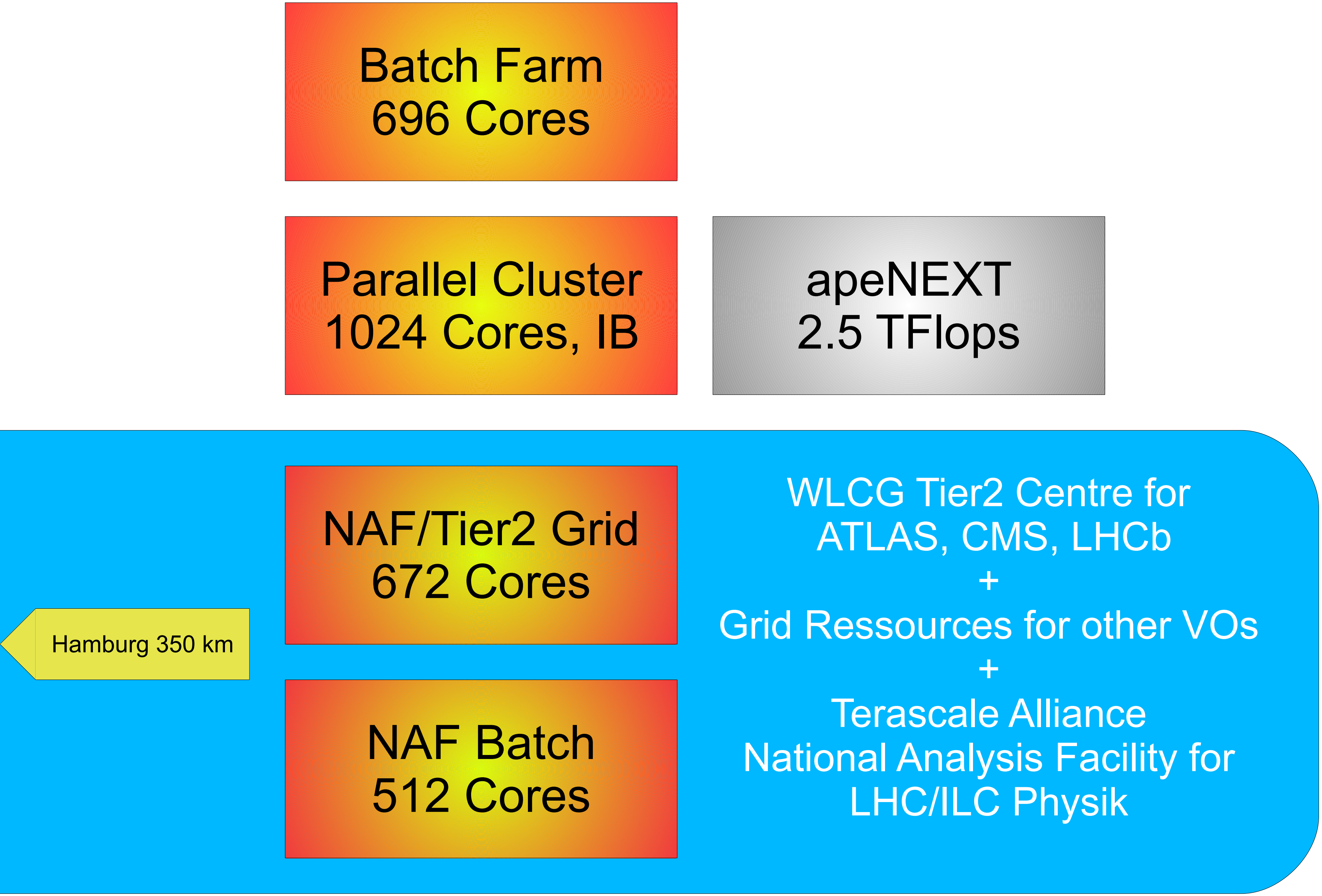
Storage in Zeuthen

Lustre, dCache, AFS

Stephan Wiesand
DESY – DV –

Zeuthen, 2010-05-17

Computing in Zeuthen



Batch Farm
696 Cores

Parallel Cluster
1024 Cores, IB

apeNEXT
2.5 TFlops

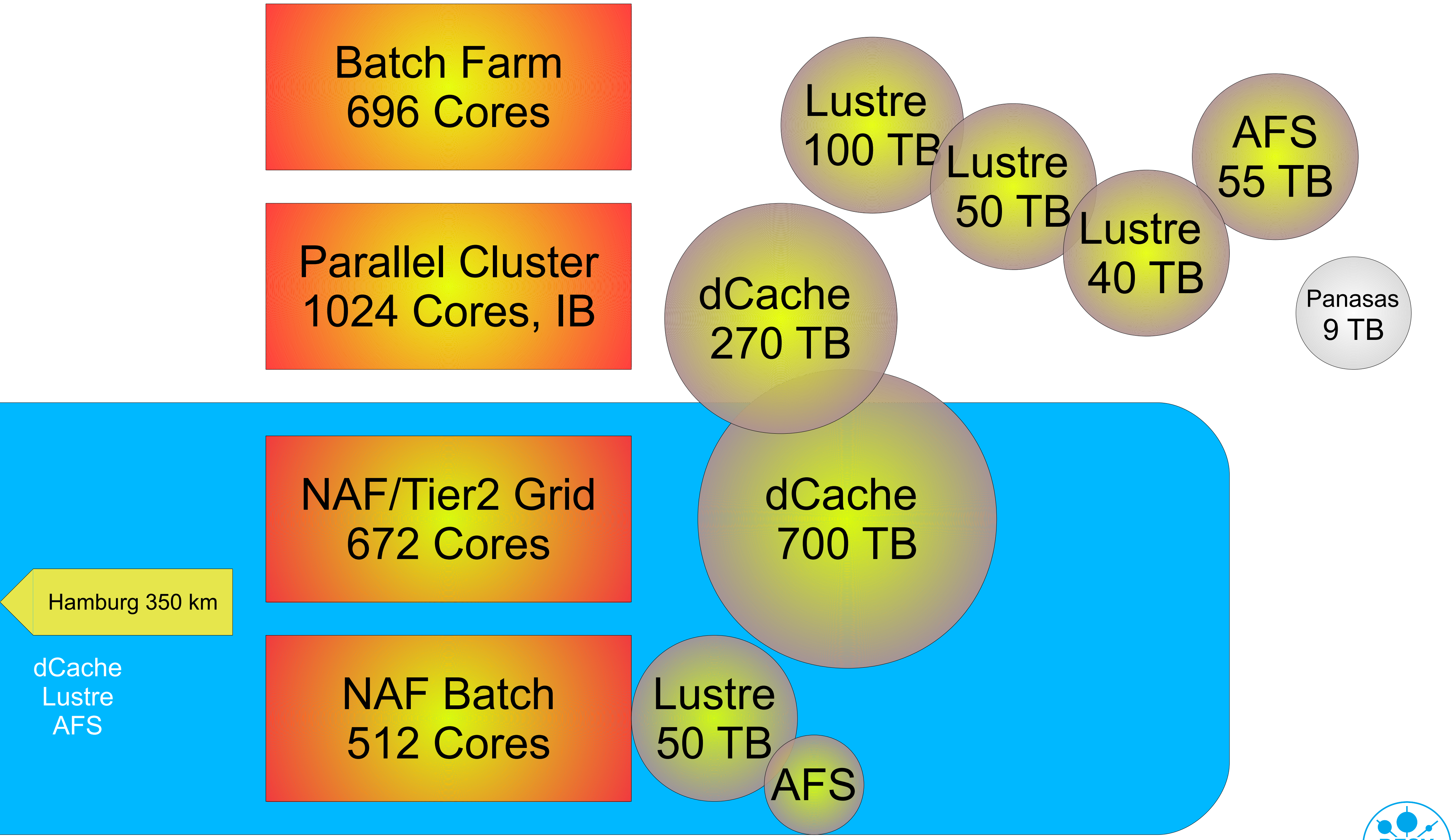
NAF/Tier2 Grid
672 Cores

NAF Batch
512 Cores

Hamburg 350 km

WLCG Tier2 Centre for
ATLAS, CMS, LHCb
+
Grid Ressources for other VOs
+
Terascale Alliance
National Analysis Facility for
LHC/ILC Physik

Computing + Disk Storage in Zeuthen



Batch Farm
696 Cores

Parallel Cluster
1024 Cores, IB

NAF/Tier2 Grid
672 Cores

NAF Batch
512 Cores

Lustre
100 TB

Lustre
50 TB

Lustre
40 TB

AFS
55 TB

Panasas
9 TB

dCache
270 TB

dCache
700 TB

Lustre
50 TB

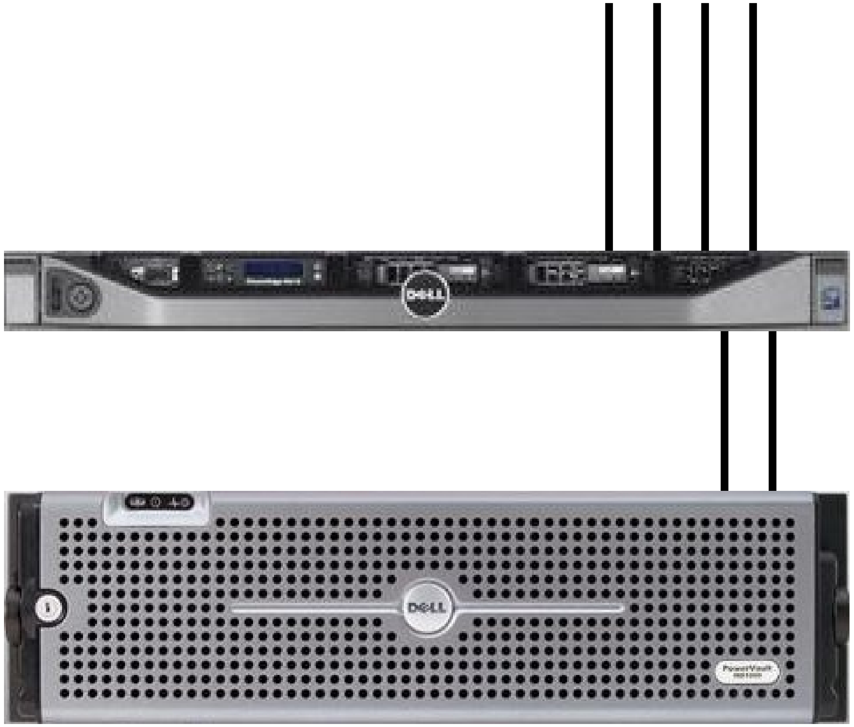
AFS

Hamburg 350 km

dCache
Lustre
AFS

The Storage Brick

- Direct Attached Storage. Typical configuration:



1-4 x GbE
IB (DDR)
10GbE

OSS / Pool Node / Fileserver
RAID6 Controller

4x3 Gb/s SAS, x2 (redundant)

JBOD
15 x 2 TB SATA
15 x 600 GB SAS

- OS: S5L 64-bit
  - automatic, central installation, configuration, monitoring
  - just as for the compute nodes

AFS

Volume Location Database cluster (at application level)

Fileservers

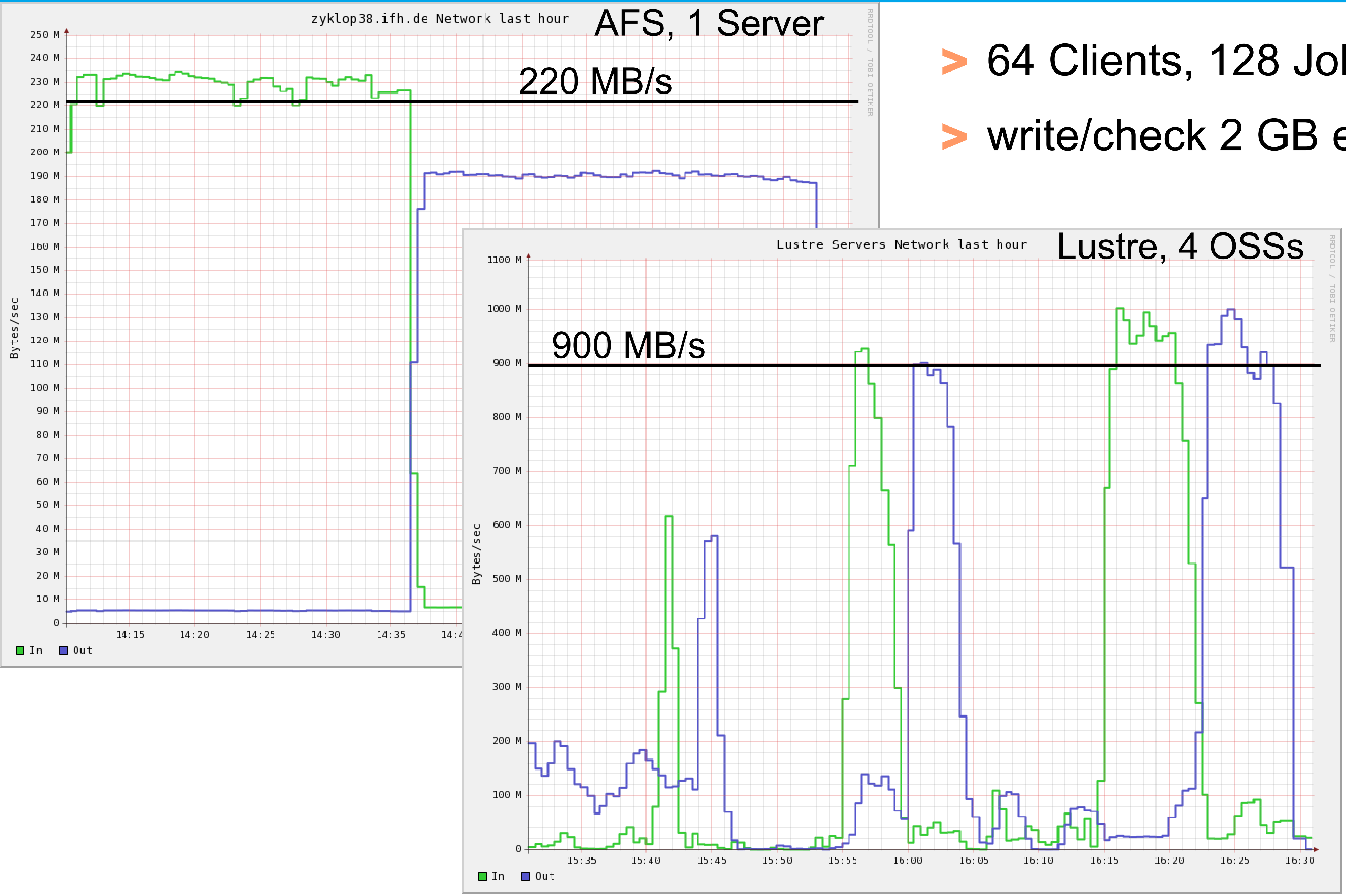
- volume location data is a very small amount
- other metadata is part of the volume and stored on fileservers
  - filename, size, owner, mode bits, a/c/m-time
- => scales well for small files
- data not distributed automatically
  - volumes are confined to a single fileserver partition

Lustre

- all metadata kept on MDS
  - filename, size, owner, mode bits, a/c/m-time
  - stripe locations on OSTs
- => does not scale well for small files
- data distributed automatically
  - => scales well for concurrent I/O
    - but not for concurrent metadata operations
      - lookup, open, create, delete, rename

Performance: AFS vs. Lustre during Burn In

- 64 Clients, 128 Jobs
- write/check 2 GB each


AFS, 1 Server

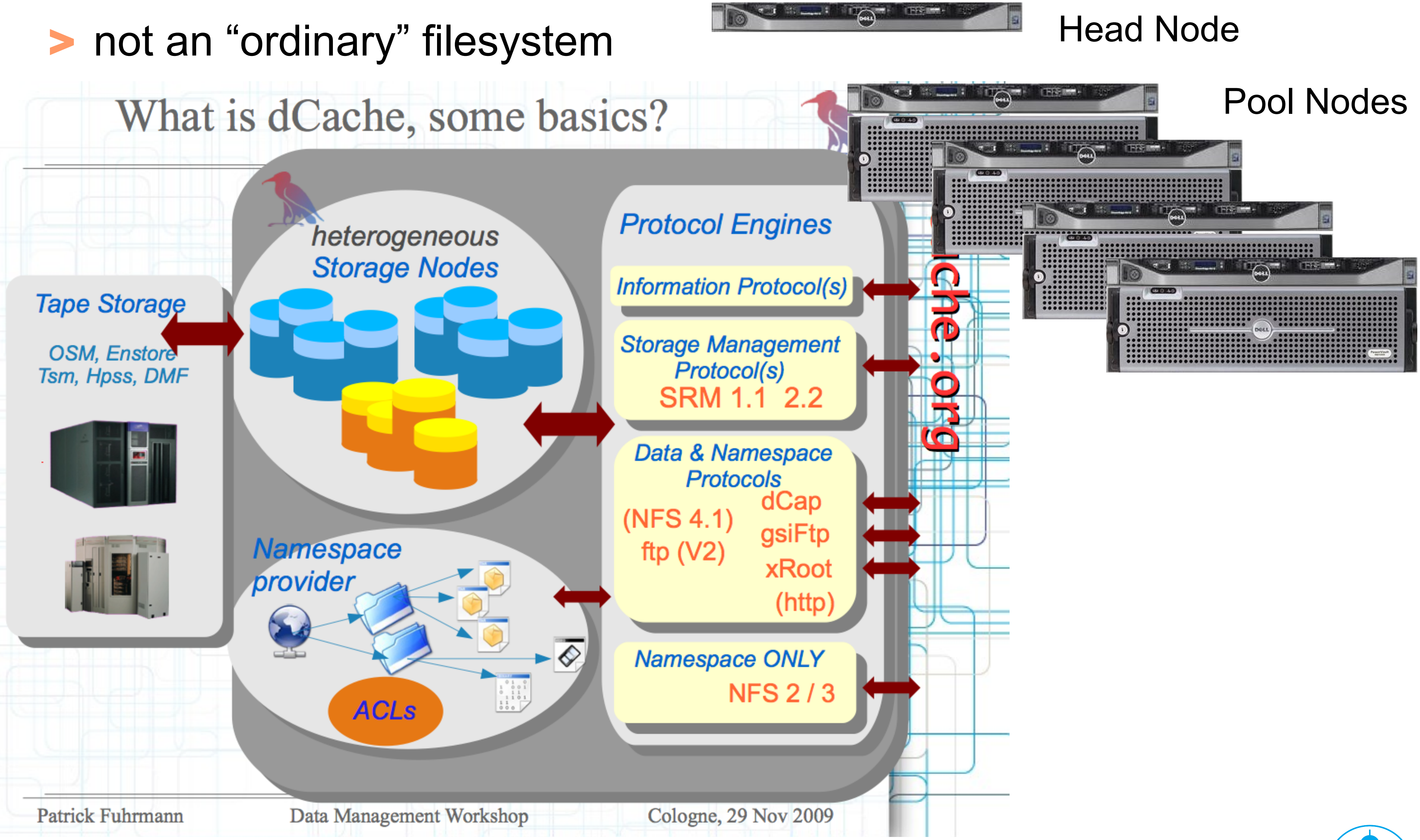
220 MB/s

Lustre, 4 OSSs

900 MB/s

dCache

- not an “ordinary” filesystem



What is dCache, some basics?

Tape Storage

OSM, Enstore
Tsm, Hpss, DMF

heterogeneous
Storage Nodes

Namespace
provider

ACLs

Protocol Engines

Information Protocol(s)

Storage Management
Protocol(s)
SRM 1.1 2.2

Data & Namespace
Protocols
(NFS 4.1)
ftp (V2)
dCap
gsiFtp
xRoot
(http)

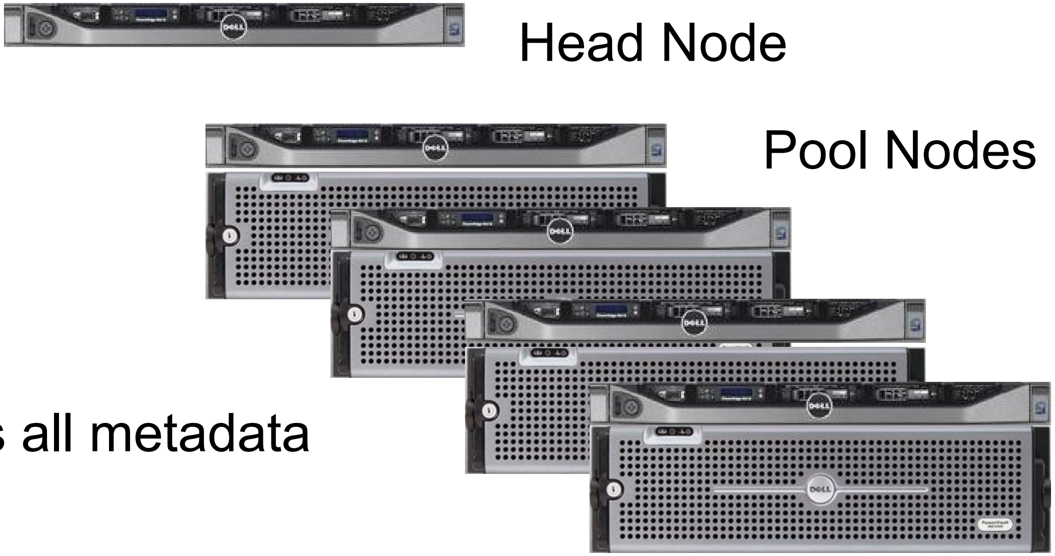
Namespace ONLY
NFS 2 / 3

Patrick Fuhrmann Data Management Workshop Cologne, 29 Nov 2009

stolen from http://www.dcache.org/manuals/20091030-storageworkshop-cologne.pdf

dCache

- not an “ordinary” filesystem
  - files cannot be modified
    - only removed an rewritten
- characteristics similar to Lustre
  - single “Namespace Provider” holds all metadata
  - data + I/O distributed automatically
- small files are especially problematic on tape!



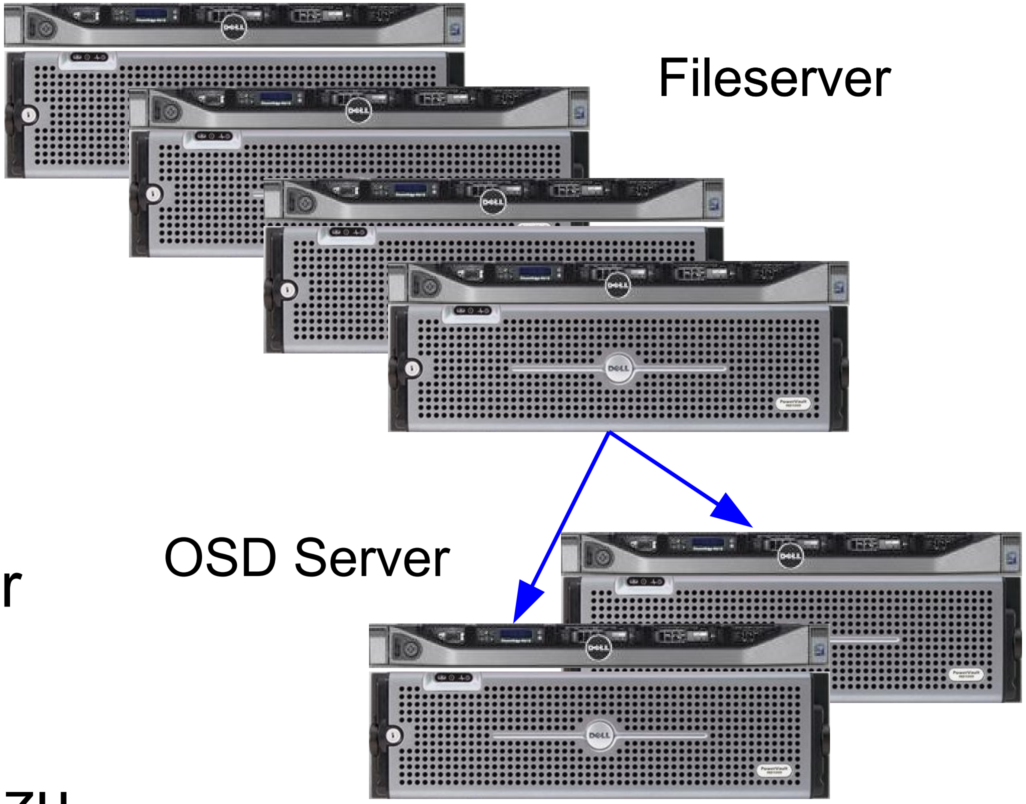
Summary

- Lustre, AFS, dCache are mutually different
  - exclusive (mis-)features:
    - Lustre
      - best single client performance, especially when using IB
      - no (r/o) replication, most volatile if things go wrong
      - hardest to manage, no simple way to migrate data to new hardware
    - dCache
      - is accessible as a grid storage element
      - optional tape backend (LRU based reuse of disk pools)
      - cannot modify files
      - dcap does not use the OS cache
    - AFS
      - suitable for small files
      - accessible from workstations, WAN
      - not well suited for large scale parallel I/O
- choosing the right filesystem for a certain task is essential

Backup Slides

AFS + OSD - vielversprechende Entwicklung

Volume Location Database
Cluster auf Applikationsebene



Fileserver

OSD Server

- Volume basiert
  - eingebette Mountpoints ergeben den Namespace
  - R/O Replizierung, asynchron
  - Transparente Migration
  - Quotas
- kleine Dateien auf dem Fileserver
- große auf den OSDs (+ Striping)
- Client greift ggf. direkt auf OSDs zu
  - ggf. direkt auf das Backend-Filesystem (z.B. Lustre, GPFS)
- http://www.rzg.mpg.de/projects/hsm-afs