dCache – Inter-Disciplinary Storage

Distributed Computing and Grid Technologies in Science and Education (GRID'2021)

Tigran Mkrtchyan for the dCache collaboration

DESY. Fermilab neic NORDIC E-INFRASTRUCTURE COLLABORATION

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

About dCache

- A distributed petabyte-scale storage system for scientific data
- Joint effort between DESY, FNAL and NeIC
- Supports standard and HEP specific access protocols and authentication mechanisms

dCache: Motivation

- Data never fits into a single server
  - Multiple nodes
  - Offload to tape
- Growing number of client CPUs
  - Horizontal scaling with number of data servers
- Control over HW/OS selection
  - Better utilization of local expertise

Scientific Data Challenges

Ingest

- High data ingest rate
- Multiple parallel streams
- High durability
- Effective handling of large number of files

Analysis

- High CPU efficiency
- Chaotic access
- Standard access protocols
- Access control
- Local user management

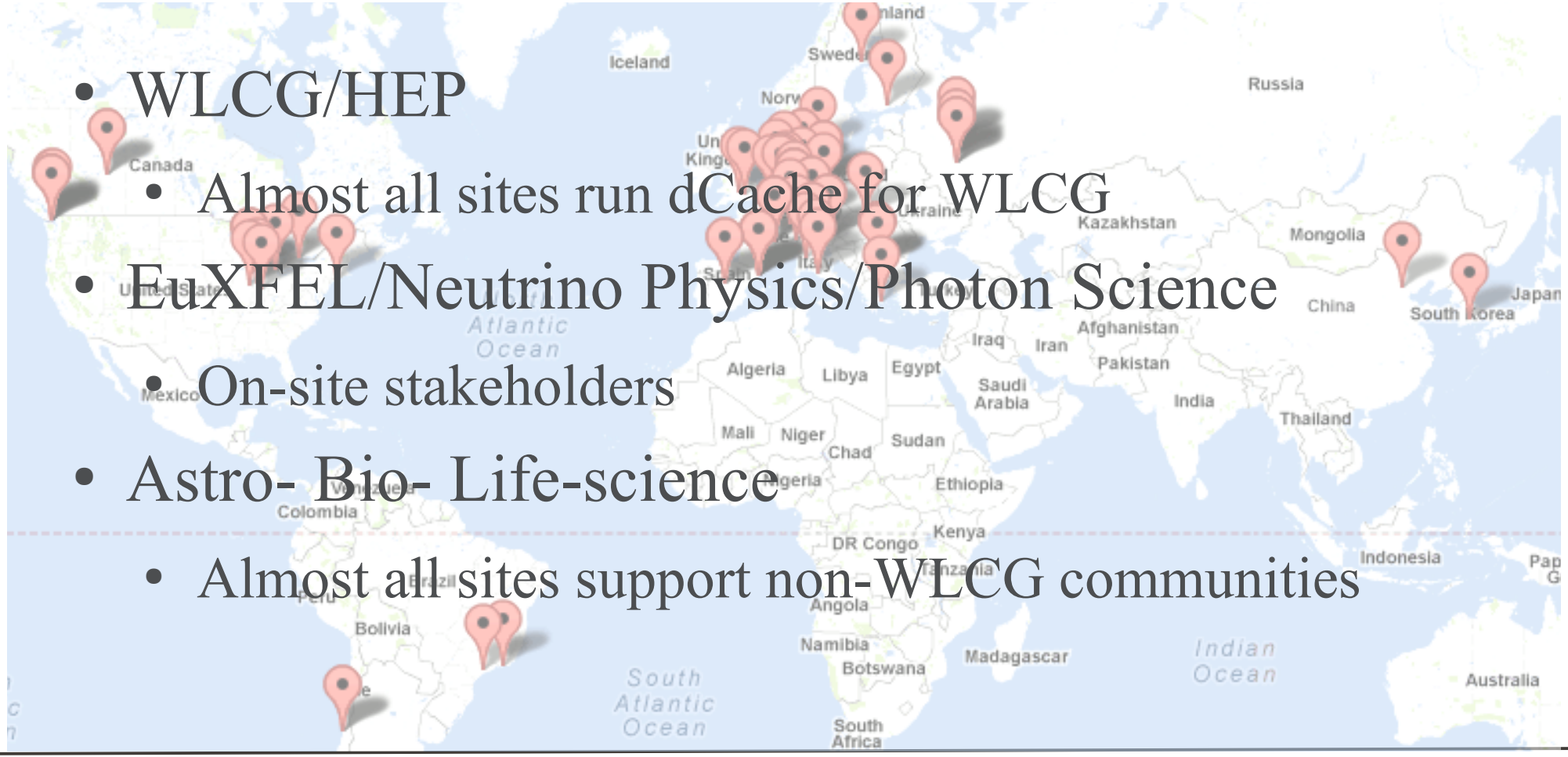
Sharing & Exchange

- 3rd party copy
- Effective WAN Access
- In-flight data protection
- Identity federation
- Access control

Long Term Preservation

- High Reliability
- Self-healing
- Automatic technology migration
- Persistent identifier

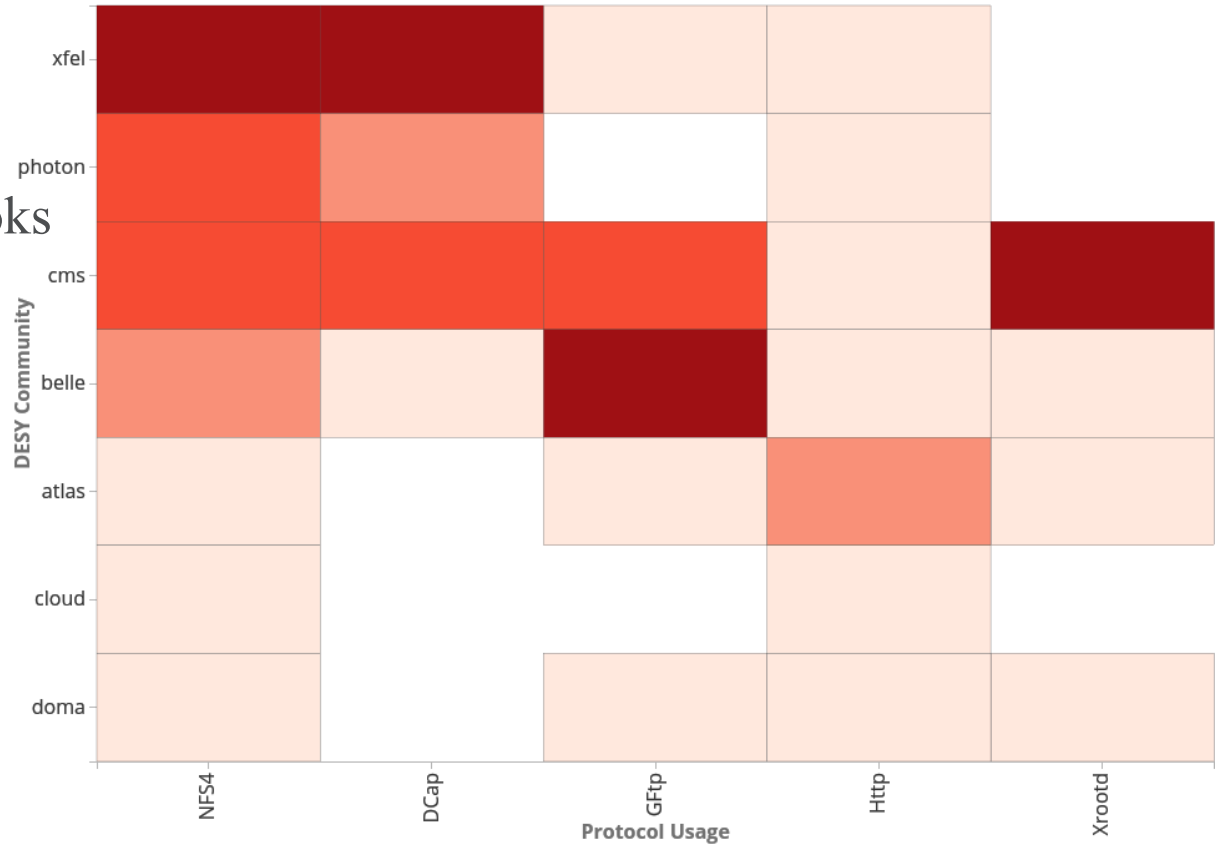
Strategic Communities



- WLCG/HEP
  - Almost all sites run dCache for WLCG
- EuXFEL/Neutrino Physics/Photon Science
  - On-site stakeholders
- Astro- Bio- Life-science
  - Almost all sites support non-WLCG communities

Data Access Variety

- ROOT-IO
- Non-HEP tool chain
  - Active use of Jupyter Notebooks
  - Non-ROOT data formats
- Industry standard AuthN
  - Tokens based authentication
  - Federated IdP
- Use of private clouds
  - Data access from a container
- Use of HPC resources



Design Goals

- Single-rooted namespace, distributed data
  - Client talks to the namespace for metadata ops only
- Bandwidth and capacity grows with the number of data nodes
- Standard access protocols
- The same data available with any protocol independent from authentication scheme

Technical Design

- Set of independent components
  - Each component does one thing
- Multiple instances of the same component can run in parallel
- Components communicate by sending messages
  - Nowadays such architecture is called *micro-services with message bus*

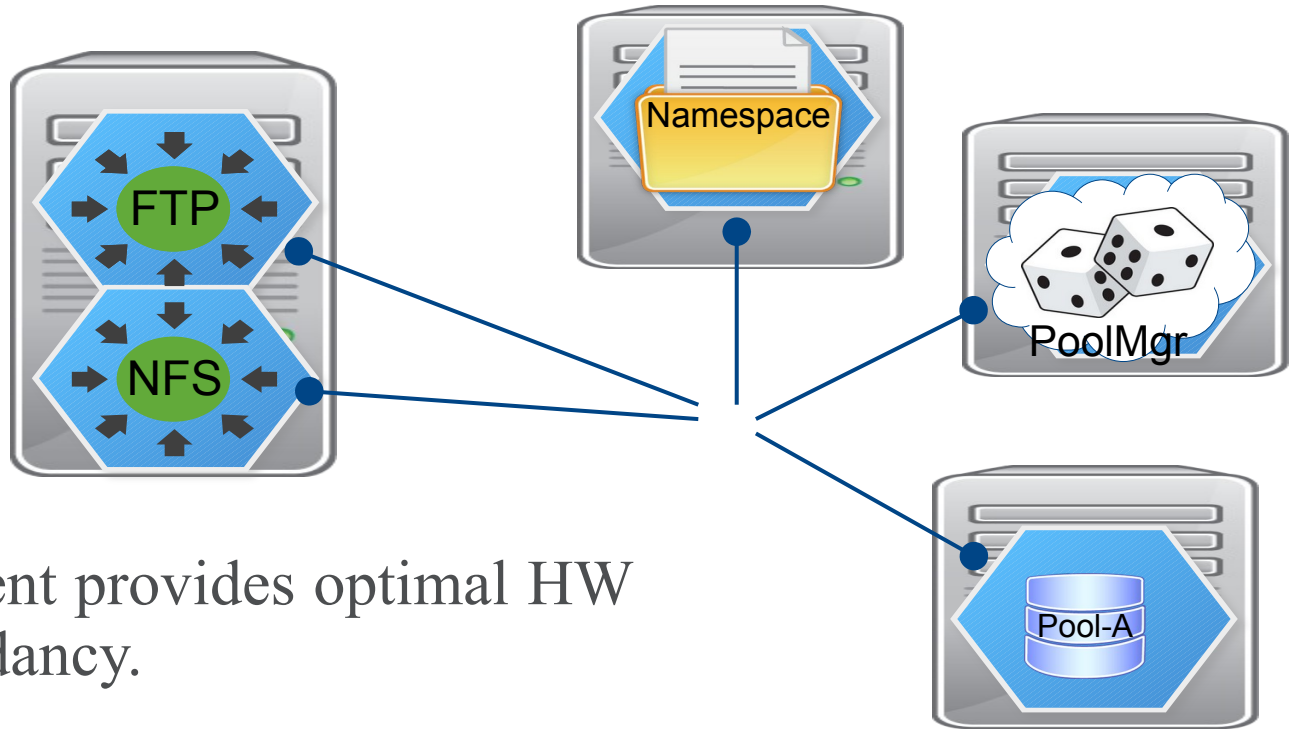
Main Components

- Namespace
  - Inventory, POSIX view layer.
- Door
  - Protocol specific user entry point (FTP, HTTP, NFS …).
- Pool
  - Data storage node. Talk all protocols
- PoolManager
  - Request distribution unit.

Fault Tolerance

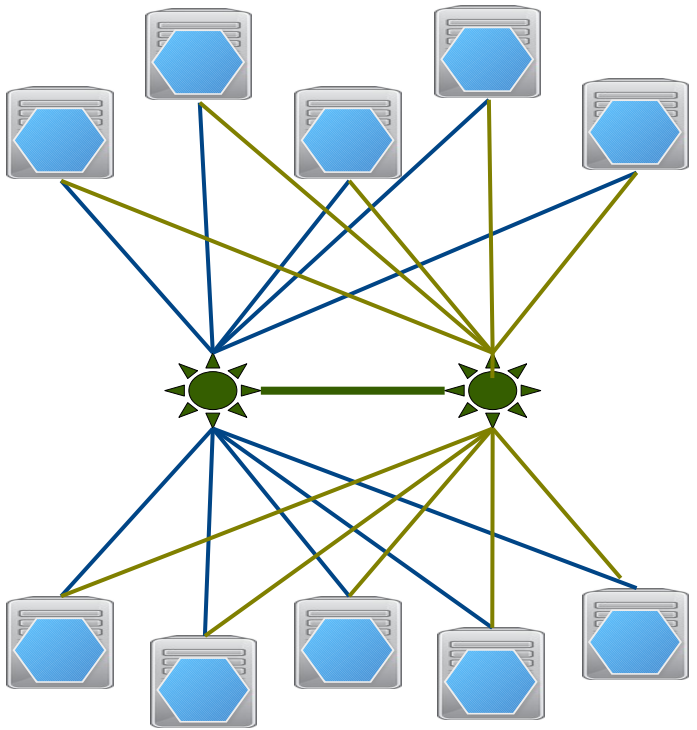
- All core services can run multiple instances
- Door/Pool restarts handled by clients
  - DCAP
  - NFSv4.1
  - Xroot
- Master-slave configuration of namespace database
  - dCache detects which node runs in master mode

Component Deployment



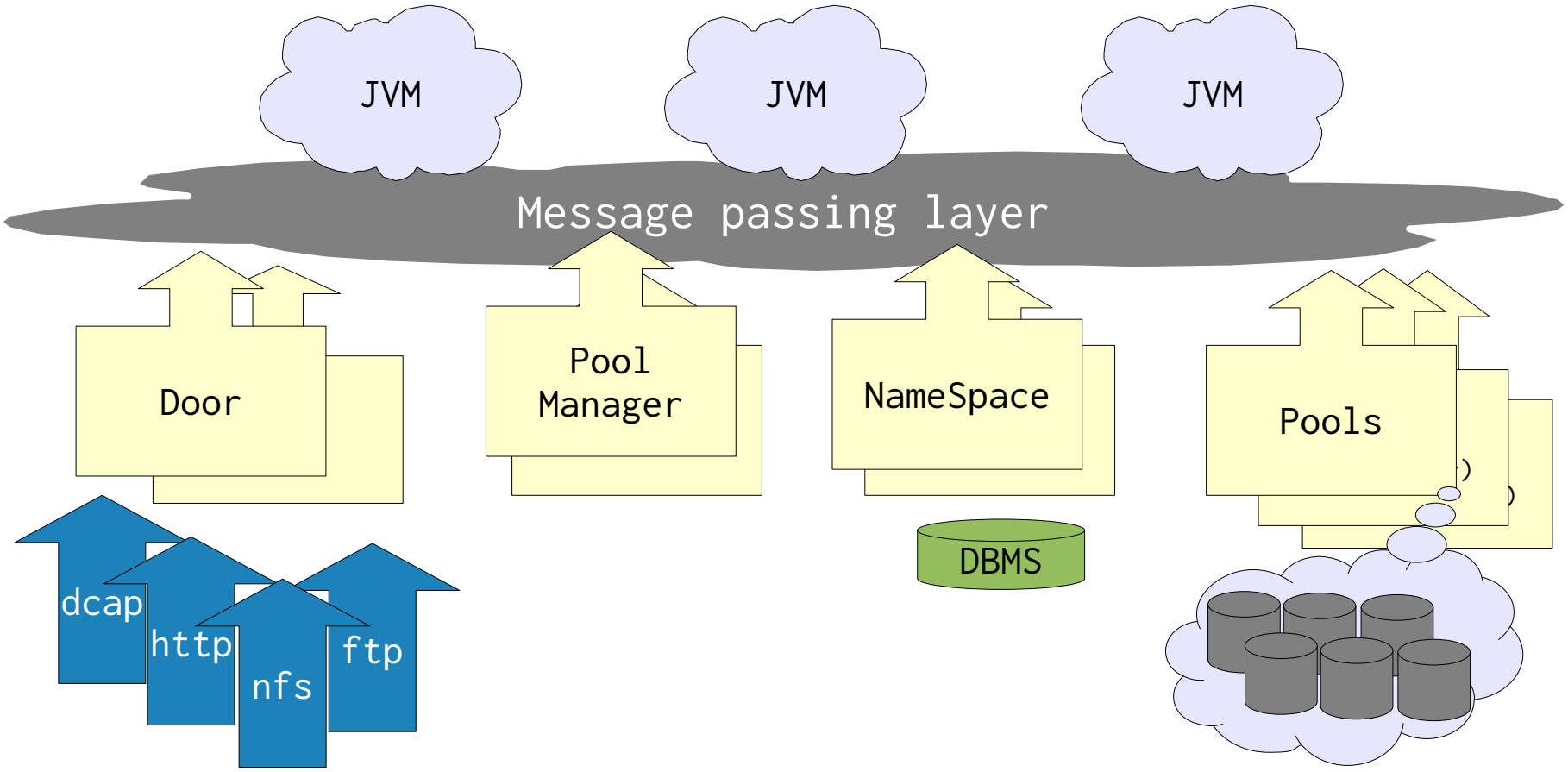
Multi-node deployment provides optimal HW utilization and redundancy.

Internal Messaging

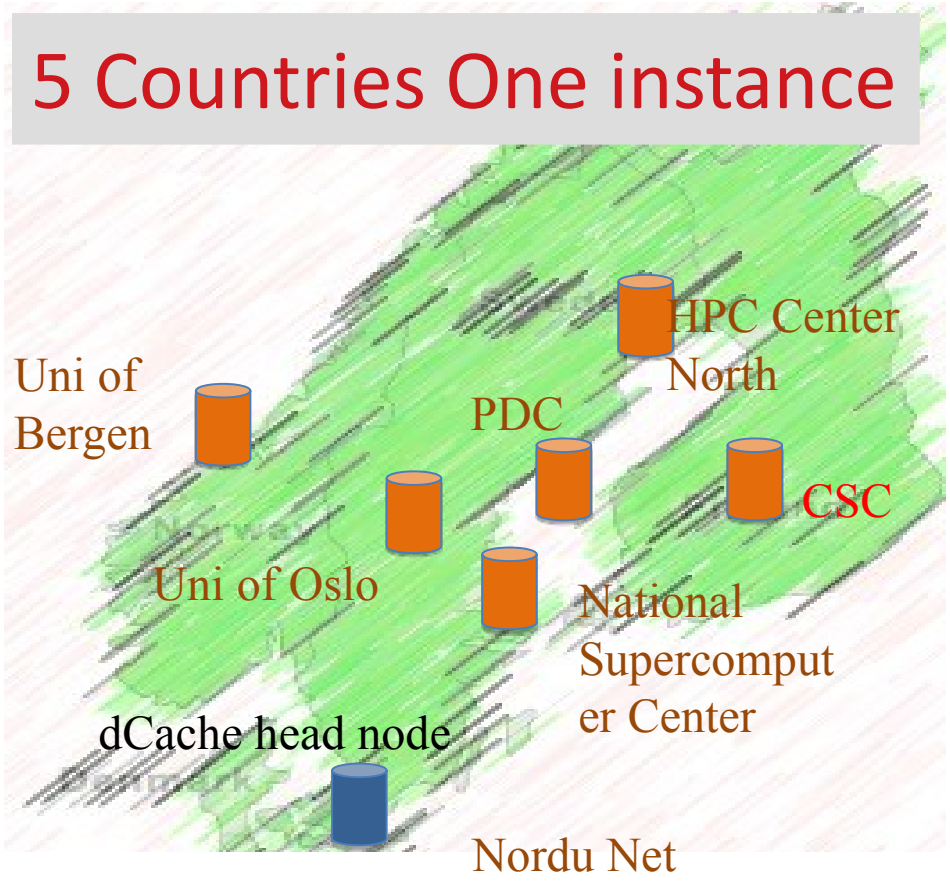


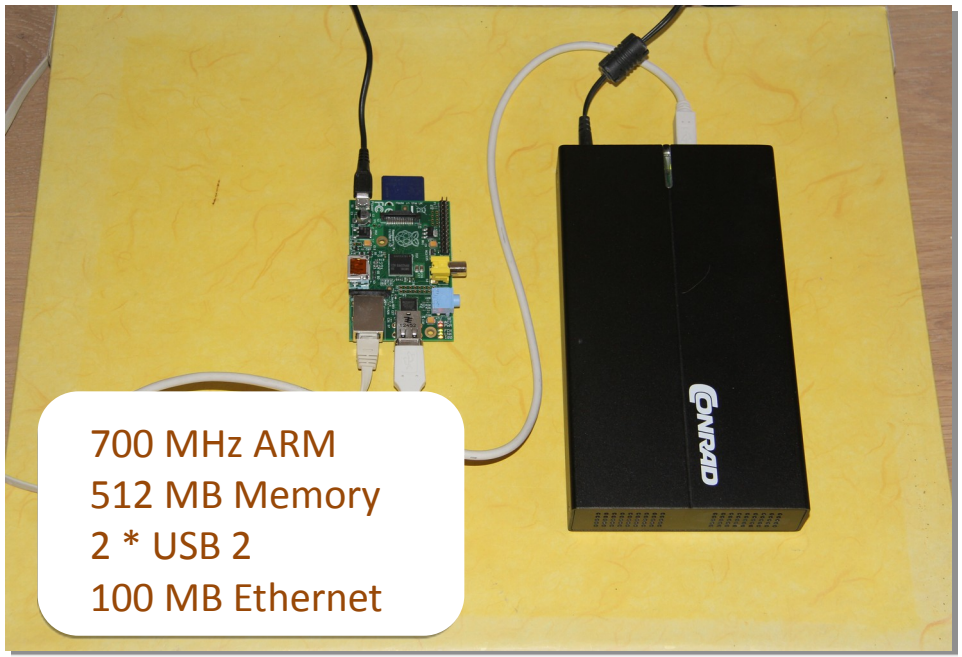
- Star-like topology
- Selected node configured as a hub called **CORE** domains
- Others called **SATELLITE**
- All communication goes through **CORE** domains
- Multiple **CORE** domains makes communication fault tolerant

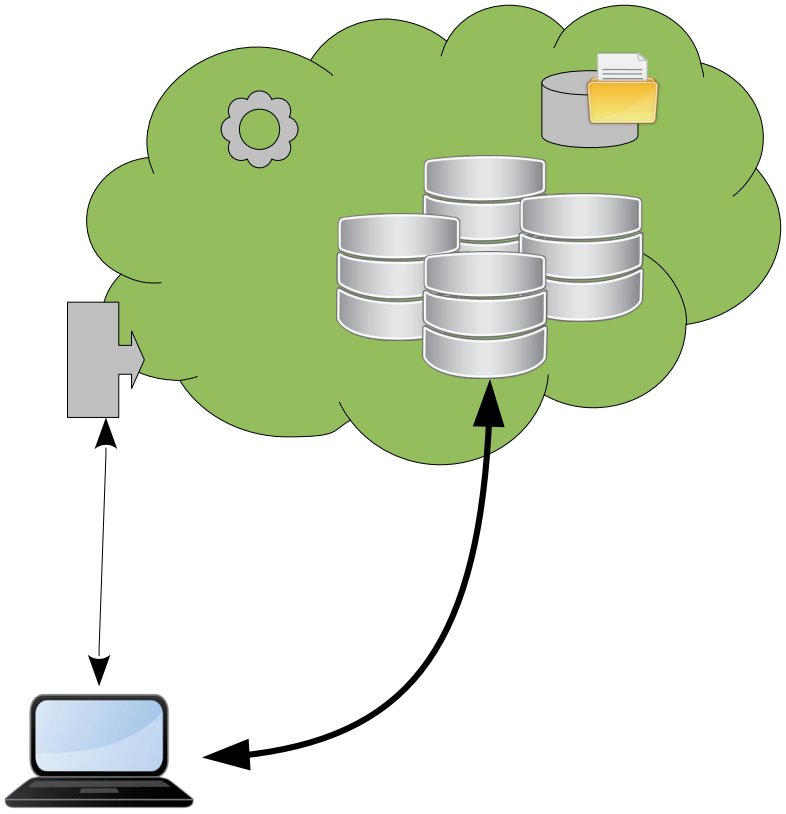
dCache on One Slide

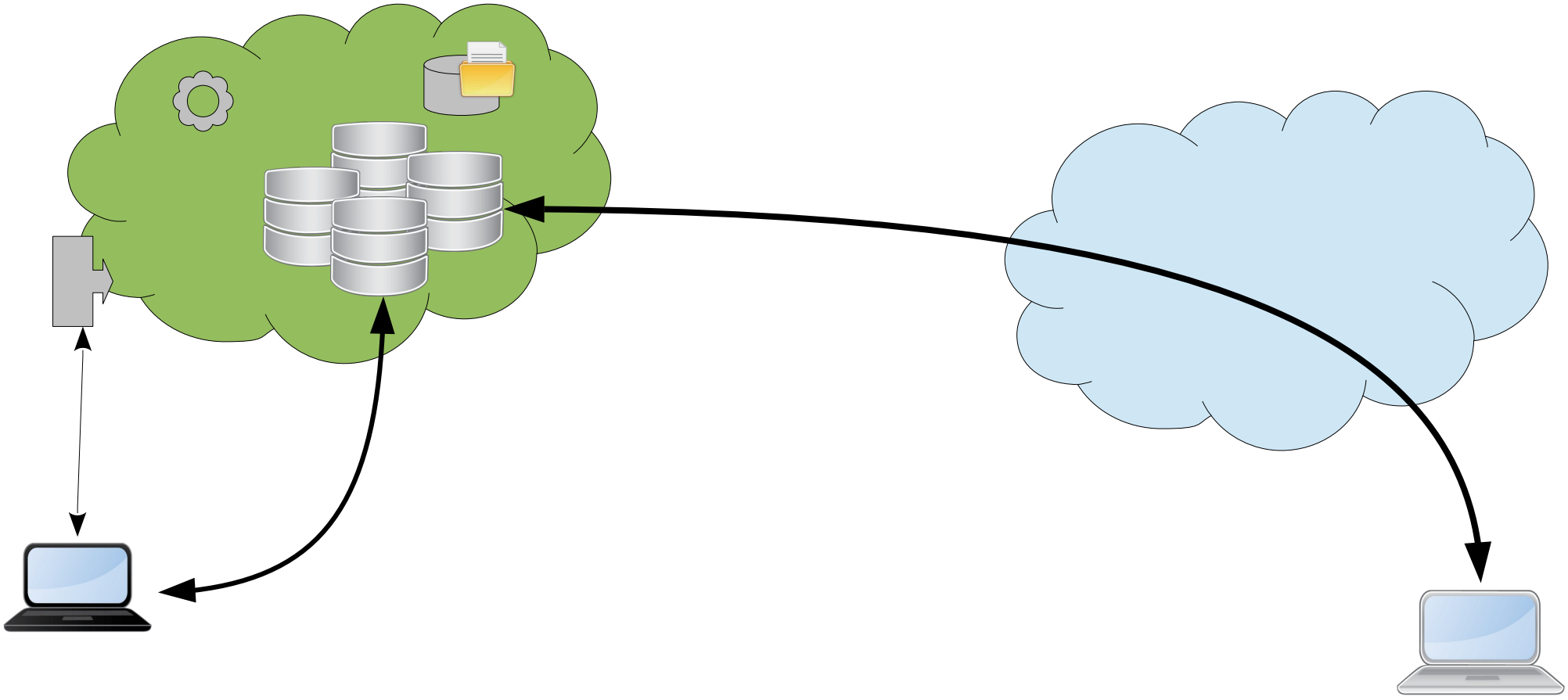


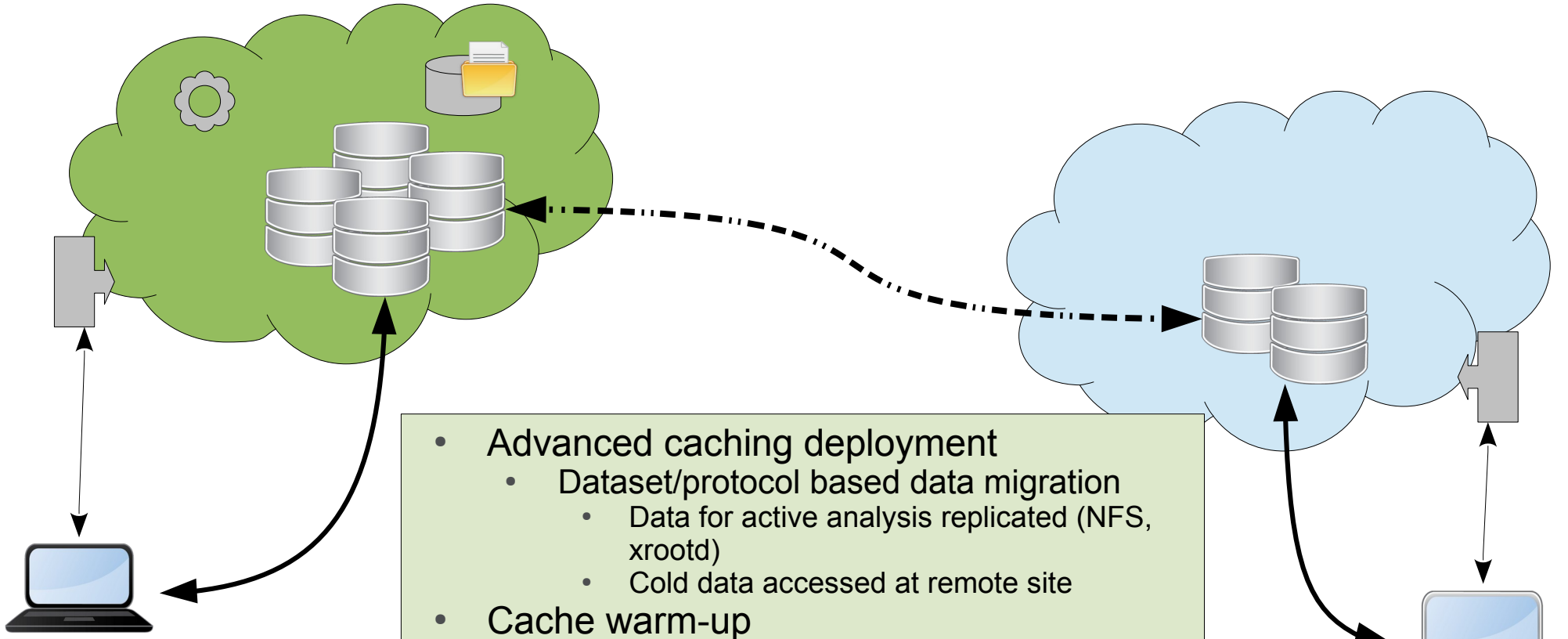
From Tiny to Huge



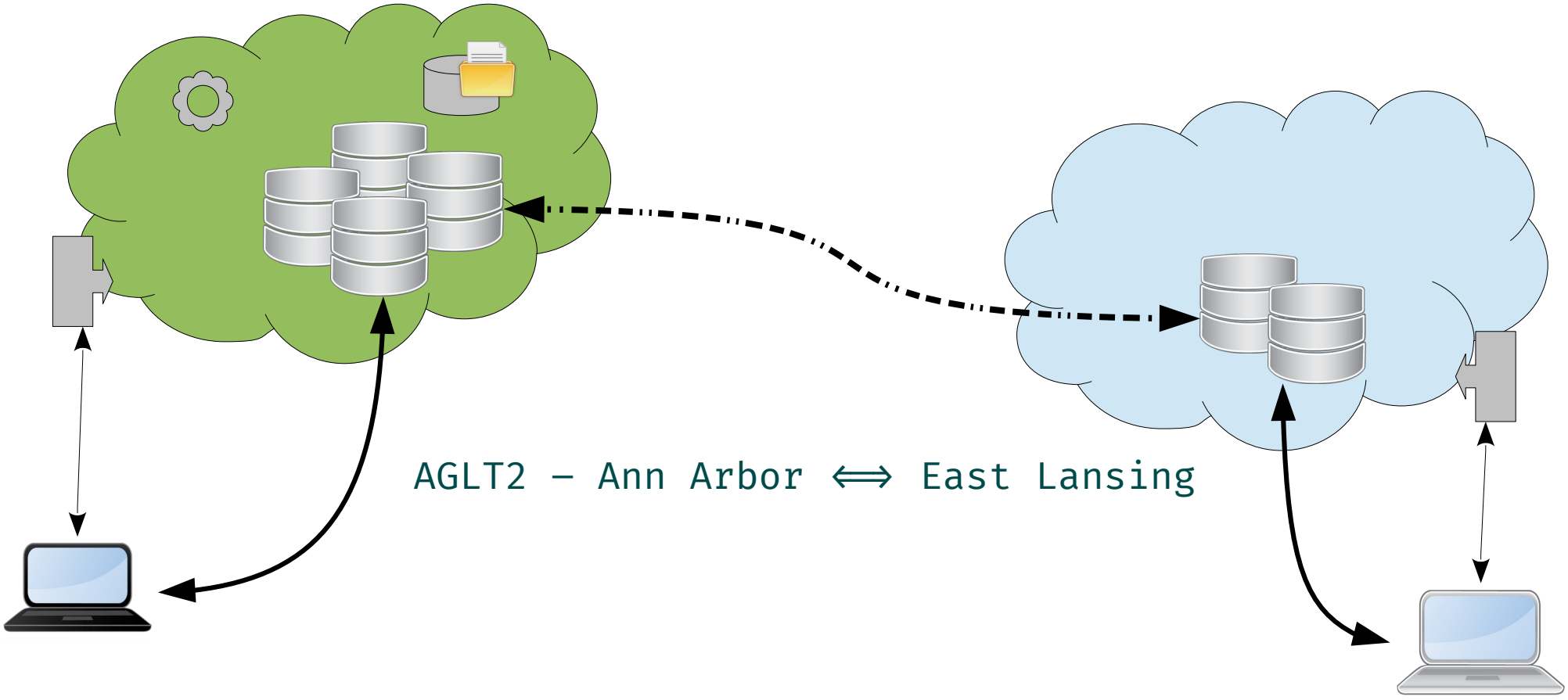


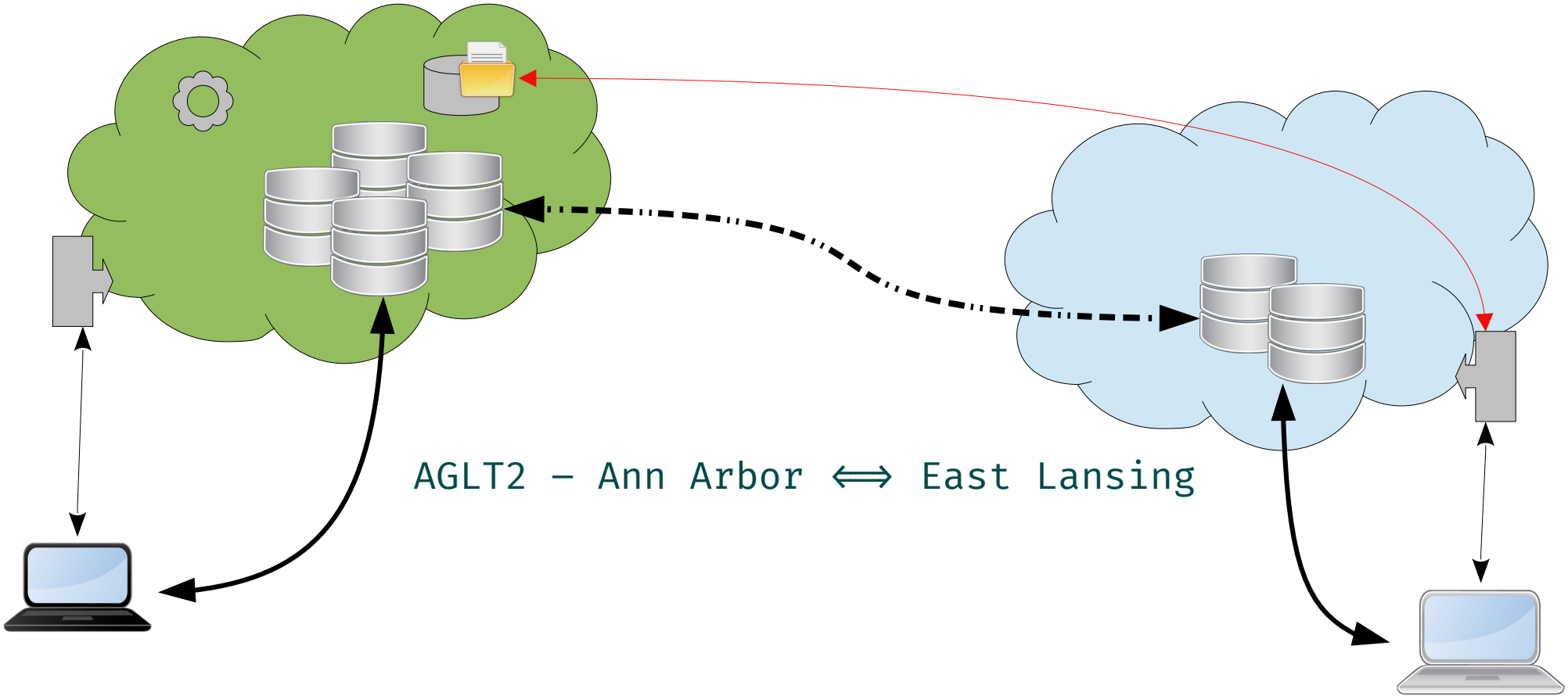


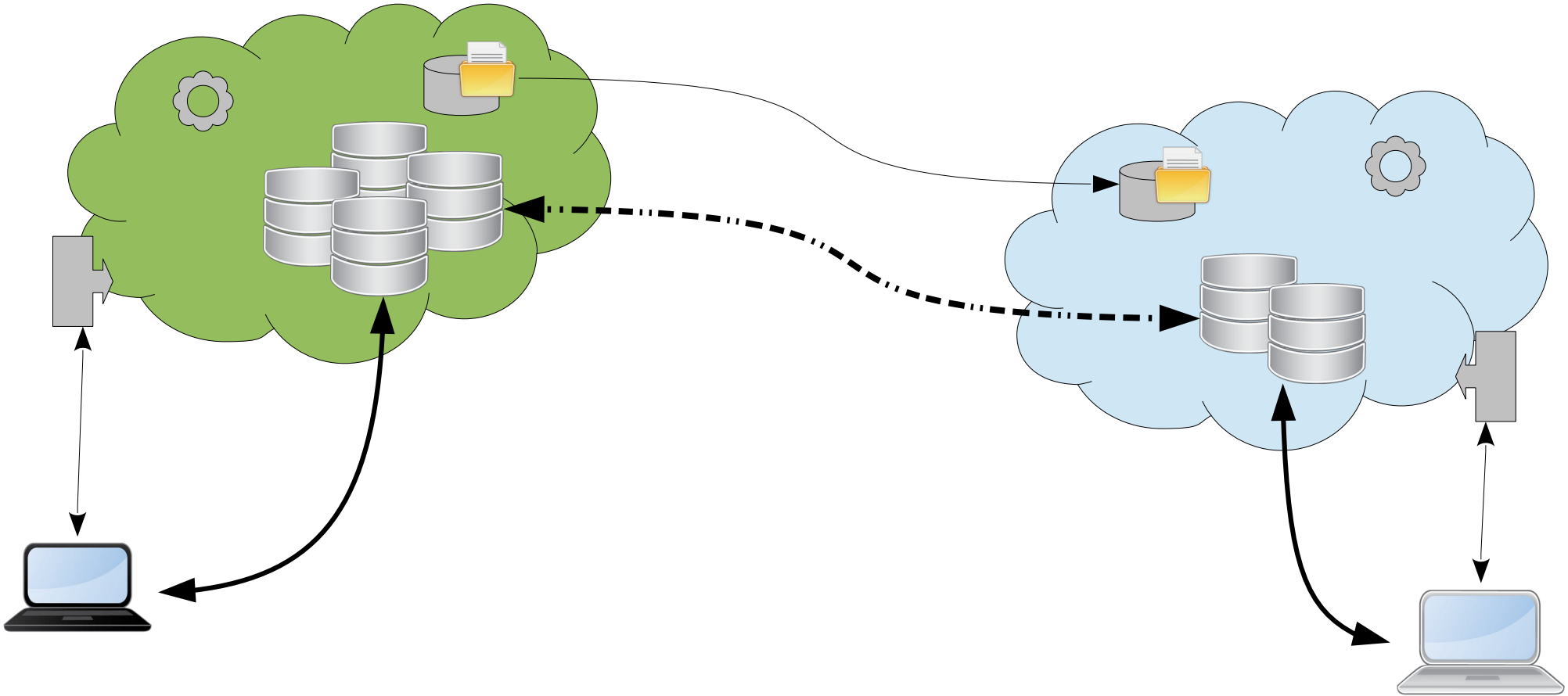




- Advanced caching deployment
  - Dataset/protocol based data migration
    - Data for active analysis replicated (NFS, xrootd)
    - Cold data accessed at remote site
- Cache warm-up


AGLT2 – Ann Arbor ⟺ East Lansing

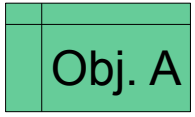

AGLT2 – Ann Arbor ⟺ East Lansing



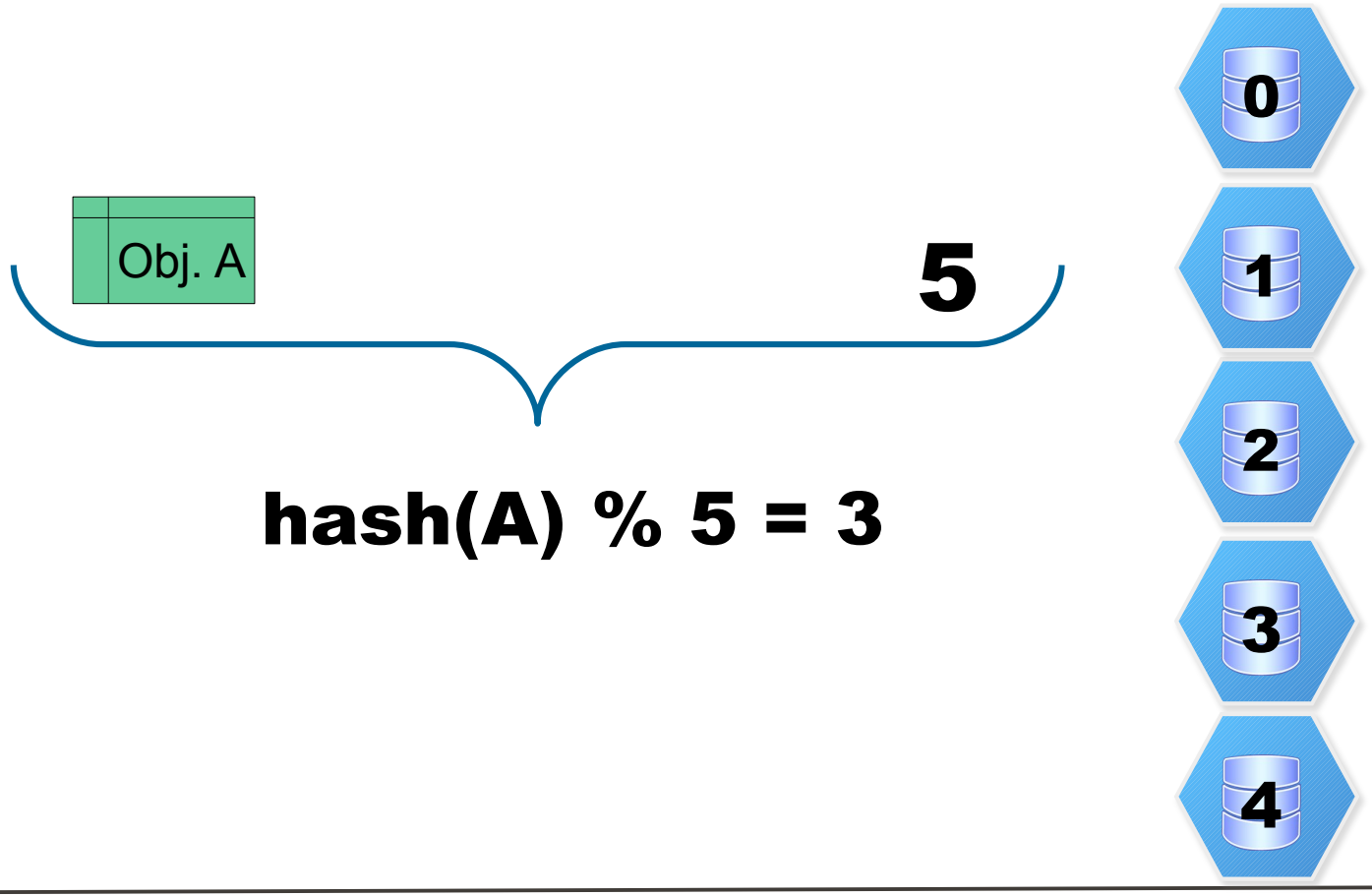
Zones: Geo-location

- Geo-location aware unit
- Dynamically groups services together
- Available in replication rules
- Network topology aware internal communication
  - Always prefer local resources
  - Disconnected operation **(CAP violation!)**

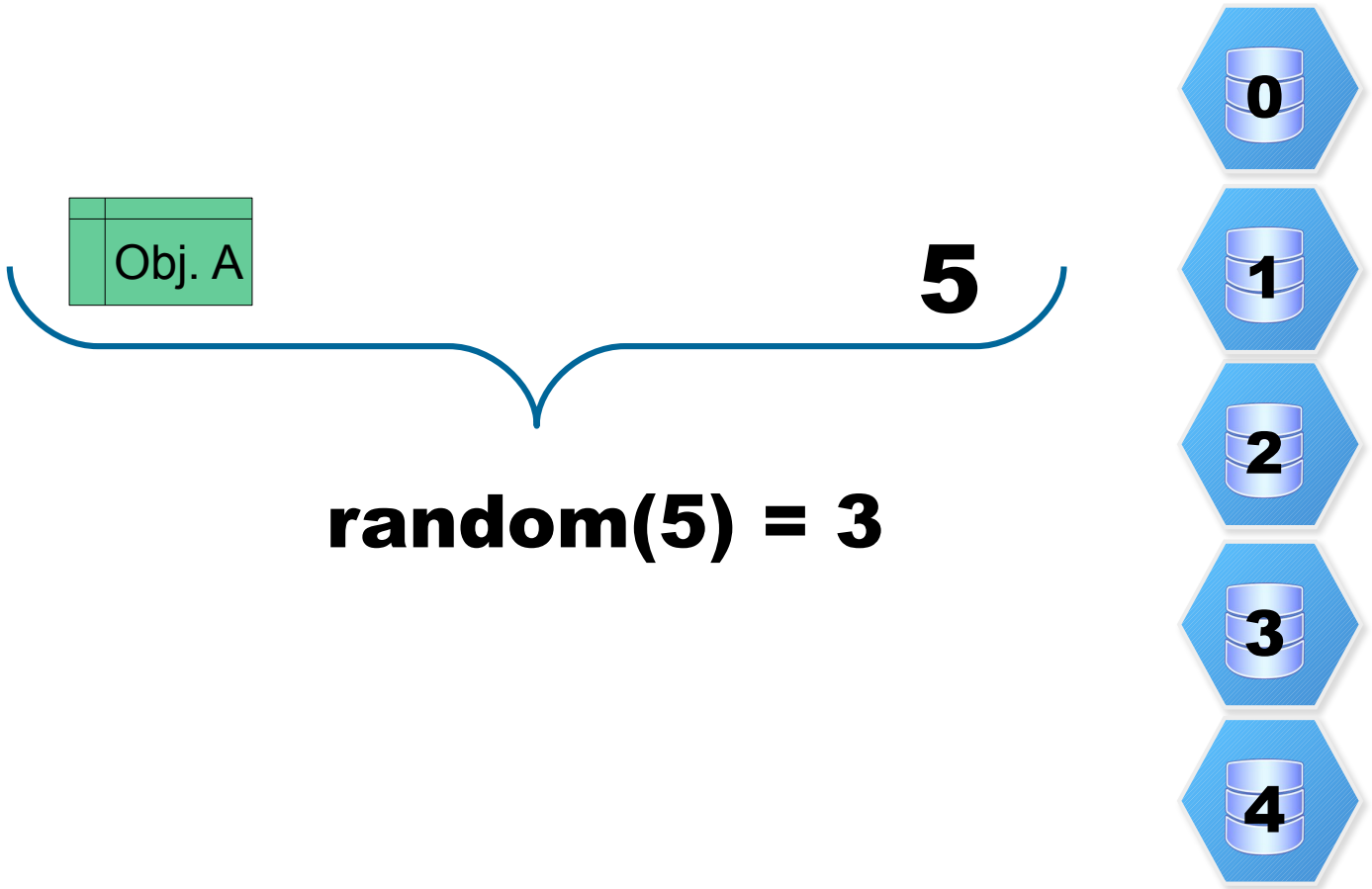
File Placement



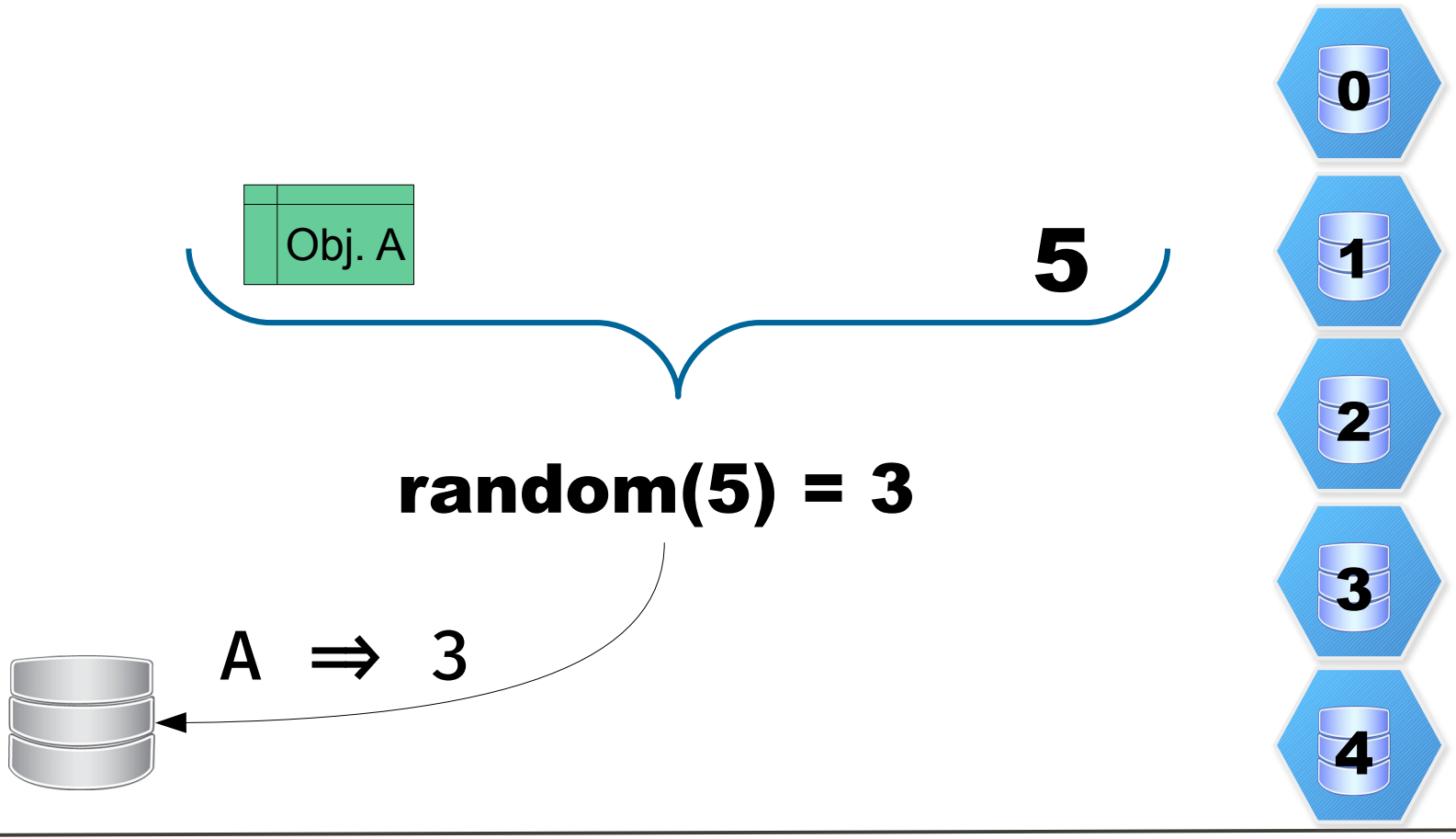
File Placement



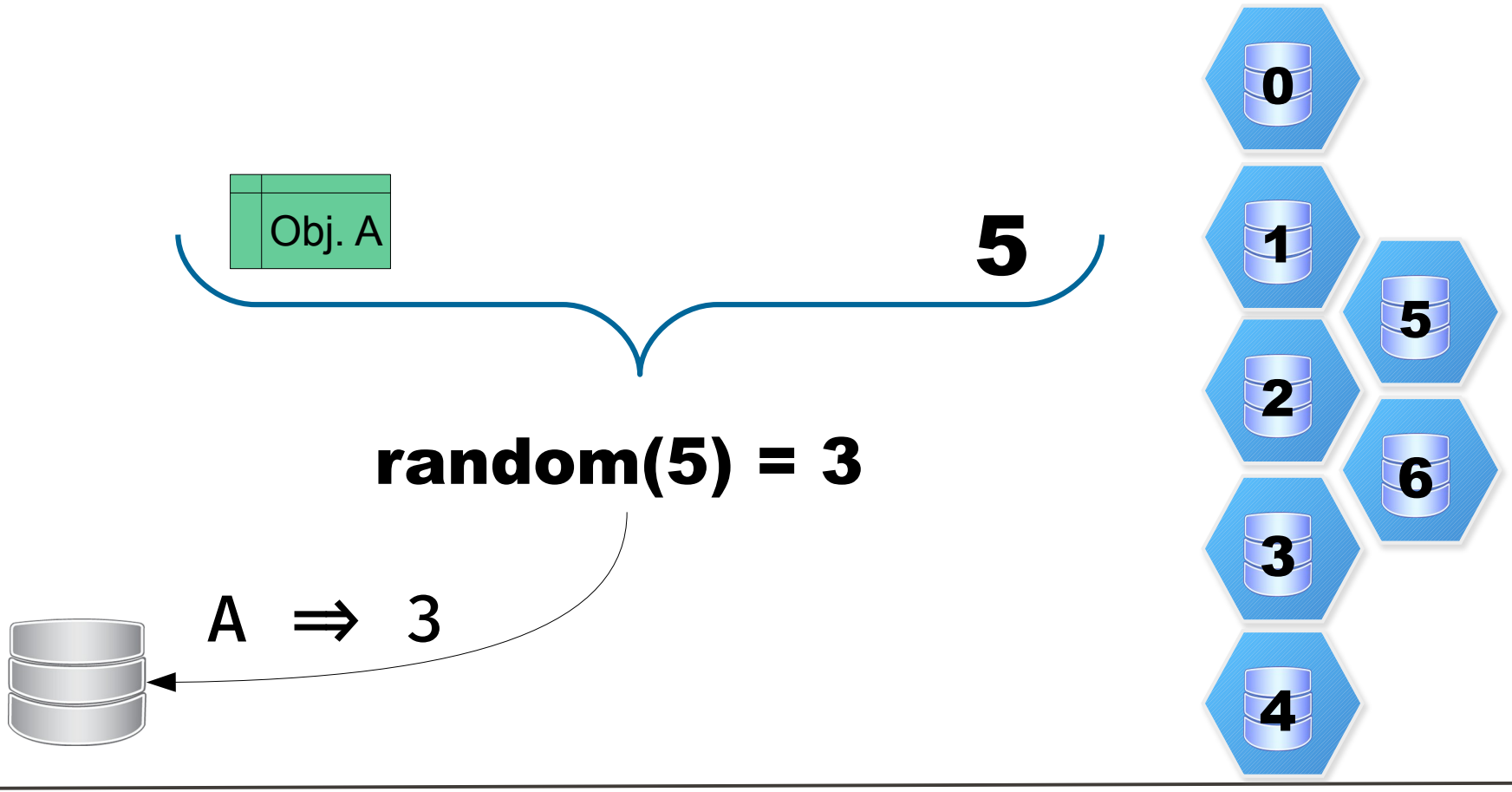
File Placement



File Placement



File Placement

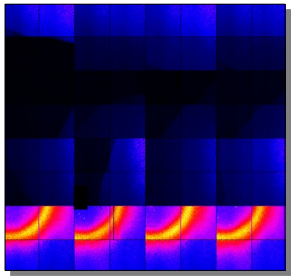


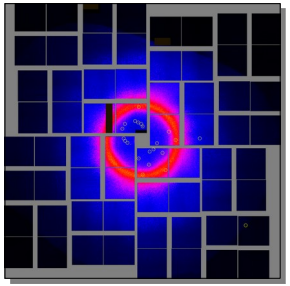
3rd Party COPY

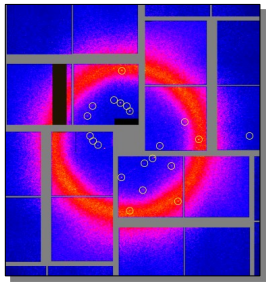
- XROOTD
  - Source/destination support
  - GSI + delegation, SciToken, TLS
  - Inter-op with SLAC xrootd client & server (DPM, EOS)
- HTTP
  - Source/destination support
  - 3rd vendor HTTP server as destination
  - X509, Macaroon and SciToken support

Data is Not Homogeneous

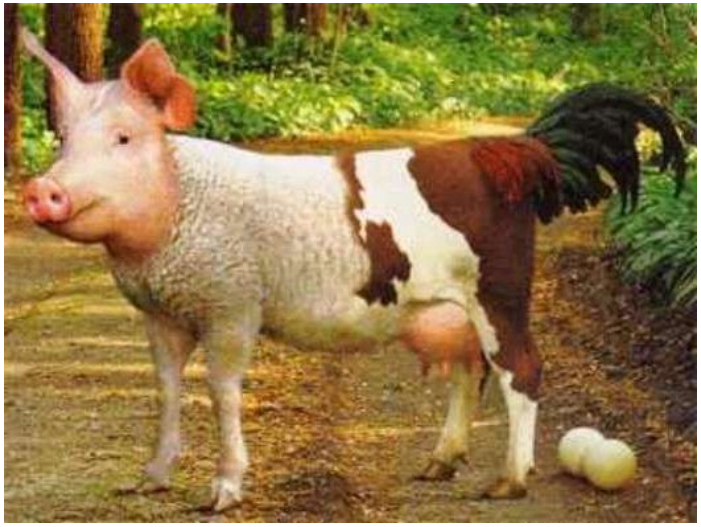
- Precious, expensive to re-create
- Derived
- Popular
- Transient, easy to re-create
- Mix-of all above







Quality of Service (in Storage)

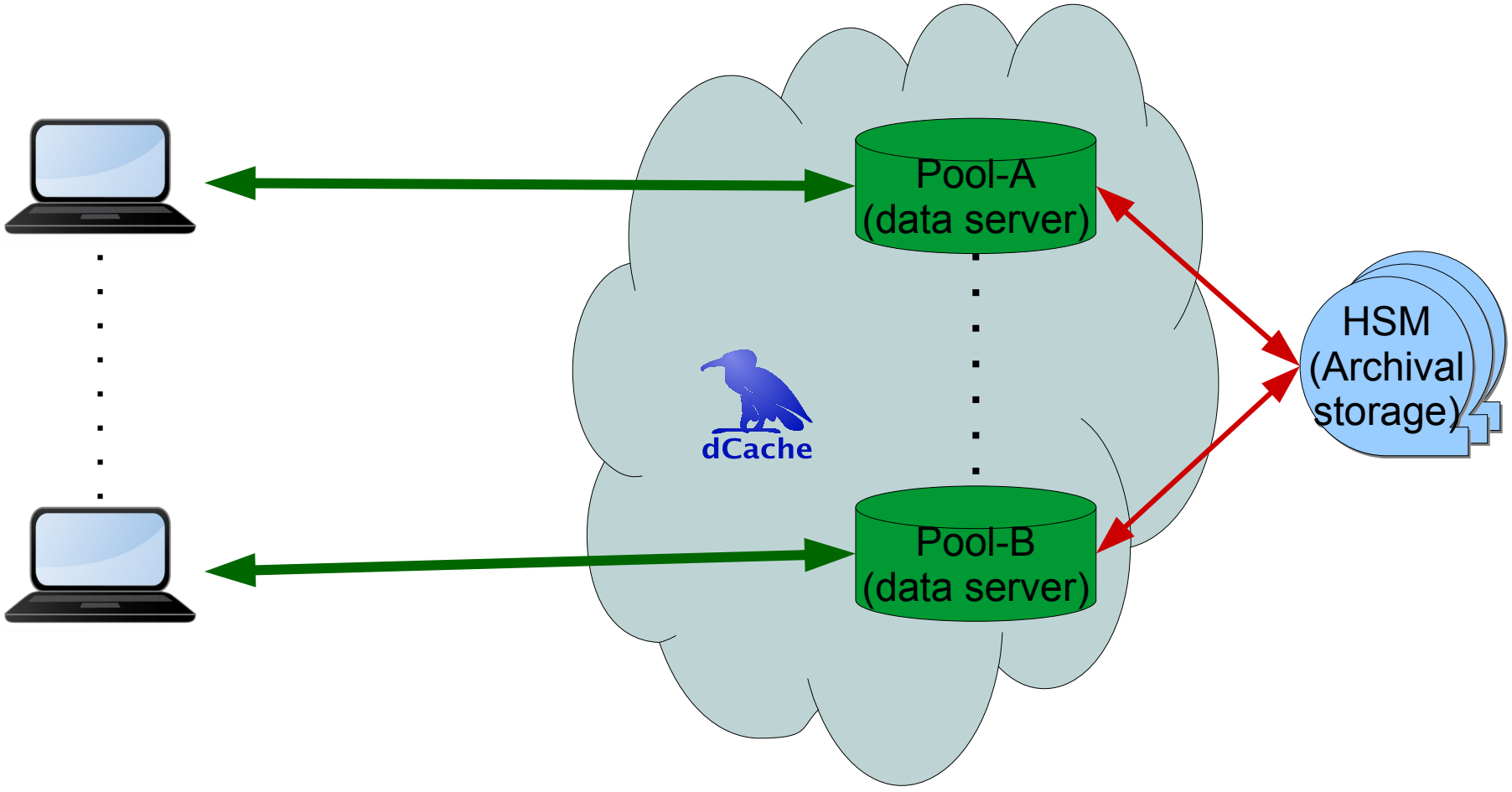


- Let users to define their needs
  - Access latency
  - Durability/Probability of data loss
  - Budget
- Let Storage system to deal with implementation
  - RELIABLE
    - Multiple Disk/Tape copies, Erasure encoding
  - FAST
    - SSD or multiple copies
  - FAST+RELIABLE
    - SSD + Tape
  - Site may decide how to provide required QoS based on know-how and budget

Tape connectivity

- Native to dCache
  - The essential part of the design
- Write-back/Read-through -like behavior
  - Transparent for the end users
- Stage protection
  - User/protocol based
- Supports multiple HSMs on a single instance
- Provides full functionality with/without HSM
  - Tape and disk files can be mixed on a single data server

Tape Connectivity



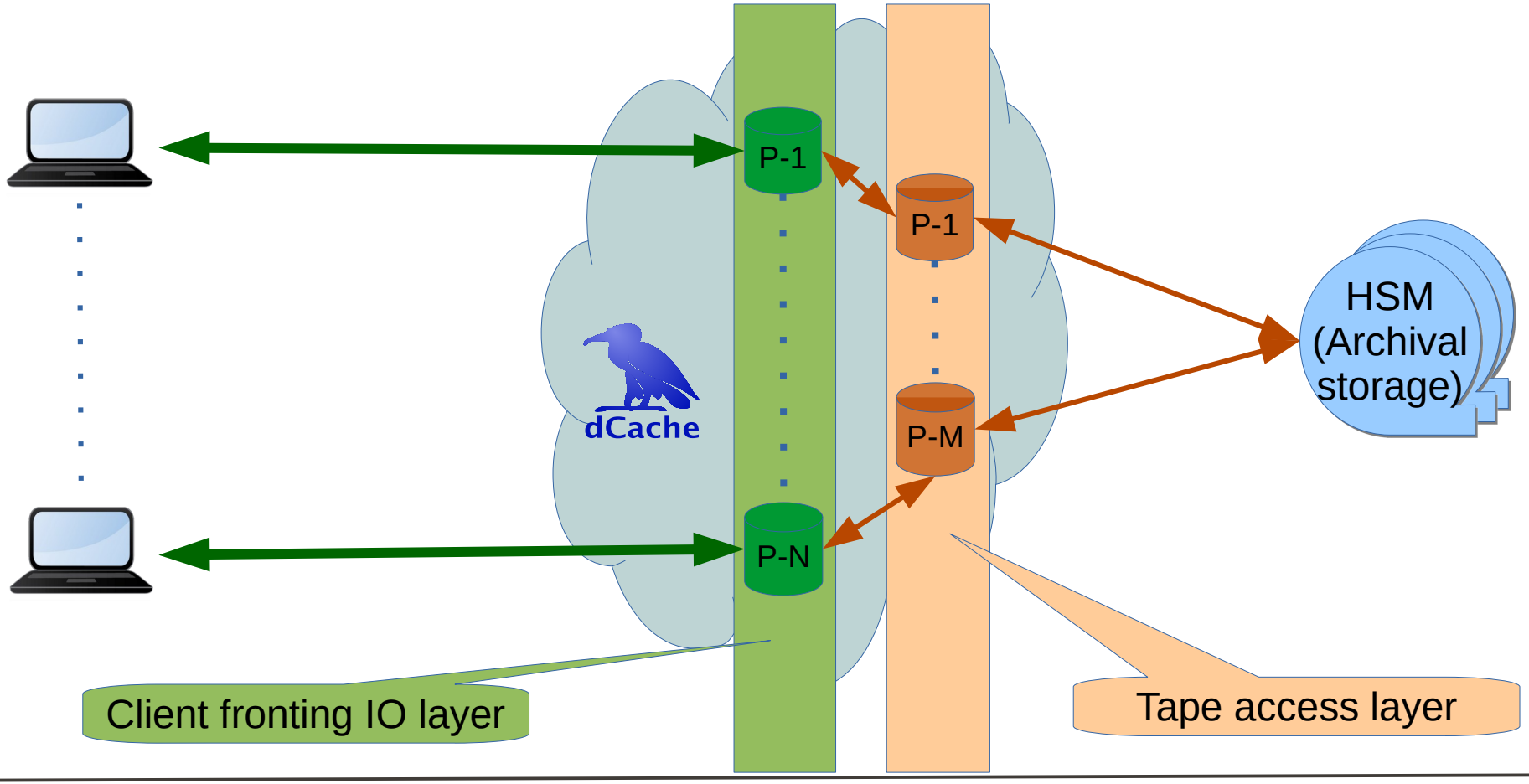
HSM, Tape, QoS

- ATLAS “Tape carousel” ⇒ WLCG “Data carousel”
  - Share the best practices
- High data volumes by EuXFEL
  - ~1PB/week
- High number of small files by Photon Science
  - ~4MB, $10^{\wedge 6}$ files per directory
- Multi-media copy guarantees

Three Directions to Address

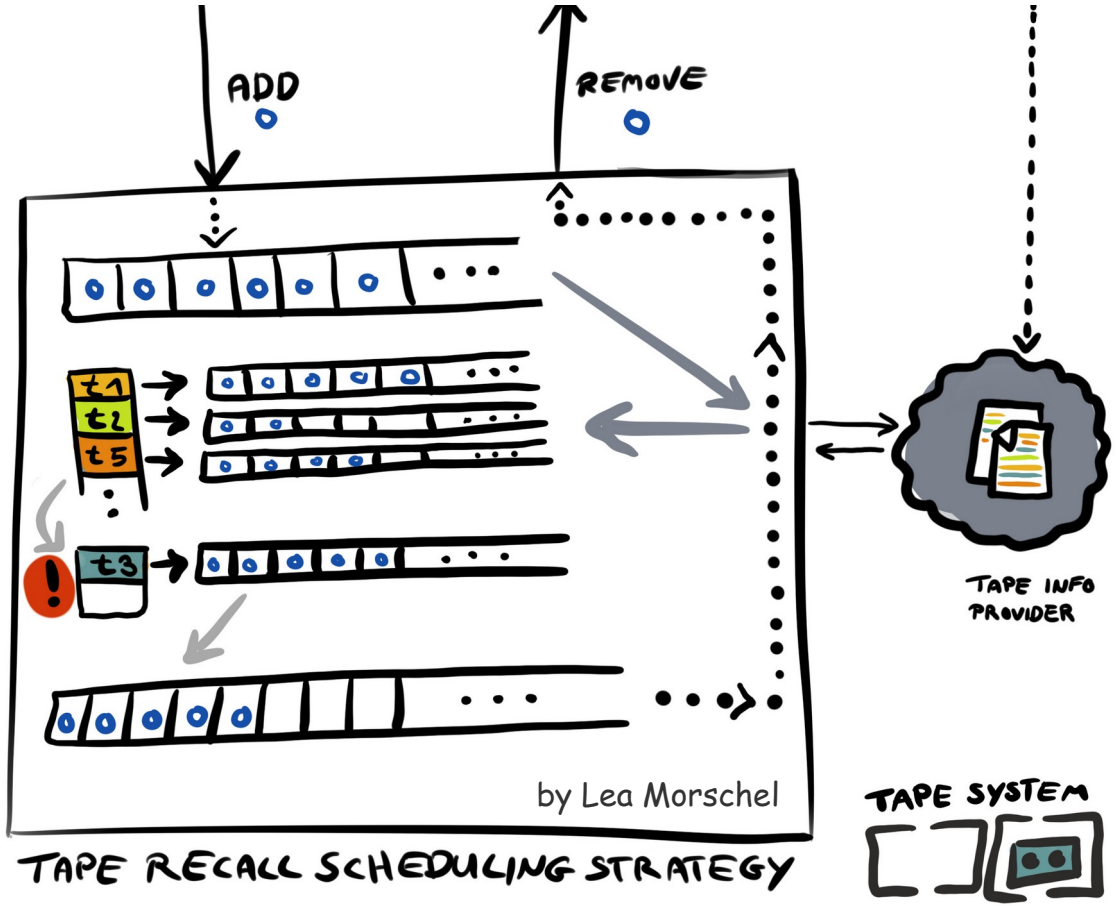
- Better HW split on tape/disk pools
  - Some nodes can be optimized for tape access only
  - A-la QoS for hardware
- Tape recall grouping by tape
  - Collect request for a single tape
  - Prototype in SRM
- **Sapphire** - small file aggregation
  - dCache native HSM driver

Layered Pools Model



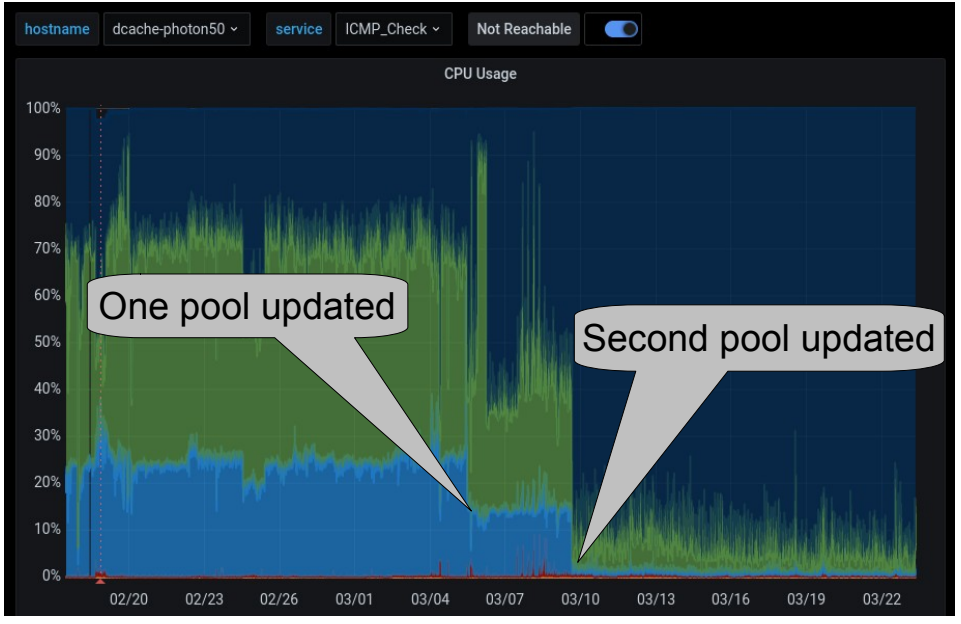
Tape recall grouping

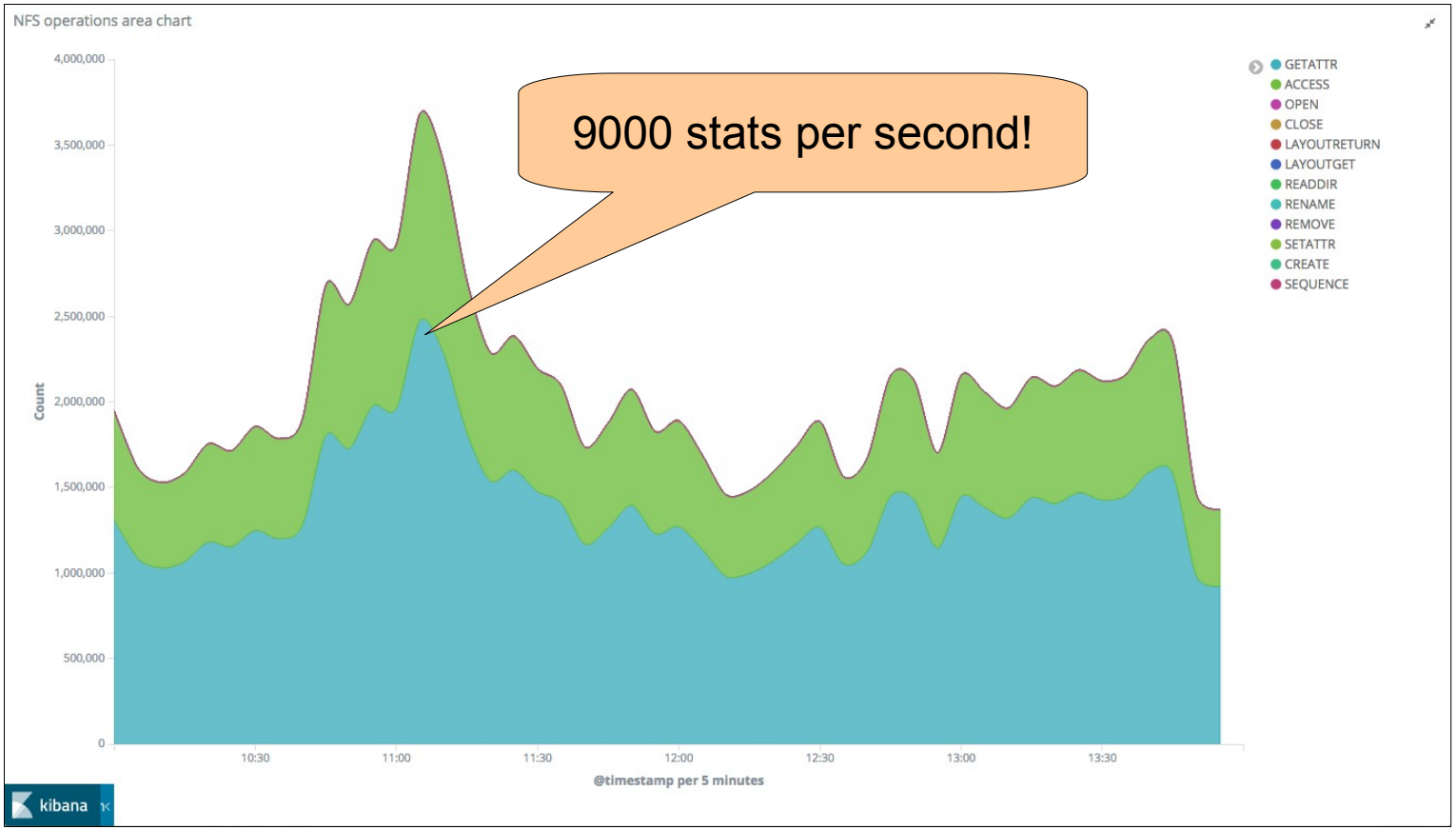
- Group requests by tape
- Recall triggered by
  - Size
  - Max idle time
- Number of parallel recall based on number of tape drives



Sapphire (small file plugin)

- Evolution of *Small-file-plugin*
  - Addresses discovered limitations
- In-dCache HSM driver
  - Full access to metadata
  - No external script
  - Stateful
- Better resource utilization



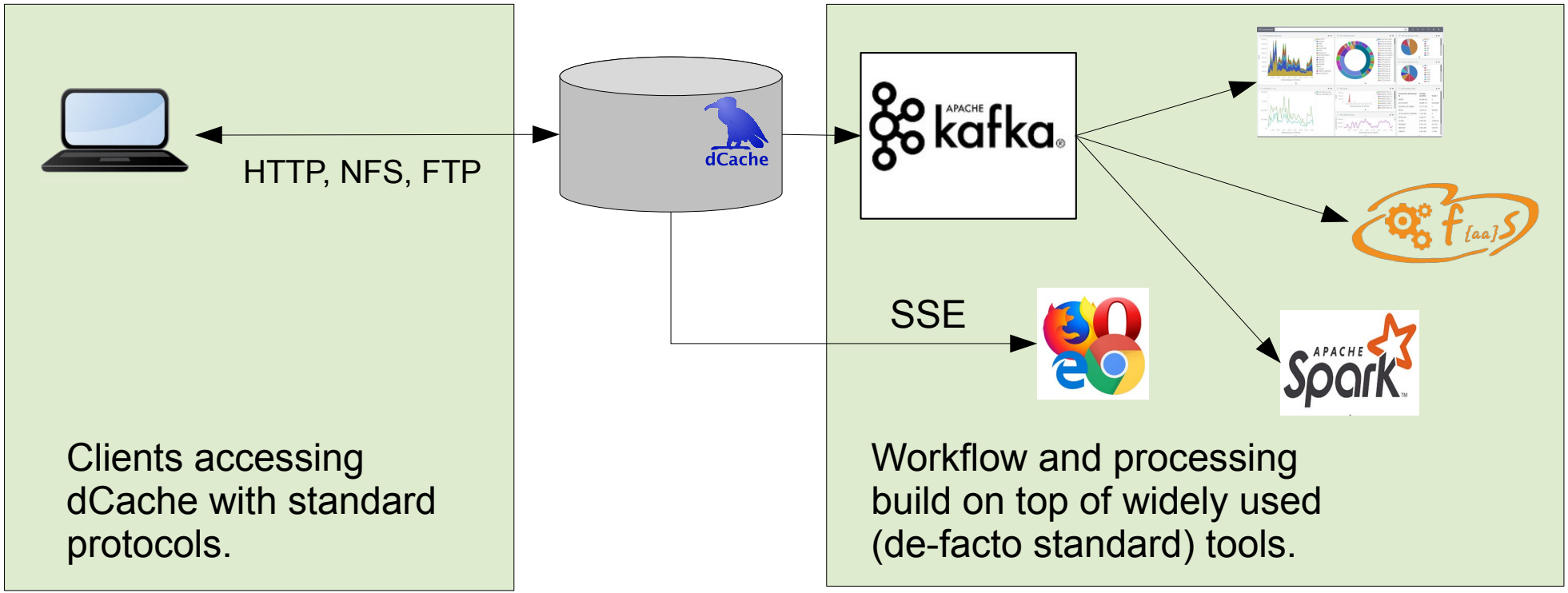

NFS operations area chart
9000 stats per second!
4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Count
10:30
11:00
11:30
12:00
12:30
13:00
13:30
@timestamp per 5 minutes
GETATTR
ACCESS
OPEN
CLOSE
LAYOUTRETURN
LAYOUTGET
READDIR
RENAME
REMOVE
SETATTR
CREATE
SEQUENCE
kibana


ARE WE THERE YET !?!
MATT GROENING

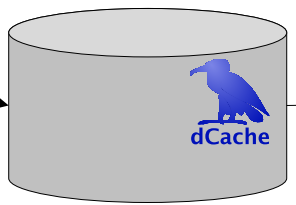
Storage Events

- Trigger actions on user activity
  - **Stop polling, Please!**
- Storage system becomes a workflow engine
- Producer-consumer model
- For infrastructure
  - Apache Kafka
- For individuals
  - Server Sent Events

Standards Everywhere…



Automatic Metadata Population



Event: new data

extract metadata

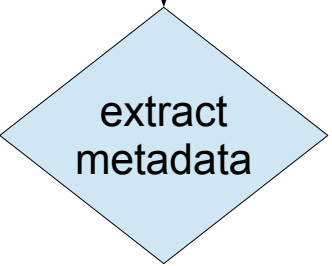
Automatic catalog update

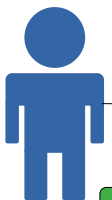
metadata catalog

#Anna Bertha Ludwig

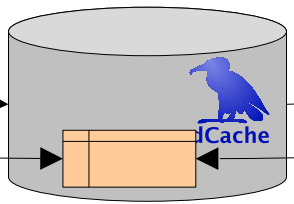
#hand #1895 #x-ray

Metadata Population





Manual Labeling



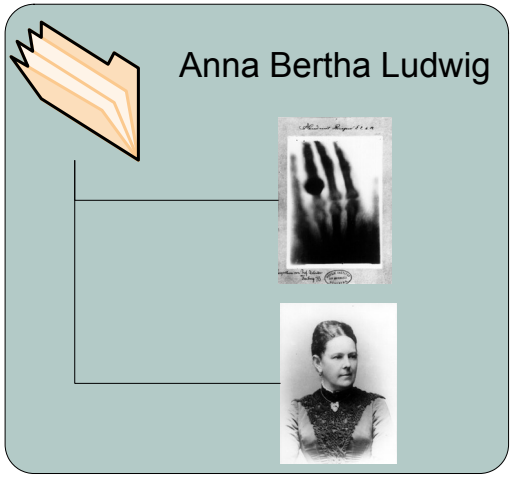
Automatic labeling

#hand #1895 #x-ray

#Anna Bertha Ludwig

User Metadata/Labeling in dCache

- Extended attributes
  - Exposed via NFS, WebDAV, REST
- Label-based virtual **read-only** directories (WIP)
  - List all files with a given label
- dCache rules applies
  - Visible through all protocols
  - Respect file/dir permissions



Multiple Identities Problem

- x509 (grid)

  /C=DE/O=GermanGrid/OU=DESY/CN=Tigran Mkrtchyan

- Kerberos

  tigran@DESY.DE

- LDAP

  uid=tigran,ou=people,ou=rgy,o=desy,c=de

- Unix ID (uid)

  3750

GPLAZMA – Plugable AuthN

- Pam -like system
- Allows to combine multiple plugins
- Supports many standard and custom authentication plugins
  - From ActiveDirectory to gridmap file

Federated IdP & Jupyter Notebooks

PaNOSC

Laser-Driven Proton Acceleration from Cryogenic Hydrogen Target

Description
2D particle-in-cell simulation of the interaction of high-intensity laser pulse (parameters are relevant to L4 laser) with a cryogenic hydrogen target. Only protons with energy above 300 MeV at the end of the simulation are tracked and their position and energy are visualized. Two different groups of protons accelerated by different mechanisms can be distinguished from each other in space: Protons originated from the target interior and from the target rear side.

Citation	Dana Scully; (2020), Re-polarization of the aft quantum plasma collector, DOI:10.9563/if.2015.87.012
Keywords	X-ray excited optical luminescence,
Type	Proposal
Author	Laima Reinhold
Other	Stuff

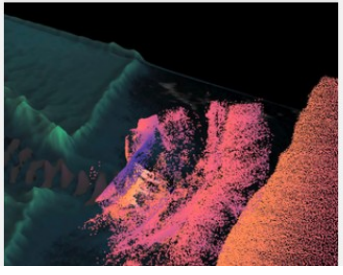
Stolen from Michael Schuh

Datasets

PaNOSC Test Dataset 11
HEIMDAL @ ESS

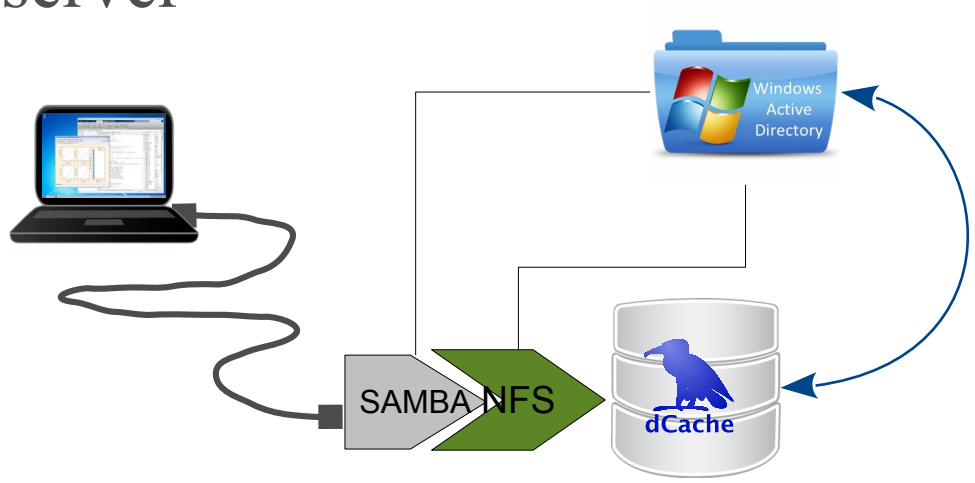
Name
my-environment
Descritption
re-produce research
Flavour
jupyter_small
Spawn

Preview Visualization



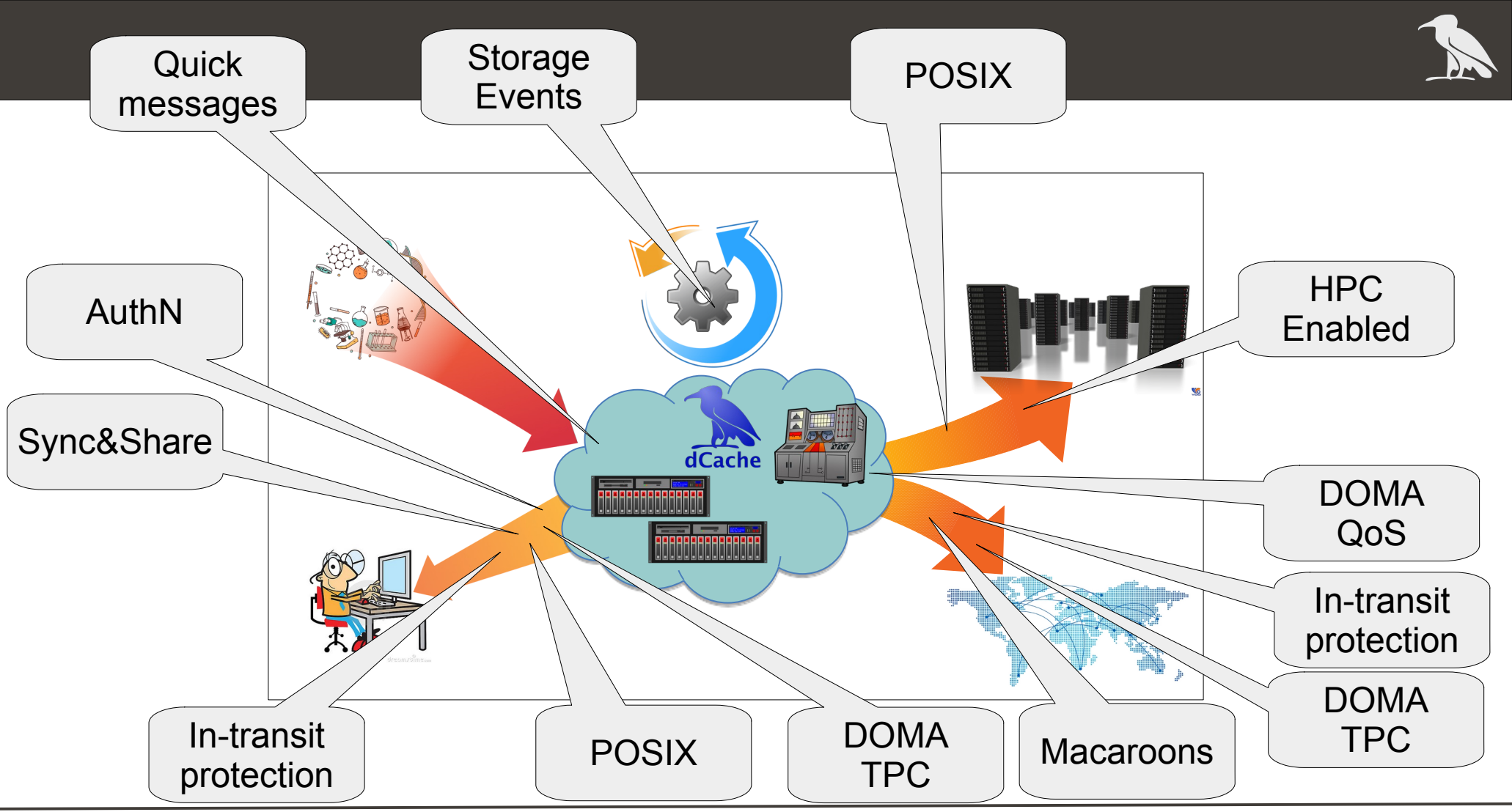
dCache+SAMBA

- Re-exporting dCache mount with SAMBA
- User permissions and identity preserved
  - dCache uses AD as LDAP server



Summary & Conclusions

- The dCache team has been providing a reliable software to manage scientific data for over 20 years.
- Seamless integration into the site's infrastructure makes dCache a natural part of any data center.
- Multi-protocol and authentication scheme capabilities allow to support multiple communities even on a single instance.
- In close cooperation with experiments we address today's and future data management challenges.


Quick messages
Storage Events
POSIX
AuthN
HPC Enabled
Sync&Share
dCache
DOMA QoS
In-transit protection
DOMA TPC
In-transit protection
POSIX
DOMA TPC
Macaroons

Thank You!

More info:

https://dcache.org

To steal and contribute:

https://github.com/dCache/dcache

Help and support:

support@dcache.org, user-forum@dcache.org

Developers:

dev@dcache.org

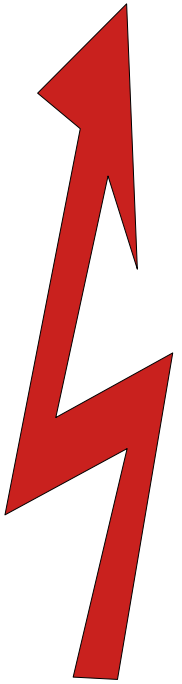
The MISSION

“... to provide a system for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual filesystem tree with a variety of standard access methods.”

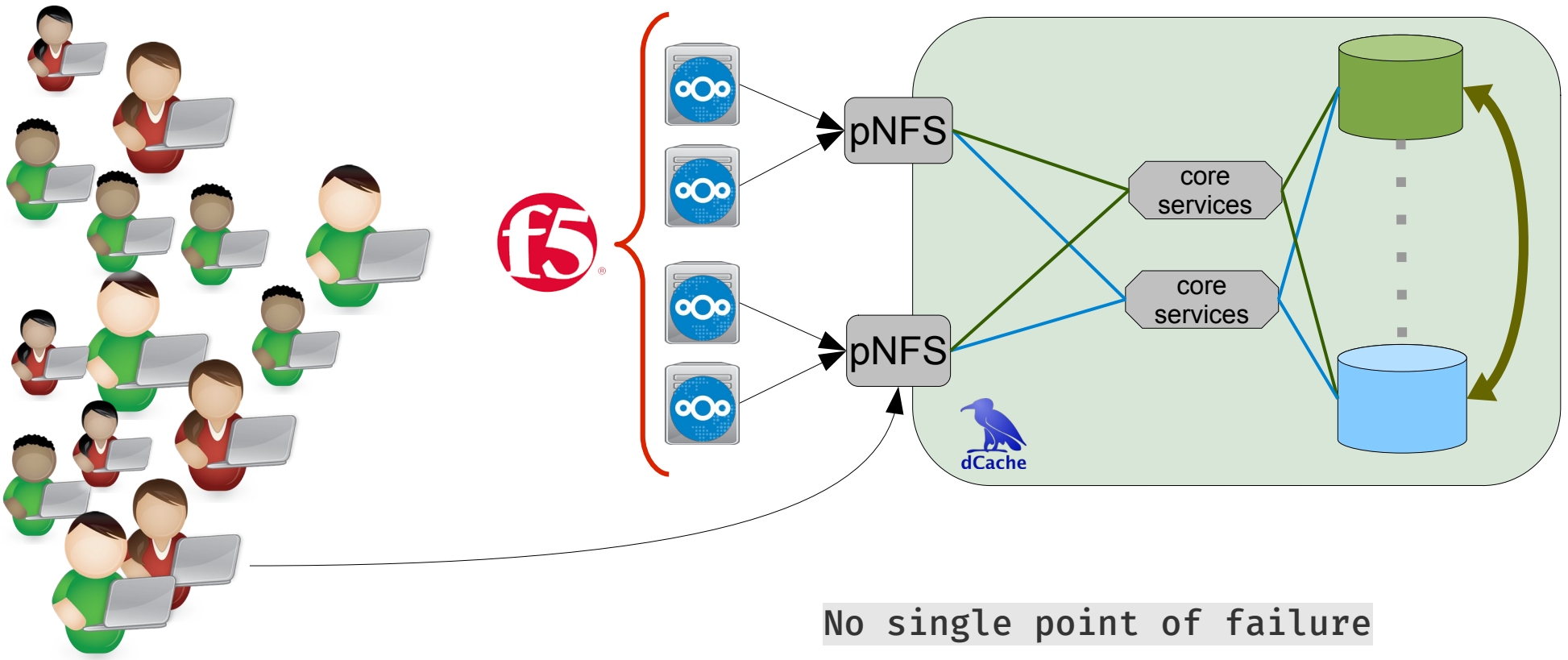
https://dcache.org/about/

dCache and CAP Theorem

- dCache provided consistency over availability*.
- All clients will see the same data at the same point in time.
- A timeout or error will be returned, if consistency can’t be guaranteed.



NextCloud Instance @ DESY



No single point of failure

dCache as a Storage Backend

- PB-scale storage system
- No changes in Nextcloud required
- Unique functionality
  - Tape integration
  - File ownership preservation
  - NFS export to selected users
  - Storage events
  - Data visible by all protocols and security flavors

v1 Bulk REST-API (like SRM, but different)

STAGE

- Request to stage many files at once

CANCEL

- Cancel bulk request

DELETE

- Cancel bulk request + clear history/status

EVICT

- unpin cached copy

PIN

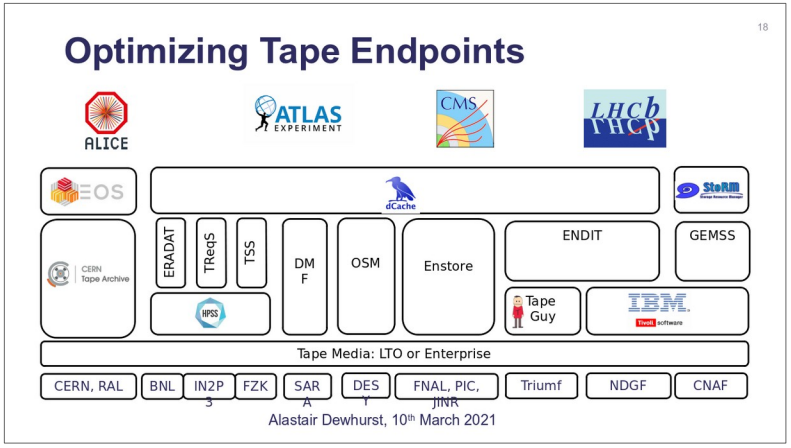
- Pin cached copies with a lifetime

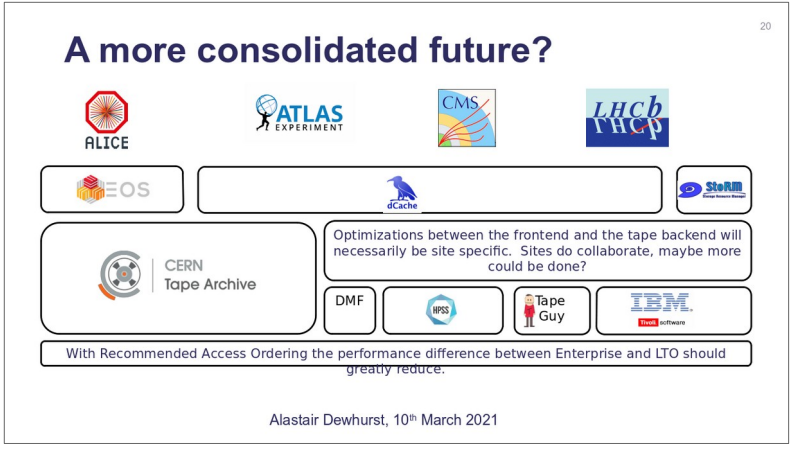
FILEINFO

- Request status many files at once (locality, checksum)

StoRM

dCache ⟺ CTA Integration



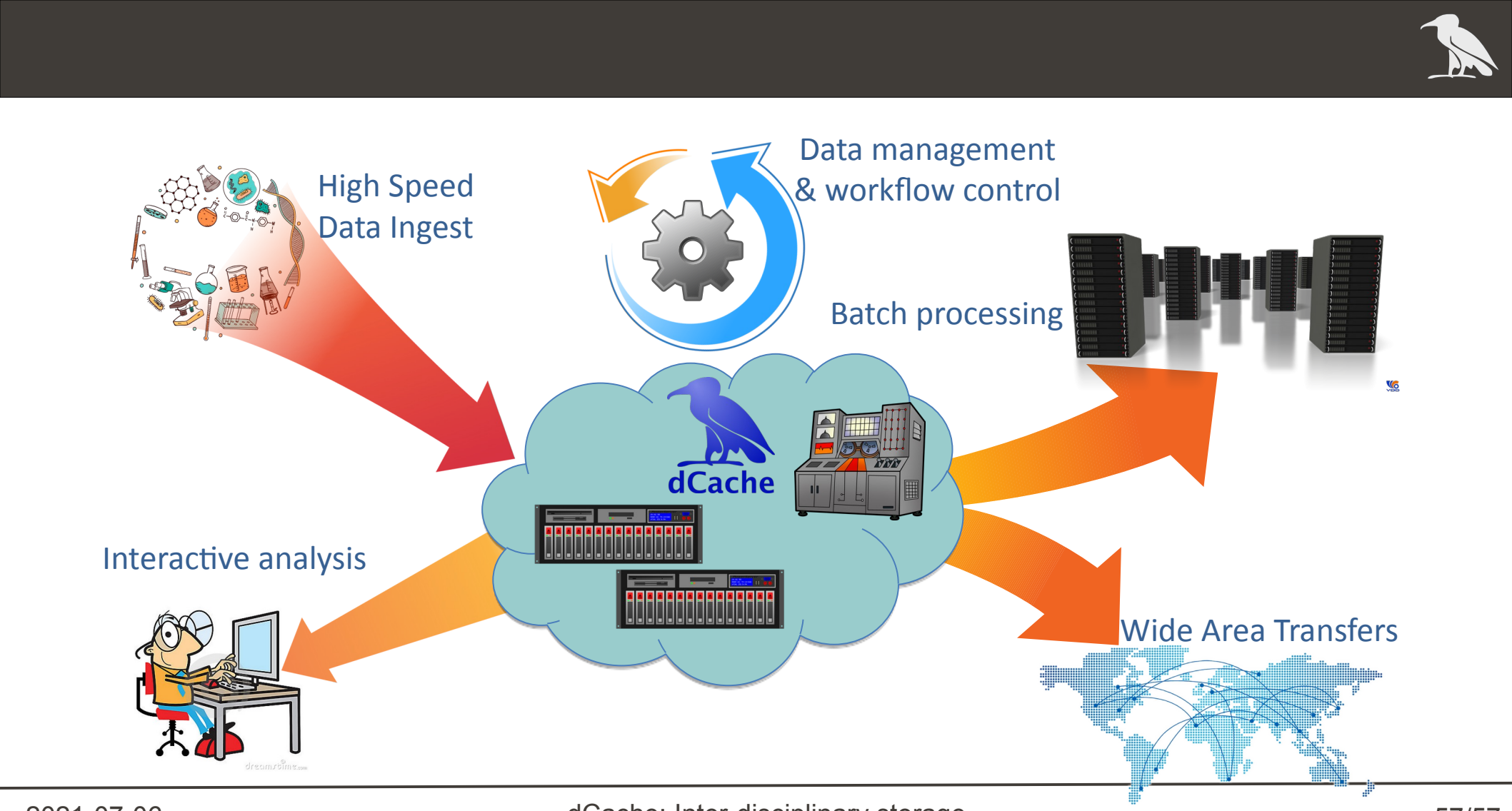


Pros

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

Cons

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


High Speed
Data Ingest
Data management
& workflow control
Batch processing
dCache
Interactive analysis
Wide Area Transfers