A National Analysis Facility for LHC and ILC @ DESY

Birgit Lewendel for the NAF team

DESY IT Hamburg & DV Zeuthen

3.6.2008

1. NAF User Committee Meeting

- Introduction
- Overview
- Technical Details
- Status

The frame for the NAF:

- **The NAF is part of the Strategic Helmholtz Alliance**
  - **More: http://terascale.desy.de/**
- **Only accessible by German research groups for LHC and ILC tasks**
  - **Planned for a size of about 1.5 av. Tier 2, but with more data**
  - **Starting as joint activity @ DESY**
- **Requirements papers from German Atlas and CMS groups**

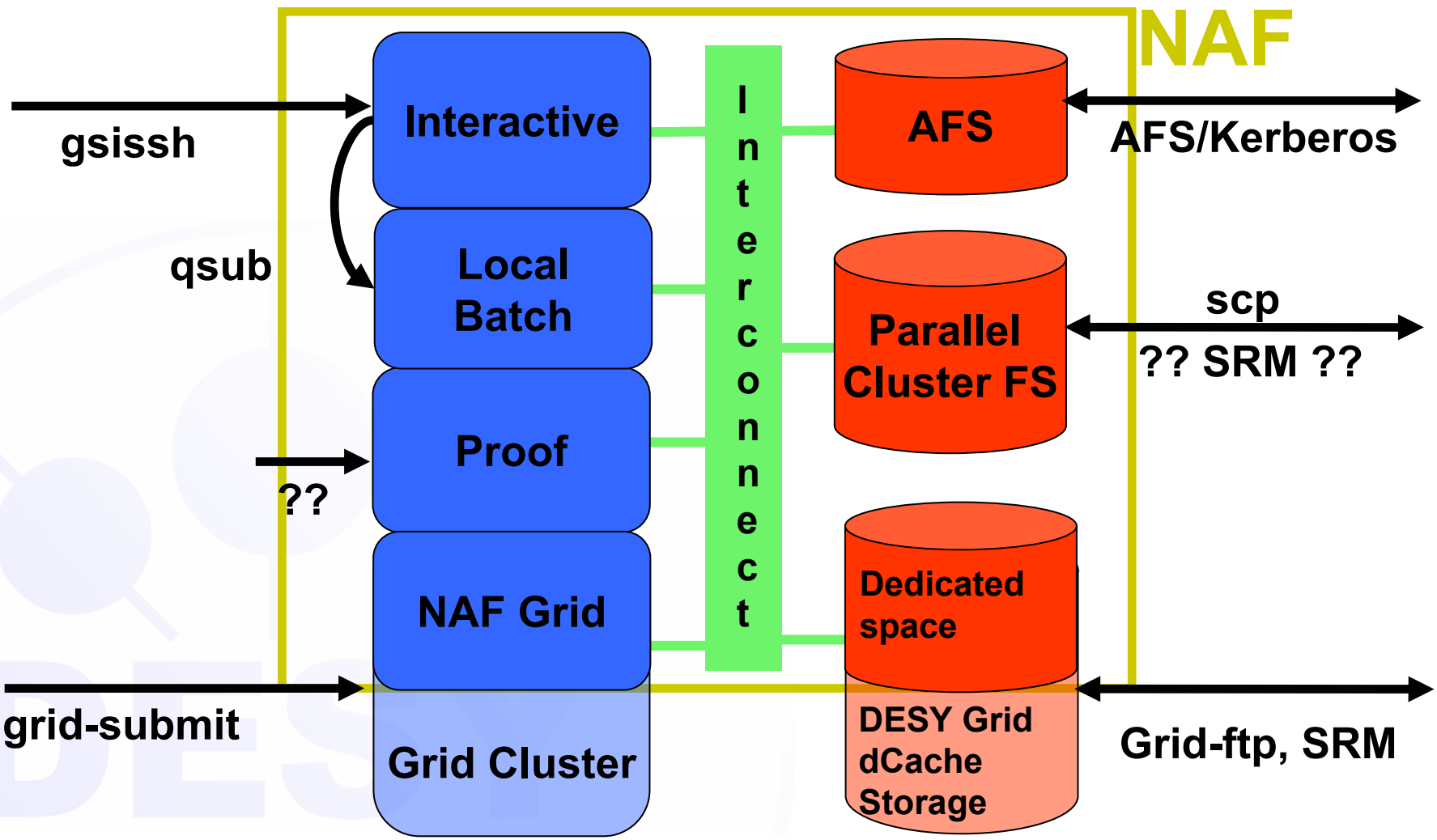
Starting with Atlas & CMS

- **Requirement papers. Some points:**

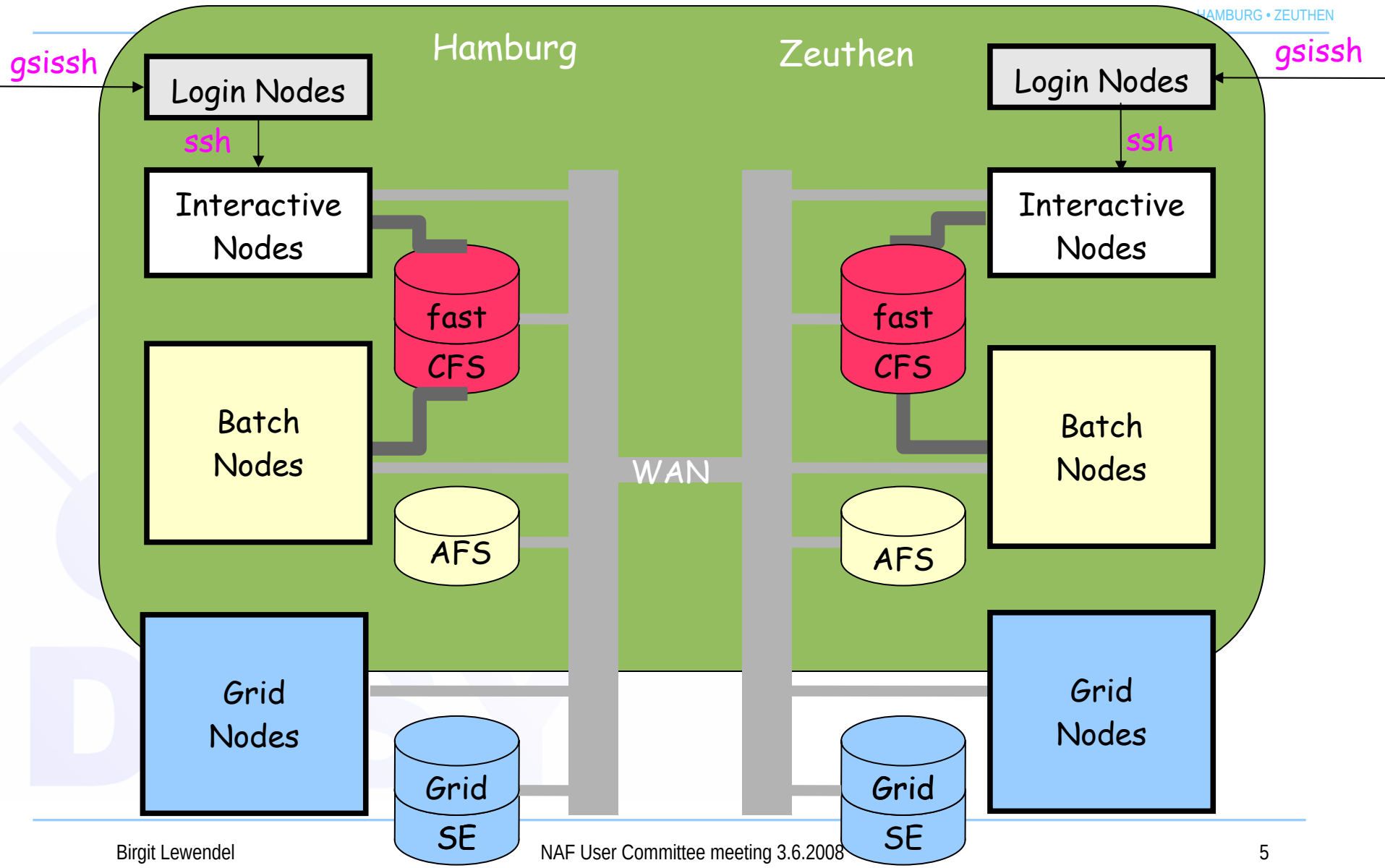
- Interactive login
  - Code development & testing, Experiment SW and tools
  - Uniform access
  - Central registry
- Personal/group storage
  - AFS home directories (and access to other AFS cells)
- High-capacity /High-bandwidth storage
  - Grid & local (with backup)
  - Grid-part: Enlargement of the T2 part

- Batch-like resources:
  - Local access: short queue, for testing purpose
  - Large part (only) available via Grid-mechanisms
  - Fast response wanted for local&Grid
- Hosted Data:
  - AODs (Full set in case for Atlas, maybe trade some for ESD?)
  - TAG database
  - User/Group data
- Additional services
  - PROOF farm, with connection to high-bandwidth storage
- Flexible setup
  - Allows reassignement of hosts between different types of services

NAF Overview



The Big Picture



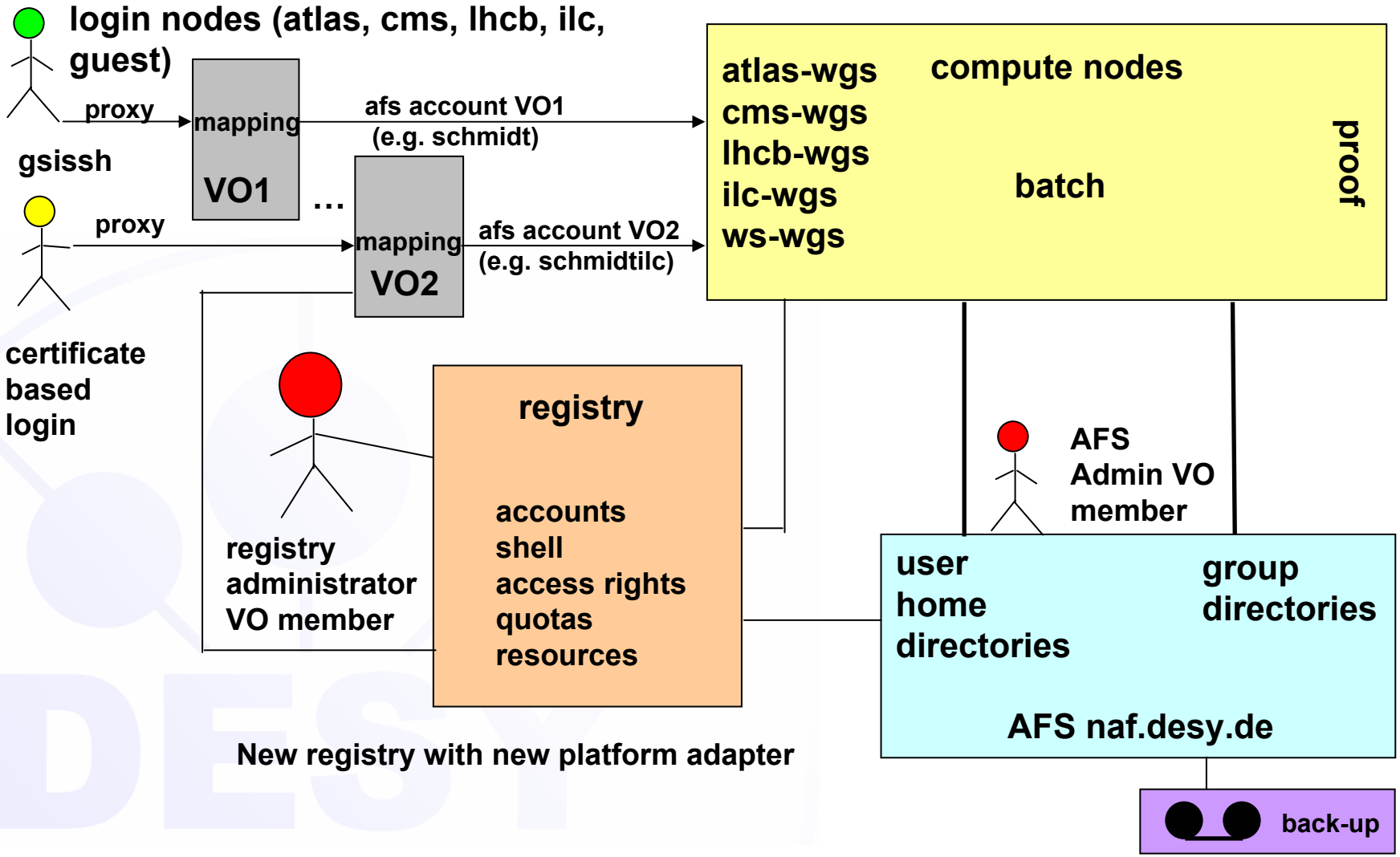
The first hardware

- Hardware Computing:
  - 6 x 16 DualCPU-Quadcore Blades (HP-Proliant BL460c)
  - 2GB RAM/core, 2x146 GB SAS HD/Blade
  - Infiniband HCA
- Hardware Storage:
  - 7 x SUN thumpers (17.5TB/box at raid 6) for dCache pools & Lustre
  - 8 x DELL Poweredge 2950 with 8x146 GB SAS Disks for infrastructure and Lustre
  - 4 SUN X4200 with Dell MD1000 Sata shelves for AFS
- Other hardware: Racks, Infiniband Switch, Infrastructure servers, …

- Hardware assignment flexible:
  - Following needs of experiments

All components distributed, mostly 2:1 between Hamburg and Zeuthen.

NAF login, interactive



IO and Storage (1)

- **New AFS cell: naf.desy.de**
  - **User & Working group directories**
  - **Special software area**
  - **Safe and distributed storage**
  - **Quotas (group and user) managed by admins**

  Status : in use

- **dCache**
  - **Well-known product and access methods**
  - **Central entry point for data import and exchange**
  - **Special space for german users in storage elements**

  Status : in production

  Under discussion: no pnfs mount in NAF in the future

NAF Cluster Filesystem - Concept

- **With the NAF, we introduce a cluster filesystem**
- **The cluster filesystem will provide high bandwidth (O(GB/s)) access to a large (O(10TB)) 'working space' for temporary data**
- **Usage Model: Copy/write your data to the working area provided by the CFS, process, then save valuable results to wide area or GRID storage**
- **Data life time in the working area: Larger than job run, limited by available space, policy**
- **For HEP, it is quite new technology**
- **Evaluate such a filesystem in the NAF environment in cooperation with users**

NAF Cluster Filesystem - Implementation

- We chose to evaluate
  - Most promising candidate providing transport via native Infiniband
  - It is on it's way to the HEP community (Sites are evaluating)
- The roadmap was considerably delayed, Version 1.8 in 2nd half of 2008
- Some features are not yet available or usable (ACLs, kerberization), ZFS->2009
- We cannot wait for 1.8, base CFS user evaluation on Lustre 1.6
  - Which is the current production version, used at several sites
  - without ACLs, Quotas, kerberized user authentification, backup
- Local access to the CFS is over Infiniband, WAN access over tcp
- There is one filesystem per VO (Atlas, CMS, ILC, misc), 16TB each.

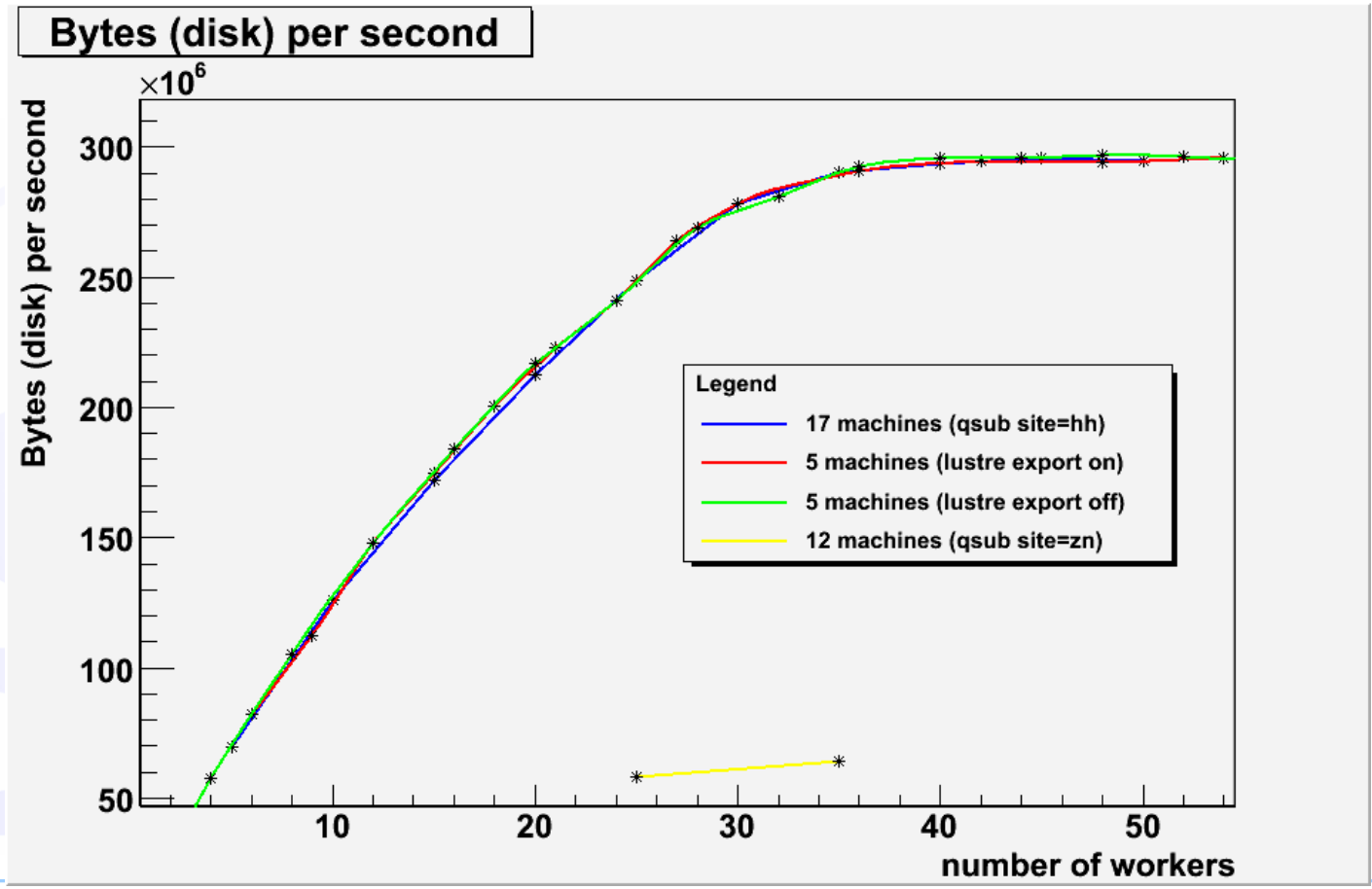
NAF Cluster Filesystem – Development, Status

- **We will set up an internal Lustre 1.8 development instance, if possible in cooperation with SUN, to**
  - **evaluate 1.8 stability, functionality, performance, test features like ZFS, ACLs, Quotas, Kerberization, Solaris based storage servers, ..**
  - **do some more hardware evaluation (DELL FS)**
  - **develop a roadmap for further use of Lustre in the NAF**

Status: temporary solution in use
infiniband mount on nodes located in Hamburg,
tcp mount on nodes located in Zeuthen.
Symmetric solution for the future.
Test with user help ongoing.

Proof at NAF

Test by UniHH on 3 dedicated proof nodes and cms wgs.
Evaluation of running proof on SGE batch nodes ongoing.

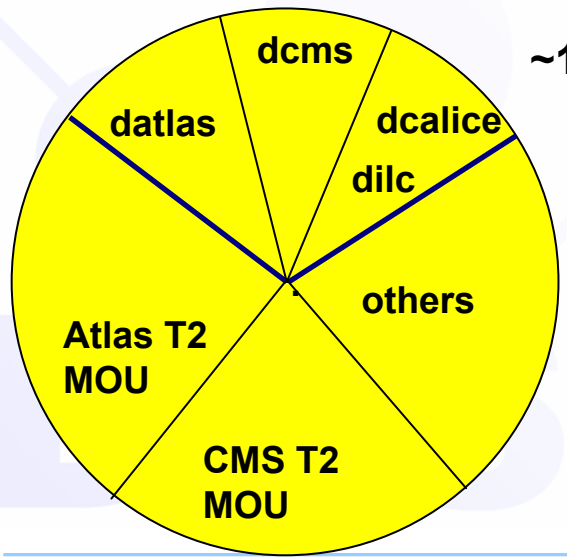


Status: test phase

no tcp buffer tuning yet

Batch @ NAF

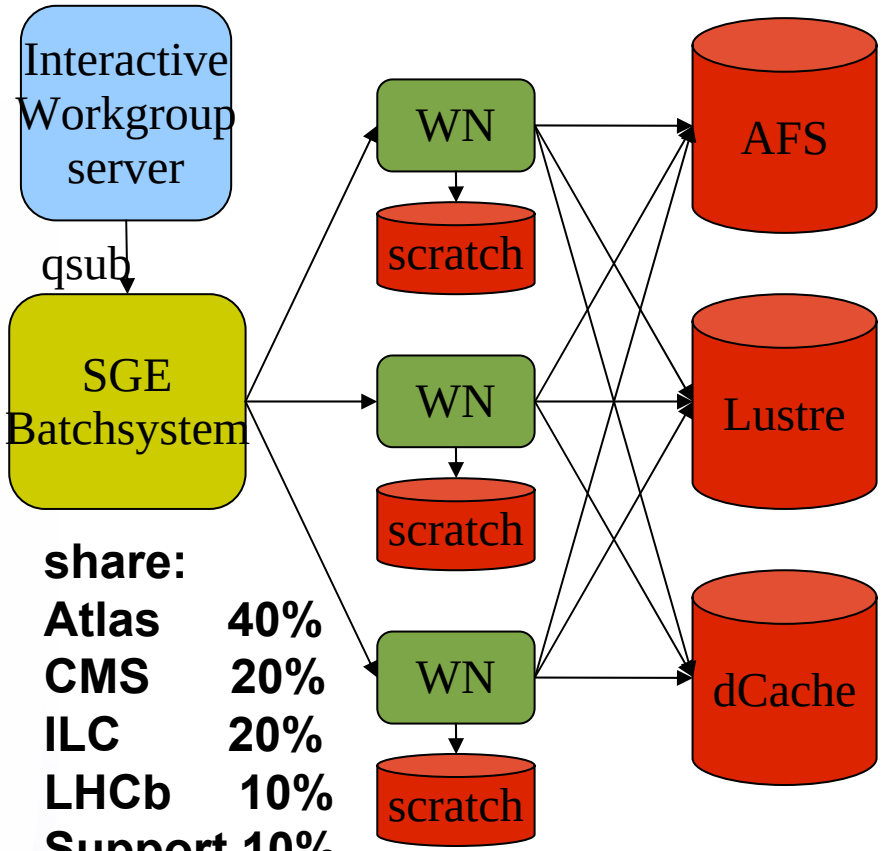
- **Grid Batch**
  - **Integrated in Desy Grid Infrastructure (grid-ce3.desy.de)**
  - **Dedicated fairshare with higher priority**
  - **Access using VOMS proxies**



~1/3 NAF of 1000 slots

- **Local Batch**

afs token from Kerberos ticket

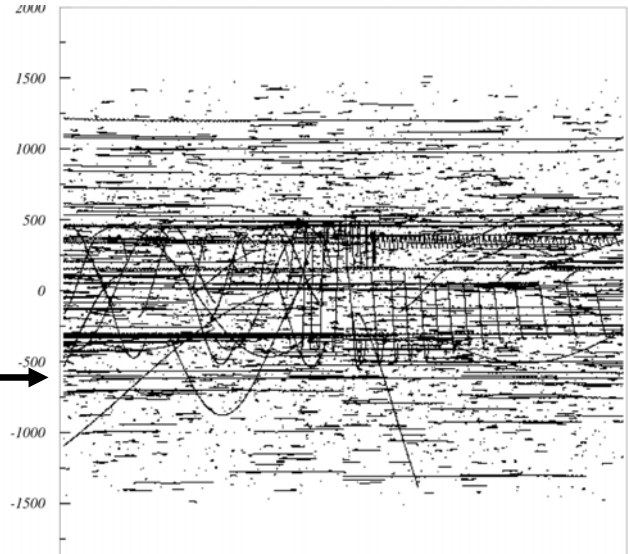


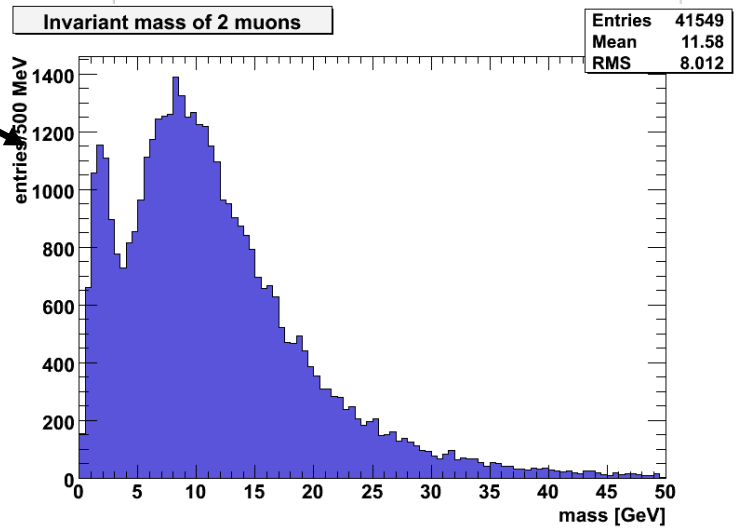
share:

Atlas	**40%**
CMS	**20%**
ILC	**20%**
LHCb	**10%**
Support	**10%**

First user experience: Feedback

- **Main users using the NAF Grid share:**
  - **Adrian Vogel (ILC): ~8k h, 10k Jobs**
    - **Machine-induced background studies, full Geant4 detector simulation**
  - **Manuel Giffels (CMS): ~70 h, 100 Jobs**
    - **Private background production (bb->2mu)**
  - **Walter Bender, Daiske Tornier (CMS):**
    - **Exotica, private Alpgen production**
- **Experience:**
  - **VOMS group works fine, no problems**



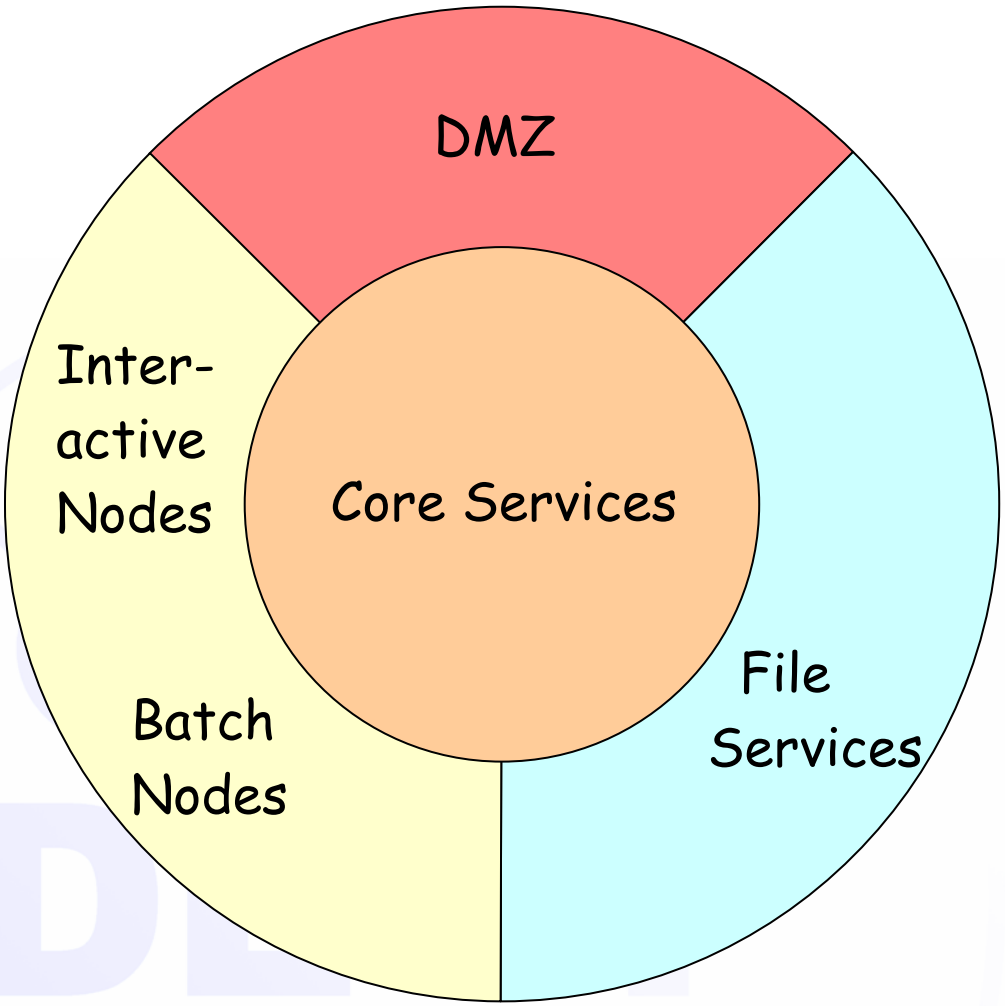


Software

- **Experiment specific software: Grid and Interactive world:**
  - **DESY provides space and tools**
  - **Experiments install their software themselves**
  - **Because of current nature of Grid and Interactive parts: Two different areas**

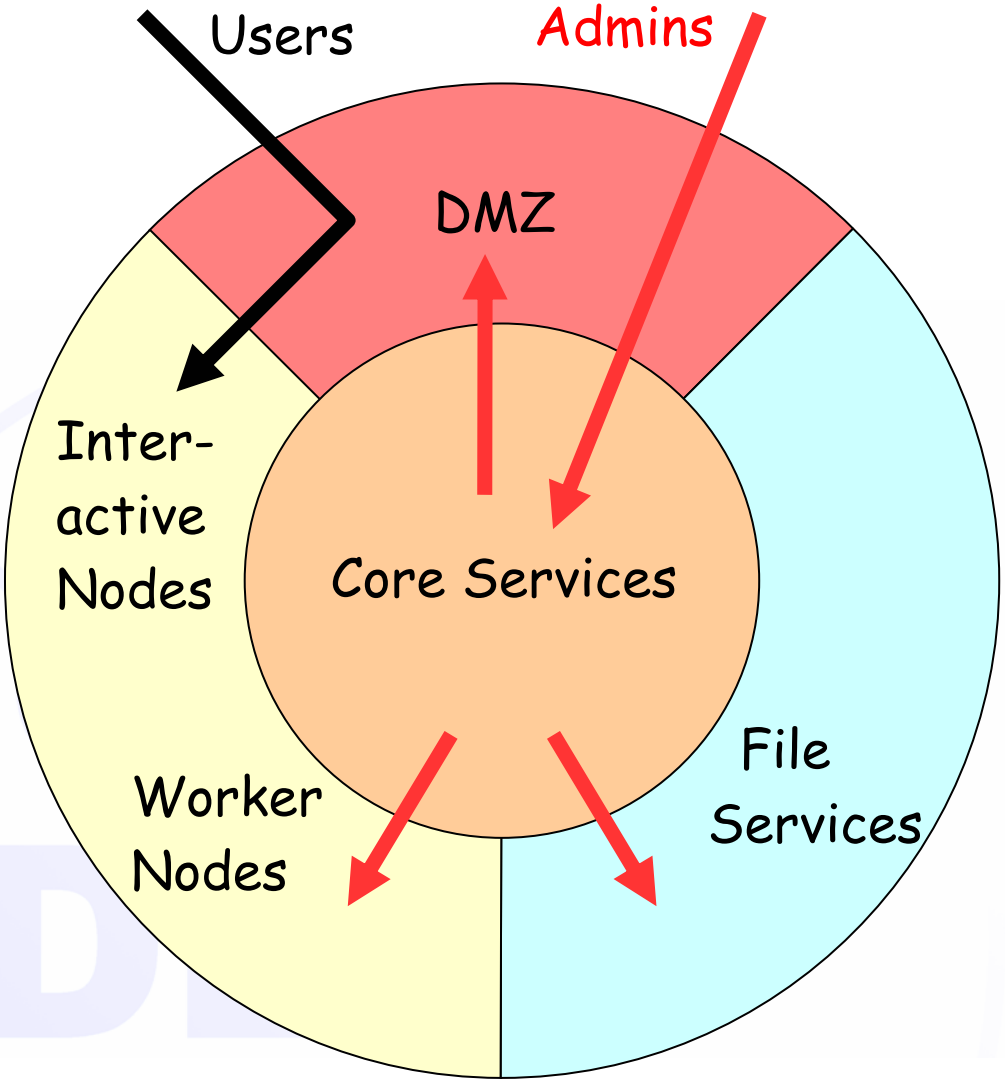
- **Common software:**
  - **Grid world: Standard worker node installation**
  - **Interactive world:**
    - **Workgroup server installation: Compilers, debuggers…**
    - **No Browser, Mailclient, ….**

naf.desy.de: Logical View



- **4 zones for different classes of systems**
- **Core services:**
  - **installation, configuration management, updates, monitoring, infrastructure (Kerberos, AFS, ...), admin access**
- **File services:**
  - **lustre**

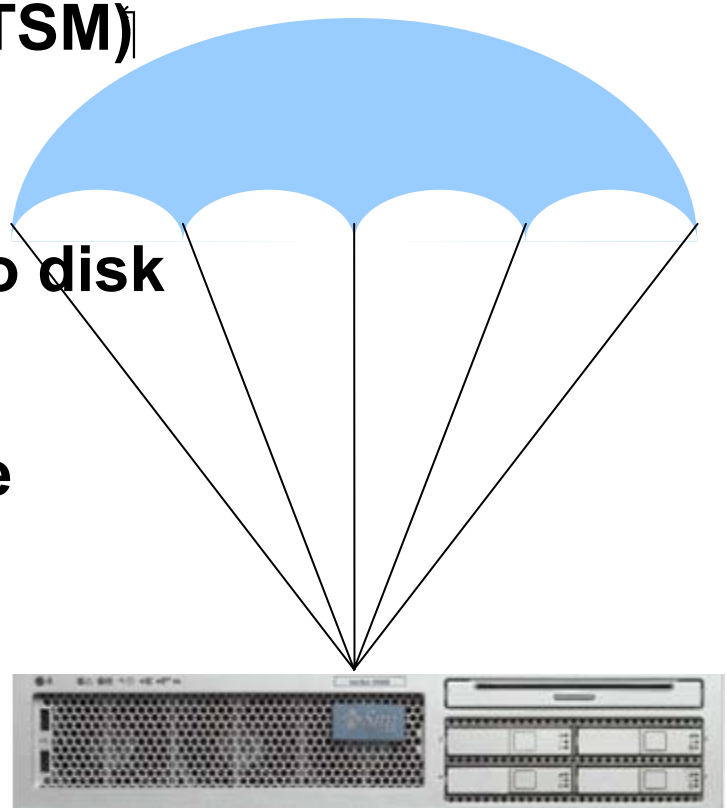
Access Restrictions -> Security



- by **default: all network ports closed on all zone boundaries**
  - where required exceptions only
  - example: arrows show *all* open ssh ports
    - admin (=root) access from few DESY systems only
- limit impact of security flaws in software
- contain breaches

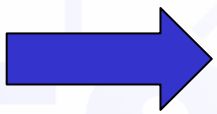
Backup

- **Core servers**
  - relevant data backed up (TSM)
- **AFS**
  - AFS backup tools dump to disk
    - butc, backup
  - TSM backs up disk to tape
- **Lustre**
  - no backup
  - fast scratch only



Support, Documentation

- **Technical aspects:**
  - **naf-helpdesk@desy.de -> UCO**
  - **naf@desy.de -> request tracker**
  - **mailing list naf-announce@desy.de (all registered users)**
- **Organisational aspects**
  - **The experiments must provide first level support**
    - **Filter user questions, several mailing lists already set up**
    - **Transmit fabric issues to NAF admins**
  - **DESY will provide second level support**
- **We NAF operators need feedback:**

from you  **NAF Users Committee**

http://naf.desy.de

Status and Time Schedule

- **Several parts already work: Grid-Batch, Grid-Storage, AFS, interactive environment (login,WGS), local batch**
- **Beta phase:**
  - **Finished for most components**
- **Public operation:**
  - **Started, how to continue -> next topic today**
- **Other components:**
  - **Lustre: test instance in operation**
  - **PROOF: under investigation**

Summary & Outlook

- **Key parts are in place**
  - **Public operation started in May**
  - **Additional parts in close cooperation with experiments**

Input from NAF User Committee welcome

Backup Slides