Data Storage

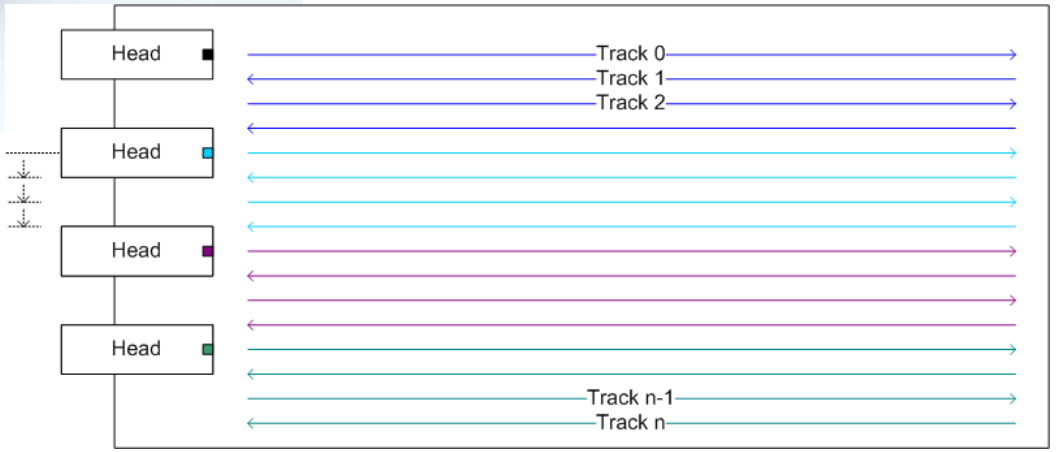
Paul Millar
dCache

Overview

- Introducing storage
- How storage is used
- Challenges and future directions

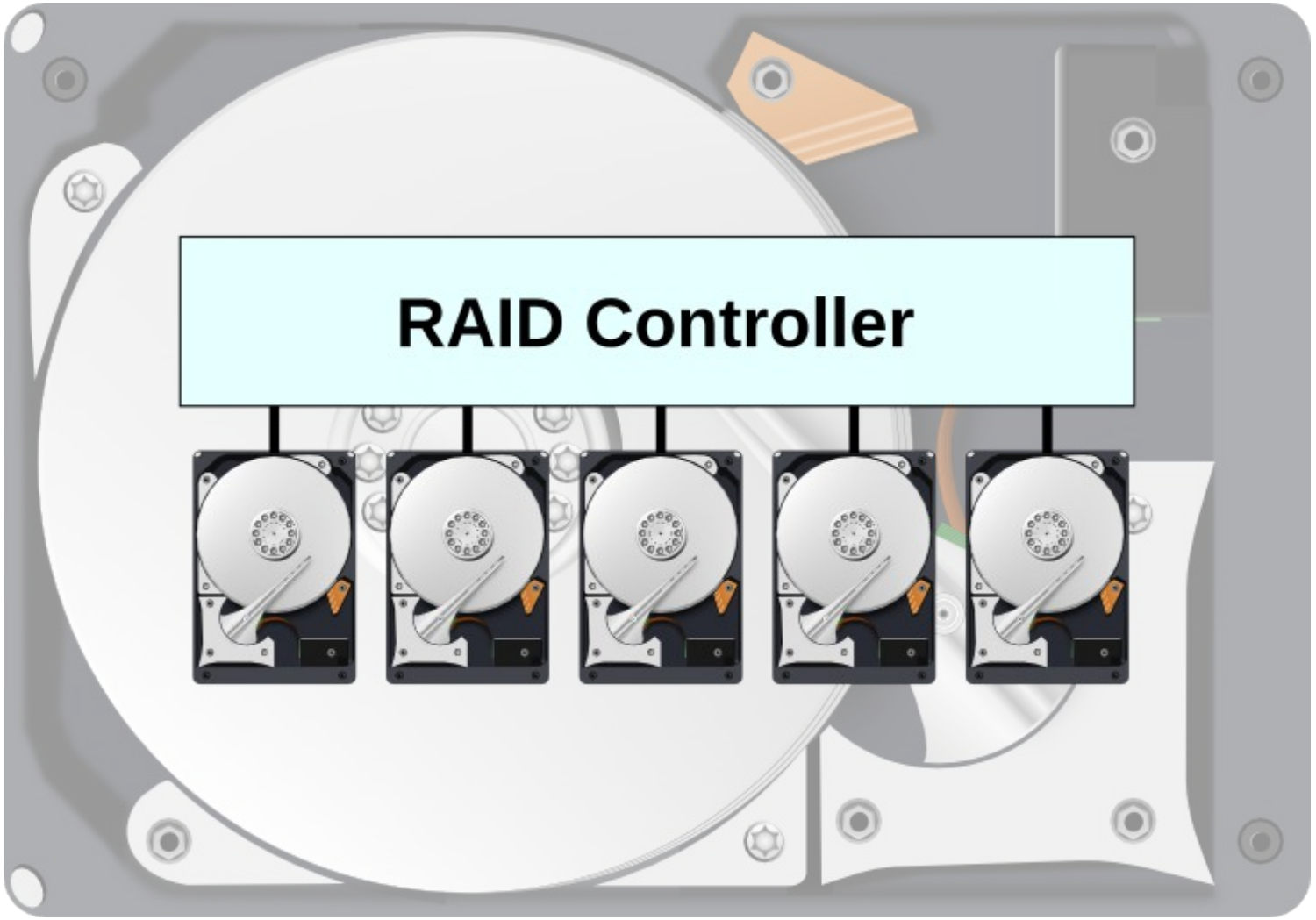
(Magnetic) Hard Disks

Tape systems

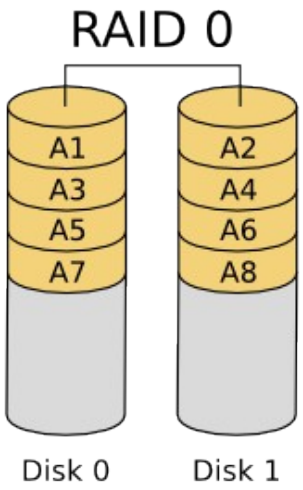


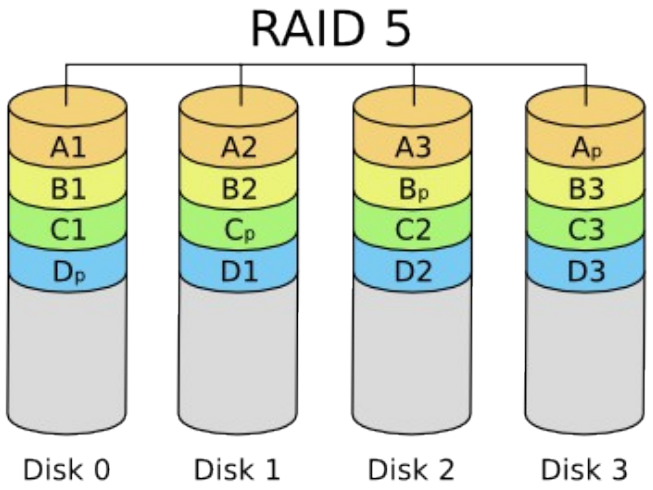
Disk enclosures

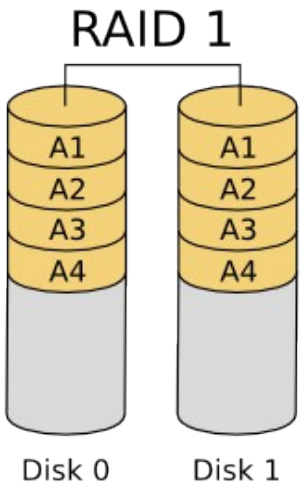
RAID systems

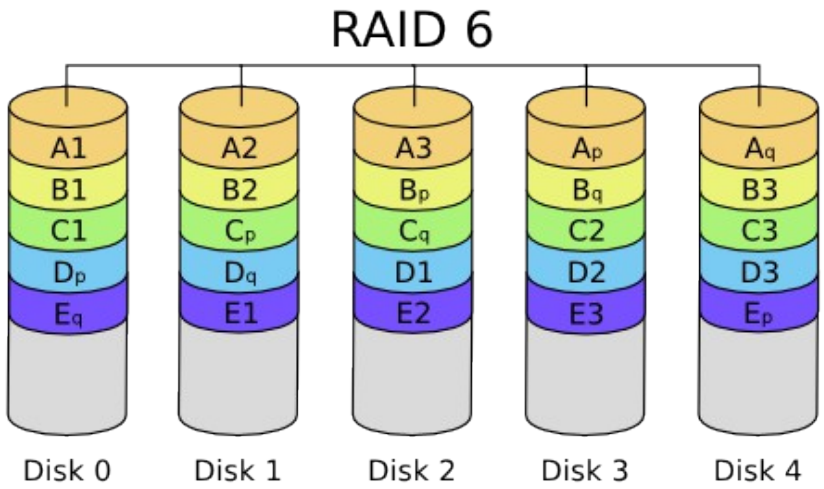


Types of RAID

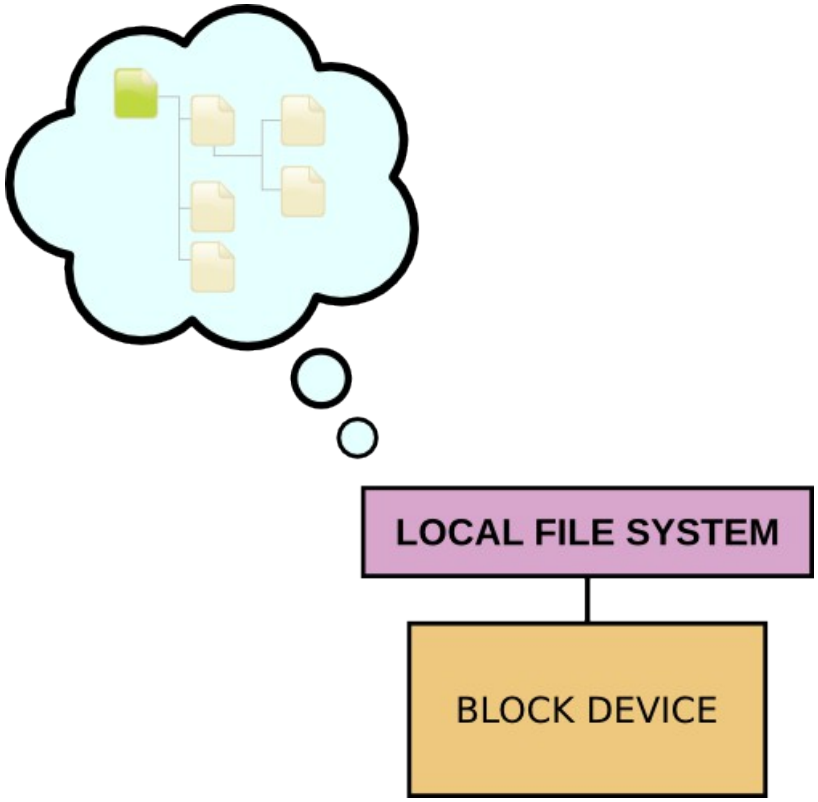






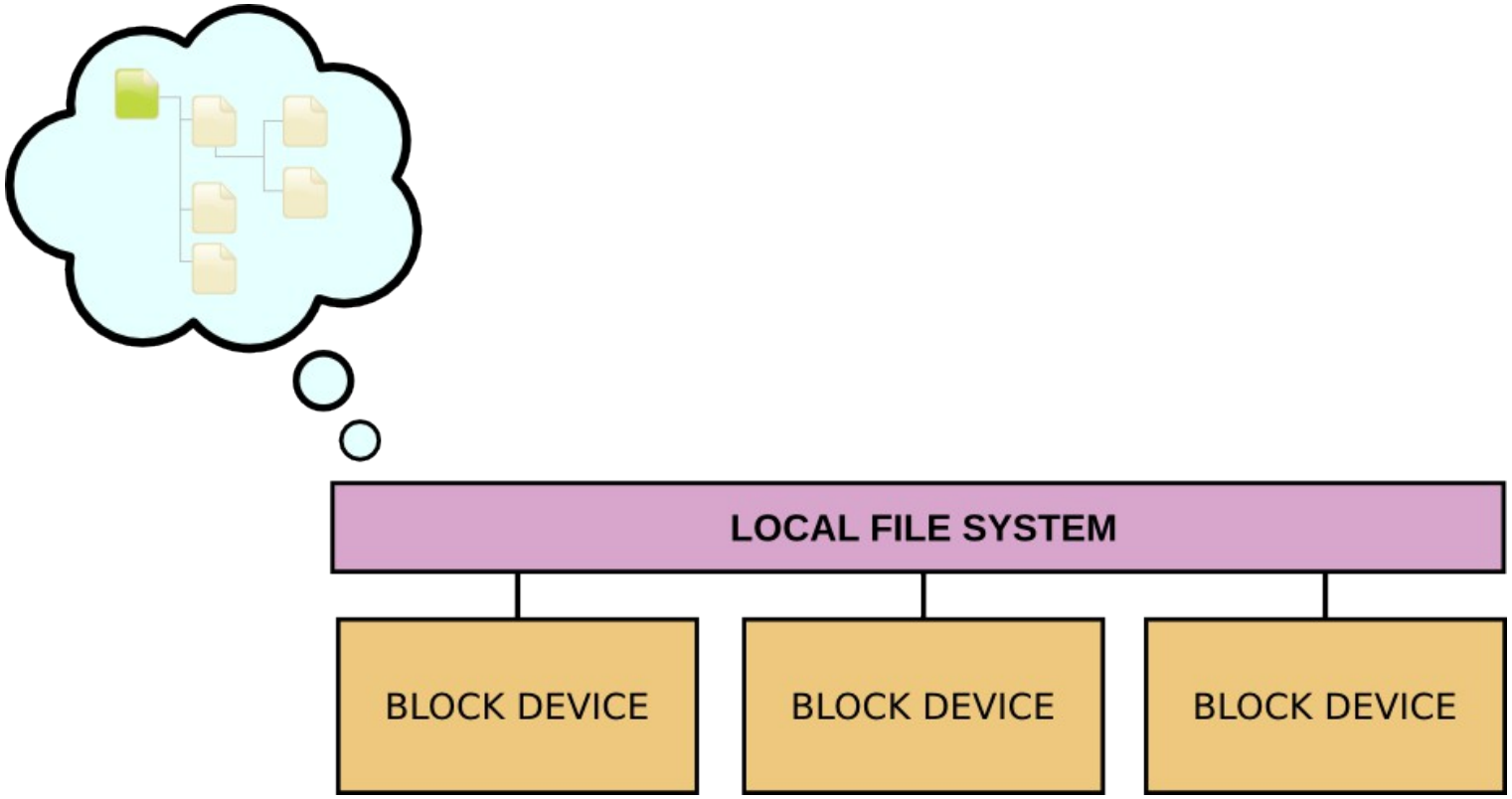


(Local) File systems



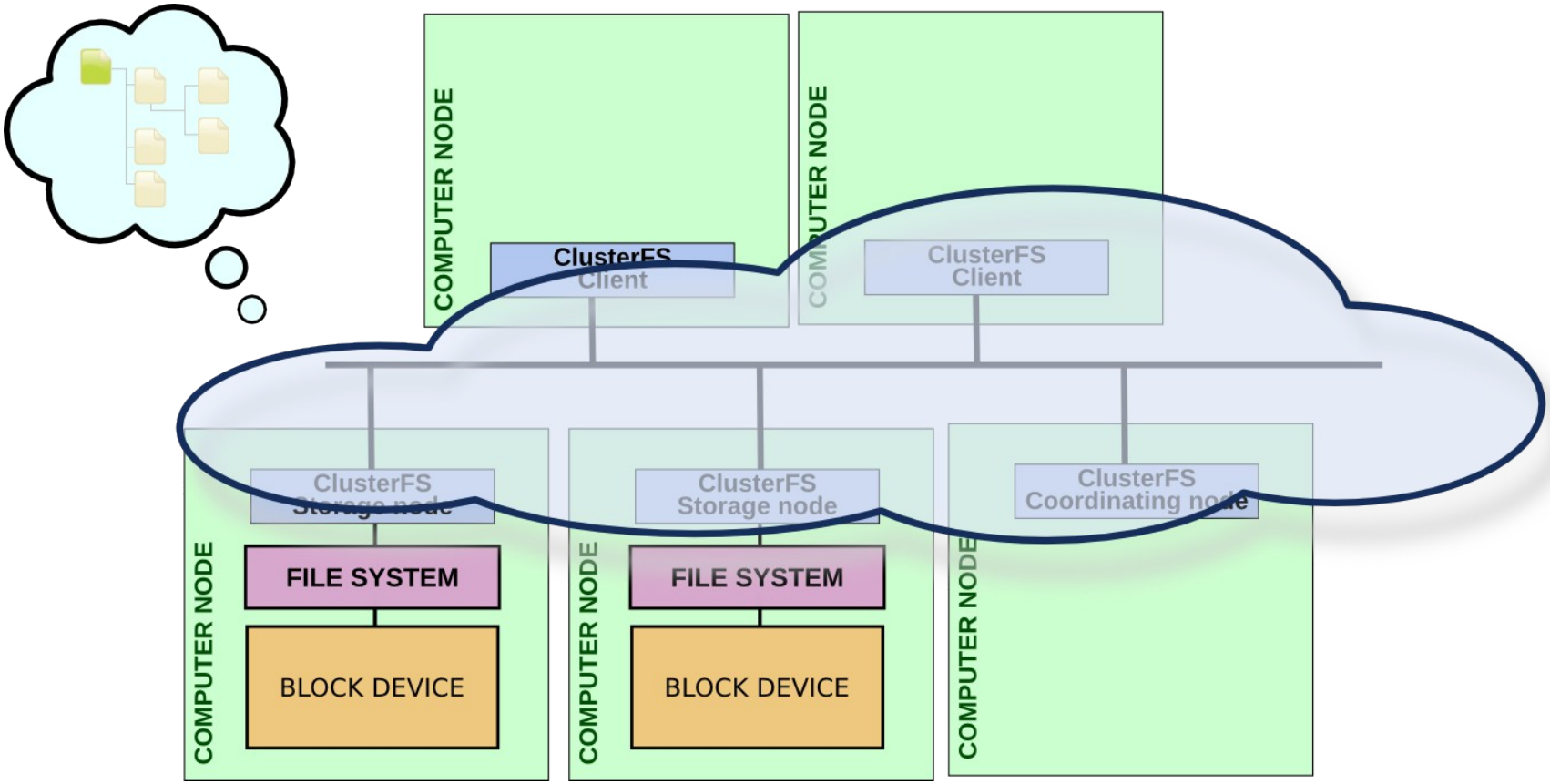
Ext3, Ext4, XFS, ...

(Local) File systems

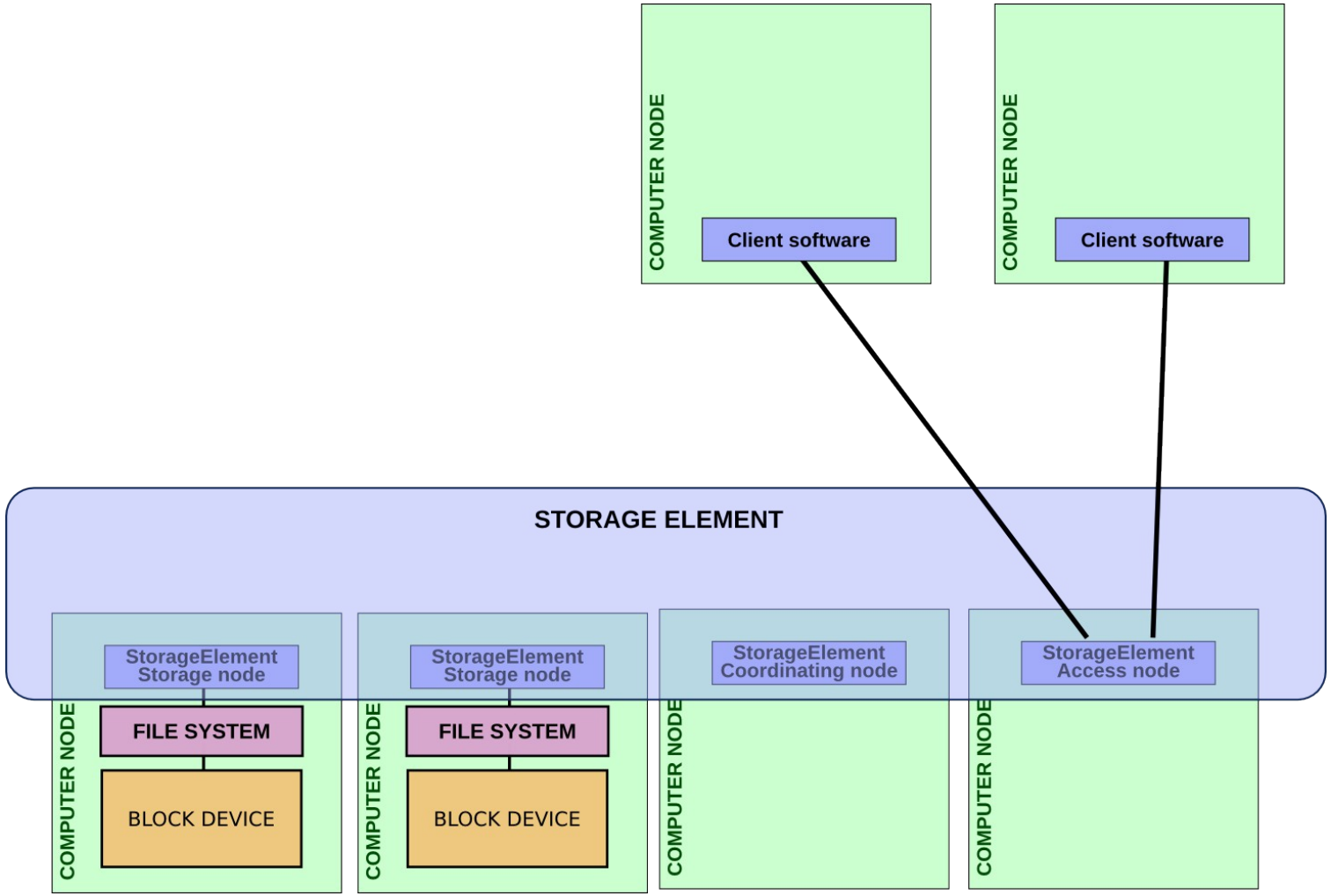


ZFS, BtrFS

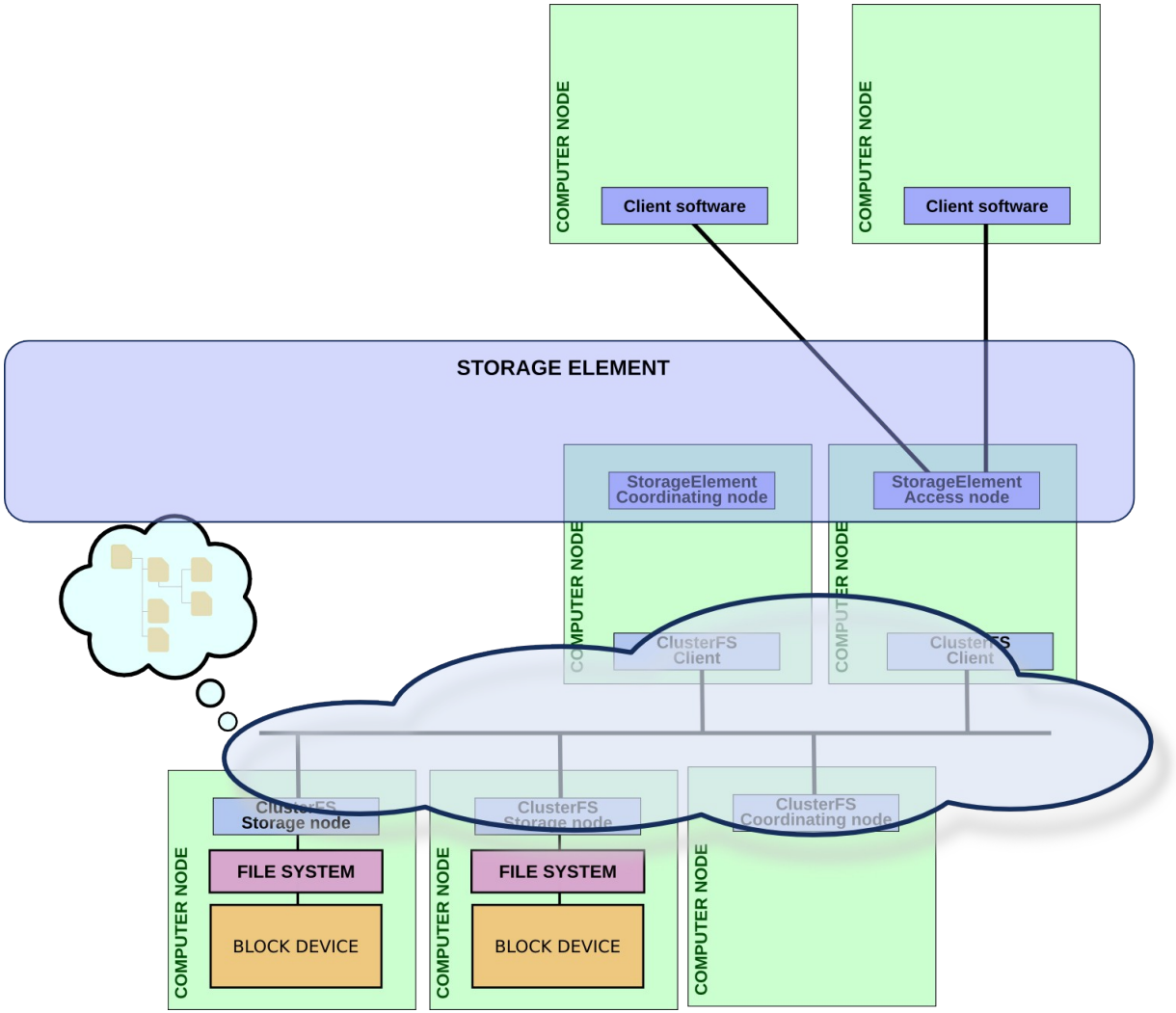
Cluster filesystems



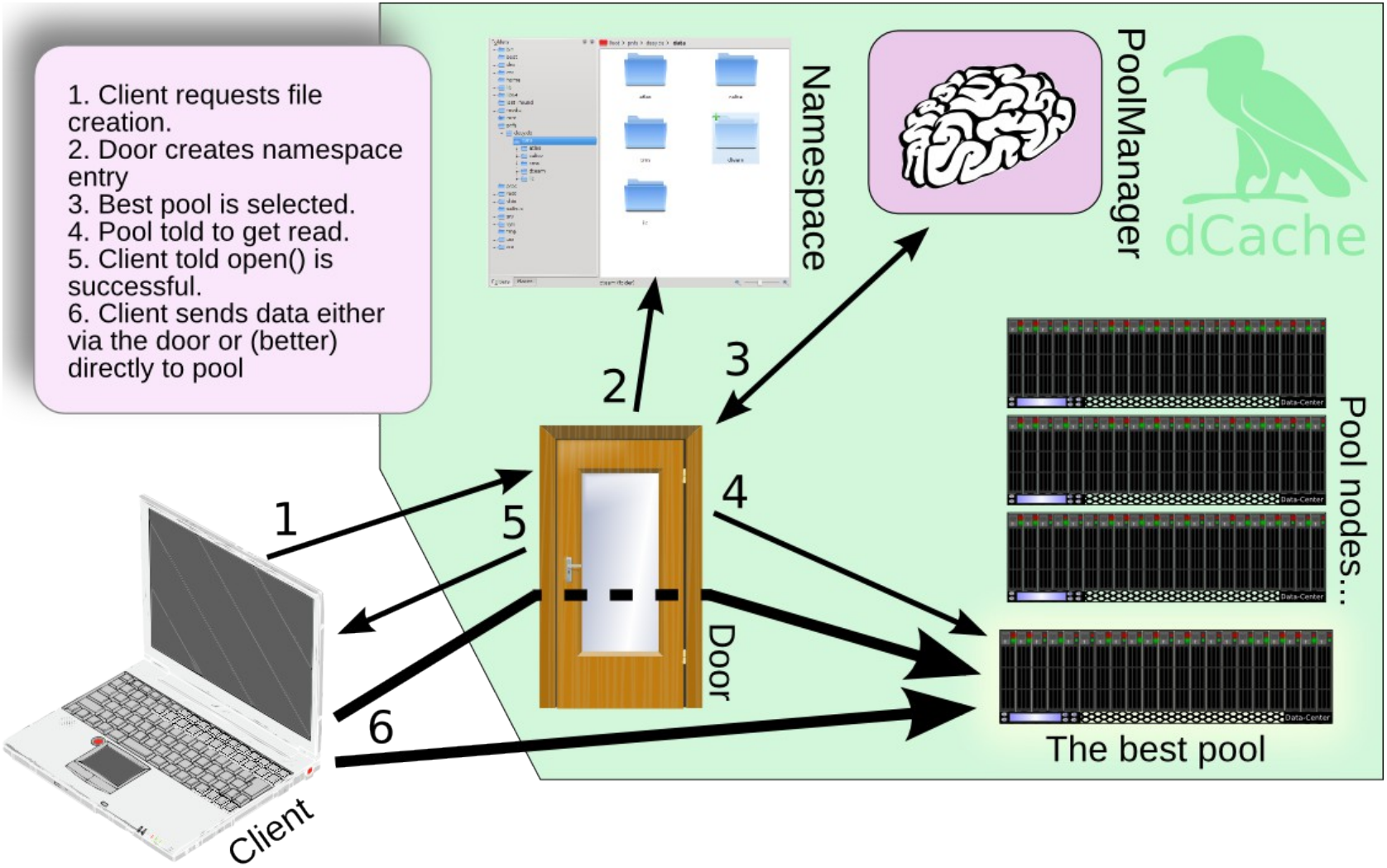
Storage Element



Storage Element



Example of redirection



Protocols

- Transferring data
  - Redirecting the client is important!
  - LAN access (for worker nodes):
    - **NFS v4.1**, dcap, rfio, xrootd, (HTTP?)
  - WAN access (for transferring data)
    - GridFTP, **HTTP**, WebDAV, (xrootd?)
- Management
  - SRM v2.2
- Standardisation:
  - GSI vs SSL/TLS

Grid storage

- Lots of sites (so, lots of SEs)
- Data appears in multiple locations
- Current Grid-level services:
  - **FTS**: moving data
  - **File Catalogues**: finding the files
- Experiment provides:
  - File grouping (data sets)
  - Access framework (software)
    - Unfortunately it adds layer of indirection between end-users and sites(!!)

Grid Storage: catalogues

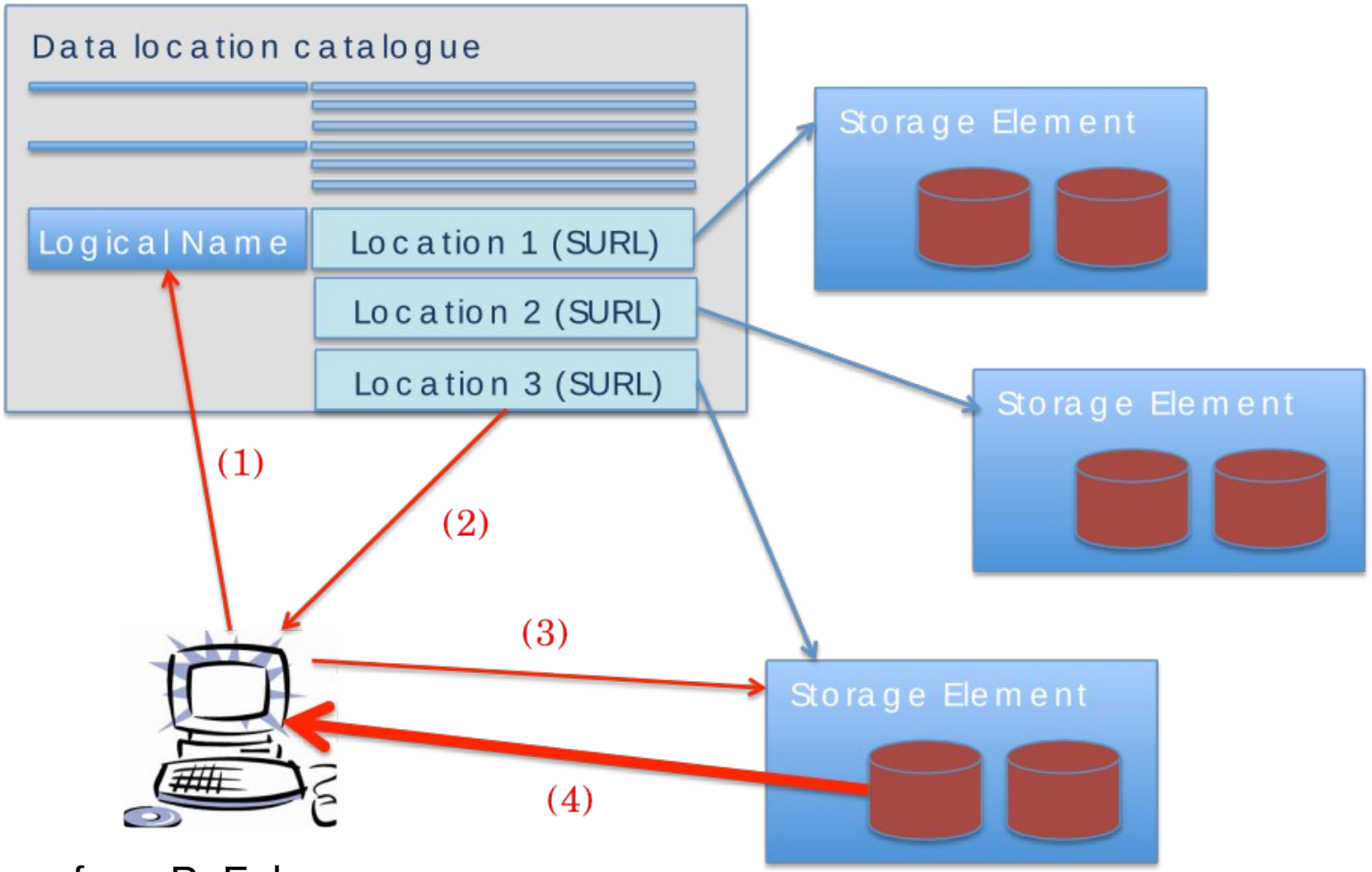


Diagram from P. Fuhrmann

Grid problems

- Communication:
  - VOs have many storage provides
  - Sites (typically) have many VOs
    - VOs have many users
- Diagnosing problems is hard
  - A networking problem could involve:
    - end-user and VO,
    - src and dest storage elements (the sites),
    - FTS, catalogue(s), network providers, ..
- Use of non-standards doesn't help!

Monte Carlo

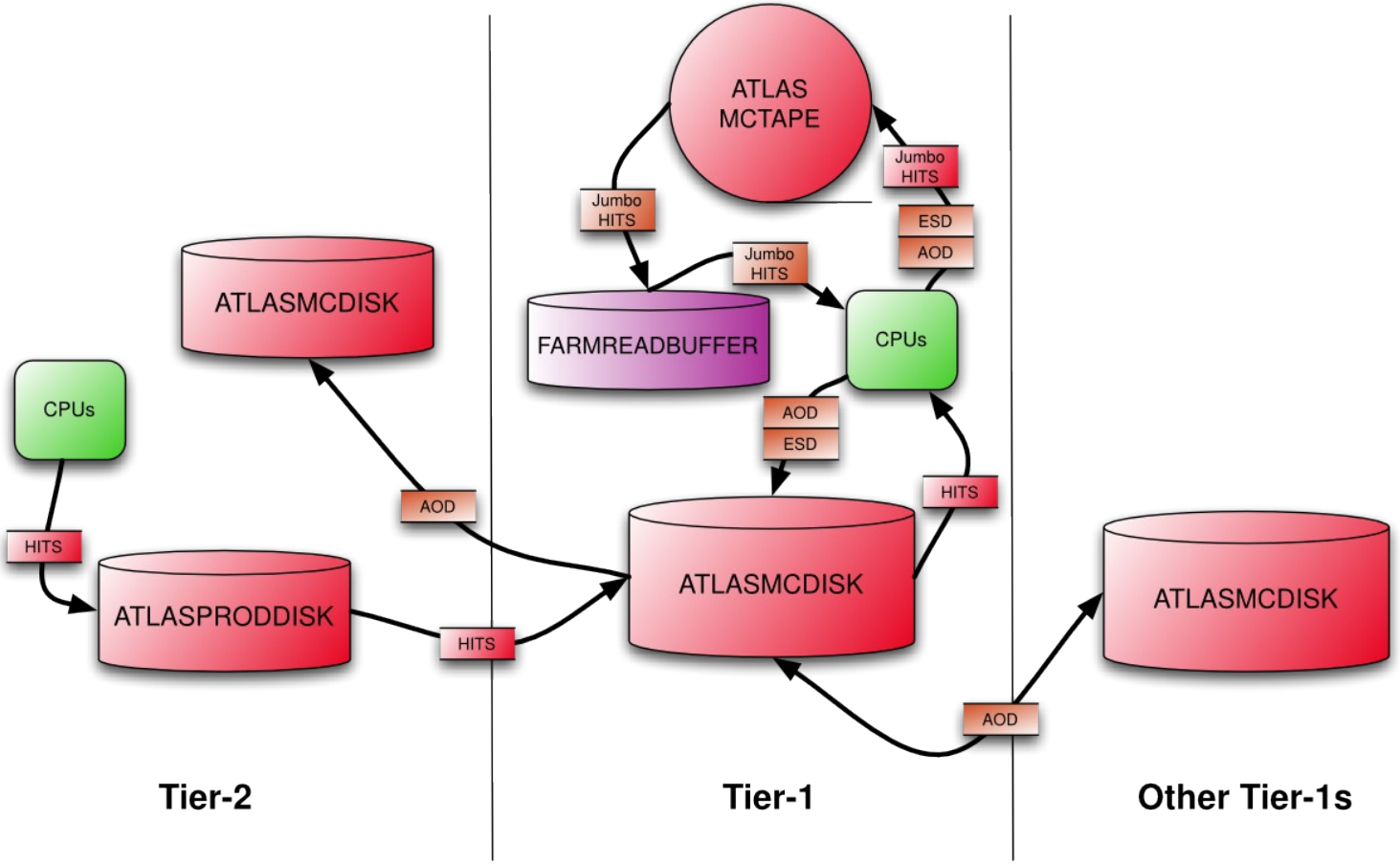


Diagram from Dr. G. Stewart

Data taking

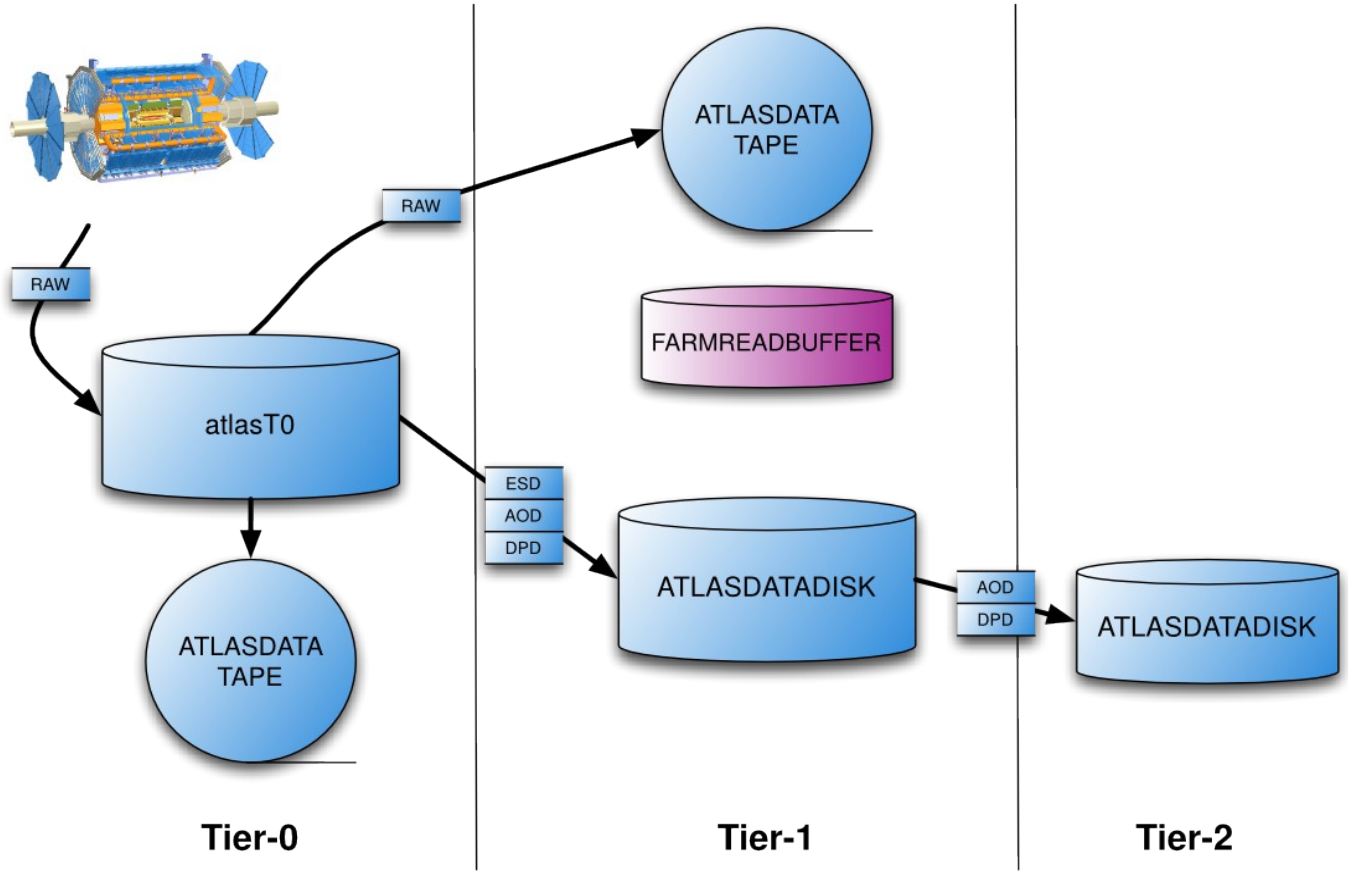


Diagram from Dr. G. Stewart

Reconstruction

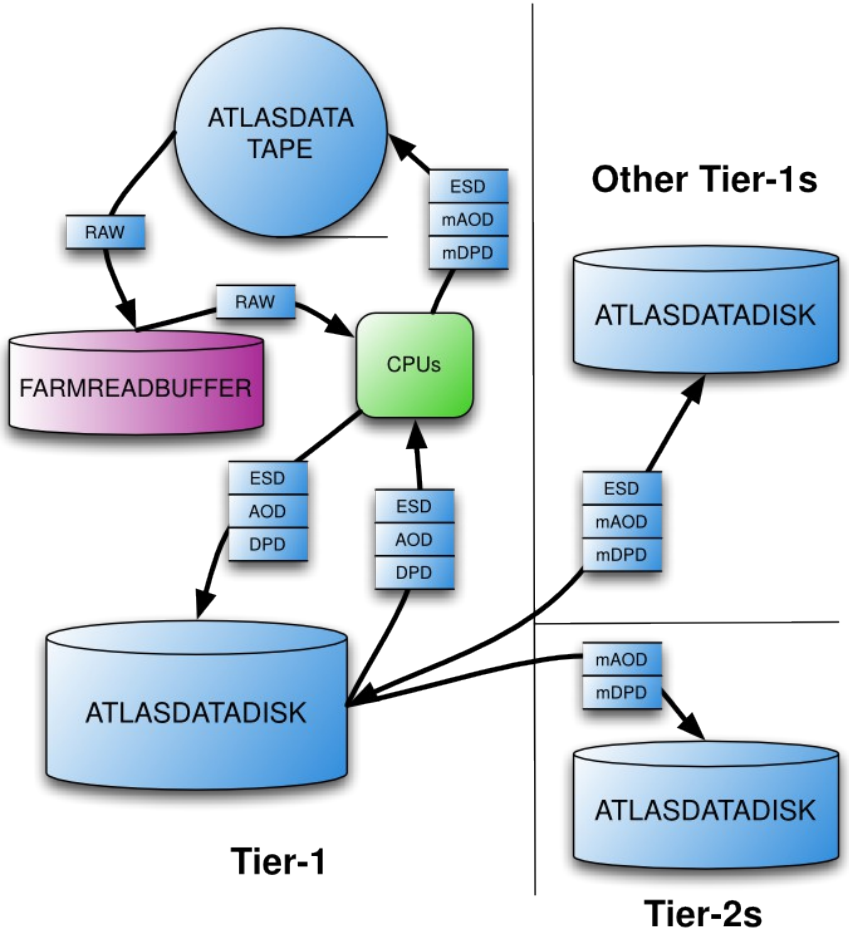


Diagram from Dr. G. Stewart

Chaotic analysis

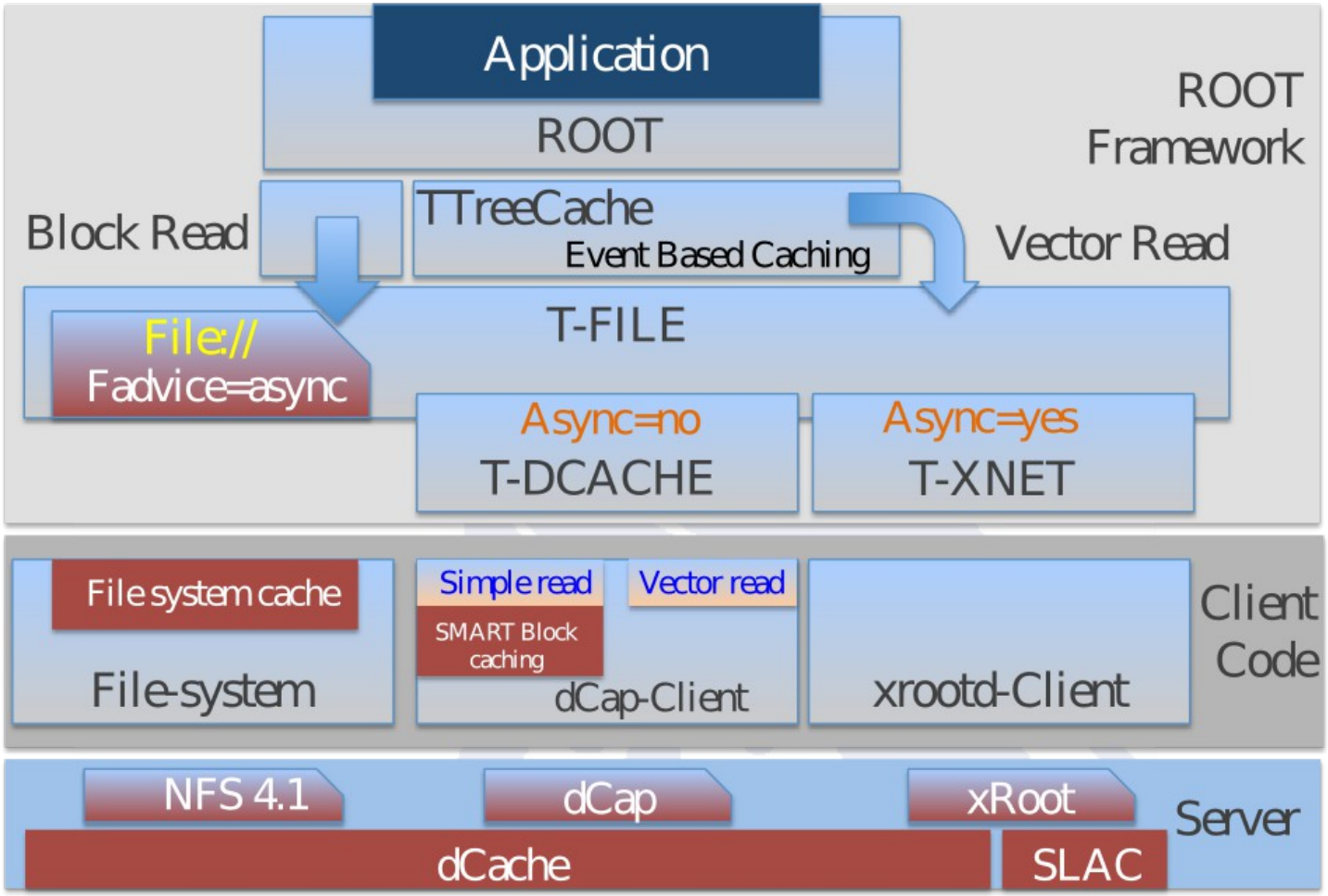
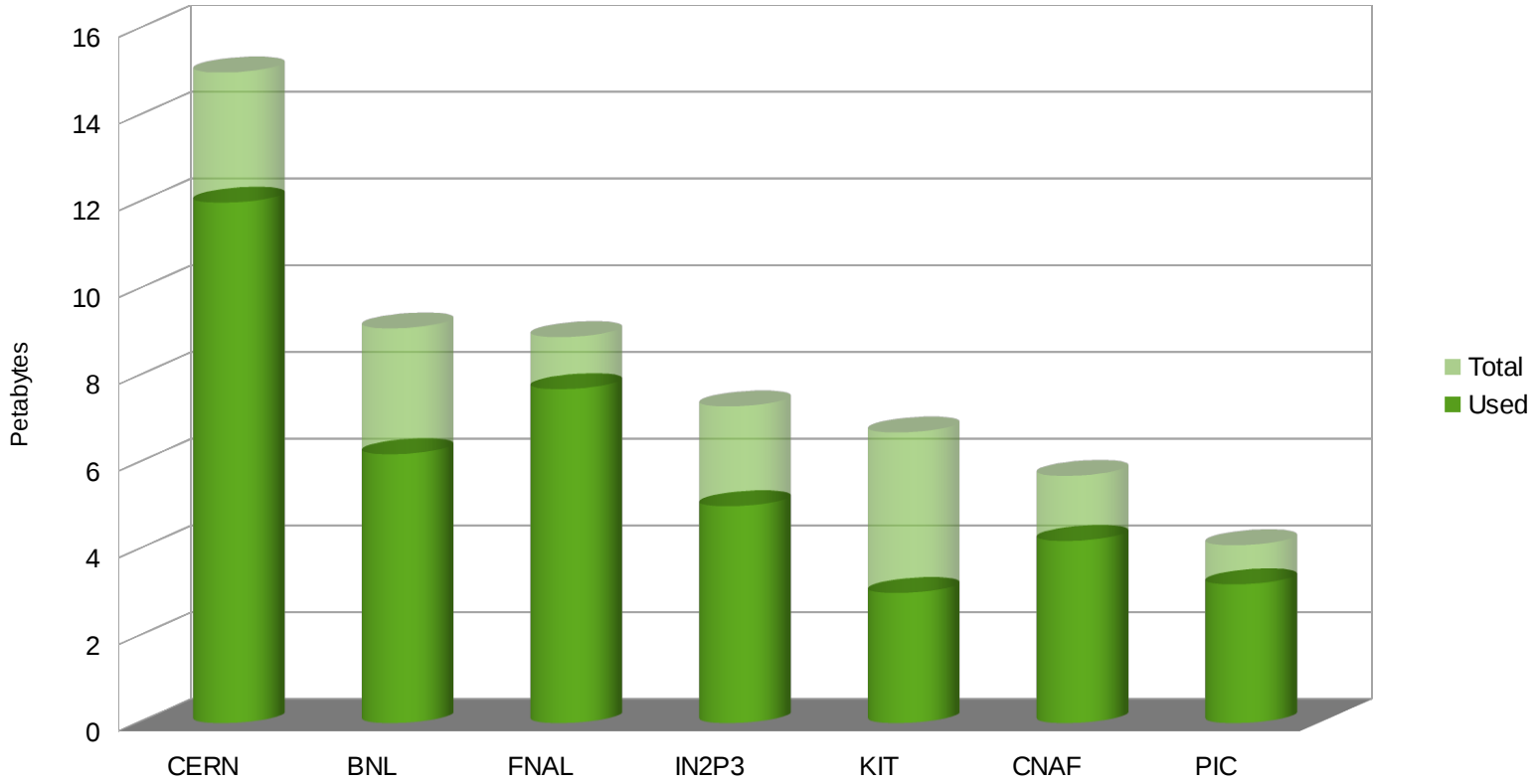


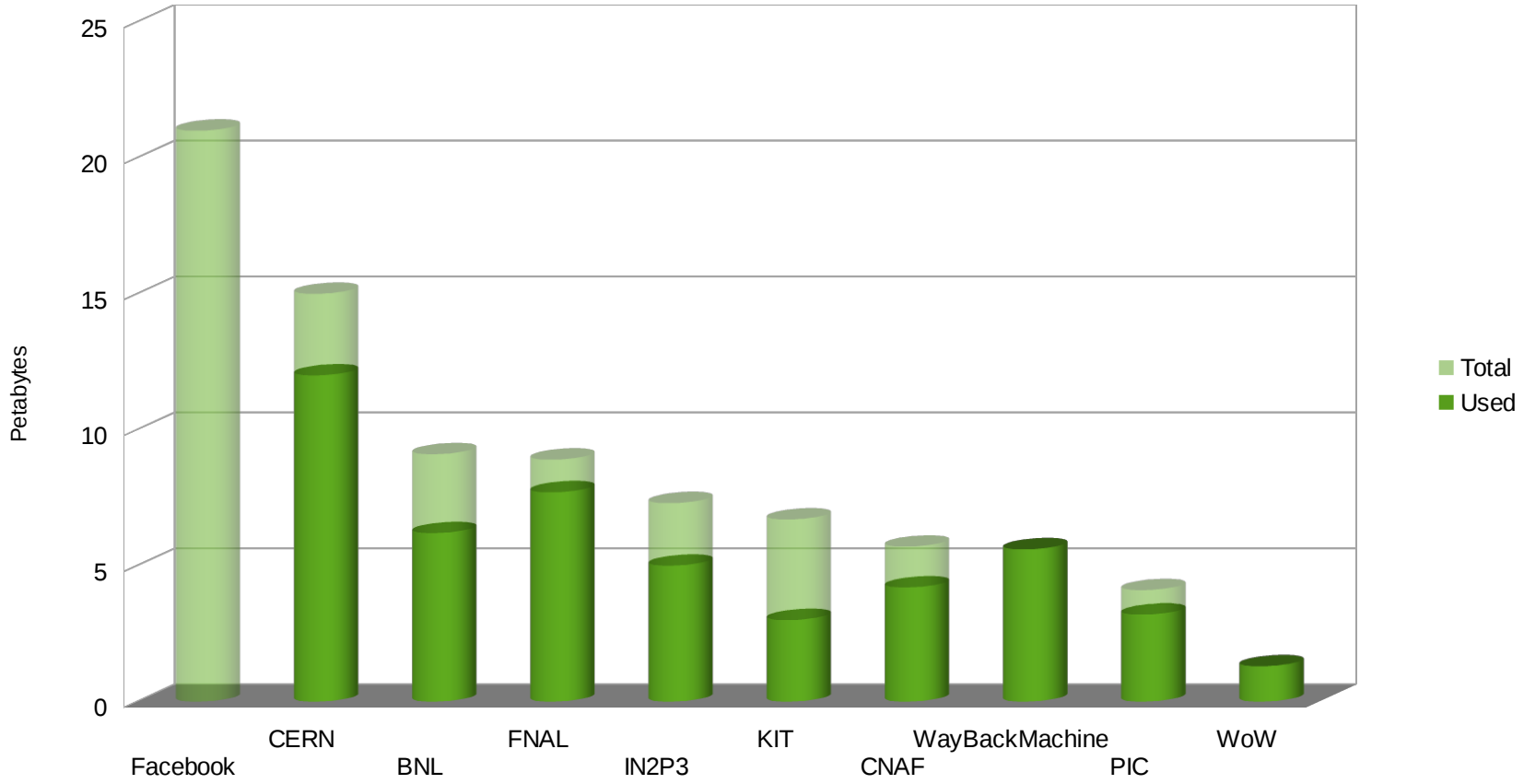
Diagram from P. Fuhrmann

Grid storage in context

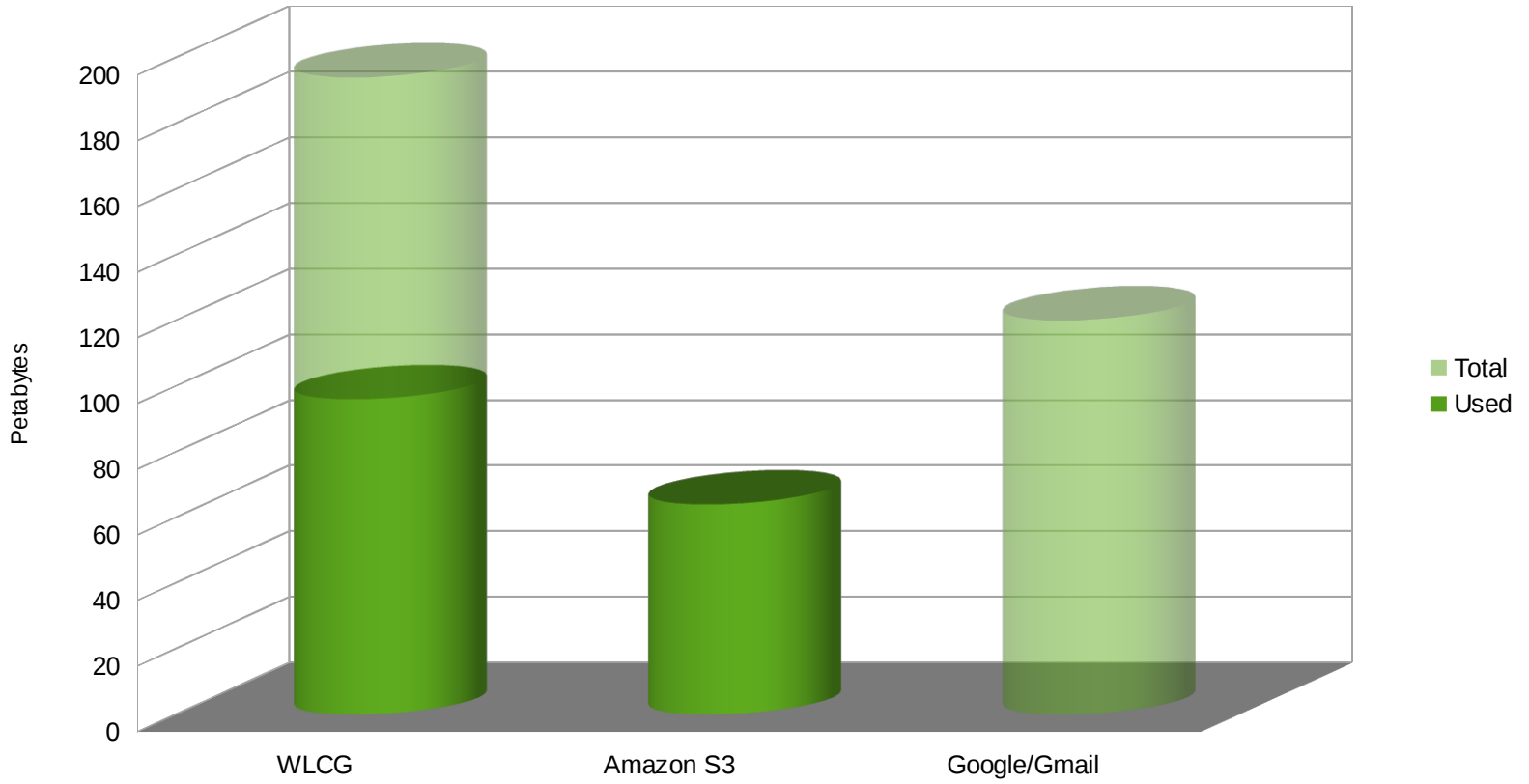
WLCG site storage capacity



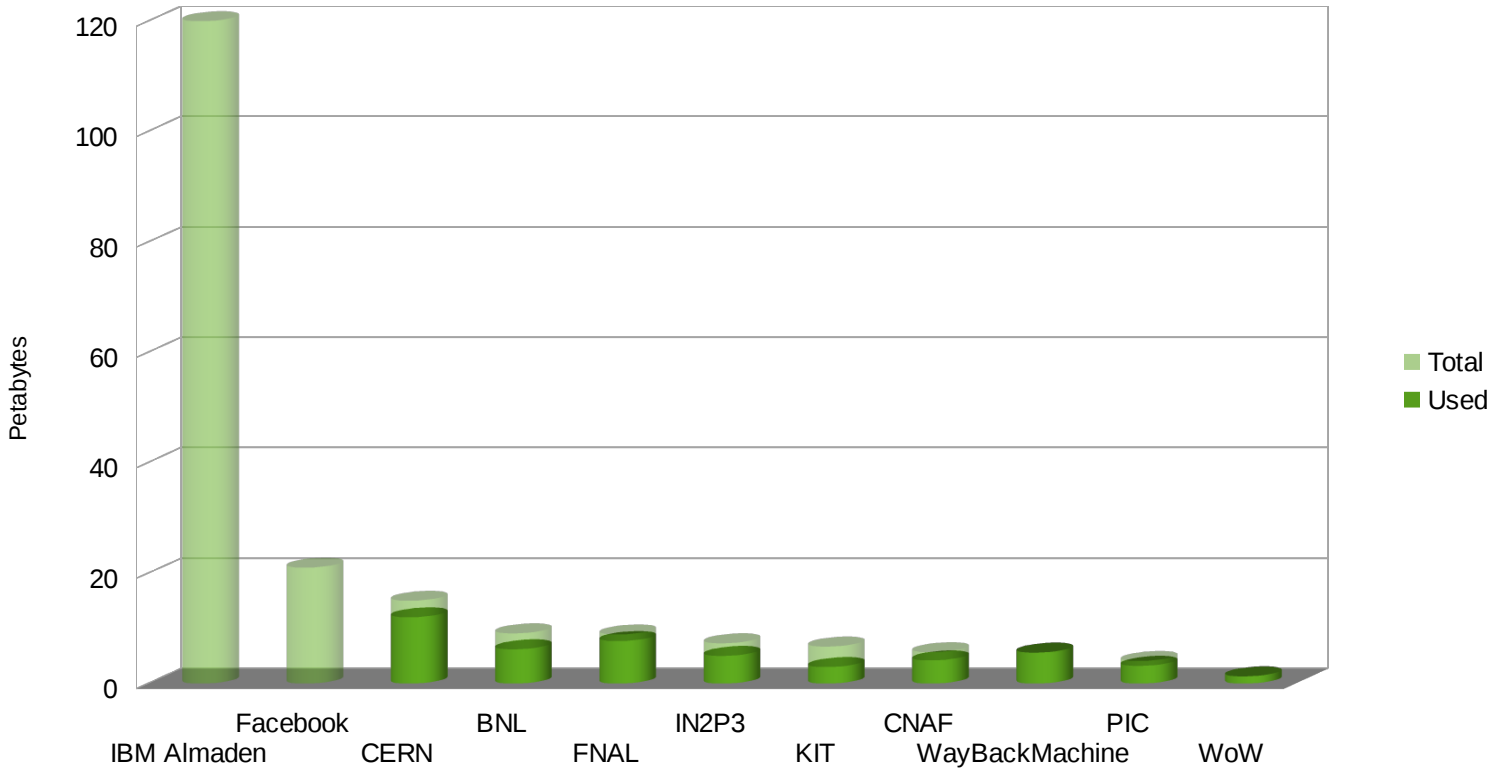
+ some non-WLCG sites



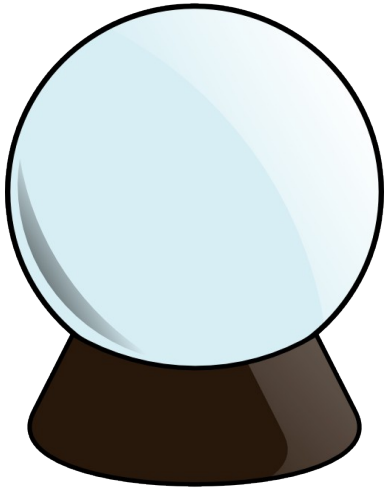
Distributed storage



Sites + IBM Almaden



Current challenges and Future directions



Dynamic data placement

- Example from ATLAS
  - Data was copied based on what people thought would be useful
  - Turns out they didn't know!
    - Lots of data copied but never read.
- Try replicating based on use:
  - Example policy:

    When a T2 pulling in a file from T1, make two additional replicas elsewhere.
  - So far, working pretty well.

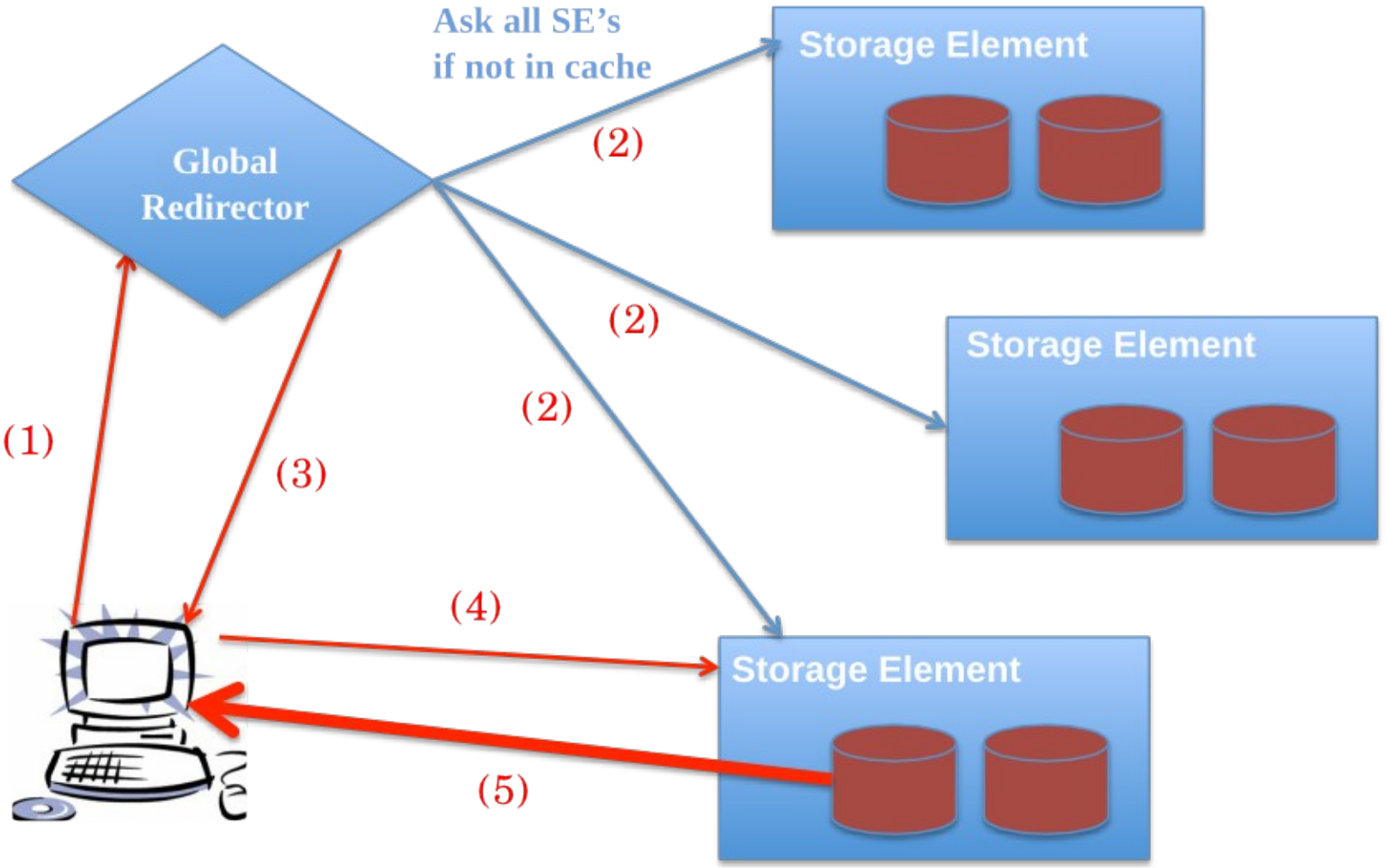
Standardisation

- HEP storage requirements aren't that enormous any more.
  - Others are finding solutions, don't reinvent the wheel!
- EMI: we're switching from Grid-specific protocols to standards
  - GSI to SSL/TLS,
  - GridFTP to HTTP/WebDAV,
  - LAN custom protocols to NFS v4.1

The death of Monach

- Monach is a rigid Tier structure.
  - T0, T1, T2.
  - Rational: network will be a bottleneck
- Reality:
  - Prolifically of classifications:
    - Non-geo. T1, “Large” T2, T3, Exp. “Clouds”
  - Backbone network isn't a bottleneck
- Gradual relaxing of rules
- Eventually: any file from anywhere.

Global namespace



Future of tape

- Only really HEP that uses TAPE storage in-band.
  - elsewhere used for archiving data.
- Still need tape for archive, but..
  - Data processing move to (almost) completely on disk
  - Fetching from tape will be like a copy
  - Tape will be “write once, read never”

Disks: where are SSDs?

- SSDs are FLASH memory in a block-device format
  - Much faster than Mag. Disks for reading (writing is slower)
  - Predicted introduction in data centres hasn't happened (yet)
- Why?
  - Errors are sudden, unpredictable.
  - They're still expensive
  - Software support isn't here (yet)

Satellites

- Structure storage
  - SSDs for random access (analysis)
  - Mag. disks for “archival storage”
- Support?
  - In filesystems: ZFS
  - In cluster filesystems: GPFS
  - In storage systems: EOS
    - and dCache (soon)

Disks: new technologies

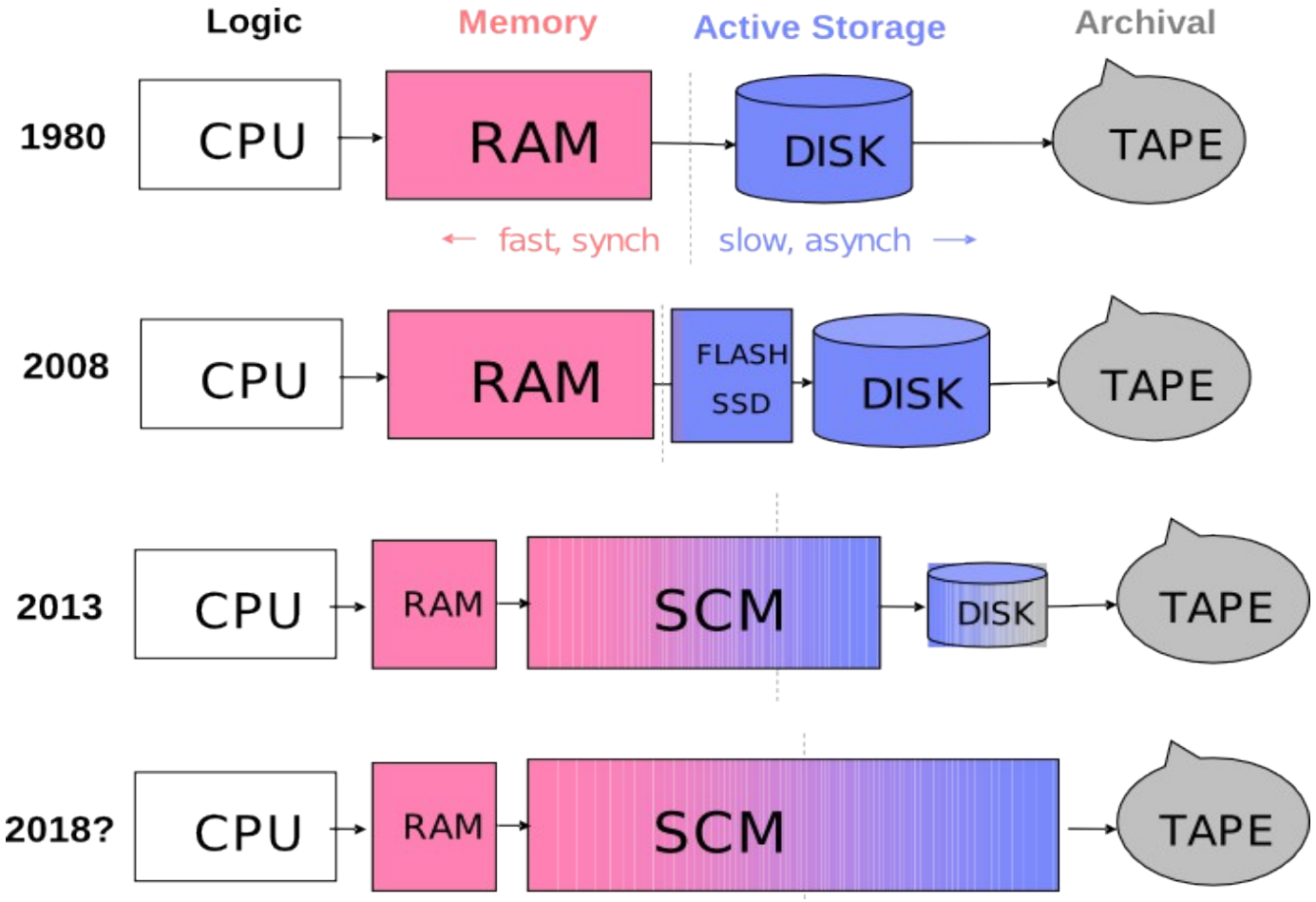


Diagram from "Storage Class Memory, Technology, and Uses" David A. Pease, IBM Almaden Research Centre.

Data integrity

- More data means more likely to see corruption
- Detecting corruption:
  - Disk (T10 DIF)
  - RAID systems (scrubbing)
  - Filesystems (ZFS, BtrFS)
  - Storage Element (file-level checksums when uploading; scrubbing)
  - Tape: (proposed)

What is EMI?

- EMI is an EU-funded project to provide Grid software
  - Combines four technologies (ARC, dCache, gLite, UNICORE)
  - Single responsibility allowing
    - Mix-n-match usage.
    - Consolidation.
- First major release, EMI-1, is now available

Thank you!

EMI is partially funded by the European Commission under Grant Agreement RI-261611

(Magnetic) Hard disks

- Block device (addressable units of fixed size)
- Characteristics
  - Streaming is fast
  - Random access is slow
  - The more concurrent activity, the poorer the overall throughput
- Failure modes are well understood
  - J-curve bath-tub (wrong!)
  - See Google Con
  - SMART!

Storage: from small to big

- Disk
- RAID
- Filesystems
- Tape
- Cluster filesystems
- HSM
- Storage element
- Grid

Tape / disk separation

- Motivations:
  - Avoid “accidental staging”
- Want clear separation (separately addressable) between disk and tape.
  - Store data is either disk or tape, never both
- Part of a move away from including tape as part of normal data-flow.

HSM storage

- Files migrate to slower media
- Based on policies or explicit commands
- Commercially available (TSM, SAMFS/QFS, ..)
-

Networking

- 10G is now che
  use
  - Sites are (or h
  - Needs CAT6a (
- 40G and 100G
  - too expensive
  - could be used