Implementation of the National Analysis Facility at

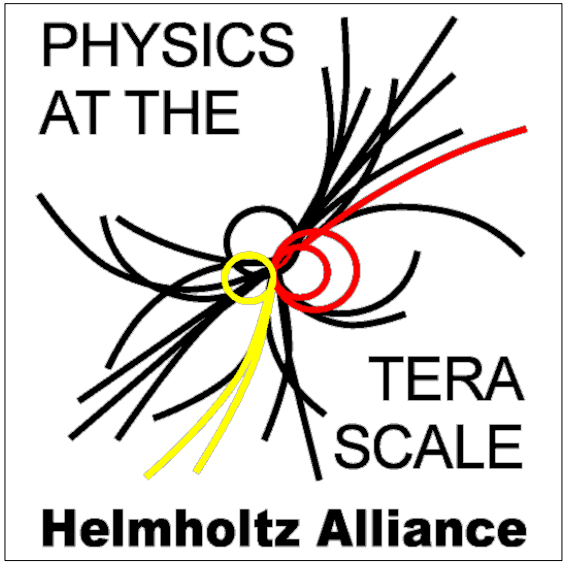
2008-05-08

Stephan Wiesand
DESY - DV -

for the NAF team

The National Analysis Facility

- one of the work packages in the proposal for the terascale alliance
  - see http://terascale.de
    - 17 german universities
    - 2 Helmholtz centers
    - 1 associated Max Planck Institute
  - NAF participating partners: DESY, FZK, University of Karlsruhe, University of Göttingen, LMU Munich, HU Berlin
- additional resources for batch & interactive work
  - for german physicists working on LHC & ILC physics
- starting out at DESY's sites in Hamburg and Zeuthen
  - to be distributed over more sites later if and as required
    - depending on advances in network infrastructure



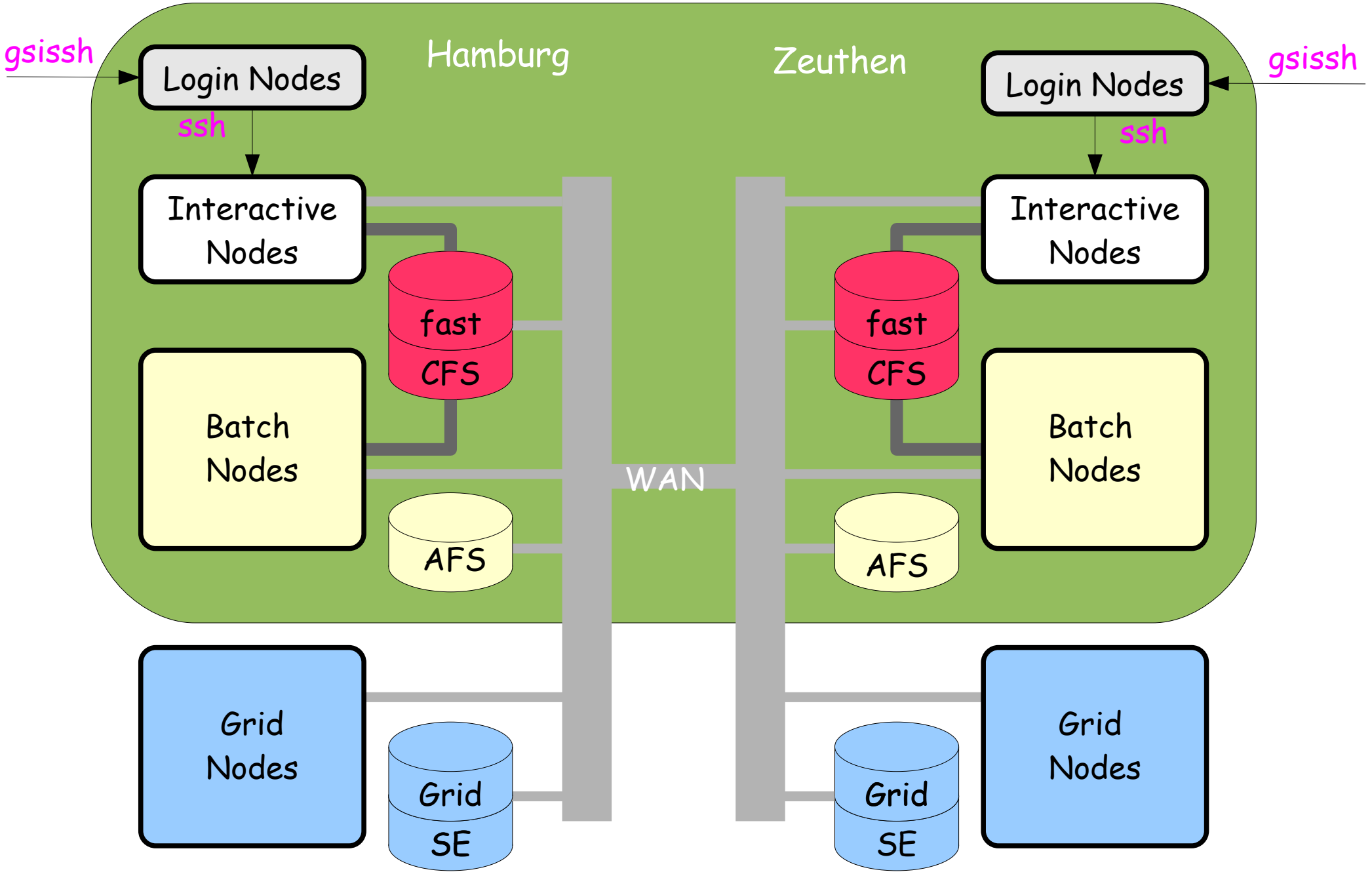
Requirements

- papers received from german ATLAS and CMS groups
- items include:
  - interactive login
  - batch resources
    - grid + "plain old" batch system
  - user & group storage, AFS client
  - high capacity / high bandwidth storage
  - AOD, some ESD; access to all data stored on Tier2
  - TAG database
  - facility for parallel interactive analysis (PROOF)
  - **flexible assignment of resources**

Planned Resources

- about 1.5 average Tier2 centers
  - with a special focus on data storage
- starting with two distinct components:
  - grid part
  - interactive & batch part
    - fast, predictable turnaround
    - user accounts, home directories, AFS access
    - additional fast filesystem

The Big Picture



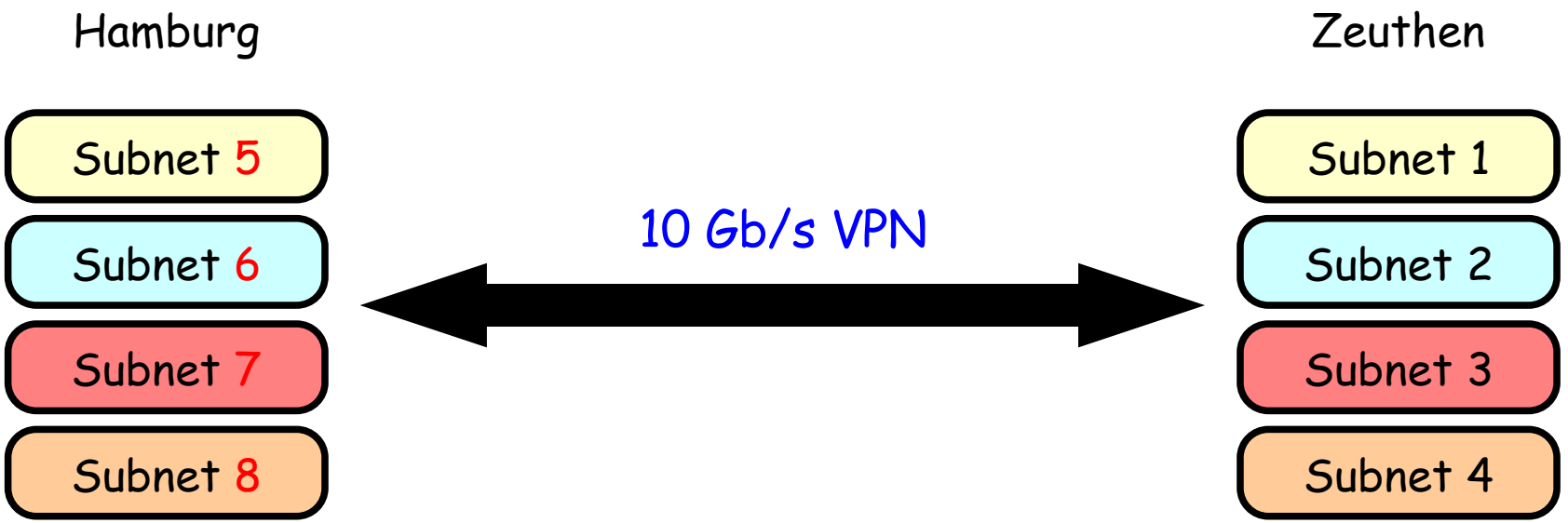
The Grid Part

- 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 site local administration methods
  - exactly like existing nodes, no additional features
  - dedicated shares for NAF users via VOMS roles
  - dedicated space on (dCache) SE

The Interactive/Batch Part

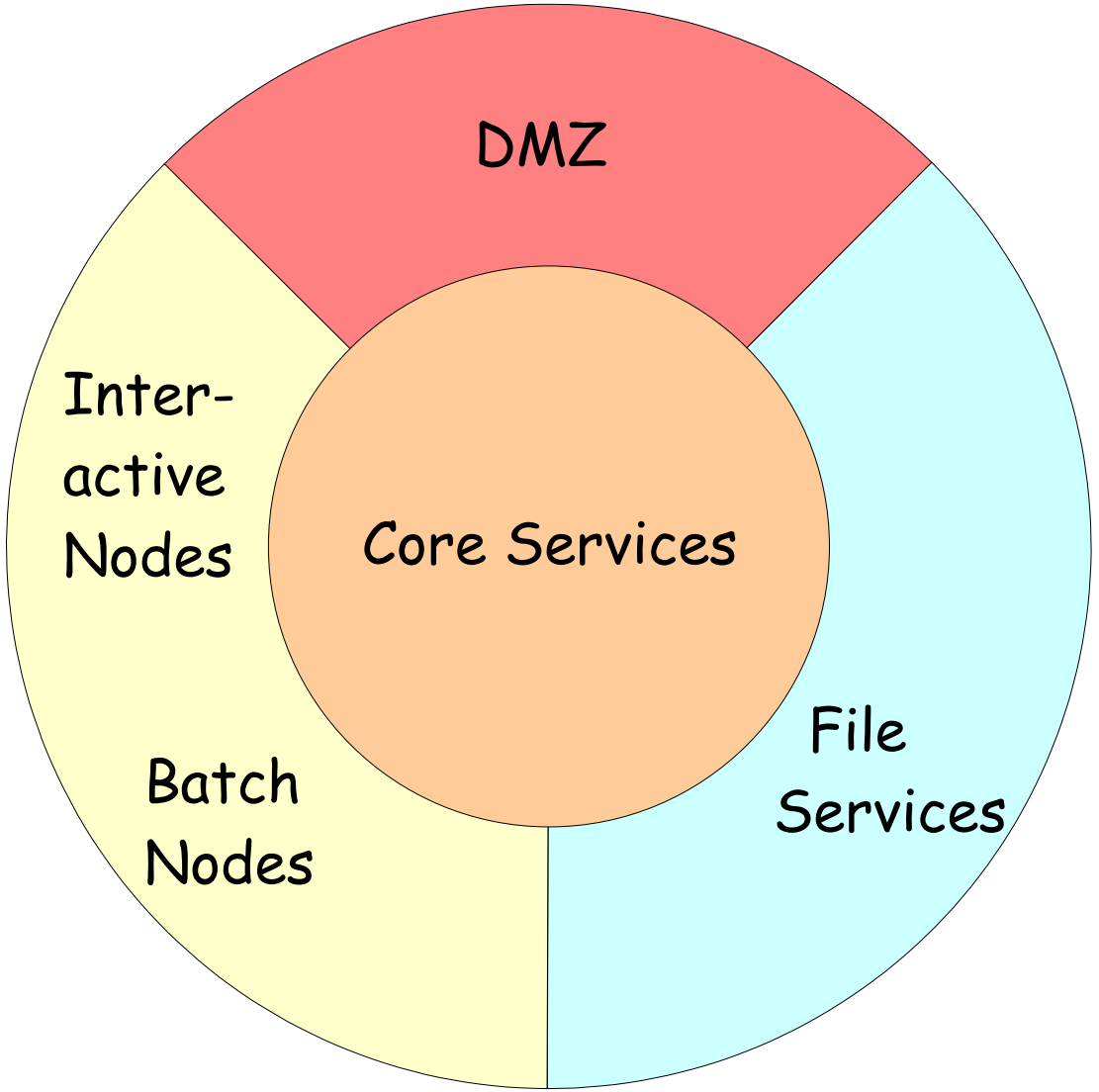
- **unique chance to build in open countryside**
  - new DNS domain naf.desy.de
  - AFS cell & Kerberos Realm with same name
  - NAF instance of DESY user registry
  - NAF platform adapter
  - SGE instance
  - Dedicated NAF resources
    - Worker/Interactive Nodes
    - AFS Fileservers (home & group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

naf.desy.de: Physical View



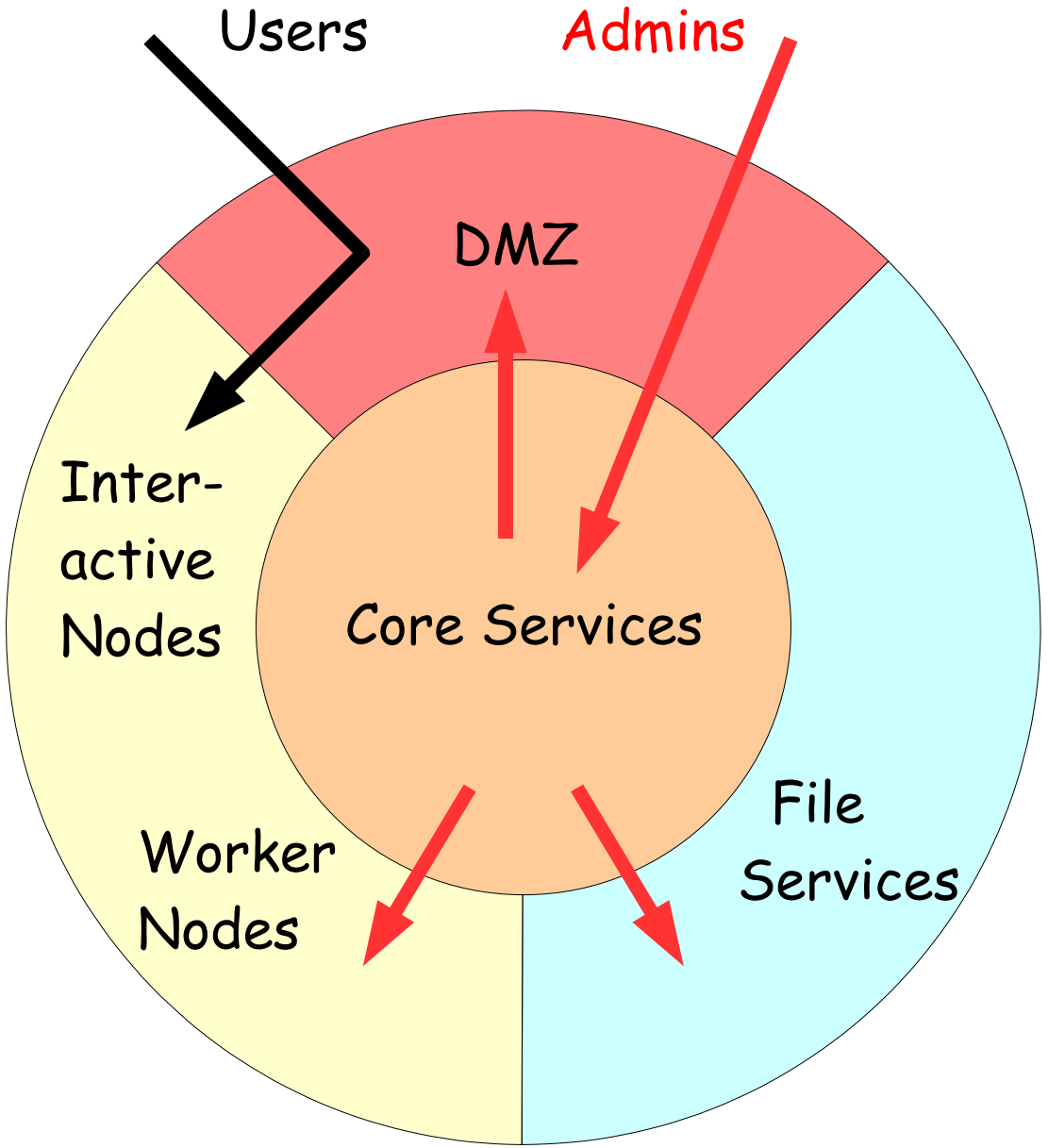
- packet round trip time: 5.3 ms
  - typical in LAN: < 0.2 ms, physical limit HH<->Zn: 2 ms
- all addresses are from 141.34.x.y (Zeuthen range), but
  - no layer2 subnets across the VPN link
  - different rules & parameters (gateway address,...)

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

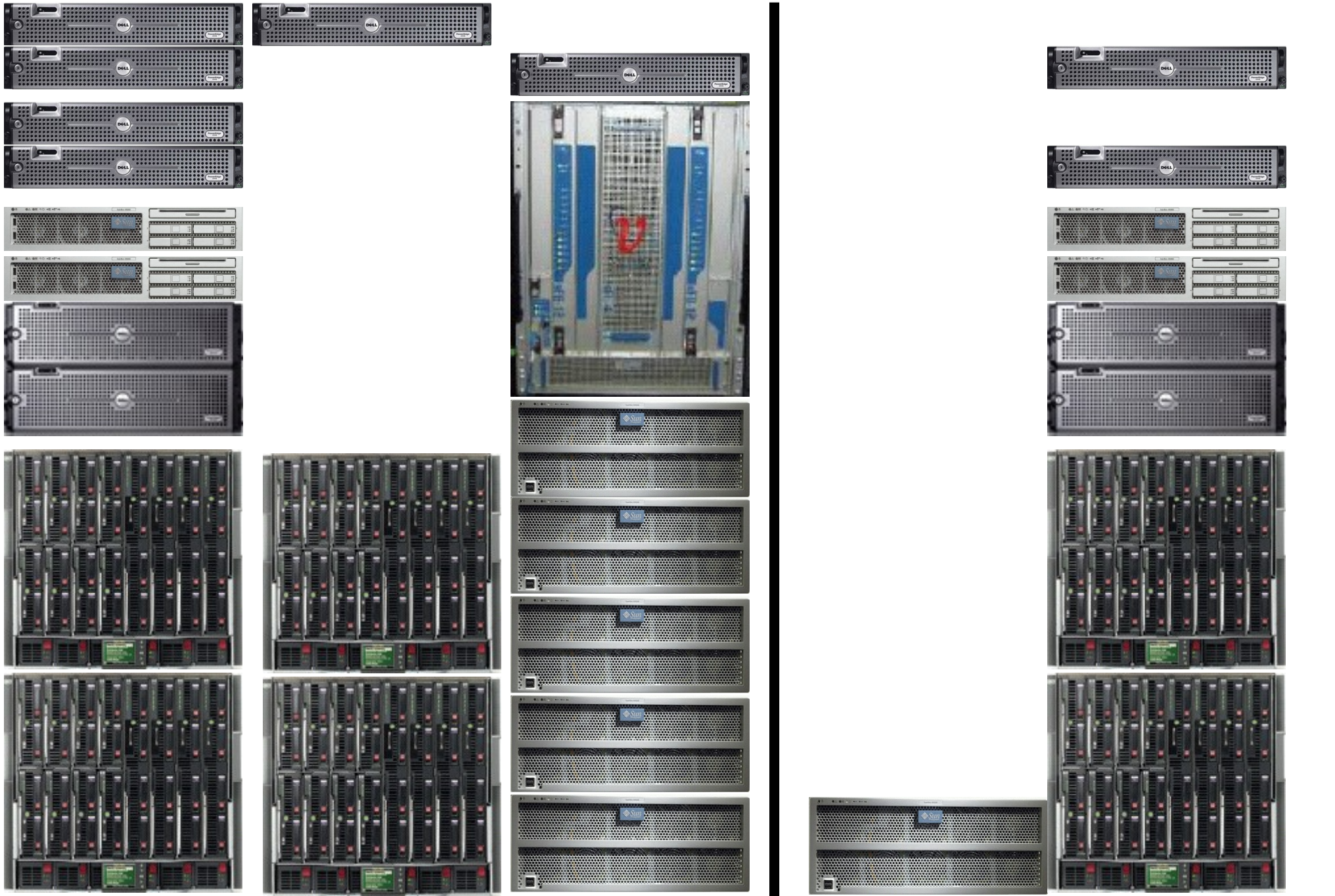


- by default: all network ports closed on all zone boundaries
  - exceptions only where required
  - example: arrows show *all* open ssh ports
    - admin (=root) access from few DESY systems only
- limit impact of security flaws in software
- contain breaches

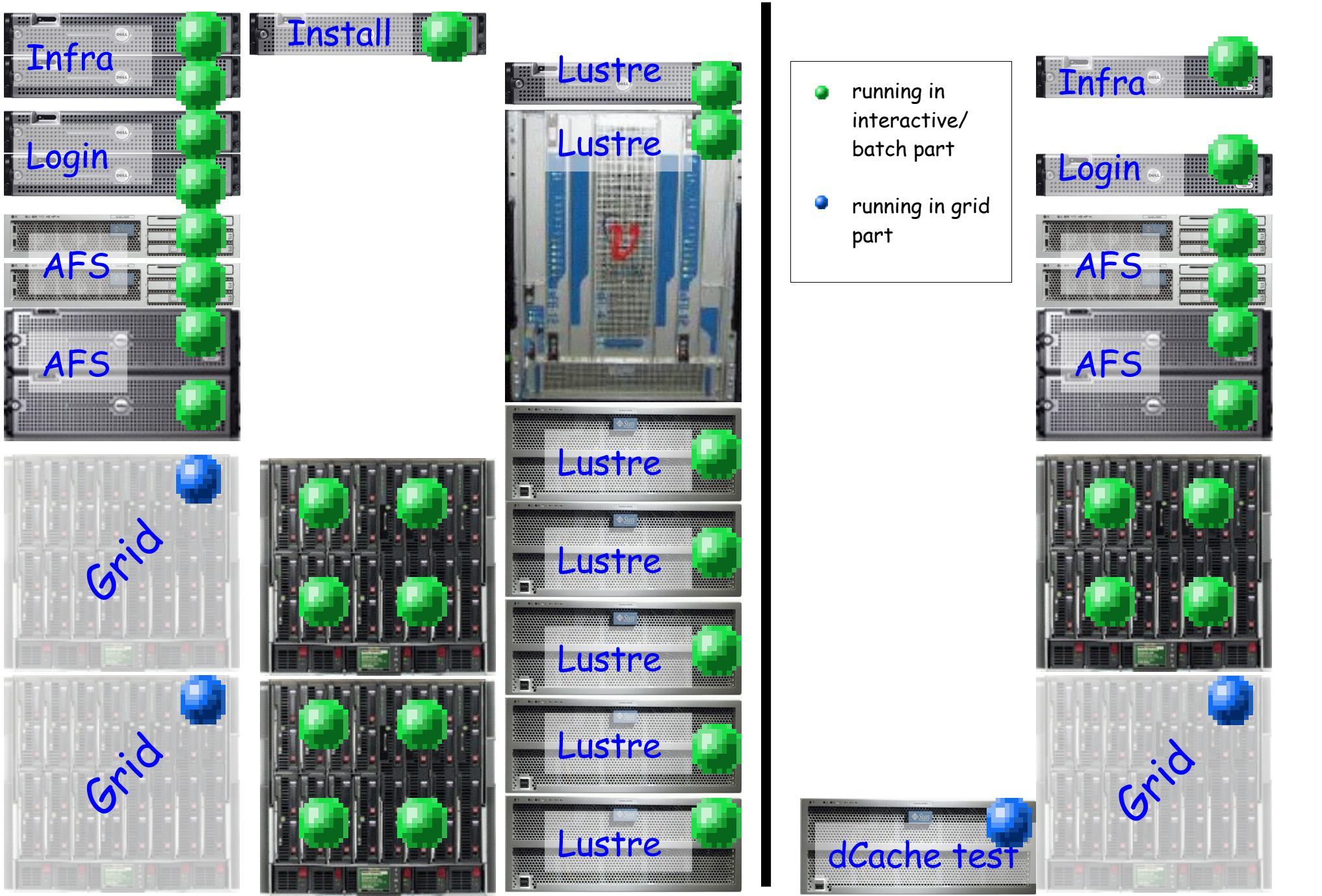
SPRING HEPiX 2008

Hardware Resources

DESY



Usage & Deployment Status



Virtualization Workhorse

- 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
- almost all services are on VMs (64-bit SL5.1 paravirt)
  - 2 + 1 (HH + Zn) Kerberos KDCs, 2 + 1 AFS DB servers
  - batch masters, monitoring servers, ...
  - DMZ login systems (each dedicated to one supported VO)

AFS File Service

- for user & group storage
- wanted: ZFS => Solaris => Sun X4200 Servers
  - 4 GB RAM, 2 x 2 Cores, LSI SAS HBA
- also wanted: SAS disks, not SATA
  - alas, no JBODs available from SUN (@ target cost/capacity)
  - => Dell MD1000 Shelves, 15 x 146 GB SAS each
- works very well
  - problem: getting the right cables took a while
  - "looking forward" to first service case...
- space administration delegated to users
  - afs_admin (by Wolfgang Friebel, see earlier HEPiX meetings)

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
- Supported: SL4 & SL5, 64-bit (32-bit not foreseen)

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

- batch: SGE 6.1u4
  - SL4 64-bit only yet
  - project membership derived from unix group membership
  - AFS token provided for all jobs (arcx/kstart)

High Bandwidth Storage

- chosen: Lustre, with Infiniband interconnect
  - Dell Poweredge 2950 as MDS
  - SUN X4500 ("Thumper") as OSSs (16 TB each)
    - many dead on arrival
    - uOSS (Usermode , Solaris) server delayed (Q4/08?)
    - initial stability problems under Linux
      - seem solved with latest kernel + patches
    - SL4 only (no SUN support for EL5) - marvell driver
  - 288 Ports Voltaire IB switch
    - no IB switch/cables & servers in Zeuthen yet
- no [concept for] (cross site) TCP access yet
  - IB+TCP on servers? routing nodes? tunnel?
    - will this work? not yet tested

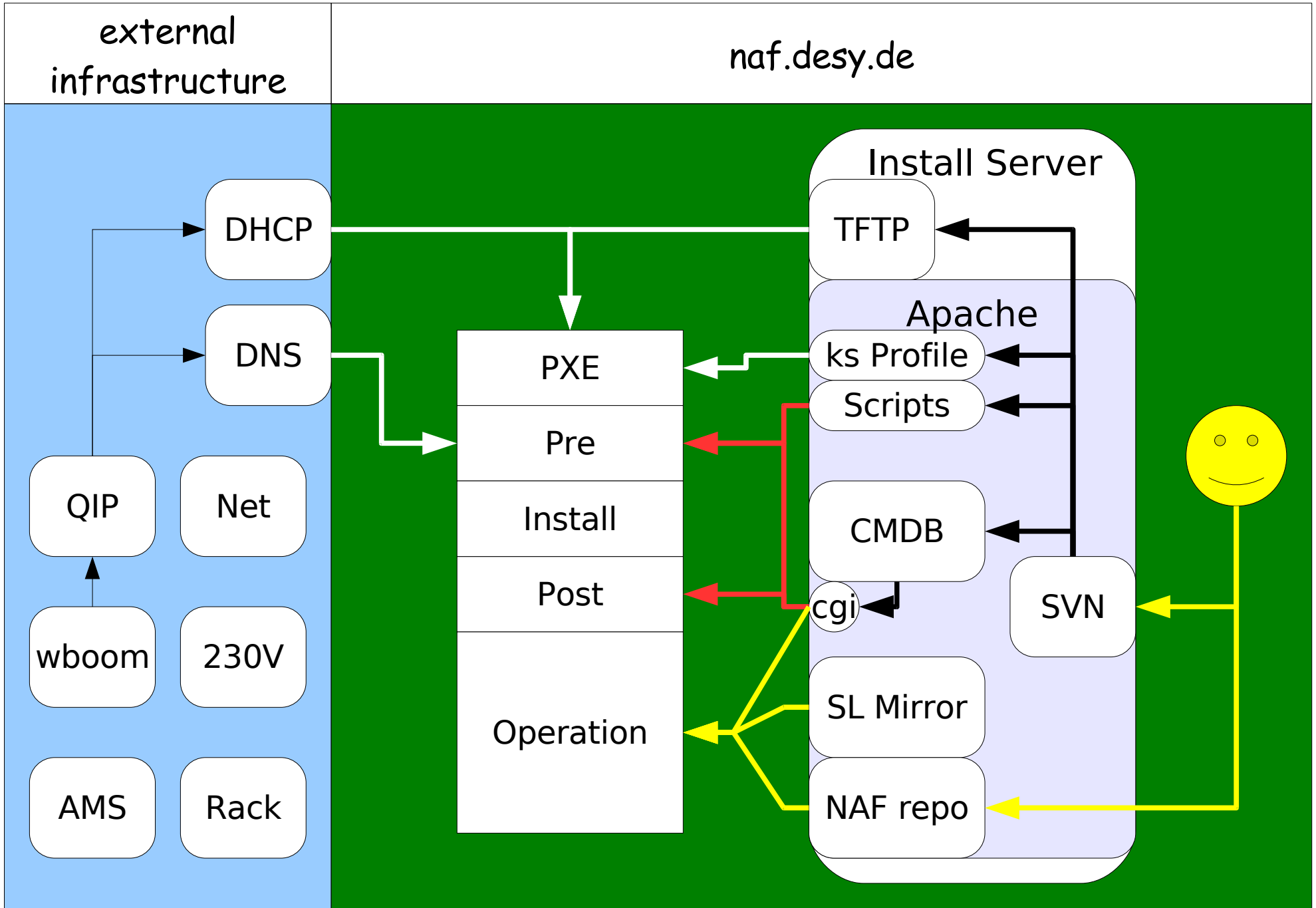
SPRING HEPiX 2008

Linux Installation & Management

DESY

- implemented from scratch
- lightweight solution, no framework used, just standard tools:
  - Subversion, Apache, RPM, YUM, and some glue scripts
  - configuration database is a simple text file in subversion
    - no overlap with existing databases
      - asset management, QIP (DNS/dhcp)
        - common key: hostname
- single install/management server for both sites
  - works very well, even dhcp & tftp between sites
- simple commands for
  - VM installation
  - Worker/Interactive Node reinstallation

Linux Installation & Management



Linux Installation & Management

- admins either deposit RPMs, or modify files in subversion
  - automatic propagation to http/tftp areas upon checkin
  - example: configuration database:

```
# 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
```

  - cgi interfaces for node data & expanded global DB

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

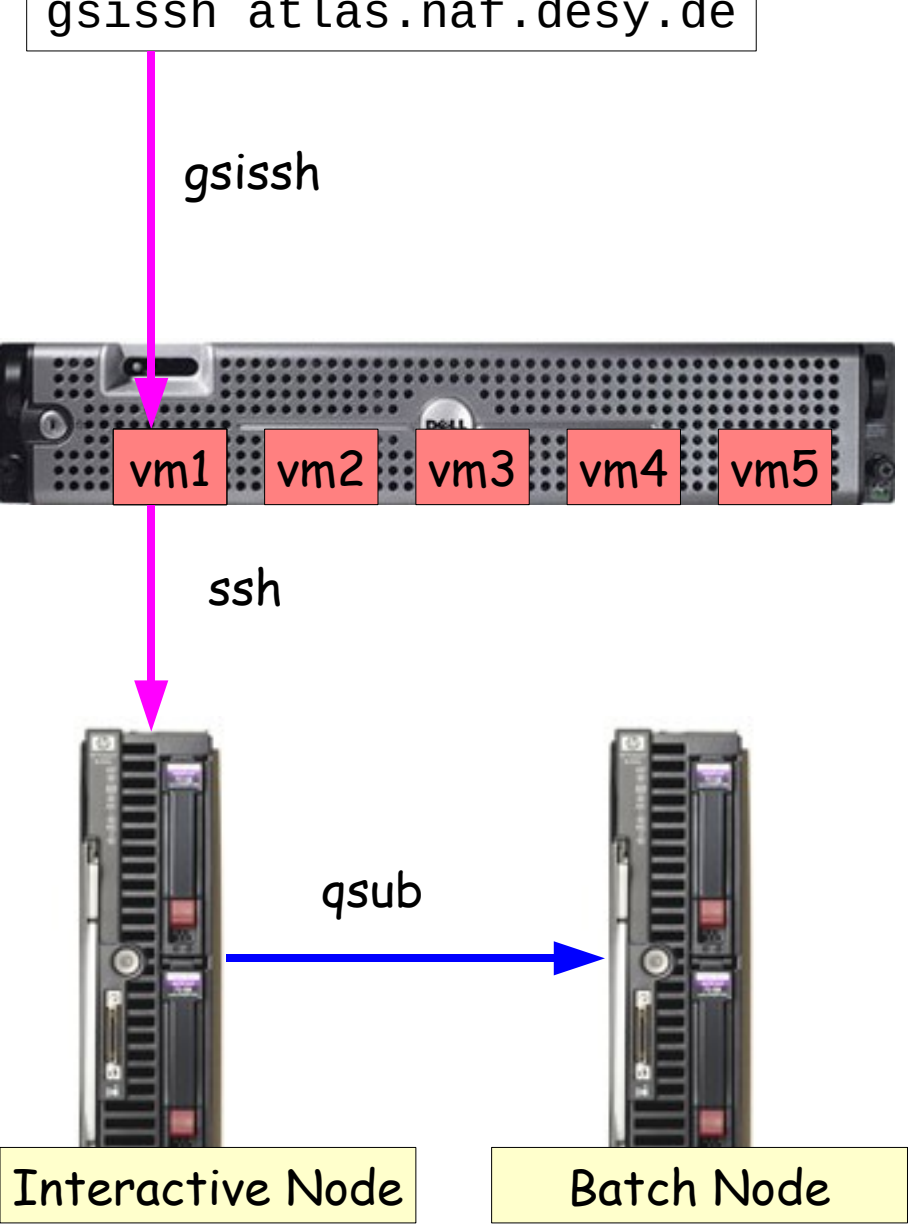
- global data prepared and distributed from install server
  - /etc/hosts, ssh_known_hosts, ...
- secure mechanism for providing secret keys to new systems
  - ssh host keys, kerberos keytabs, host certificates, ...
  - admin authorizes one-time distribution
    - prepares tarball with keys
  - cgi script 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
  - augmented by cfengine where PKG lacks functionality

User Administration

- standalone NAF 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
```



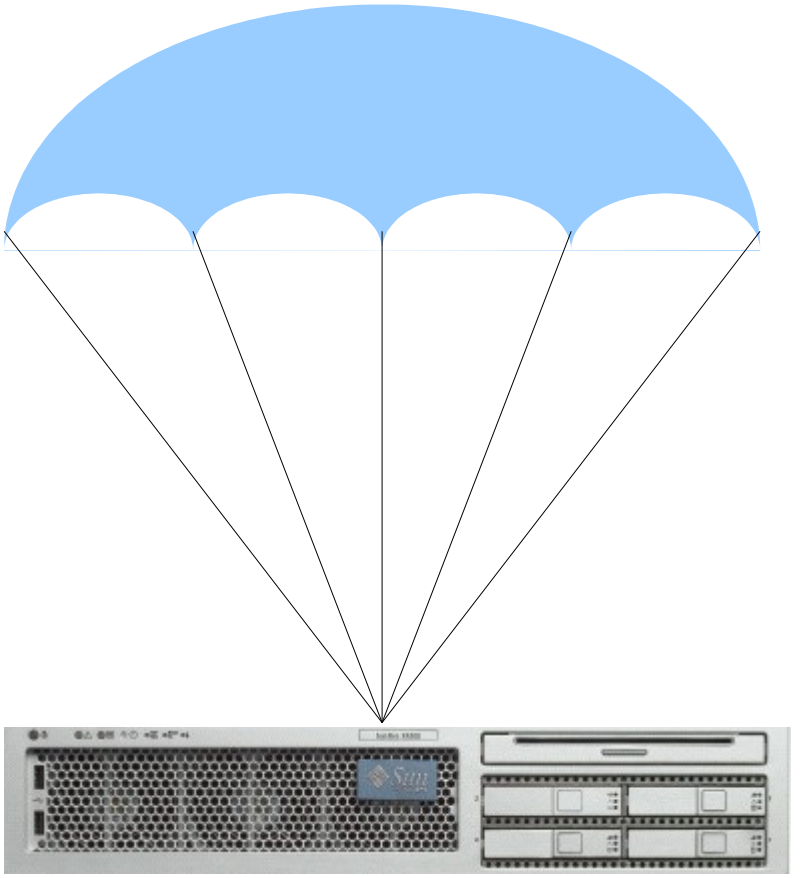
- rfc compatible proxy
  - standard with Globus Toolkit 4
    - gLite default: GT3
- Krb5 ticket & AFS Token generated from proxy certificate
  - per system, DN can only be mapped to one account
  - => 1 system/VO required
- VMs are not for work, just login
- hop to IN will eventually be automatic (->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

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
  - AFS backup tools dump to disk
    - butc, backup
  - TSM backs up disk to tape
- Lustre
  - no backup
  - fast scratch only



Summary

- implementing the NAF was a chance to design and build a significant facility from scratch
  - legacy free
  - using many new concepts & techniques
    - login with grid certificate
    - lightweight system management
    - most servers virtual, latest OS wherever possible
    - new fast filesystem
- deployment well advanced, users testing & working ("beta")
- future developments (depending on funding/HR):
  - enable cross site Lustre access via TCP
  - no idea yet how to run a sizable PROOF facility