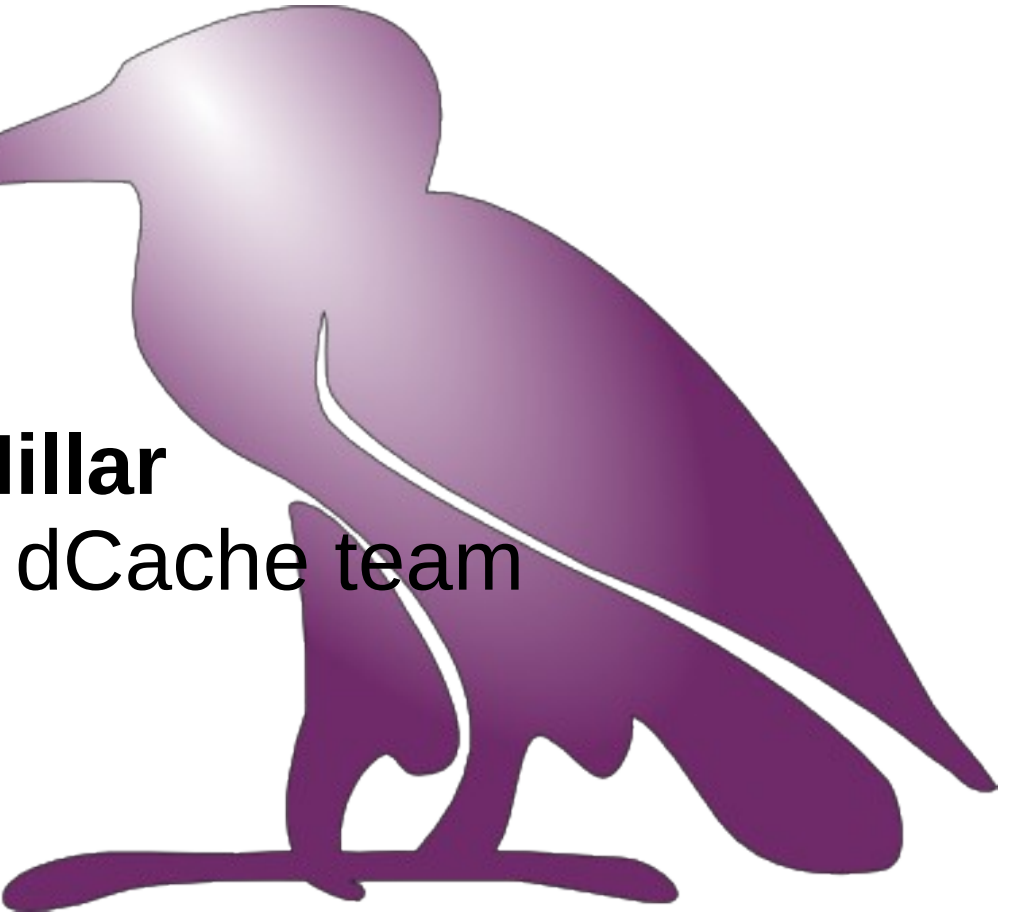
dCache

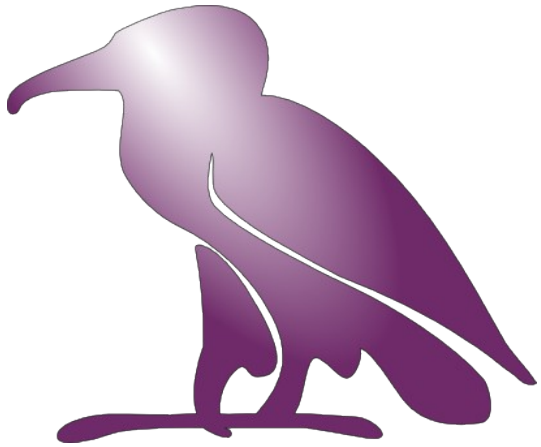


Paul Millar
On behalf of the dCache team

DESY

Overview of talk

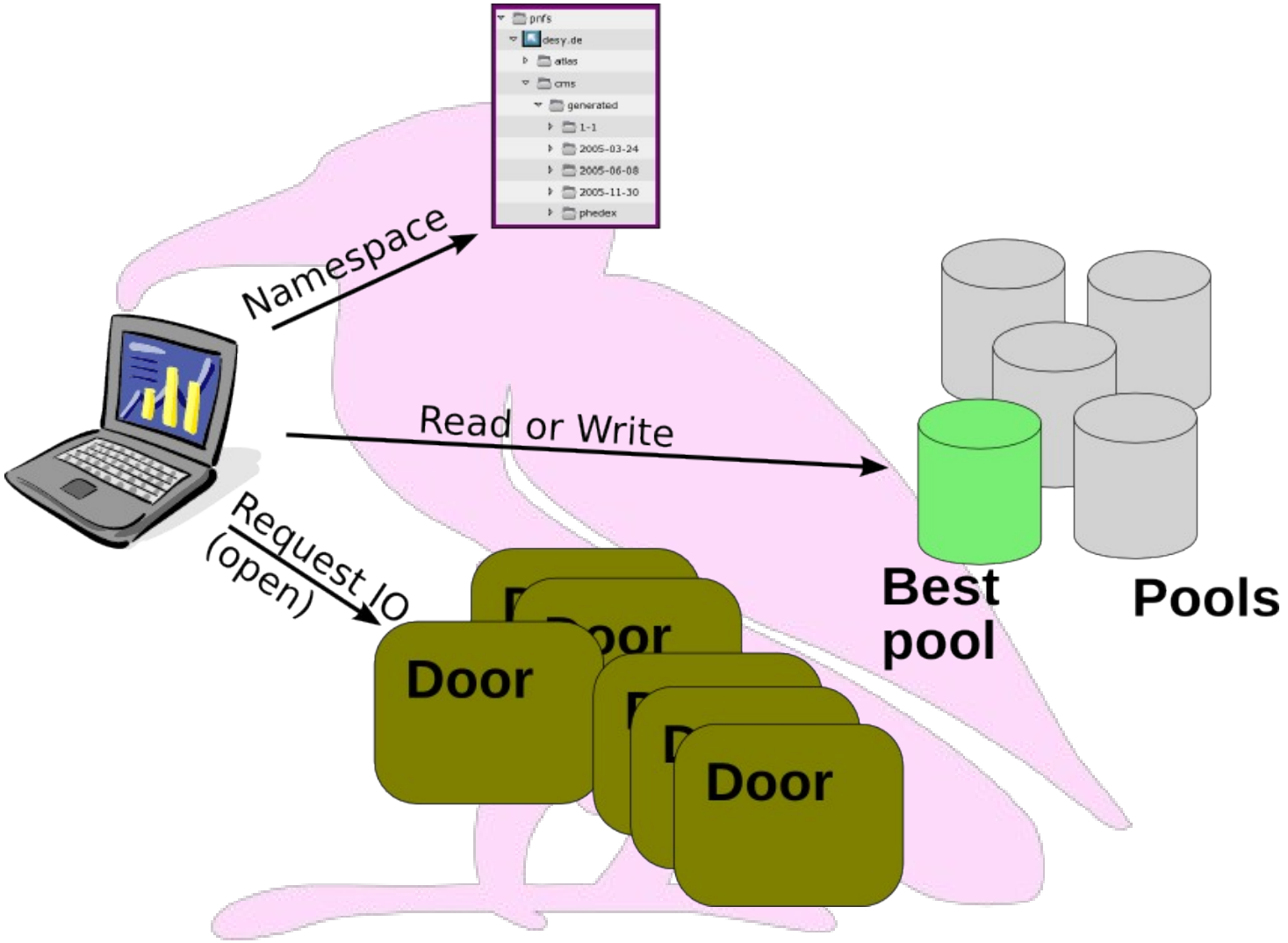
- Looking at dCache,
- We, dCache.org, ...
- Example deployments,
- dCache in a Grid context,
- Crystal ball.



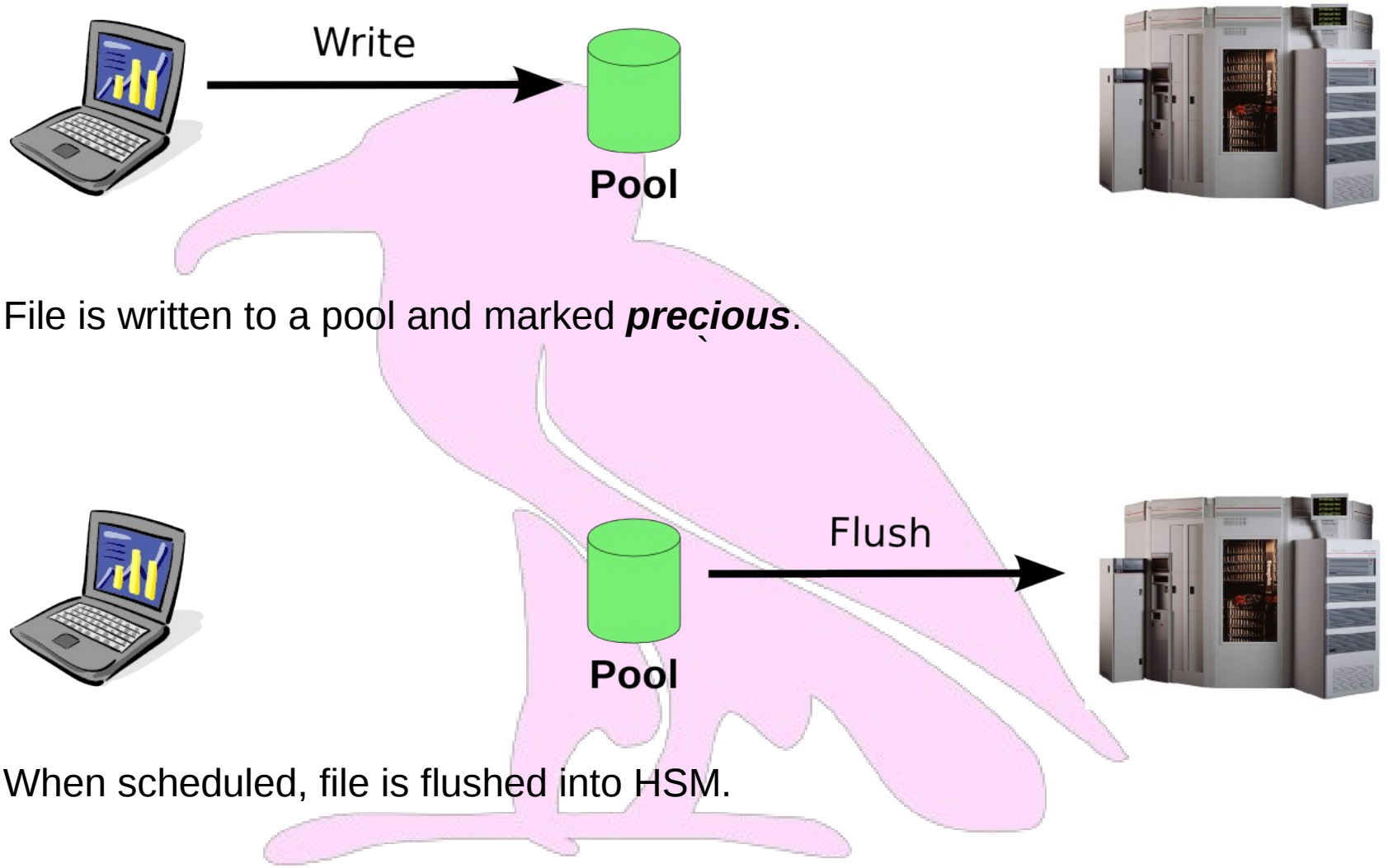
dCache at a glance

- (Grid) Storage Software,
- Can combines *lots* of heterogeneous servers
  - Only requirement: Java.
- Provides a rooted name-space
  - Completely decoupled from storage
- Support for HSM storage
  - HSM is transparent to end users
  - May have replicas internal- and external- to HSM
- Supports many file access protocols.

Disk: read/write



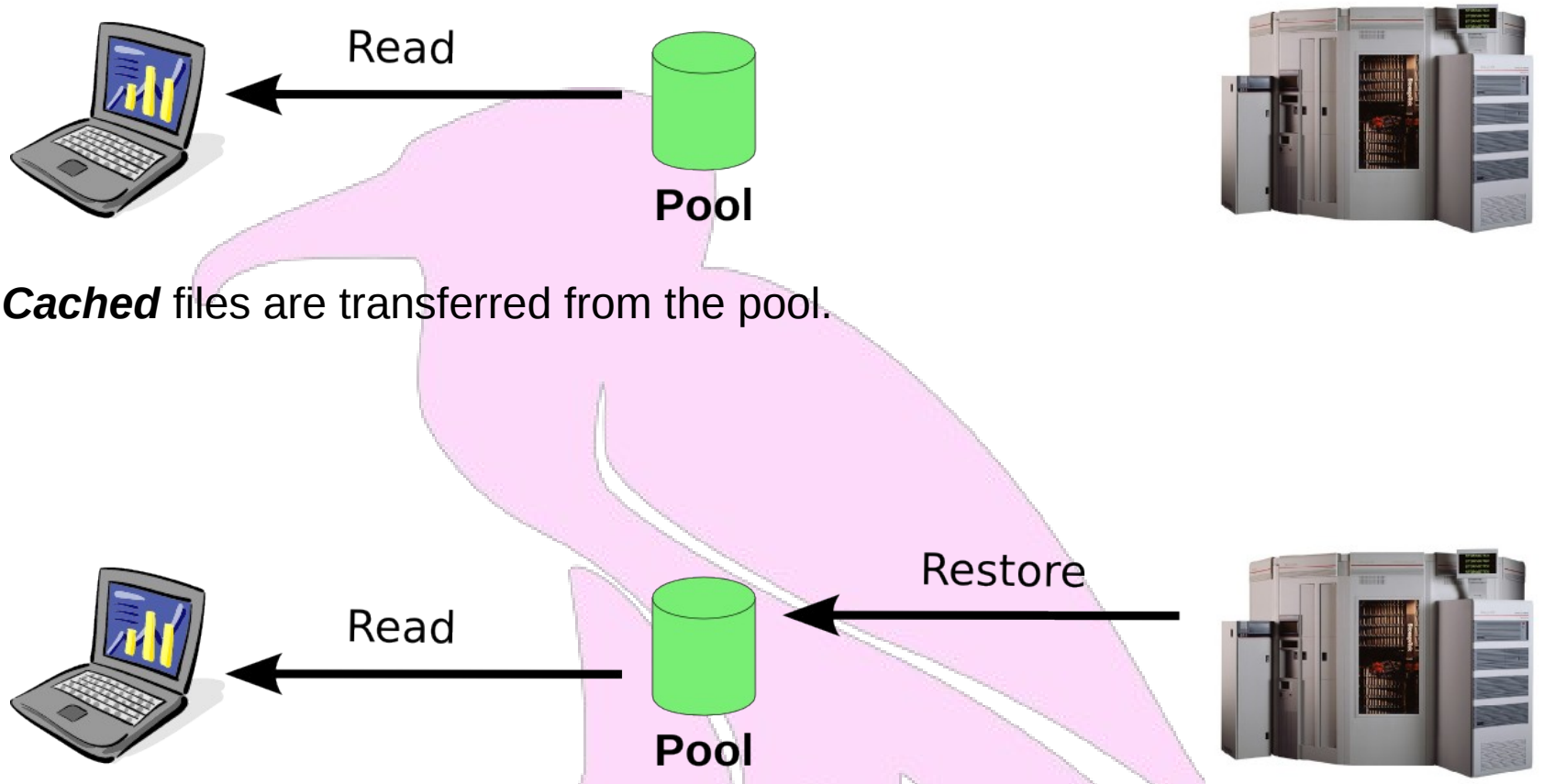
HSM: write



File is written to a pool and marked ***precious***.

When scheduled, file is flushed into HSM.

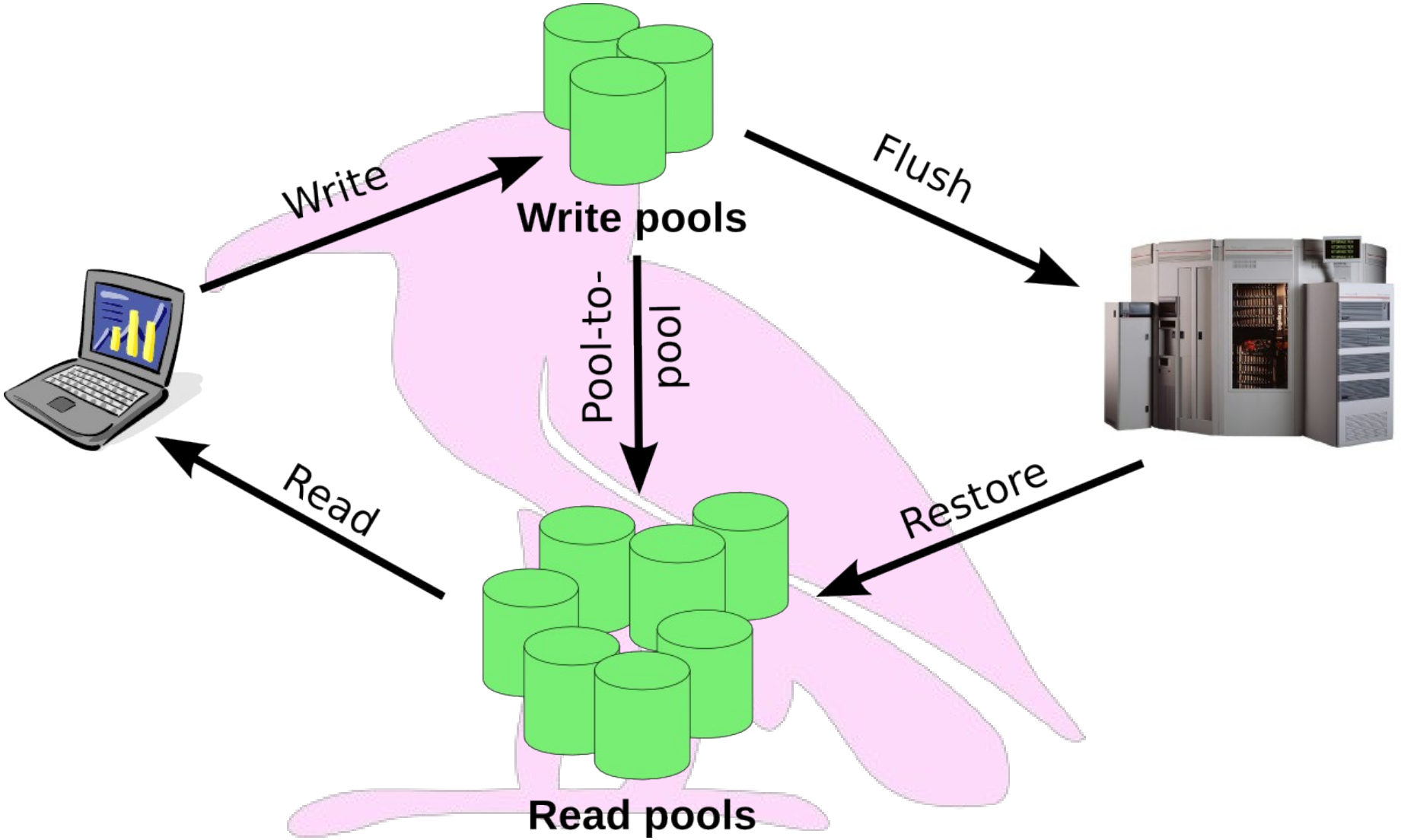
HSM: read



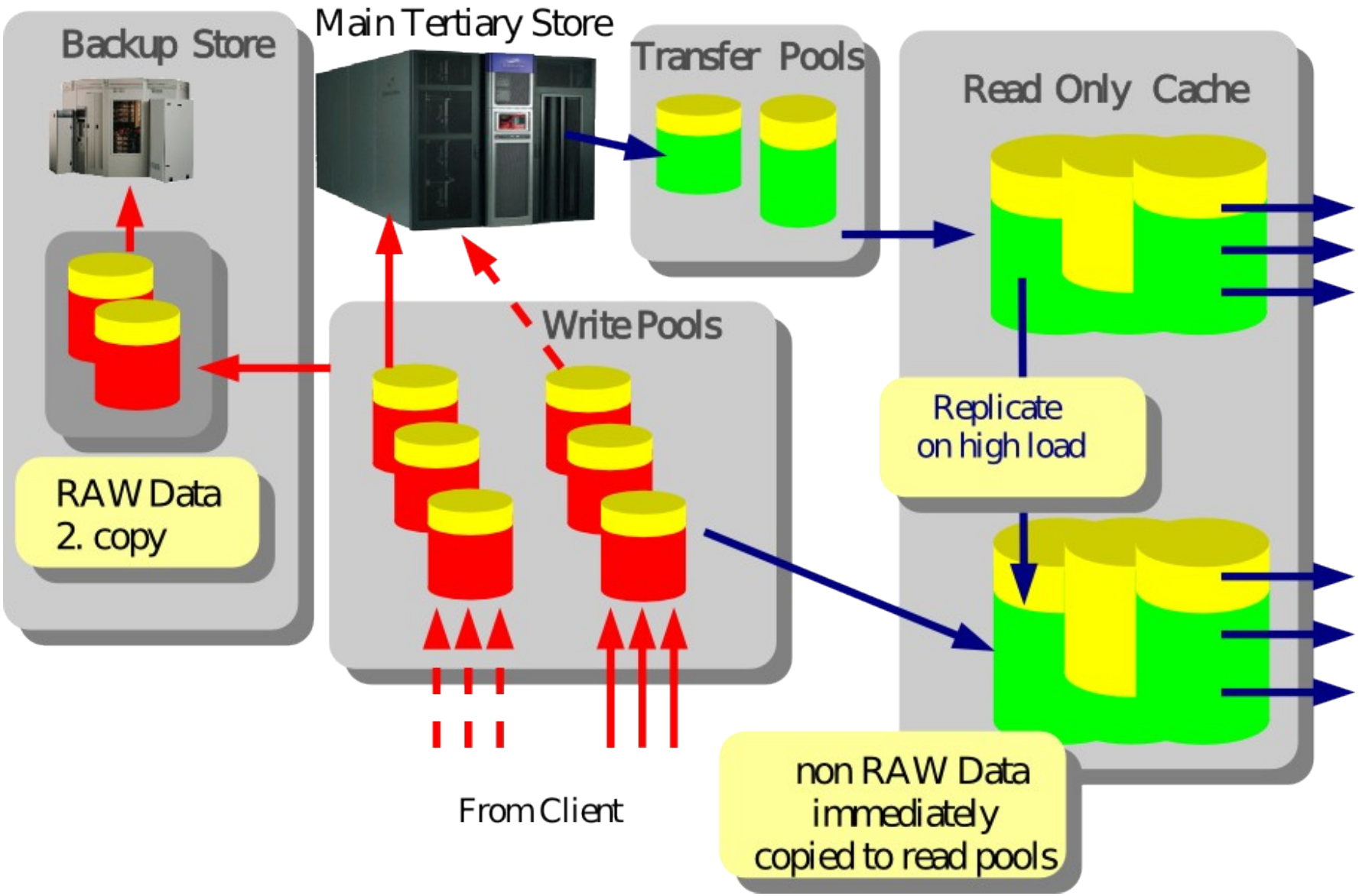
Cached files are transferred from the pool.

Otherwise first restore from HSM to create a cached file.

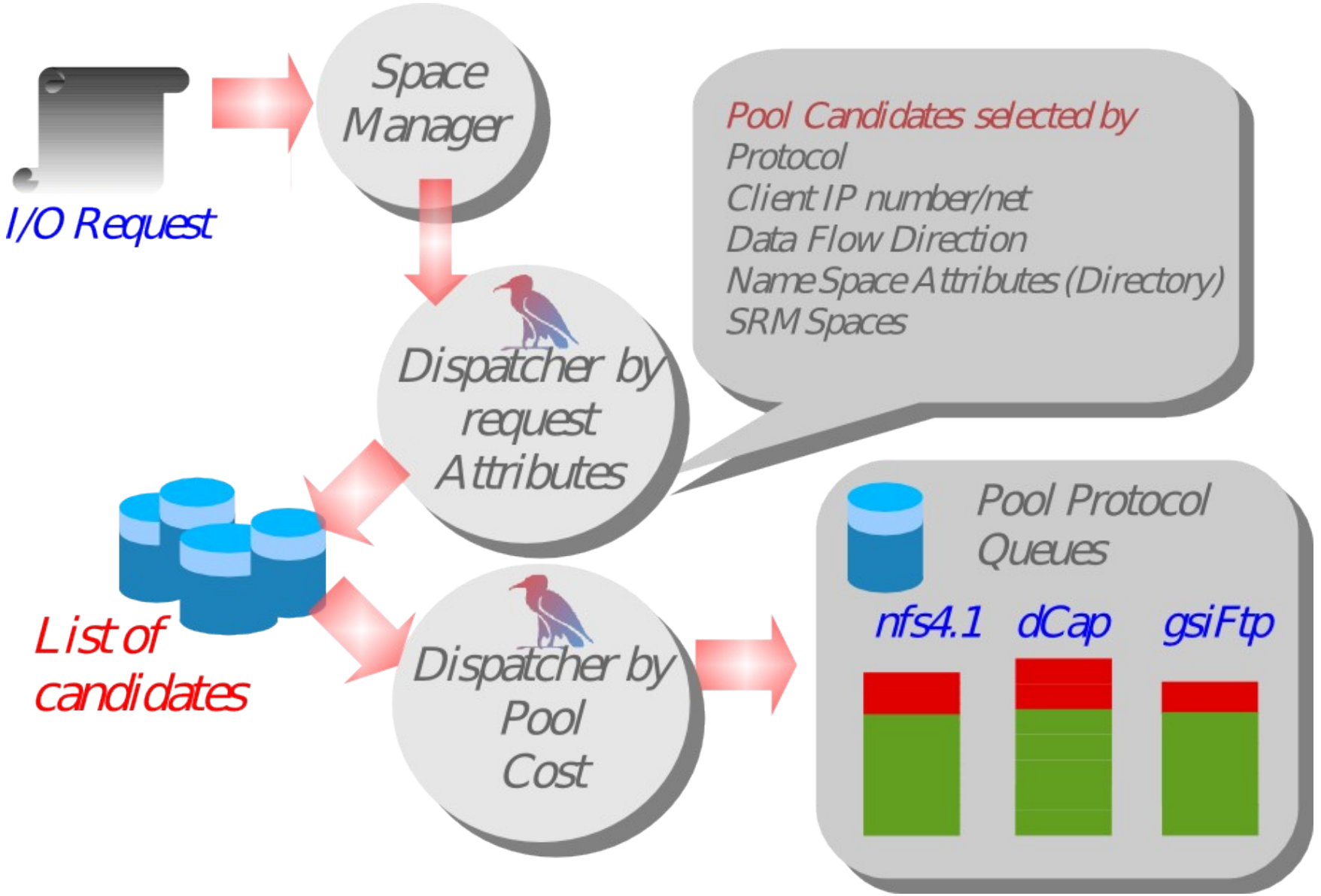
HSM: dedicated pools



File-hopping and replication

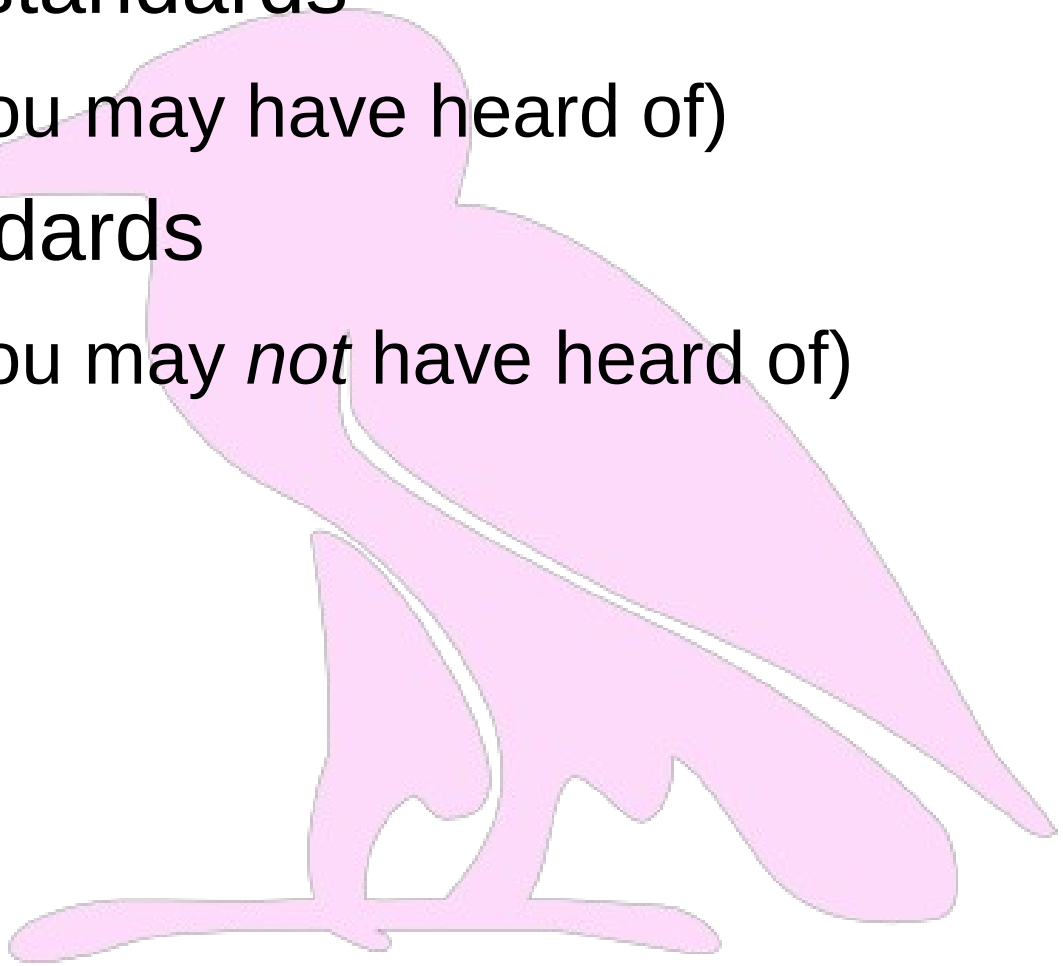


Meltdown/Overload protection



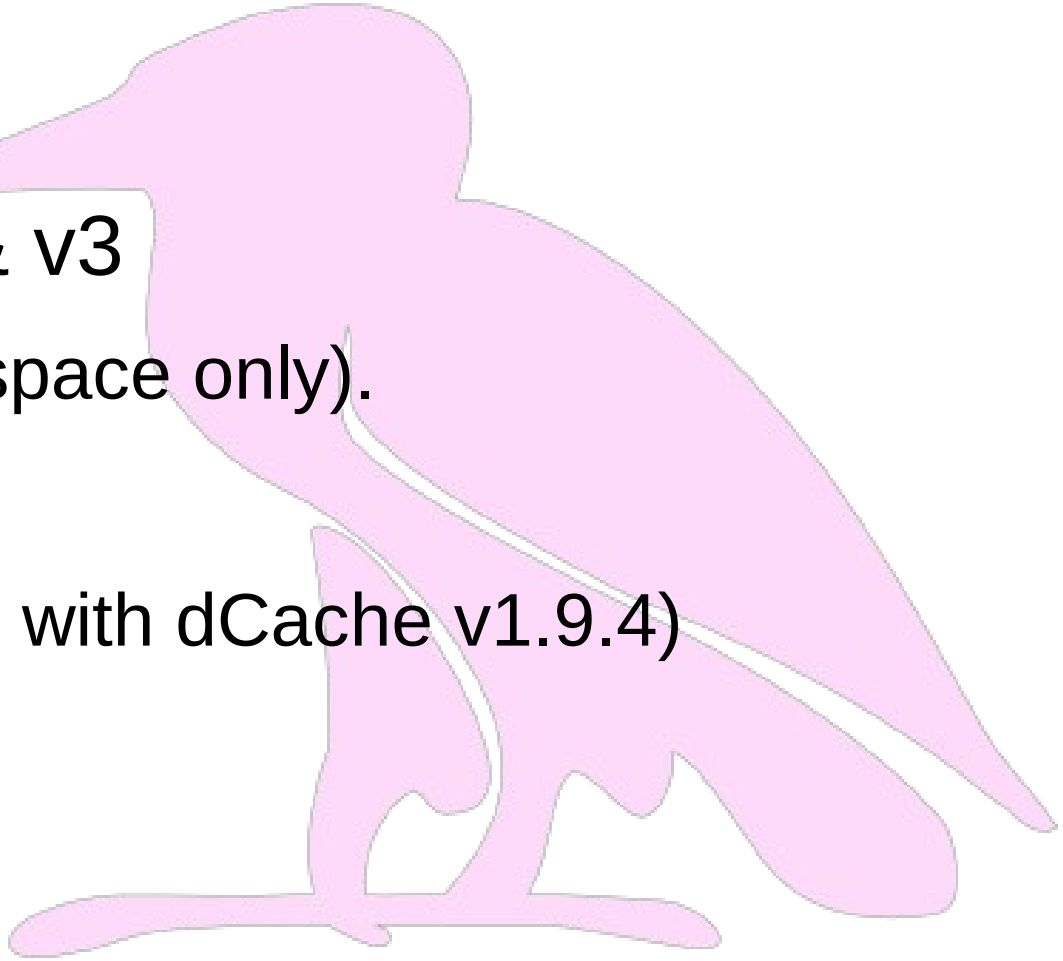
Standards

- Industry standards
  - (ones you may have heard of)
- Grid standards
  - (ones you may *not* have heard of)



Industry standards

- **FTP**,
- **HTTP**,
- **NFS** v2 & v3
  - (name-space only).
- **NFS** v4.1
  - (coming with dCache v1.9.4)



NFS v4.1

- Amongst other benefits, NFS 4.1 provides support for redirecting clients to disk pools
- dCache has been involved in “bakeathons” for many years
  - Well tested against Solaris- and Linux- kernel clients.
  - dCache works (with a few clauses)
    - No striping, no updates --- clients are happy with this.
- Anticipate support for NFS v4.1 coming to Linux kernel Q1 2009.

Grid standards

- “Community” standards:
  - **GLUE** v1.3 (WLCG)
  - **syncat** (dCache initiative)
  - **xrootd** (ROOT/WLCG)
  - **VOMS**
- OGF standards:
  - **GLUE** v2.0
  - **GridFTP** v2
  - **SRM** (v1.1, v2.2)

The Standards marked green are where dCache has had strong involvement.

SRM

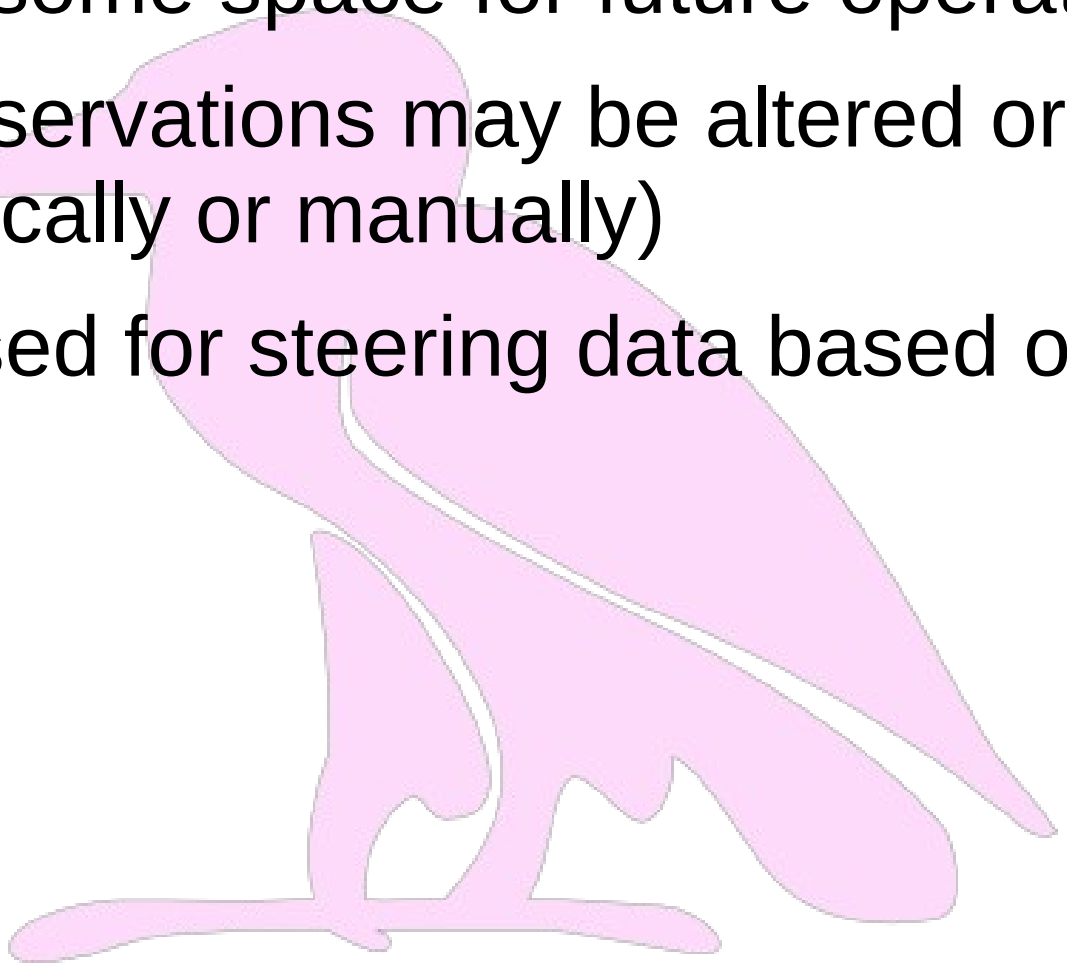
- Data transfer
- File pinning / unpinning
- Space management (reservations)
- Directory operations (e.g., ls)
- Permission management

SRM: storage attributes

- Two properties: Access Latency and Retention Policy
  - Access Latency is: **online**, **nearline**, (**offline**)
  - Retention Policy is: **replica**, (**output**), **custodial**.
- WLCG understand this as:
  - Replica/Online => D1T0 (stored on disk)
  - Custodial/Online => D1T1 (disk & tape)
  - Custodial/Nearline => D0T1 (tape only)

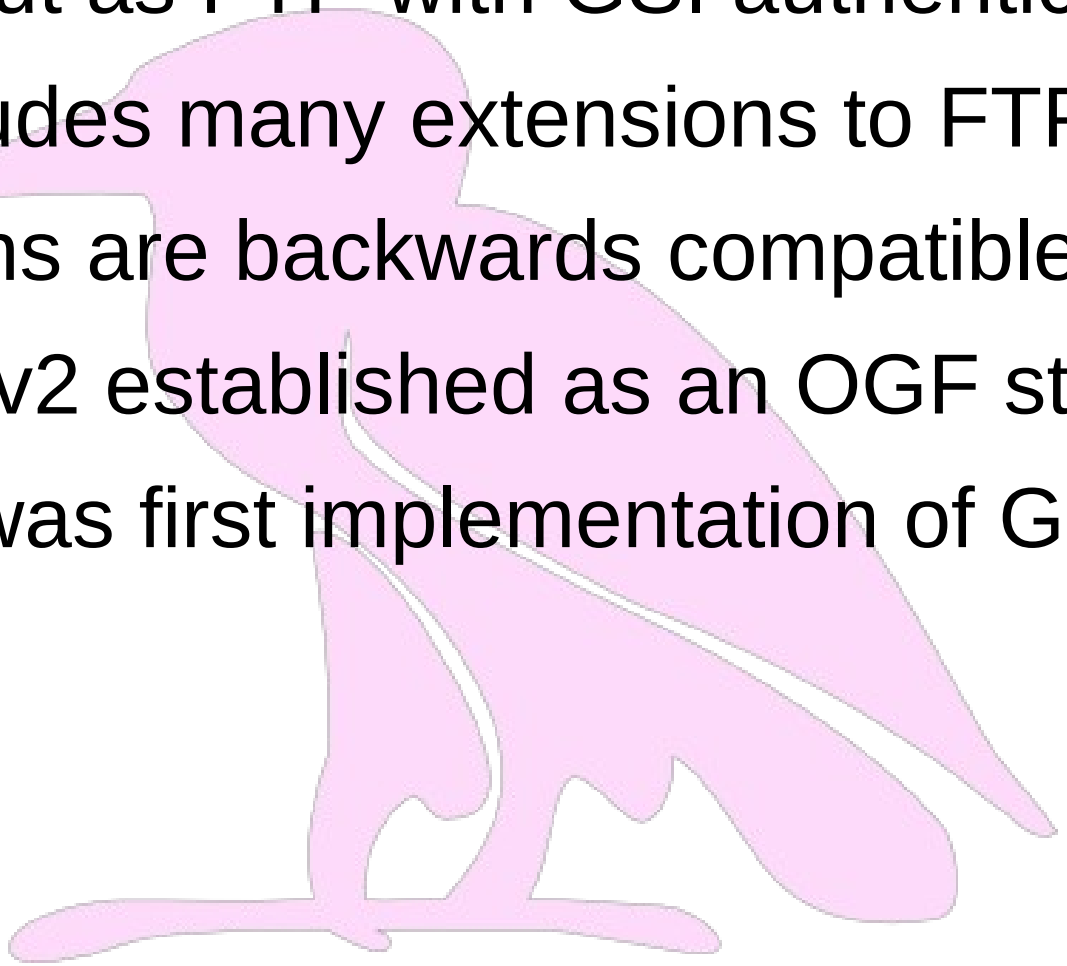
SRM: space reservations

- Reserve some space for future operations
- Space reservations may be altered or released (automatically or manually)
- Mostly used for steering data based on storage attributes



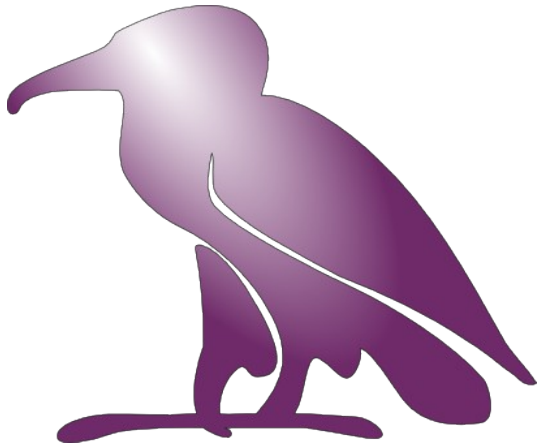
Grid-FTP

- Started out as FTP with GSI authentication.
- Now includes many extensions to FTP
- Extensions are backwards compatible
- GridFTP v2 established as an OGF standard
- dCache was first implementation of GridFTP 2



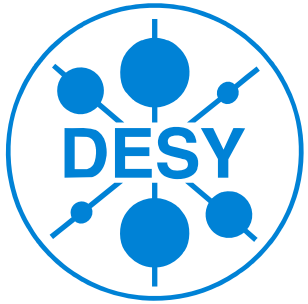
We, dCache.org, ...

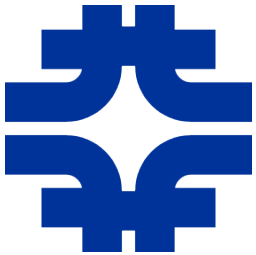
- Team structure
- How we work,
- How we're funded.



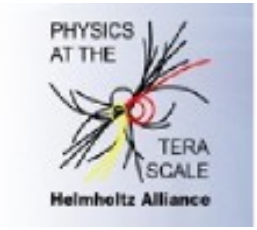
We, dCache.org, ...

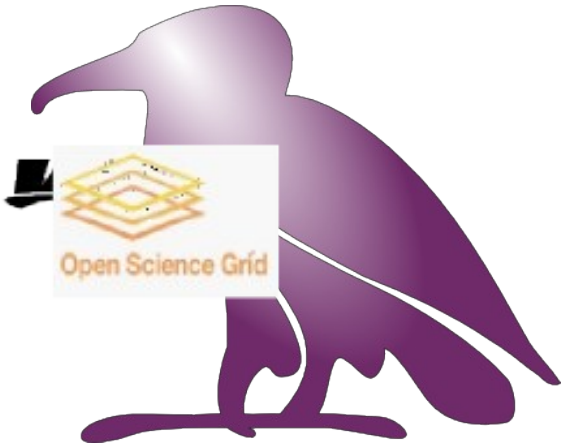
Main development team:





Additional funding, support or contributions from:





Members of the team

Head of dCache.org
Patrick Furhmann

Core team (DESY, Fermi, NDGF)
Tatjana Baranova
Andrew Baranovski
Bjoern Boettscher
Ted Hesselroth
Alex Kulyavtsev
Iryna Koslova
Tanya Levshina
Dmitri Litvintsev
David Melkumyan
Paul Millar
Owen Synge
Neha Sharma
Vladimir Podstavkov

Head of development (DESY)
Tigran Mkrtchyan

Head of development (Fermi)
Timur Perelmutov

Head of development (NDGF)
Gerd Behrmann

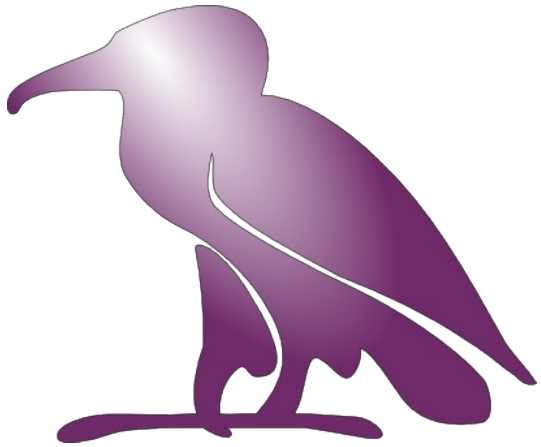
External

Development
Abhishek Singh Rana, SDSC
Jonathan Schaeffer, IN2P3

Support
German HGF Support Team
Greig Cowan, GridPP
Flavia Donno, CERN

Example deployments

- Fermi National Laboratory,
- NDGF.



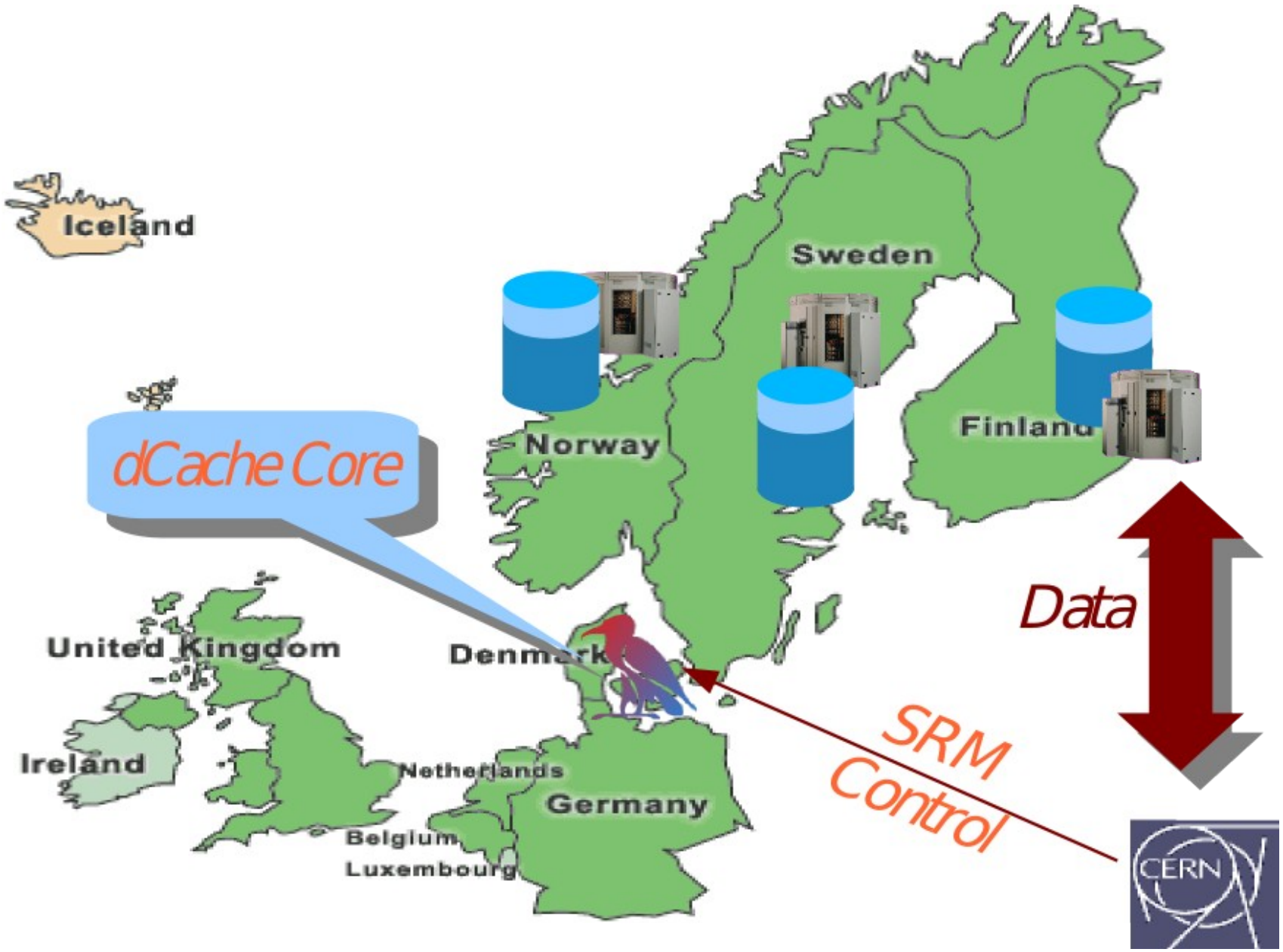
Fermi CMS Tier-1

- As of January 2009 Fermi has
  - 690 pools (distributed over 140 nodes),
  - 3.3 PB online disk storage,
    - Planning on expanding to 4.8 PB in a month or so
  - 5 PB stored on tape,
  - LAN access throughput can go up to 20 GiB/s,
  - WAN 1.2GiB/s typical (peak: 2.5GiB/s),
  - 60 TiB/day read (100,000 files!)
  - 2 TiB/day write (8,000 files)
- Fermi also run three other dCache instances.

NDGF: *really* distributed storage

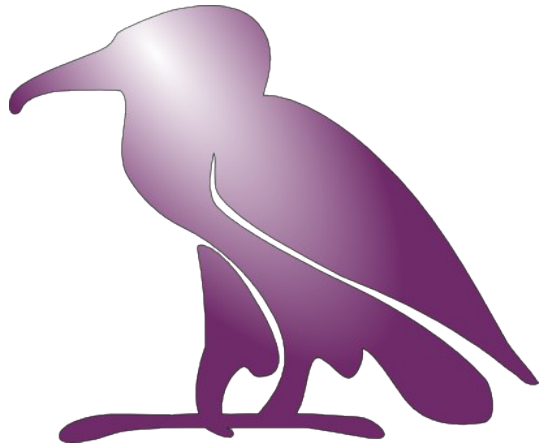
- Storage pools split over four countries (Norway, Sweden, Finland, Denmark).
- A single dCache instance
- HSM storage in each country
  - Pools in a country can see their local HSM only.
- Storing raw data onto tape should be resilient against any country going down.
- Legal requirements mean cannot be black-box deployment: local admins must do admin work.

Distributed Tier-1



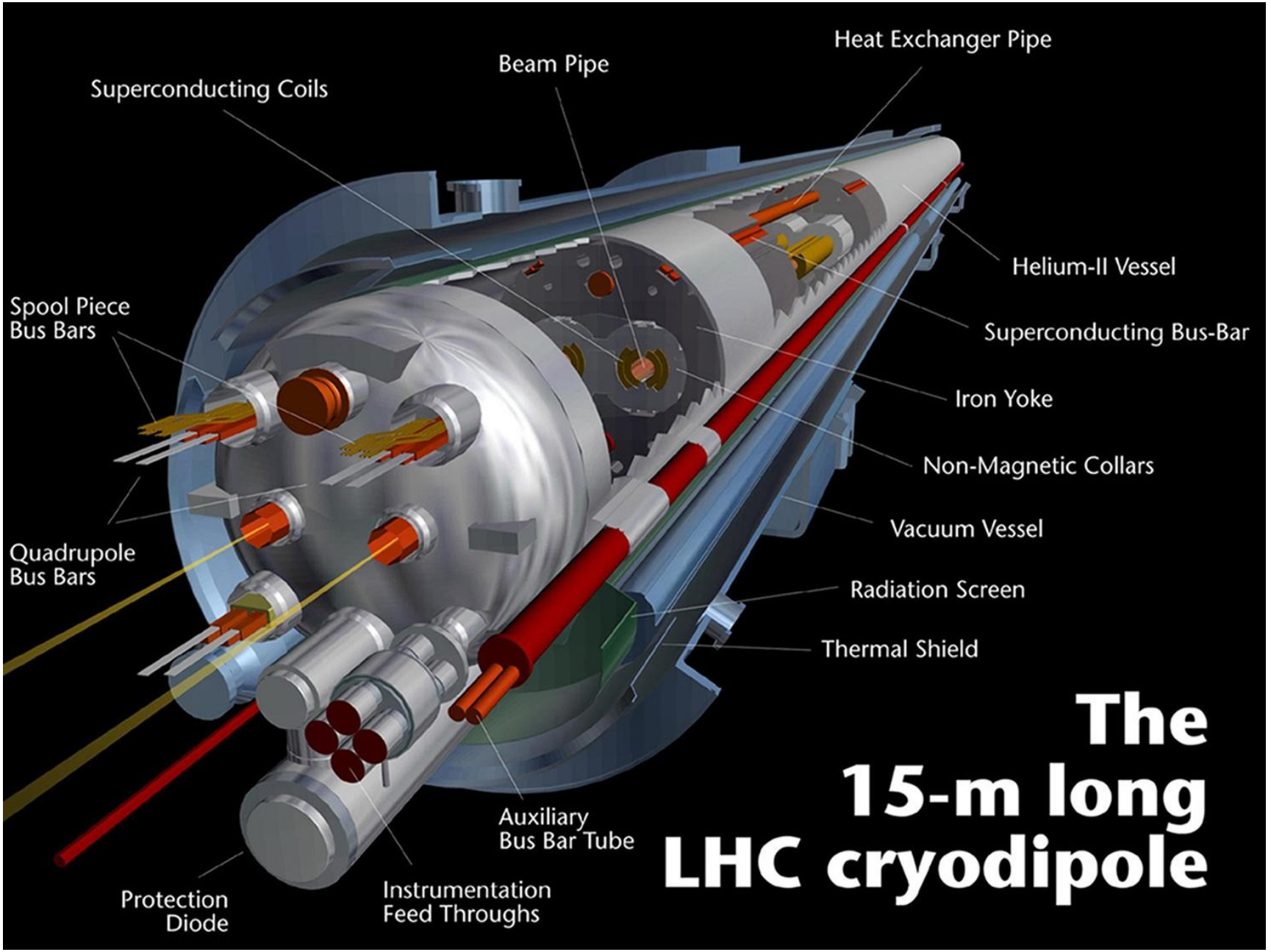
dCache in a Grid context

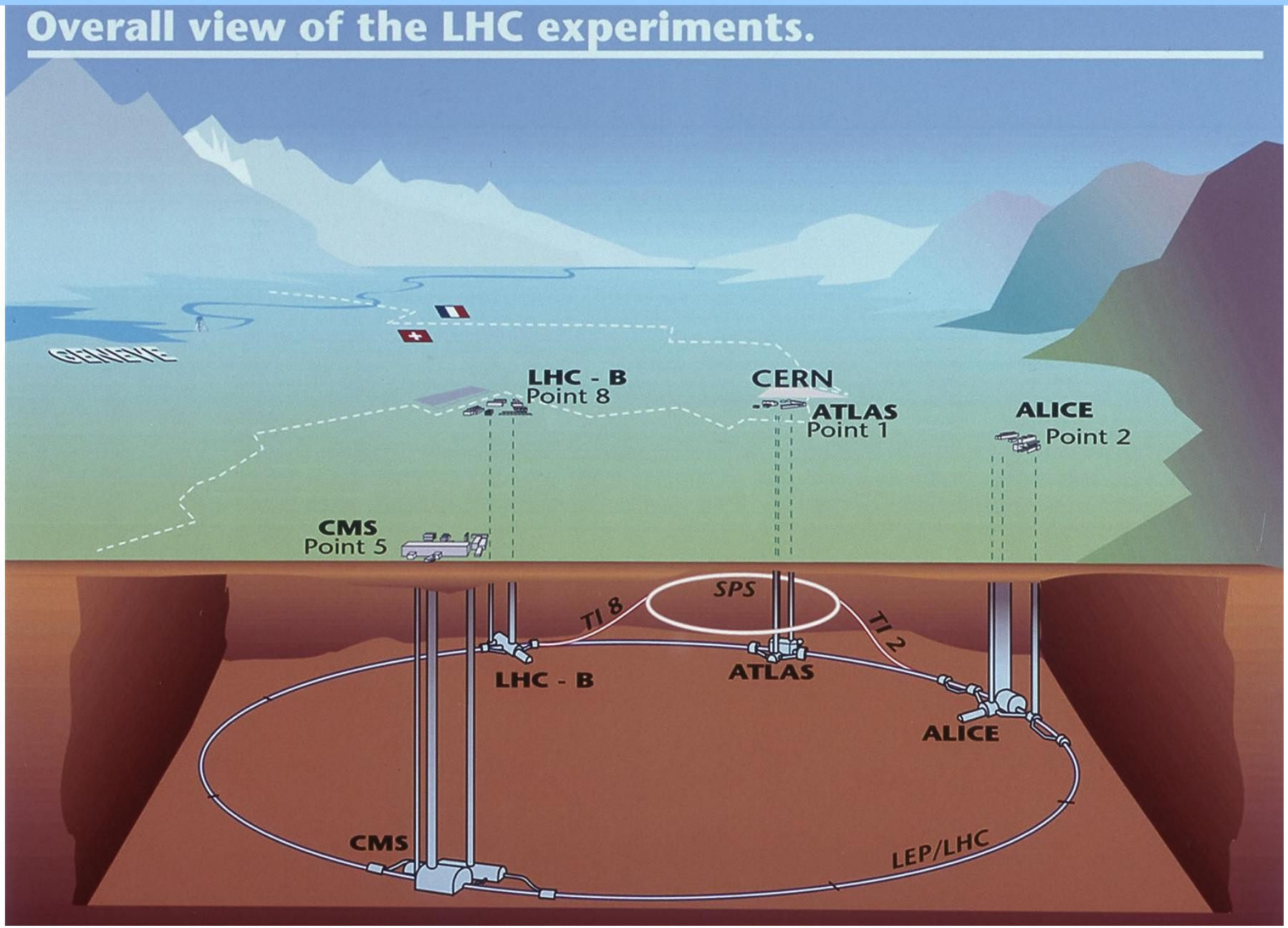
- The LHC,
- Experiments,
- (typical) HEP workflow,
- WLCG: a HEP Grid,
- How institutes contribute to WLCG?
- Workflow revisited



The LHC facility

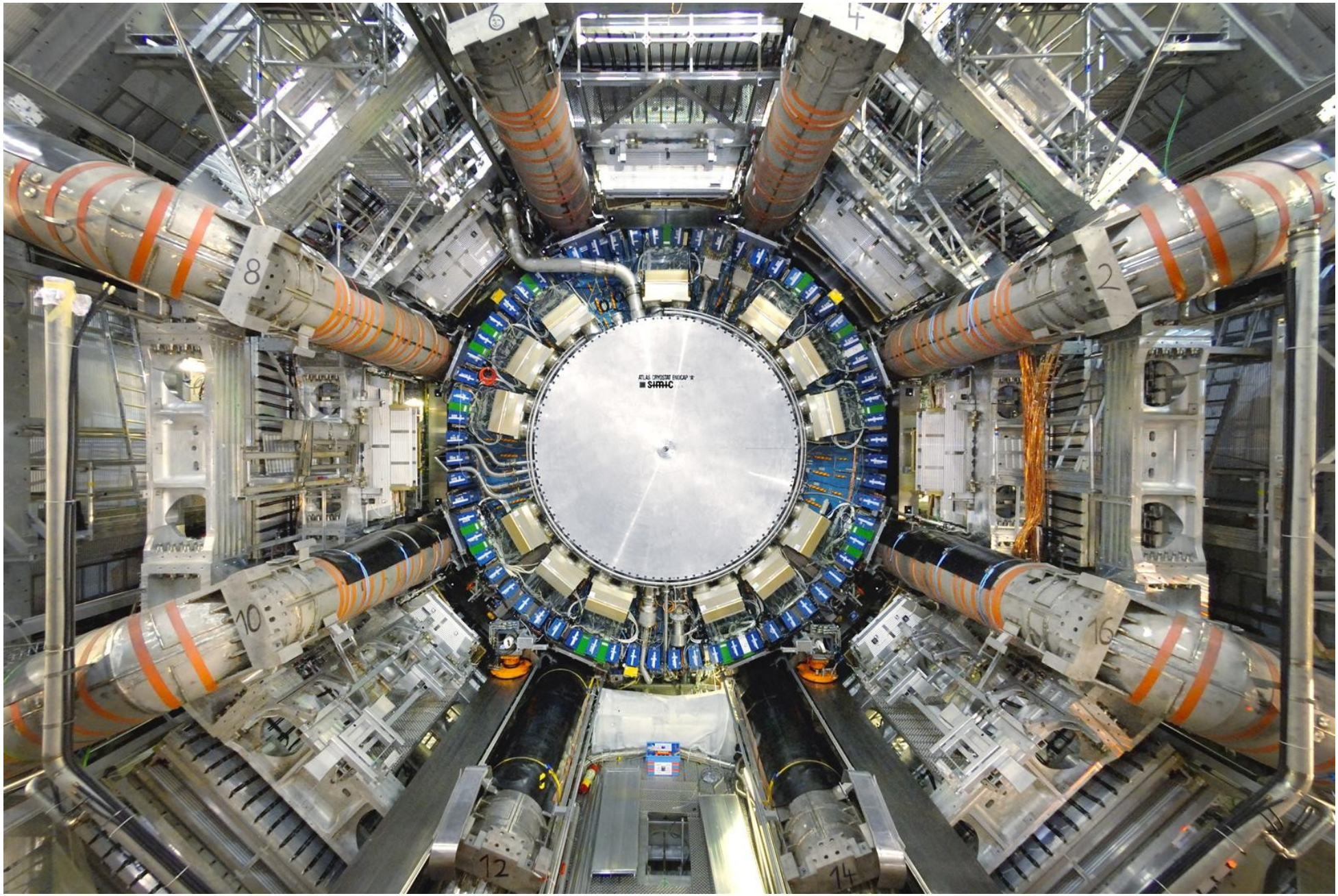
Dipole magnet



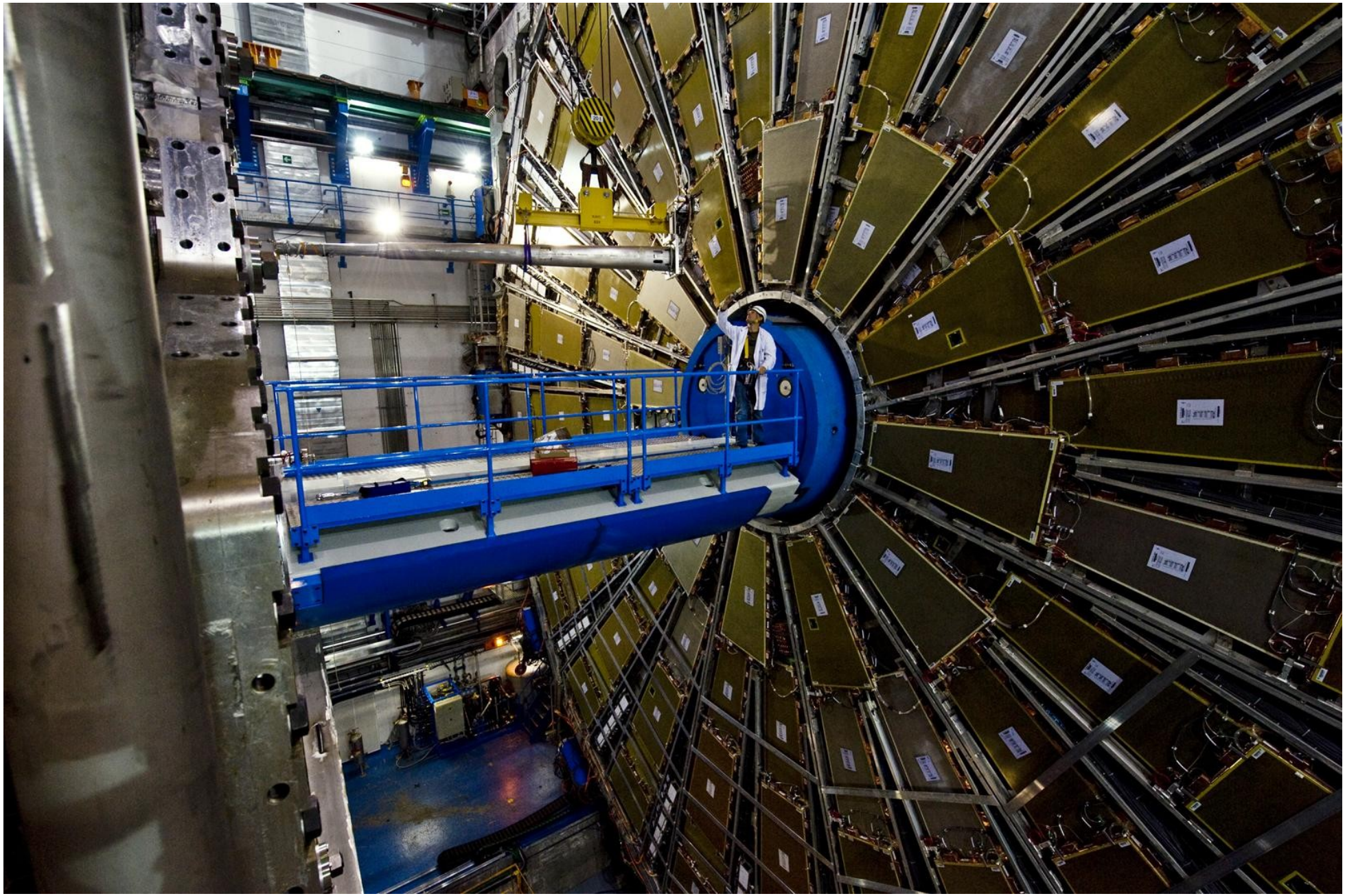

Overall view of the LHC experiments.
GENEVE
LHC - B
Point 8
CERN
ATLAS
Point 1
ALICE
Point 2
CMS
Point 5
SPS
TI 8
TI 2
LHC - B
ATLAS
ALICE
CMS
LEP/LHC

Digging experiment cavern

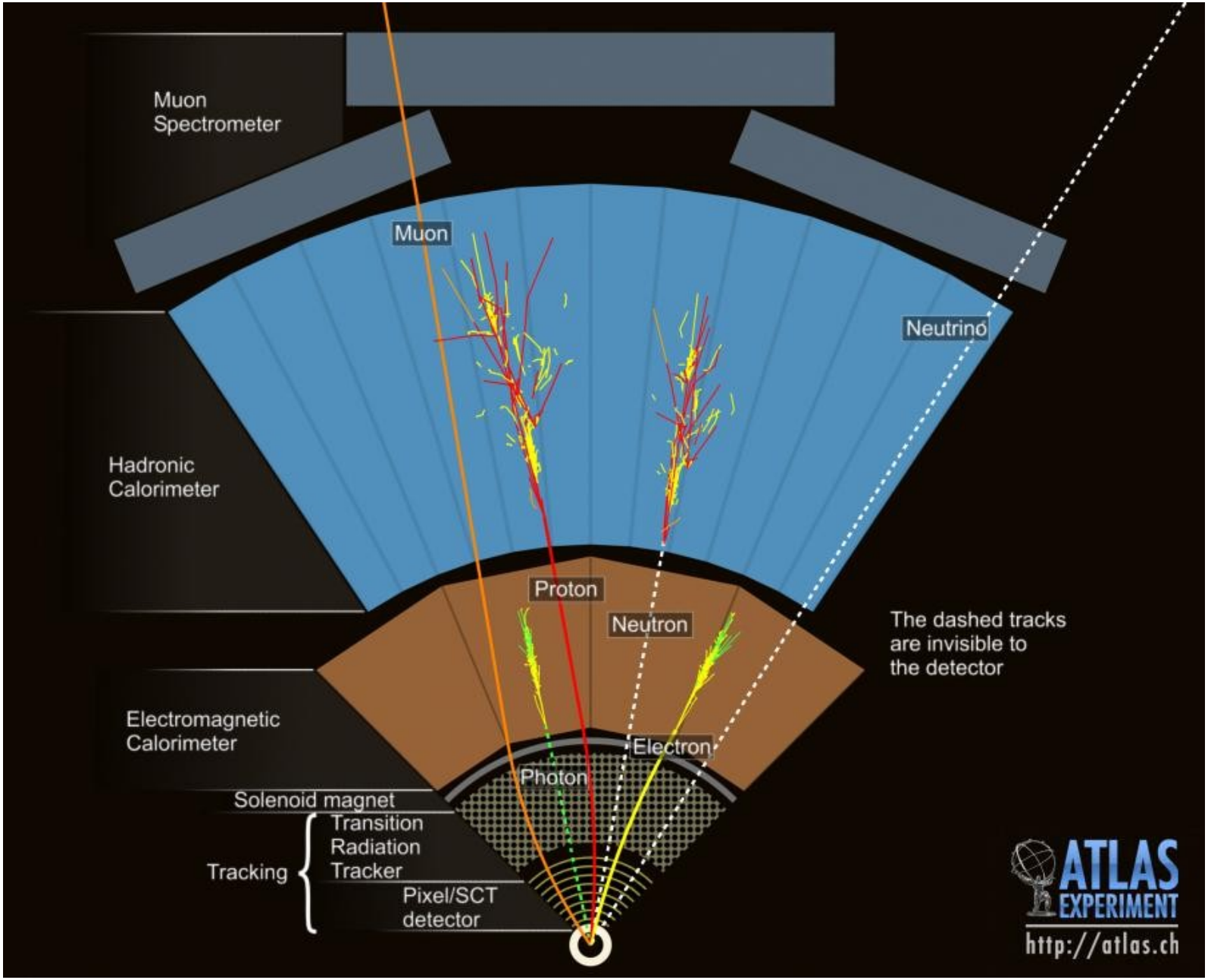
The ATLAS detector 2007



ATLAS beam pipe being fitted

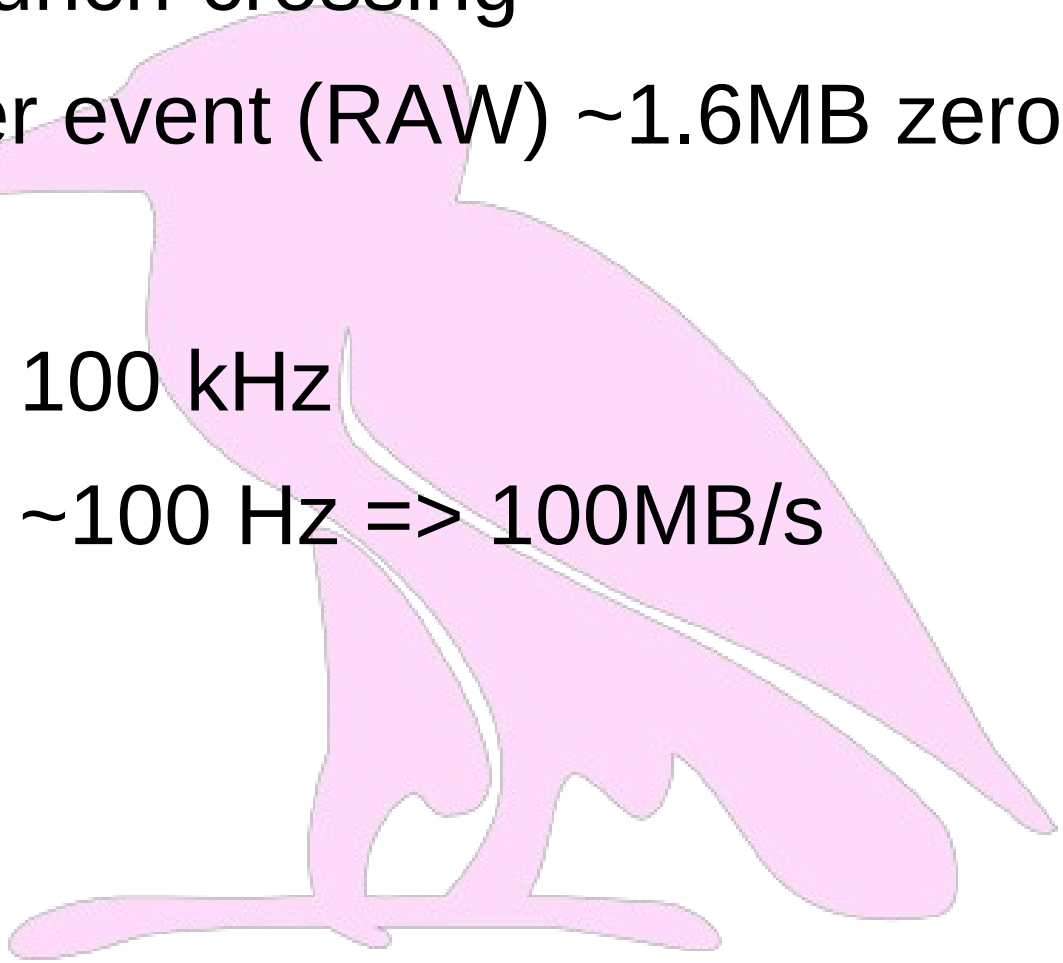


Different particles interacting with sub-detectors

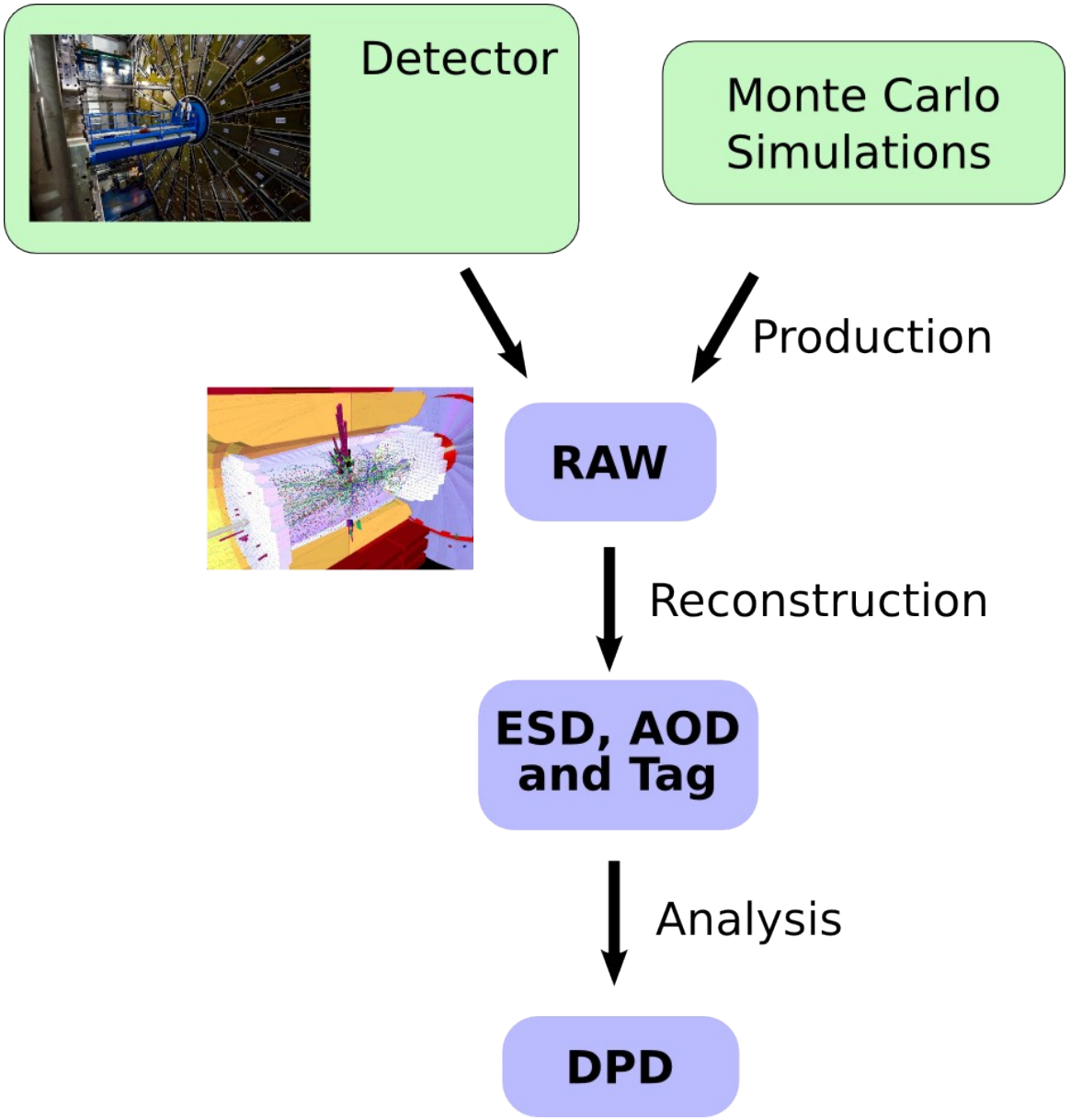


Experiments

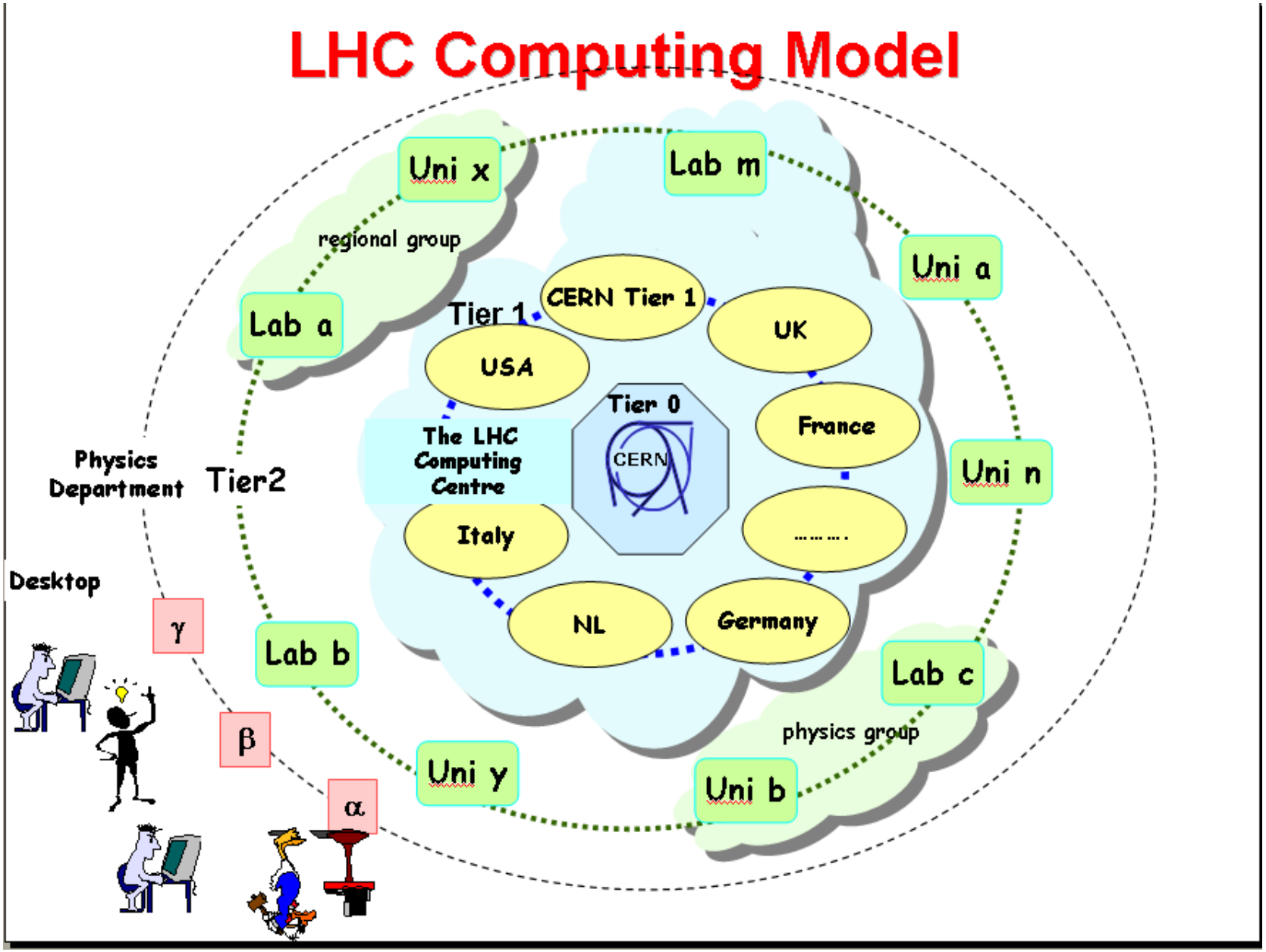
- 40MHz bunch-crossing
- 25 MB per event (RAW) ~1.6MB zero-suppress
- 1 PB/s
- Lvl-1 trig. 100 kHz
- Lvl-3 trig. ~100 Hz => 100MB/s



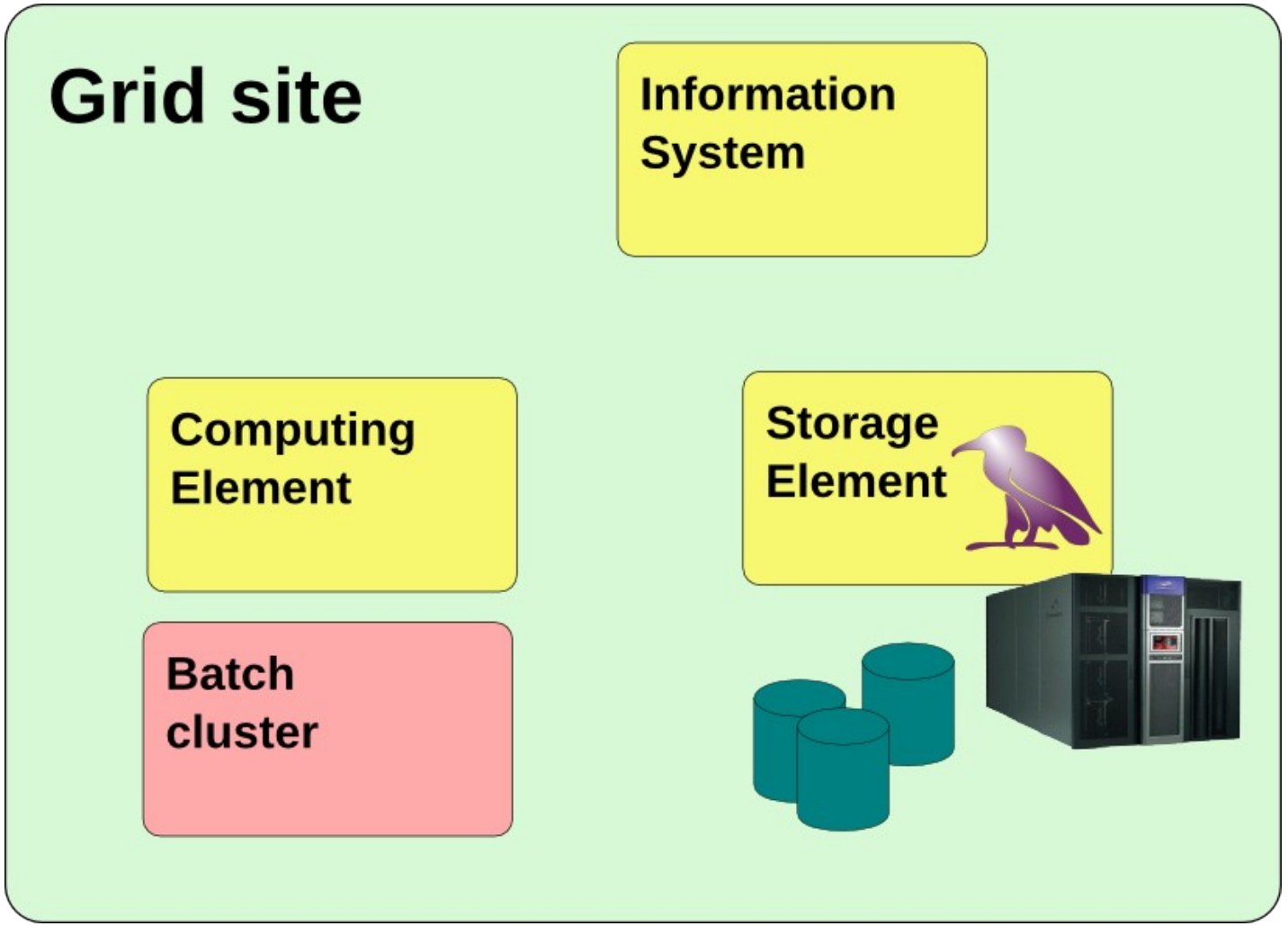
Typical HEP workflow



WLCG: a HEP Grid



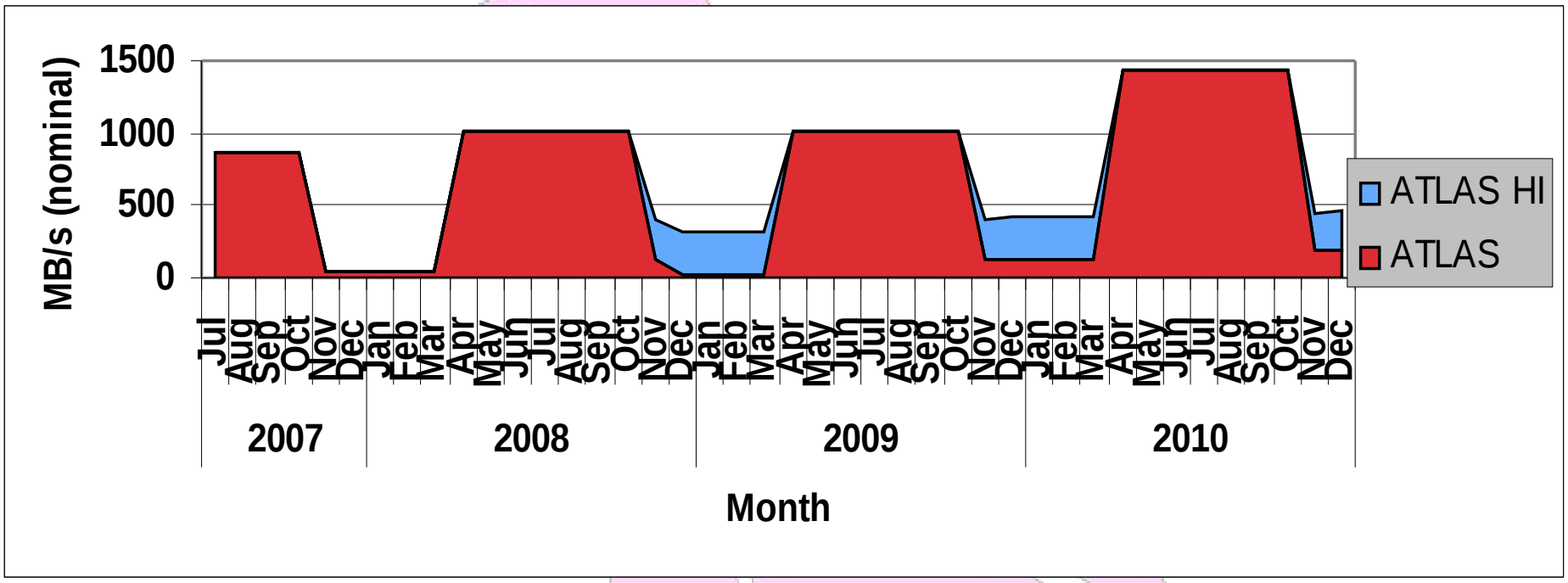
How institutes contribute



Workflow revisited

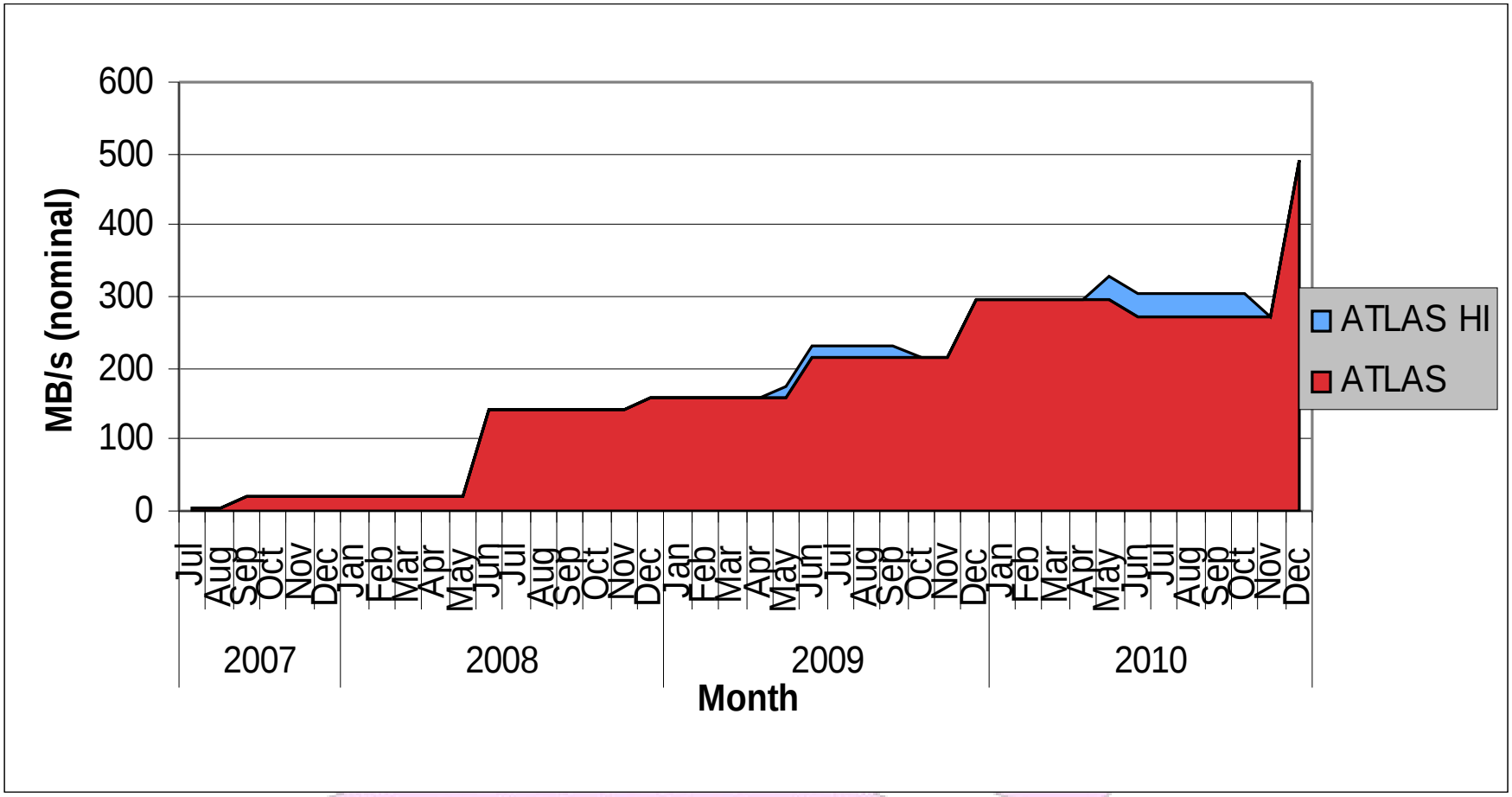
- CERN stores all RAW data and does quick ESD/AOD generation
- Tier-1s take some part of total RAW data
  - archive data on tape,
  - Reconstruction, based on better calibration
    - T1 to T1 distribution.
- Monte-carlo and Analysis happens on Tier-2s, results are:
  - stored locally
  - then sent to Tier-1 for archiving.

Predicted T1-CERN bandwidth

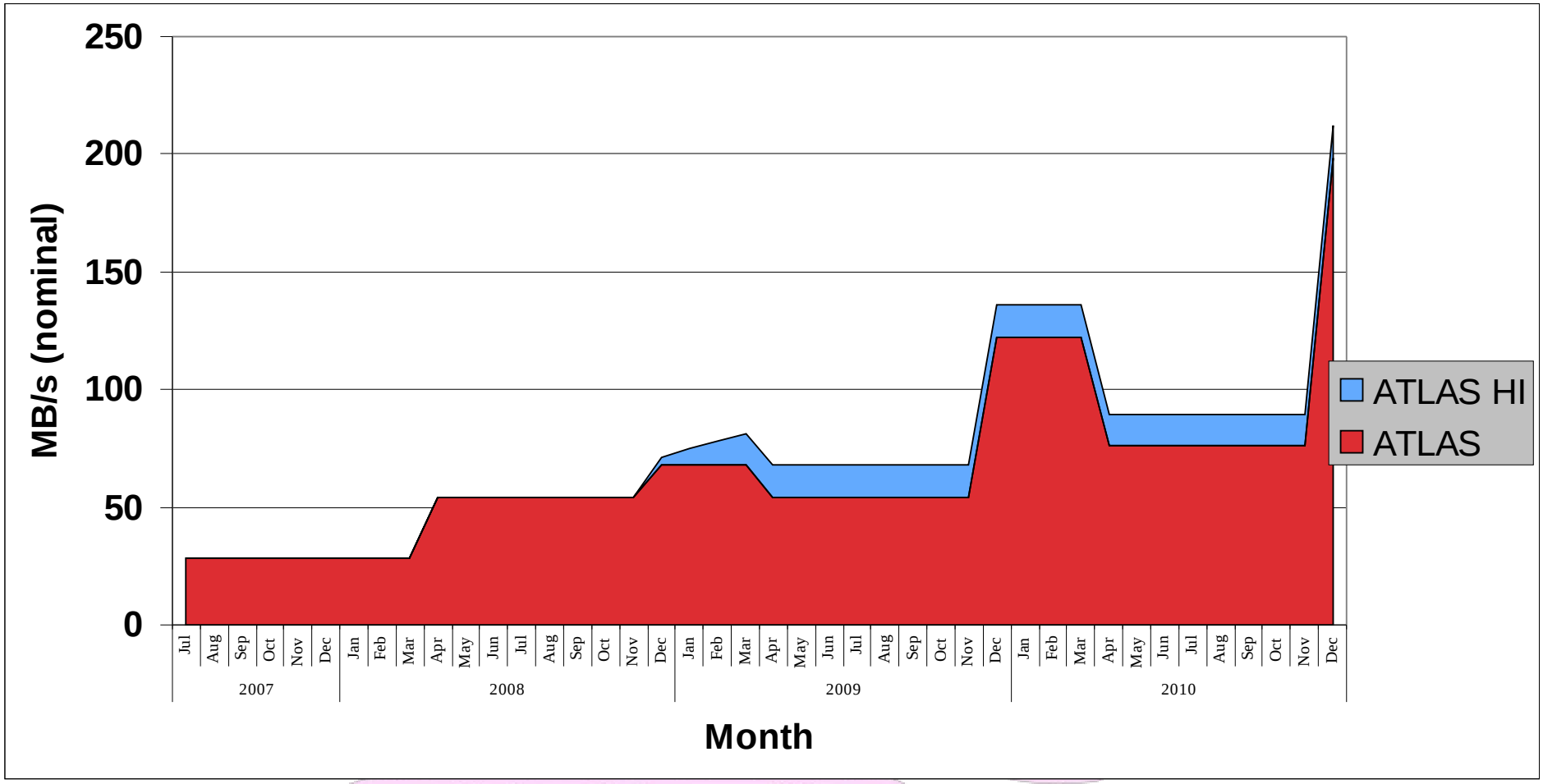


Data from Roger Jone's CHEP 2006

Predicted T1-T1 bandwidth



Predicted T1-T2 bandwidth

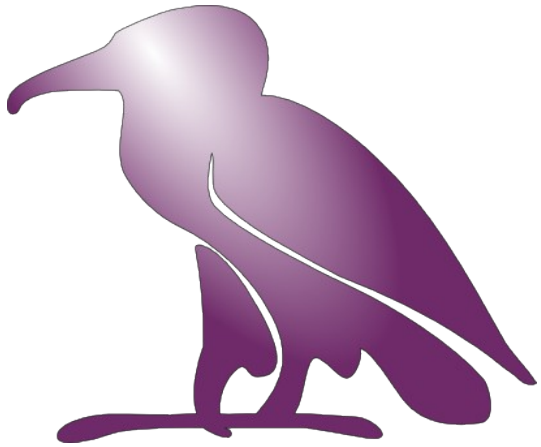


WLCG usage

- Majority of data outside CERN is stored using dCache.
- 8 (out of 11) Tier-1 centres use dCache and many Tier-2 centres.
- Storage requirements (all four experiments):
  - 2008: 40 PB
  - 2009: 80 PB
  - 2010: 120 PB
  - 2011: 160 PB

Crystal ball

- Clouds,
- Support for map-reduce,
- Increasing use of HTTP,
- Support for peer-to-peer,
- Clouds



Clouds

- Everything in cloud:
  - Have successfully run dCache using Amazon infrastructure (S3, EC)
  - Allows an Amazon S3 storage to appear as an SE.
- Hiding the cloud:
  - Can treat the cloud storage as an HSM,
  - Integration between multiple clouds,
  - Allows integrated cross-site disk usage

Map-reduce

- High-performance: move job to data.
- Very low requirements to be a dCache pool
- Build a cluster of compute- / storage- hybrid nodes.
- Using a framework, like Hadoop, to control execution.

HTTP

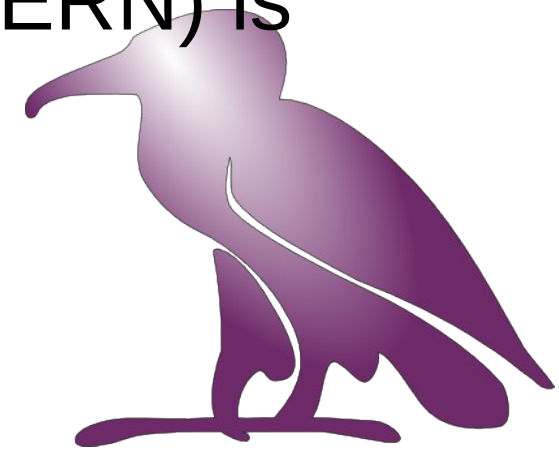
- Improve support for HTTP read
  - Add support for Range: implementing support for vector reads
- Add support for HTTP write
- Add additional functionality
  - Alternative protocol selection,
  - Load-balancing,
  - Secure 3rd party copies,

Better data movement

- Currently use a reliable transfer service:
  - A reliable transfer service schedules requests to optimise throughput.
  - Has a simple model of the network.
- Better approach may be using a peer-to-peer protocol.
- Already have a central catalogue: LFC
- Need LFC modified to be a tracker, dCache modified to be a torrent client/server

Conclusions

- (Grid) Storage software
- Supports HSM backends.
- Fault-tolerant.
- Namespace and data-storage are separated.
- Supports HTTP and (with v1.9.4) NFSv4.1
- Majority of LHC data (outside of CERN) is stored on dCache servers.



The end