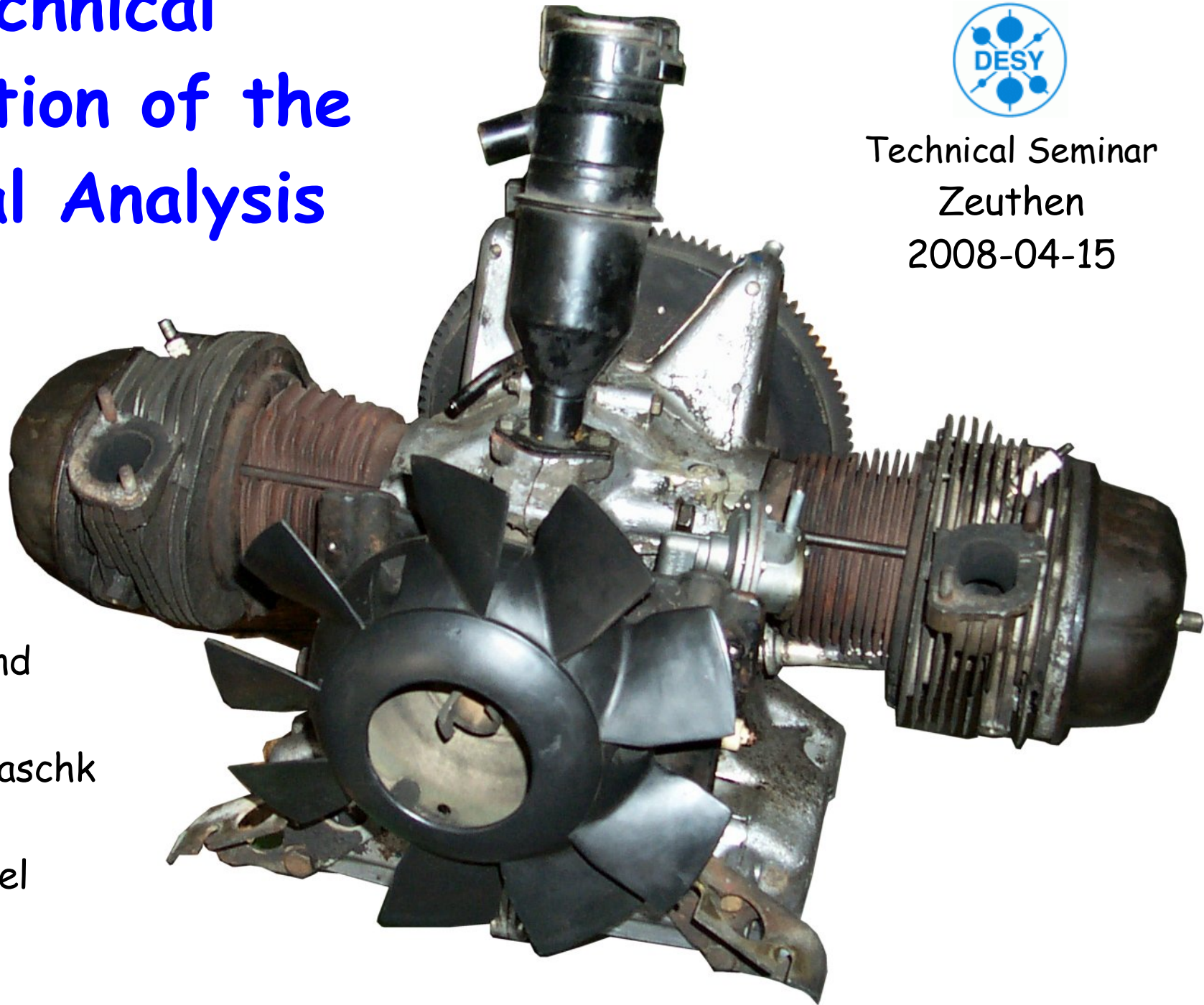


The Technical Realization of the National Analysis Facility



Technical Seminar
Zeuthen
2008-04-15



Stephan Wiesand

Waltraut Niepraschk

Andreas Haupt

Wolfgang Friebel

Kai Leffhalm

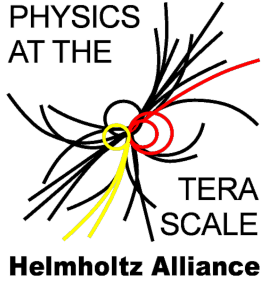
Götz Waschk

Peter Wegner

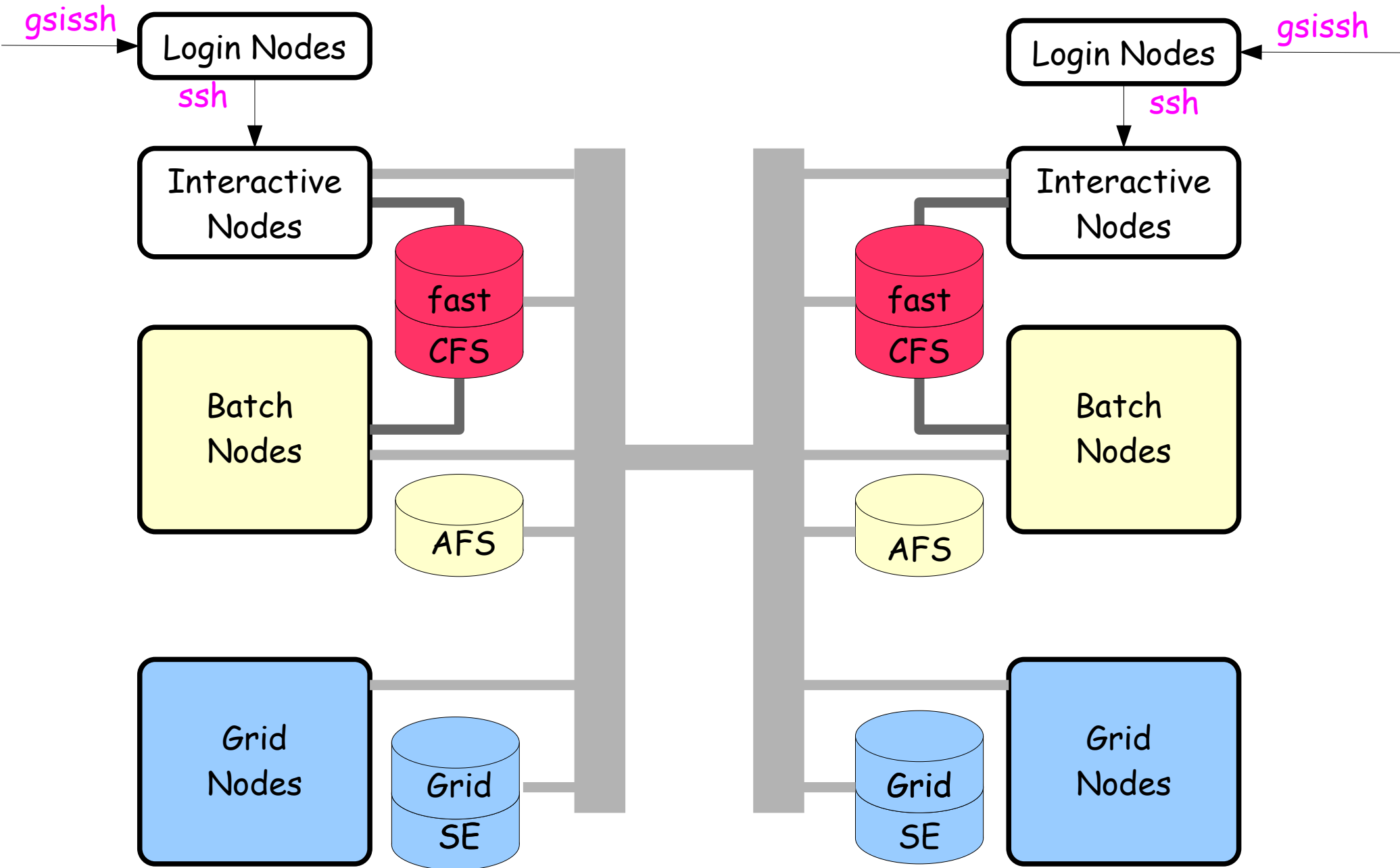
for the NAF team

The National Analysis Facility



- A data analysis facility for LHC & ILC physics
 - later: HERA
- Part of the proposal for the 
 - http://www.terascale.de/general_information/proposal
- Two components:
 - grid part
 - interactive & batch part
 - fast, predictable turnaround
 - user accounts, home directories, AFS access
 - additional fast filesystem
 - Distributed but unified facility: Hamburg, Zeuthen, ...?

The Big Picture



Excursion: The Grid Part



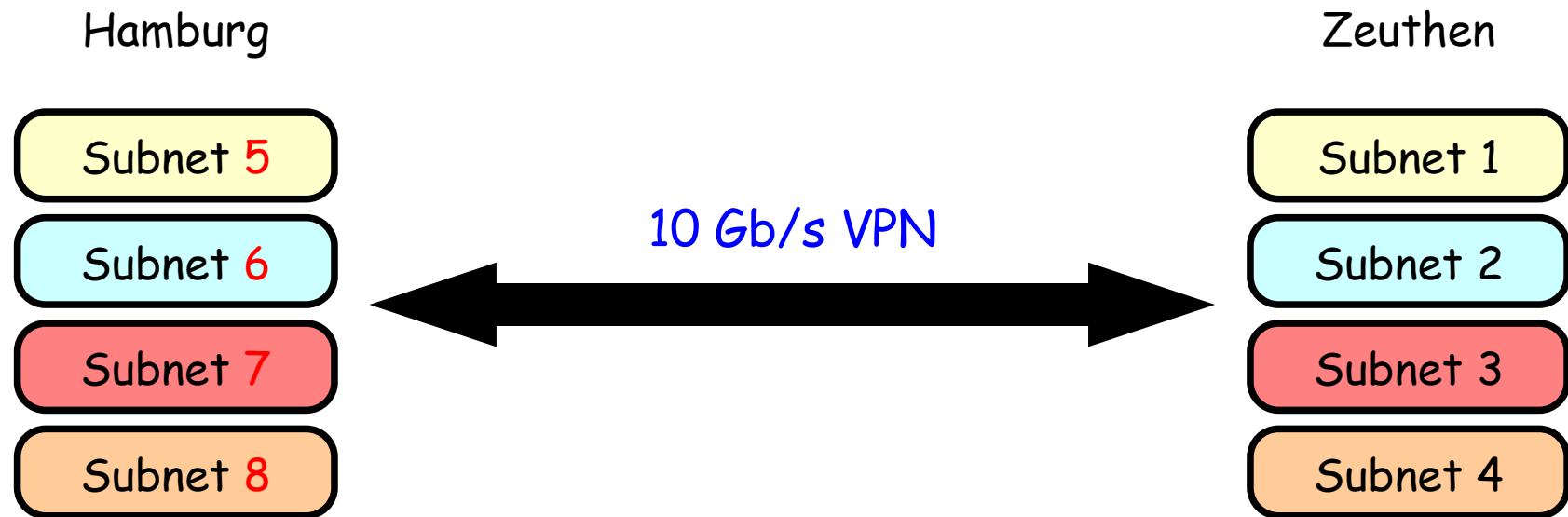
- initially planned as a separate grid site, with access to all NAF resources (file systems,...)
 - problem: would make it a yet another CE
 - problem: VO software installation, validation, tagging
 - problem: impossible to restrict access to NAF users
- => now simply an **extension to the existing Tier2**
 - using local administration methods in HH and Zn
 - exactly like existing nodes, no additional features
 - **dedicated shares** for german users via VOMS roles
 - ATLAS : CMS : ILC/LHCb = 1.5 : 1 : 0.5

The Interactive/Batch Part

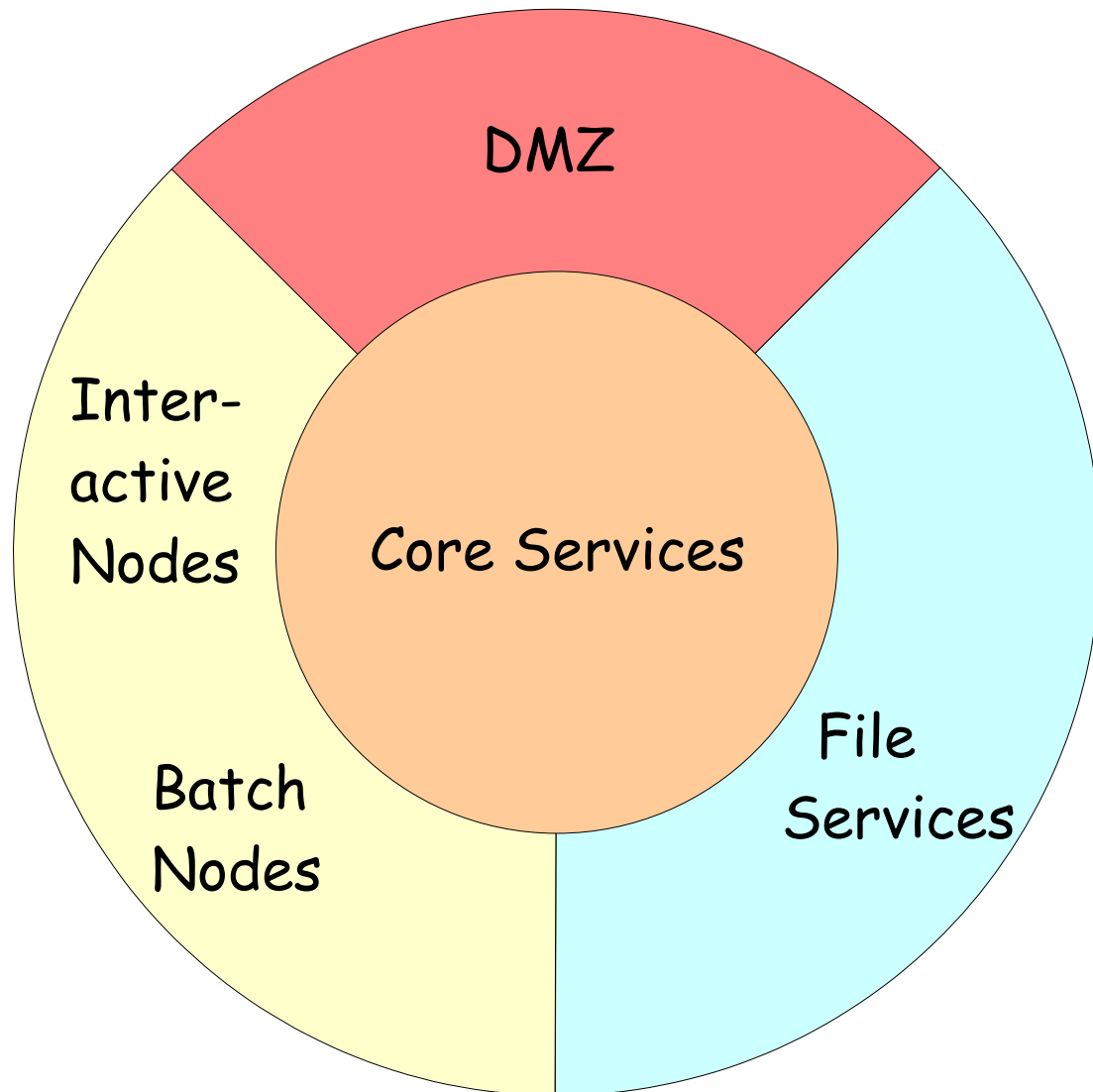


- New DNS domain naf.desy.de
- **AFS** cell & **Kerberos** Realm with same name
- NAF instance of DESY **user registry**
- NAF **platform adapter**
- **SGE** instance
- Dedicated NAF resources
 - **Worker/Interactive Nodes**
 - **AFS Fileservers** (home & group space)
 - **Lustre Fileservers** (bulk data, fast)
 - **Infrastructure servers**

naf.desy.de: Physical View

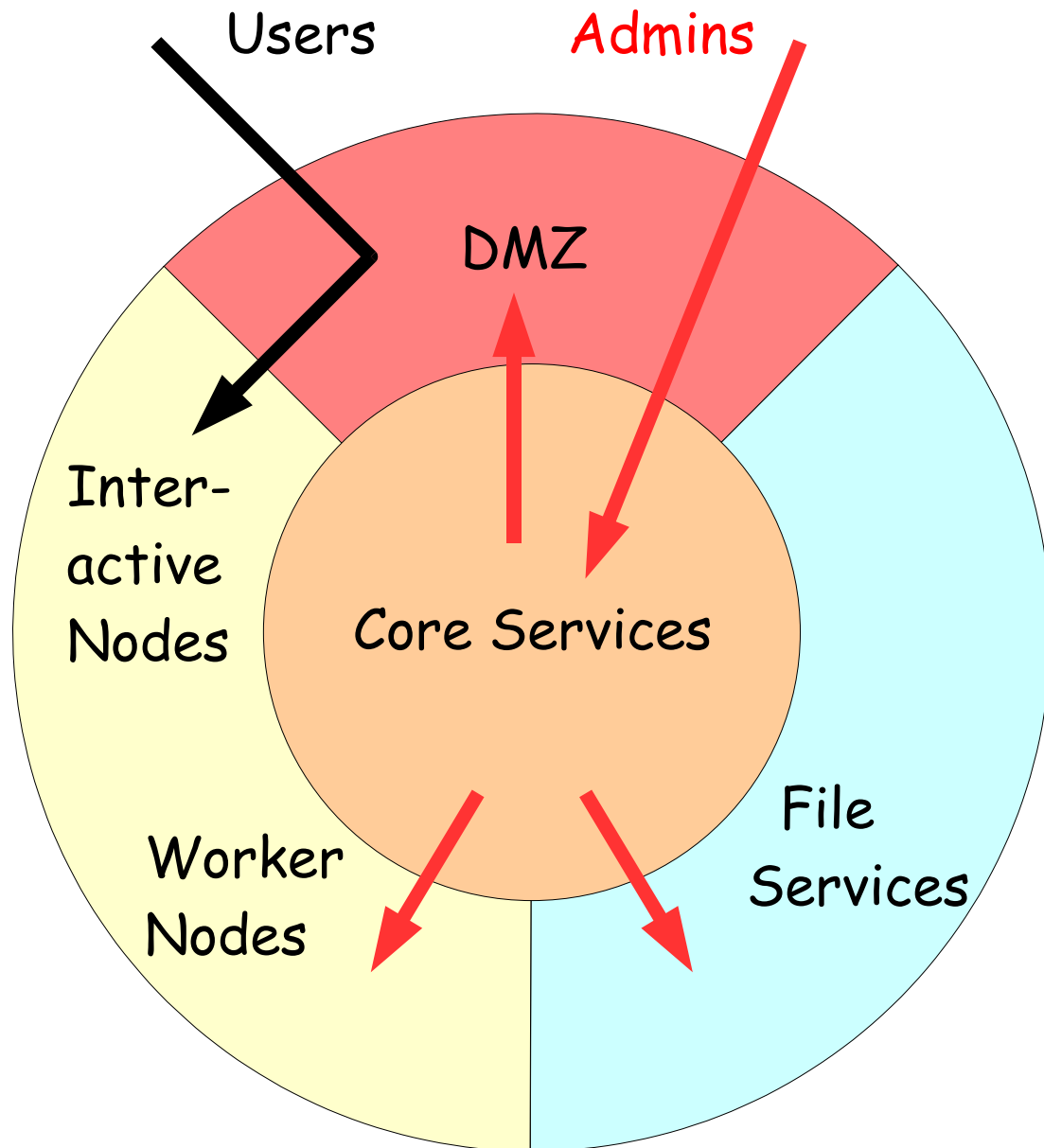


- packet round trip time: 5.3 ms
 - typical in LAN: < 0.2 ms, physical limit HH<->Zn: 2 ms
- all addresses are from 141.34.x.y, but
 - no layer2 subnets across the VPN link
 - different rules & parameters (gateway address,...)



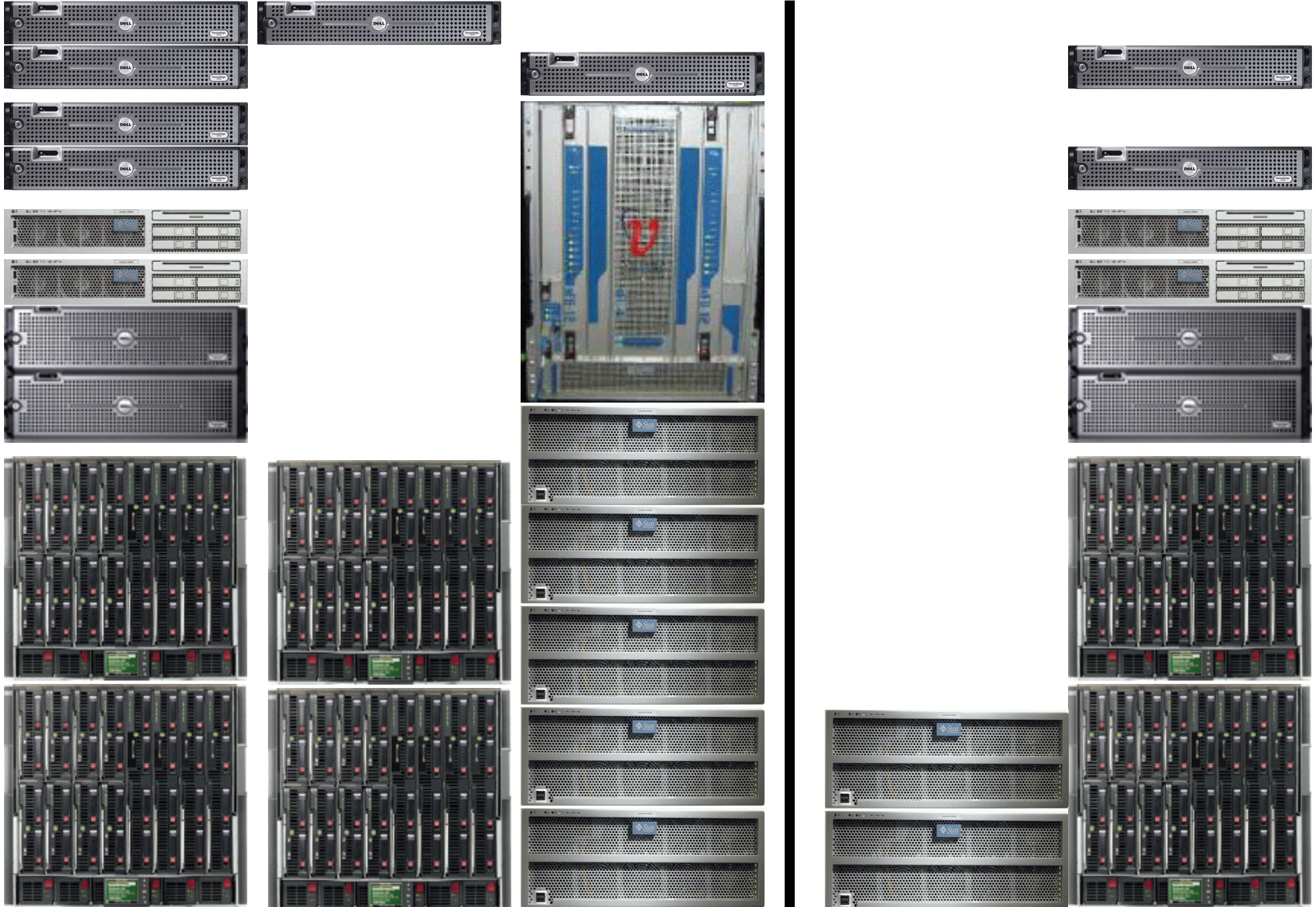
- 4 zones for different classes of systems
- Core services:
 - installation, configuration management, updates, monitoring, infrastructure (Kerberos, AFS, ...), admin access
- File services:
 - lustre

Access Restrictions -> Security



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

Hardware Resources



Infrastructure Servers



- Virtualization hosts; **all actual services are on VMs**
 - Dell Poweredge 2950
 - 2 x 4 Cores, 2.33 GHz (Clovertown), 8 GB RAM
 - 8 x 146 GB SAS Disk (2.5"), RAID-5,
 - 2 logical drives (system + data)
- SL 5.1, 64-bit, SELinux enabled, Xen Virtualization
- each server hosts **up to 5 virtual machines**
 - 2 + 1 Kerberos **KDCs**, 2 + 1 **AFS DB** servers
 - **batch** masters, **monitoring** servers, ...

DMZ Login Servers



- Hardware: identical to infrastructure servers
- actual login systems are VMs again (5 per server)
- dedicated login system required for each VO
 - due to gsissh access, (explained later)
- 4 supported VOs + support accounts / guests
 - => with redundancy, at least 10 systems required
 - not really feasible without virtualization
- Hosts & VMs: 64-bit SL5, SELinux enabled

AFS File Servers



- wanted: **ZFS** => **Solaris** => Sun X4200 Servers
 - 4 GB RAM, 2 x 2 Cores (AMD Opteron), LSI SAS HBA
- also wanted: SAS disks, not SATA
 - alas, no JBODs available from SUN
 - => Dell MD1000 Shelves, 15 x 146 GB SAS each
- **problem: getting the right cables** took a while
- "looking forward" to first service case...

Batch/Interactive Nodes



- 4 +2 HP BladeSystem c7000 enclosures, each with
- 16 HP BL460c Blades, each with
 - 2 x 4 Cores, 2.33 GHz (Intel Clovertown)
 - 2 x 146 GB SAS Disks (2.5"), RAID0 => 250 GB scratch
 - Infiniband HCA (10 Gb/s connection to fast storage)
- Supported: SL4 & SL5, 64-bit (32-bit not foreseen)



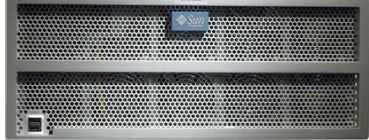
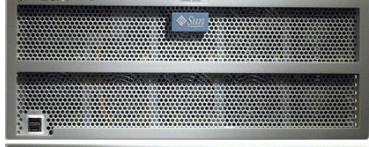
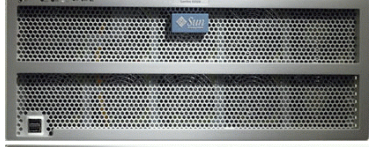
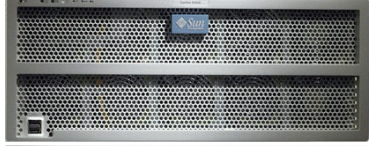
768 Cores
2 GB RAM/core
30 GB scratch/core



Storage Servers



- NAF should have a fast filesystem
 - chosen: Lustre, with Infiniband interconnect
- Dell 2950 as MDS
- SUN X4500 ("Thumper") as OSTs (16 TB each)
 - alas: many dead on arrival, problems under Linux, Solaris Lustre server delayed
 - Lustre FSs now hosted on *one* Thumper in HH available since last week, to HH nodes only
 - 1 Zn Server to be shipped back to HH (as dCache pool), 1 to be used for testing distributed dCache
 - to be replaced?
- 288 Ports IB switch, HH only

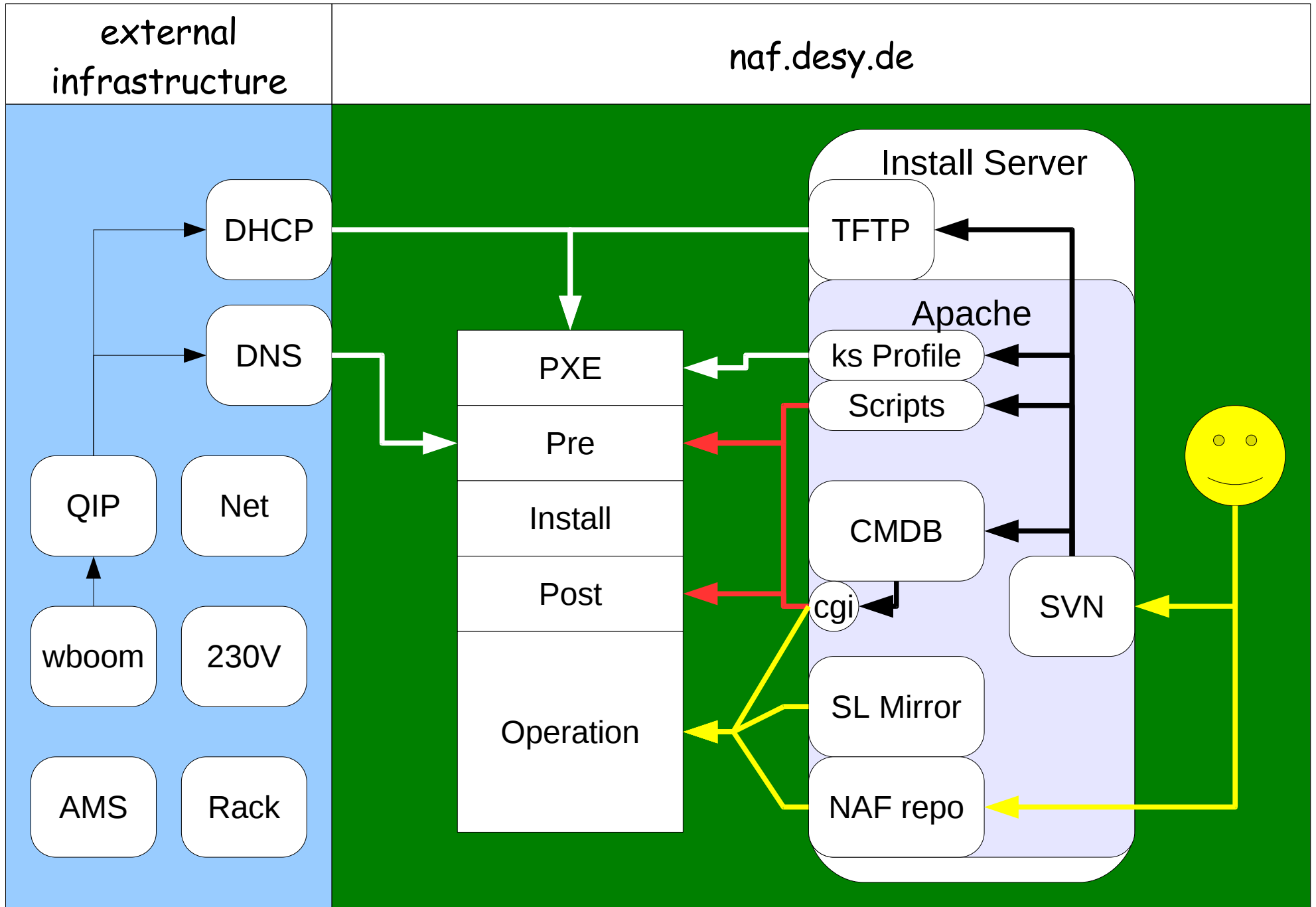


Linux Installation & Management



- another Dell 2950 (not virtualized)
 - installation services
 - SL repositories (mirrors)
 - NAF package repositories
 - Subversion repository
 - configuration database
 - installation scripts and configuration files
 - Apache for serving all this
- Linux management was implemented from scratch
 - lightweight solution, no framework used, just standard tools:
 - Subversion, Apache, RPM, YUM, and some glue scripts

Linux Installation & Management



Linux Installation & Management



- Admin interaction:
 - either deposits RPMs, or modifies files in subversion
 - automatic propagation to http/tftp areas upon checkin
 - the **configuration database** is just a text file:

```
# servers
tcsh{1..3} SL5.1_64 xenhost # Xen hosts for KDCs,...
tcsh7 SL5.1_64 is tsm cfe # install server

# external login systems (DMZ)
tcsh{5..6}-vm1 SL5.1_64 xen entrance @atlas # ATLAS
tcsh{5..6}-vm2 SL5.1_64 xen entrance @cms # CMS

# worker/interactive nodes
tcx{03..04}0 SL4.5_64 in ib lustre @atlas # alias atlas-wgs0{1,2}
tcx{03..04}1 SL4.5_64 in ib lustre @cms # alias cms-wgs0{1,2}
tcx035 SL5.1_64 in ib lustre @nafAfs # alias sl5-64 (public)

tcx0[36..3f] SL4.5_64 wn ib lustre @nafAfs # farm nodes
tcx0[45..4f] SL4.5_64 wn ib lustre @nafAfs # farm nodes
```

Linux Management with RPM



```
tcx{03..04}0    SL4.5_64  in  ib  lustre @atlas  # alias atlas-wgs0{1,2}
tcx0[36..3f]    SL4.5_64  wn  ib  lustre @nafAfs # farm nodes
```

- **Tags** define which RPMs should be installed (or not)
 - example: the WGSs above should have NAF_interactive, NAF_ib, NAF_lustre
- **RPMs** modify system configuration by
 - installing/removing files
 - running pre-/post-(un)installation scripts
 - running trigger scripts upon (un)installation of other RPMs
- RPMs may consult the tags (cached on system)
 - NAF_accounts cares for @atlas, @nafAfs (who has access?)
- ordinary **YUM updates** keep things current

More on System Management



- **global data** prepared and distributed from install server
 - /etc/hosts, ssh_known_hosts, ...
- **secure mechanism for providing secret keys** to new systems
 - ssh host keys, kerberos keytabs, host certificates, ...
 - admin authorizes one-time distribution
 - prepares tarball with keys
 - cgi script delivers to the correct IP address only
 - and then deletes the tarball
- **Solaris** management (not yet finished) uses all these as well
 - except, obviously, RPM - instead: native **PKG**
 - complemented by **cfengine** where PKG lacks functionality

Monitoring



- Nagios

- by HH operating
- ping and ssh
- physical systems only, no VMs

- Hobbit

- inside NAF
- much more data
- automatic configuration
- on its way

- Host based

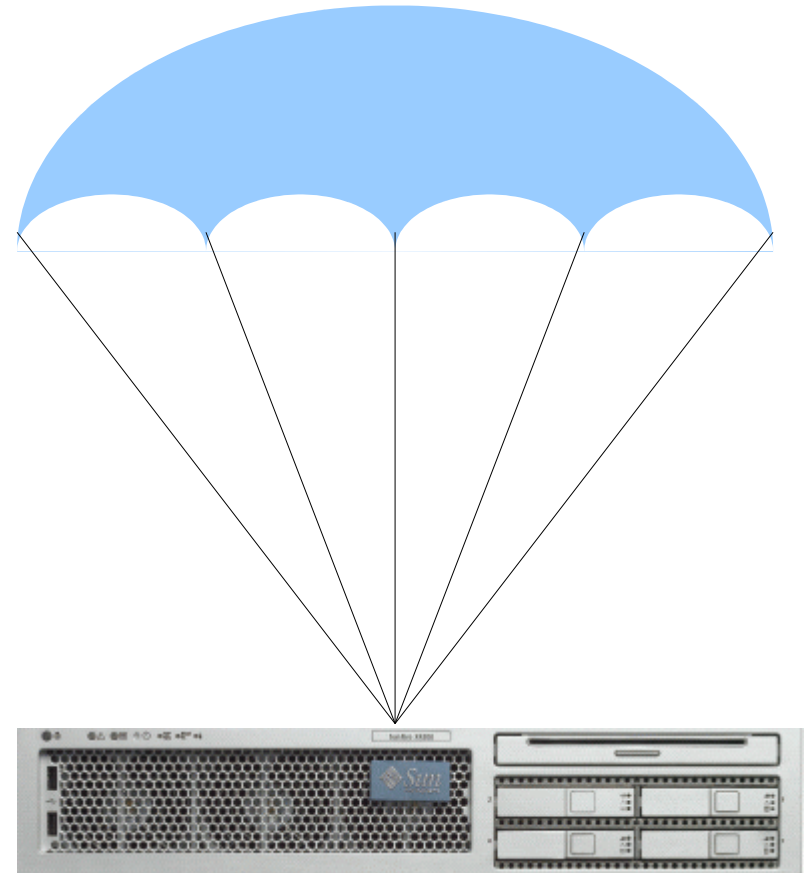
- hardware, RAIDs, ZFS monitored with (vendor) tools



Backup



- Core servers: relevant data backed up (TSM)
- **AFS**: planned, not yet implemented
- **Lustre**: not planned (?)



User Administration



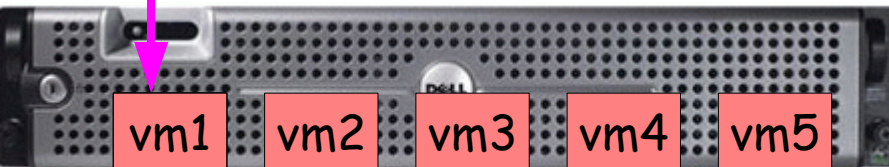
- standalone registry instance
 - identity: from user's grid certificate
- platform adapter provides account data in native format
- all account data stored locally on each system
 - no directory service (like NIS, LDAP)
 - authorization: derived from host's tags in configuration db
- authentication: Kerberos 5 - no passwords (needed)
- inside NAF: passwordless ssh using GSSAPI
 - AFS token from Kerberos 5 ticket
- ssh login from outside: passwordless gsissh

NAF Login with gsissh



```
grid-proxy-init -rfc  
gsissh atlas.naf.desy.de
```

gsissh



ssh



qsub



Interactive Node

Batch Node

- rfc compatible proxy
 - standard with Globus Toolkit 4
 - gLite default: GT3
- Krb5 ticket & AFS Token generated from proxy certificate
 - per system, DN can only be mapped to one account
 - => 1 system/VO required
- VMs are not for work, just login
- hop to IN will eventually be automatic (-> transparent)

Batch



- SUN Gridengine 6.1u4
 - setup similar to Zeuthen farm
 - SL4 64-bit supported only yet
 - project membership completely identical to unix group membership
 - AFS token provided for all jobs



Storage: AFS

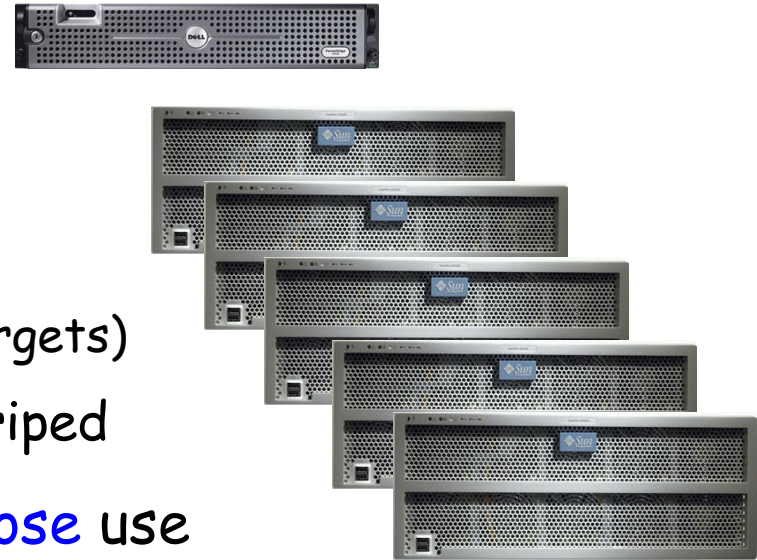


- space administration using enhanced `afs_admin`
 - volumes grouped into `volume sets` (`projects` and `subprojects`)
 - definition of storage `pools` (vice partitions) where a volume set can reside
 - => allows clustering of projects into separate partitions
 - `quota` is assigned to projects and subprojects
 - (PTS) `groups of project admins`
 - project admins can contact the AFS `admin server`
 - server grants or denies admin task according to group membership
 - `simplified AFS administration` (high level admin commands)
 - admin does not need to know about AFS details
 - `server enforces policies` (project quota, mount point convention, ...)

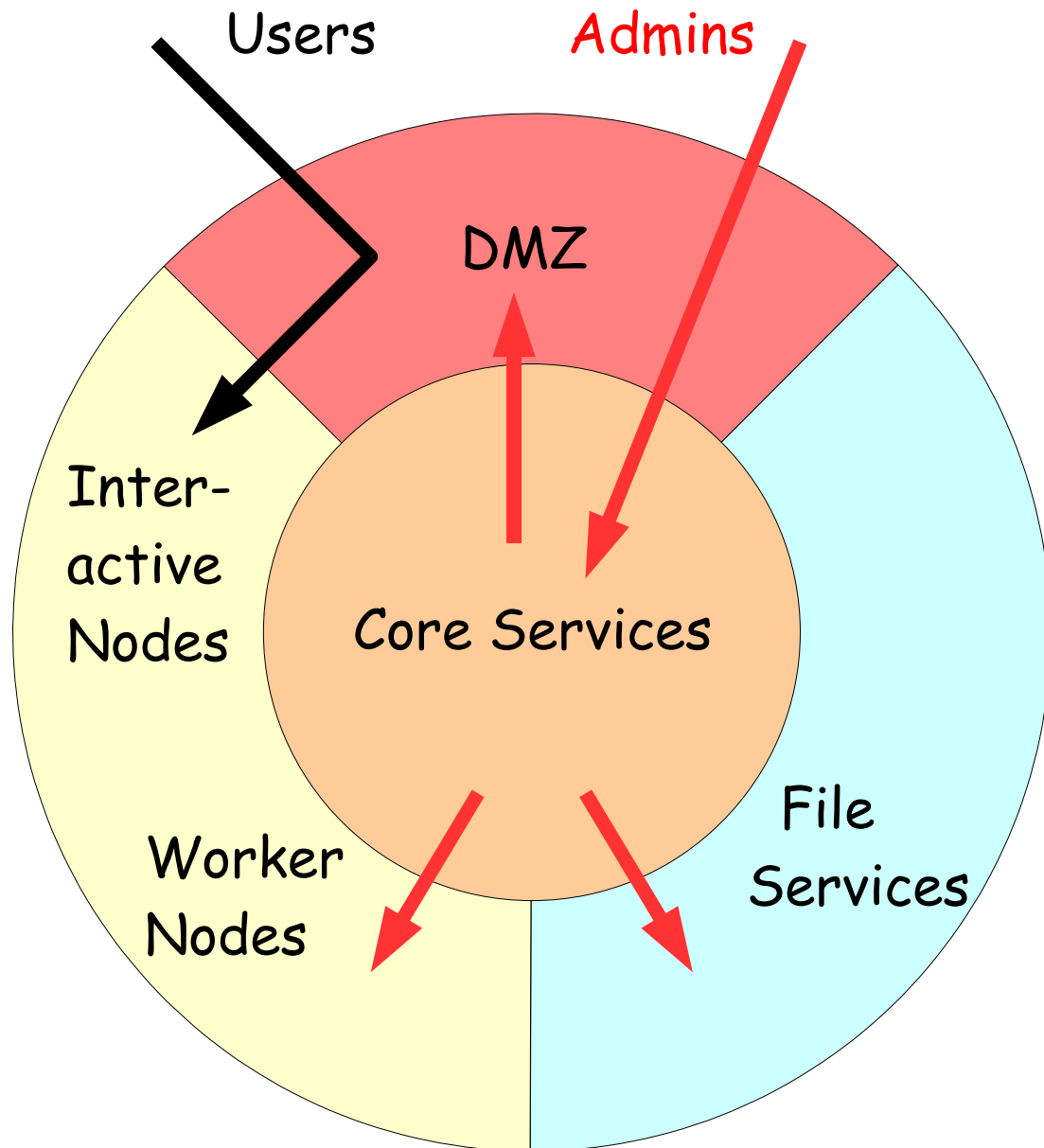
Storage: Lustre



- "Scale out" filesystem
 - 1 head node (MDS, metadata server)
 - n OSSs (object storage servers)
 - each with ≥ 1 OSTs (object storage targets)
 - files are distributed round robin or striped
- just becoming usable for general purpose use
 - no longer needs a special kernel on clients
 - now has ACLs (already usable?), Quota (usable soon?)
 - currently both disabled
- can use several interconnects (TCP, IB) - even simultaneously
- no concept for a cross site Lustre based storage architecture yet
 - installation in HH will not be available via TCP (-> in Zn)
 - expert says it's too insecure (no export restrictions)

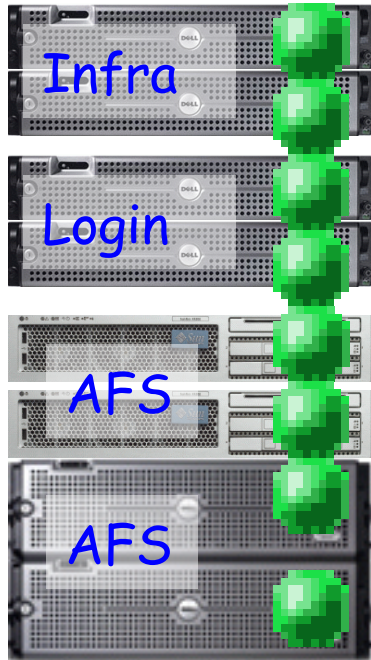


Access Restrictions -> Security

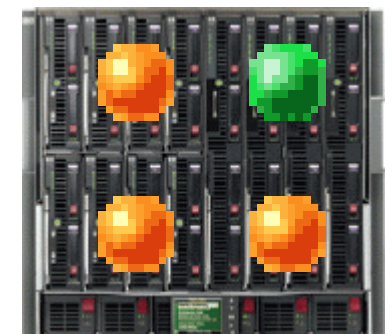


- by default: all network ports closed on all zone boundaries
 - exceptions only where required
 - example: arrows show all open ssh ports
 - admin (=root) access from few DESY systems only
- limit impact of security flaws in software
- contain breaches

Deployment Status



- running
- ready
- not ready, or no longer dedicated to NAF



Summary



- the NAF is a significant facility
 - built from scratch, legacy free
 - many new concepts & techniques
 - login with grid certificate
 - lightweight system management
 - most servers virtual
 - latest OS wherever possible
 - testing new fast filesystem
 - alas, local to site
 - problems with hardware (or hardware/OS combination)
- deployment well advanced
 - users testing

