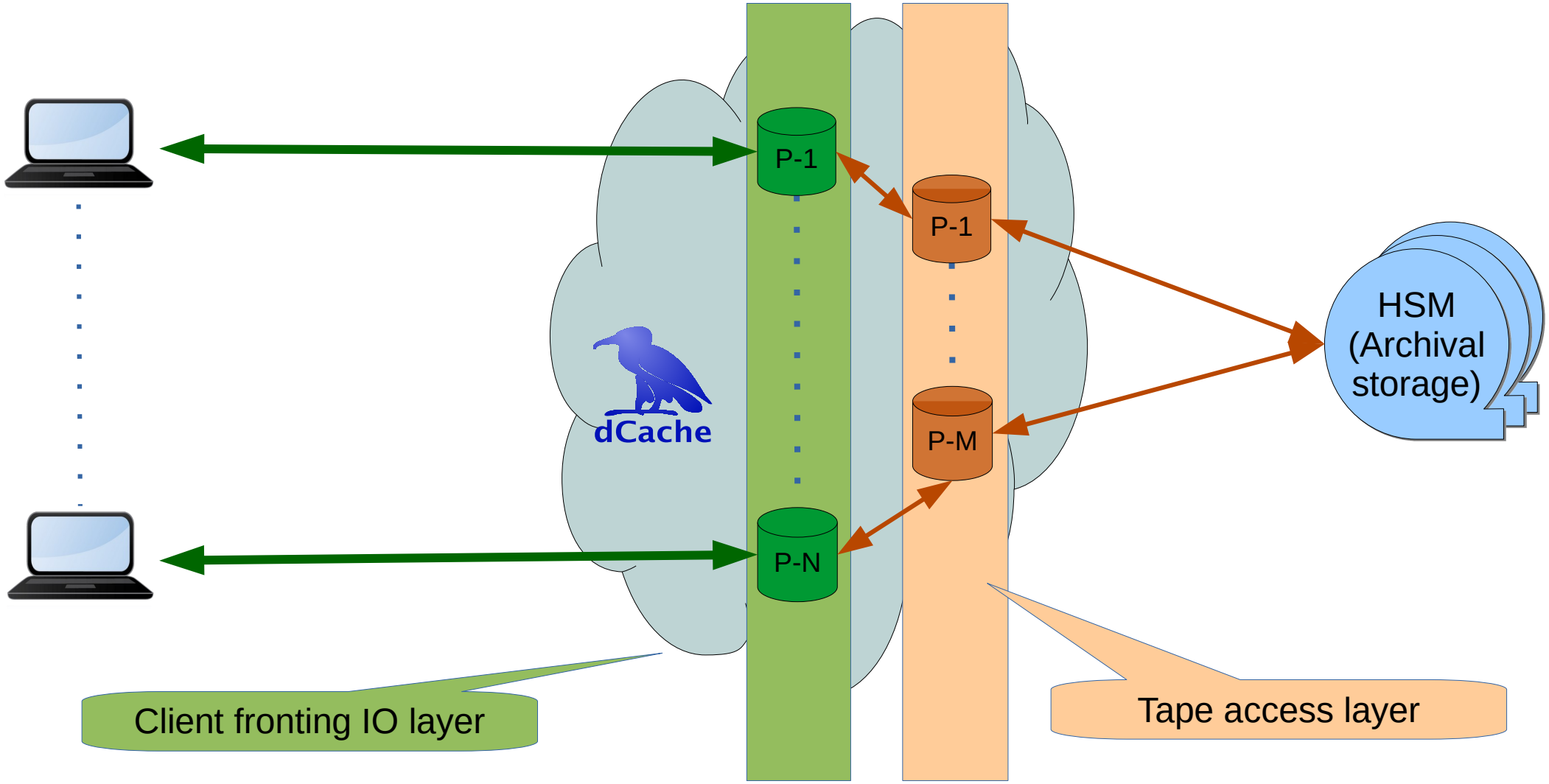
Typical Tape connectivity



Dedicated Write Pool



Layered Model



HSM Software Selection

- Maximize tape HW efficiency
  - Integration into DESY ecosystem
  - Integration with dCache tape interface
- Stable operation for a next decade
- Open Source/Open standard
  - Commercial/proprietary systems are not excluded
- Wide adoption and community

Possible candidates

- Open Source
  - Enstore
  - CTA
- Proprietary (Plan B)
  - TSM
  - HPSS

Enstore Status at DESY

Pros

- FermiLab Product
- Well tested
  - 3x Tier-1
  - 20+ years in production
- GPL (or BSD)
- Seamless integration with dCache
- Small file aggregation
- Test setup at DESY with SL8500

Cons

- Python 2
- No clear roadmap
- Non-standard access protocol

CTA Status at DESY

Pros

- CERN Product
- GPL3
- Well defined software development process
  - CI replicated at DESY
- Test setup at DESY with Virtual Tape Library

Cons

- In *early production* stage
- Orthogonal to dCache *tape awareness*
- Non-standard access protocol
- Non-standard on tape format

Summary

- Tape is an essential part of IT-Services at DESY
- dCache is the only interface to scientific data
  - Tape connectivity dominates the local development
- Enstore and CTA are evaluated as HSM solution
  - Both require on-side development
  - Commercial alternatives are not excluded !
- We expect new system to be in place in 1Q 2022
  - ~6 months to make a decision

Thank you