Grid, NAF, Lustre

The DESY Grid infrastructure: storage, batch, services

The NAF: National Analysis Facility for German LHC community

Lustre or our experience with large file store

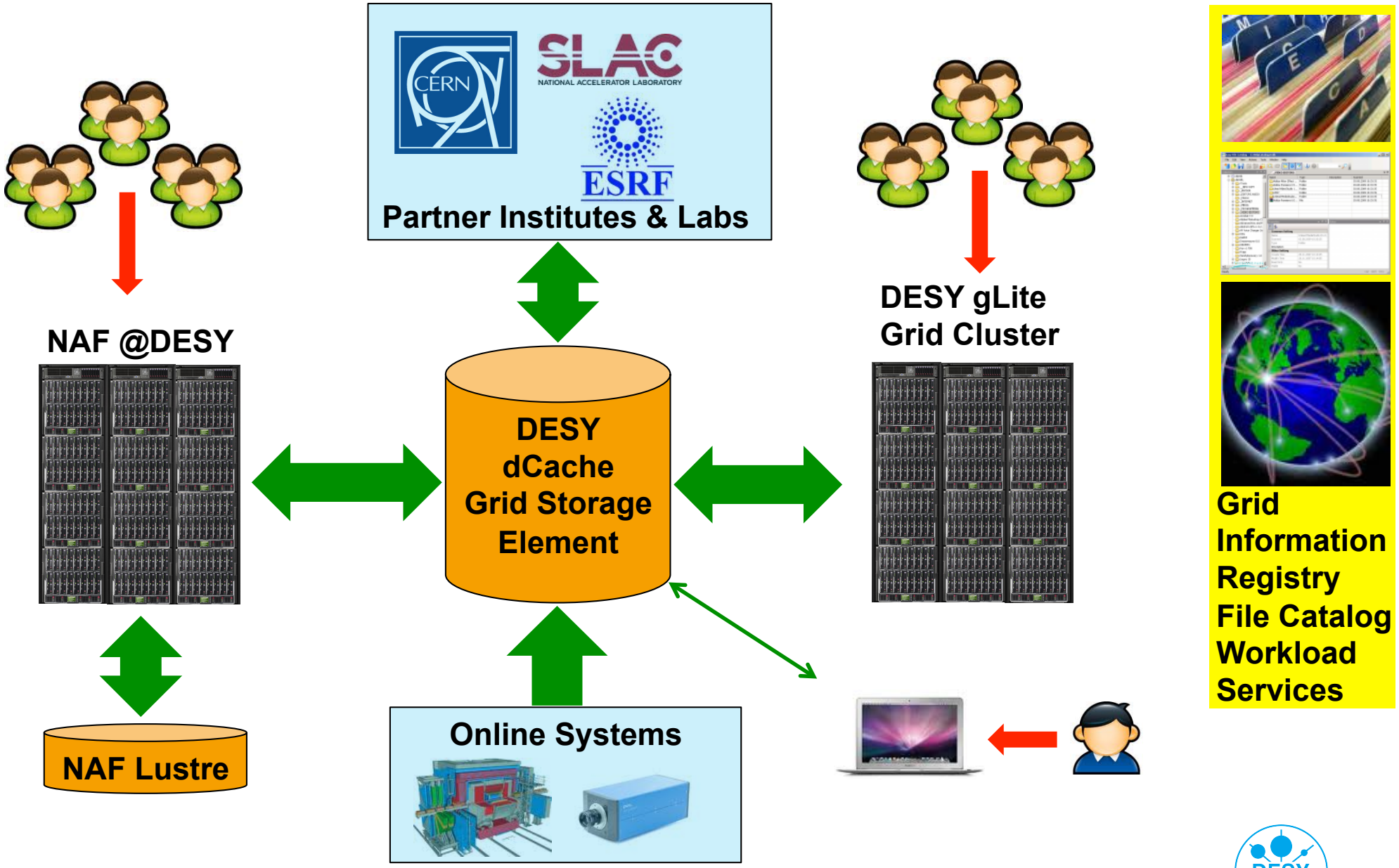
Yves Kemp, DESY IT
Photon Science Workshop 6.7.2010

Presenting work from:
-DESY Grid Team
-DESY dCache Operations Team
-NAF Team

HELMHOLTZ ASSOCIATION

DESY

A view of the DESY Grid Center Infrastructure



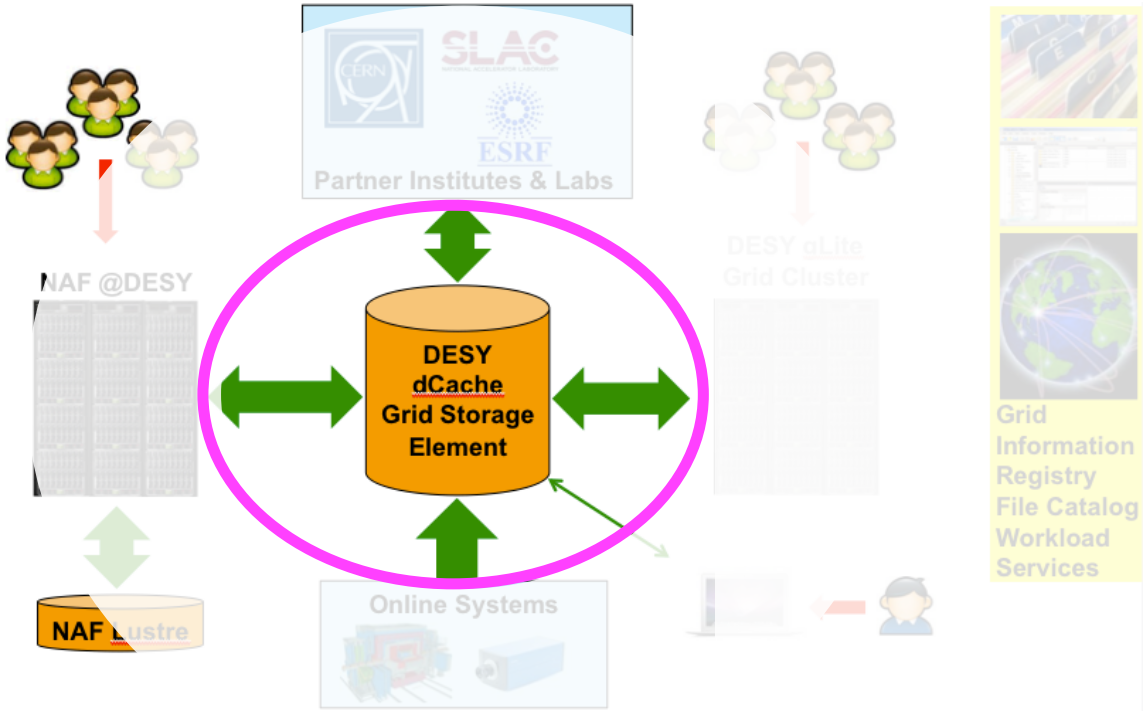
dCache: The central storage element

- dCache central data import export and exchange place
- dCache access protocols… Paul told you everything
- Bandwidth & Capacity @DESY:
  - To experiments: LAN speed
  - Grid/NAF cluster: Same switch
  - WAN: e.g. LHC dedicated VPN

    10 Gbit/s to GridKa
  - Example installations @ DESY:

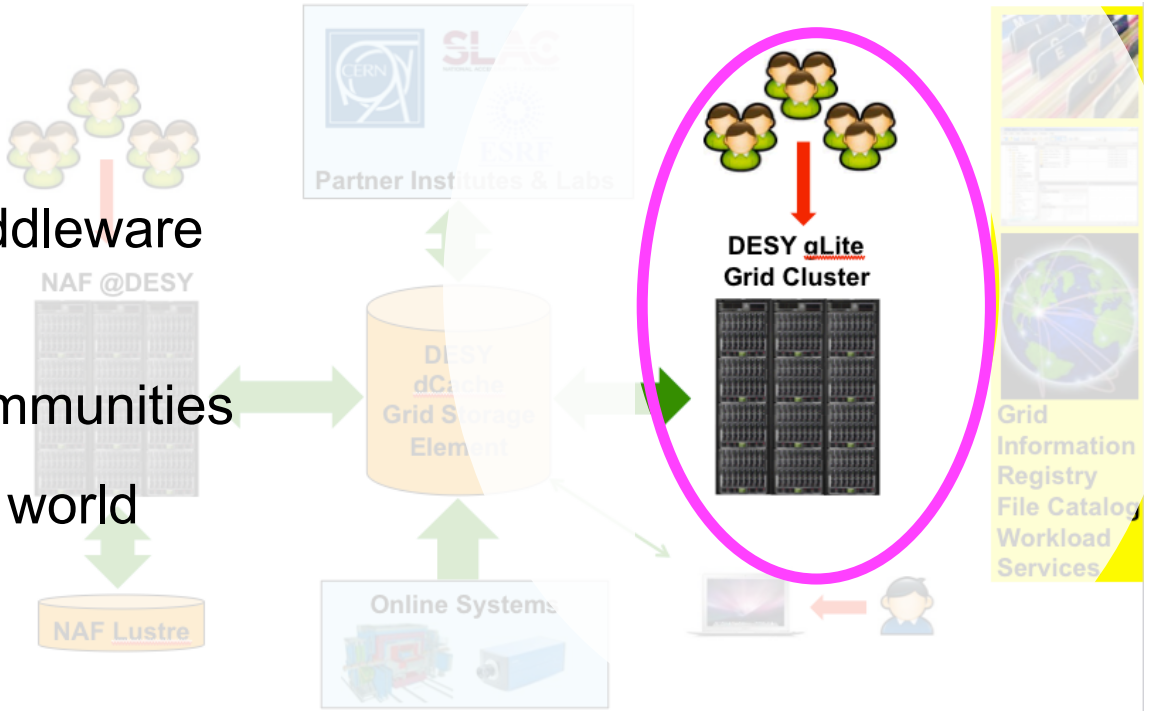
  CMS: 2 GByte/s / 450 TB Capacity

  DESY: 1 GByte/s / 250 TB Capacity
  - Can be tuned to needs
- Similar installations all over the world



DESY gLite Grid Cluster

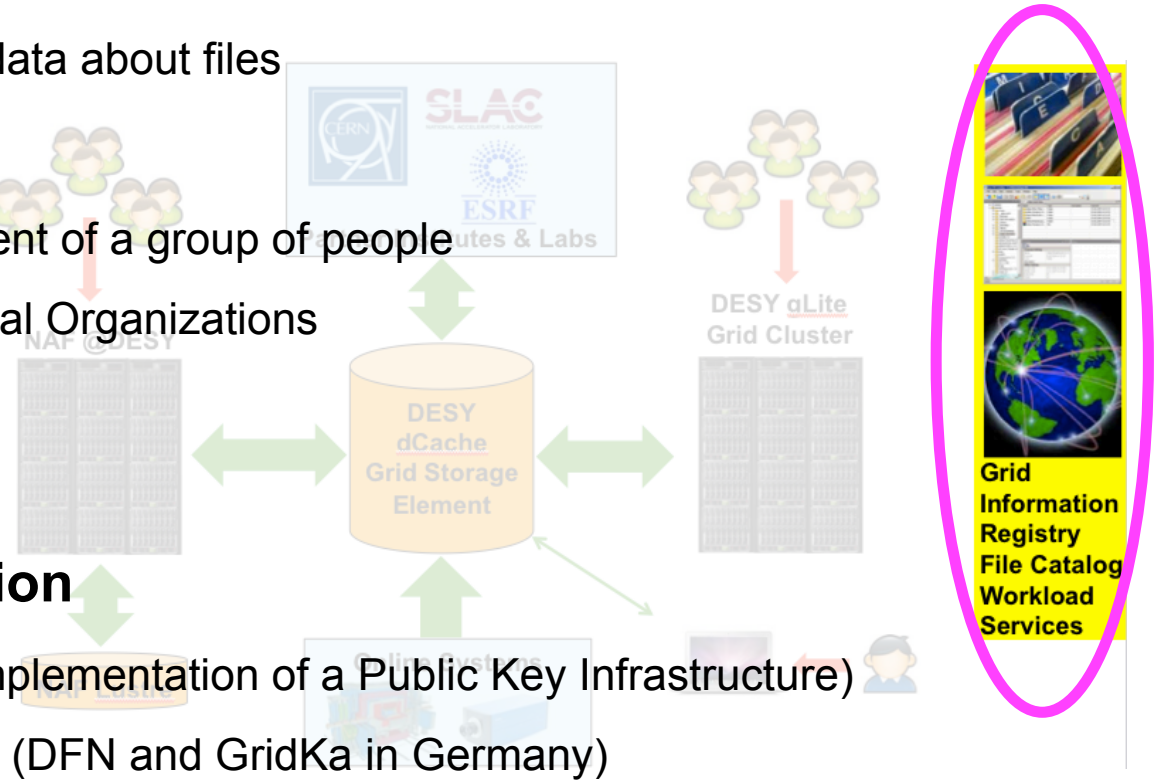
- Largest CPU batch cluster at DESY
  - ~4000 CPU cores currently
  - In production since end 2004
- Access using the gLite Grid middleware
  - “Batch submission” job type
- Serving many different user communities
- Similar installations all over the world



- DESY Grid allows:
  - Computing Grid
  - Data Grid
  - Mixture of both
- Remember: Jobs transient and Data persistent

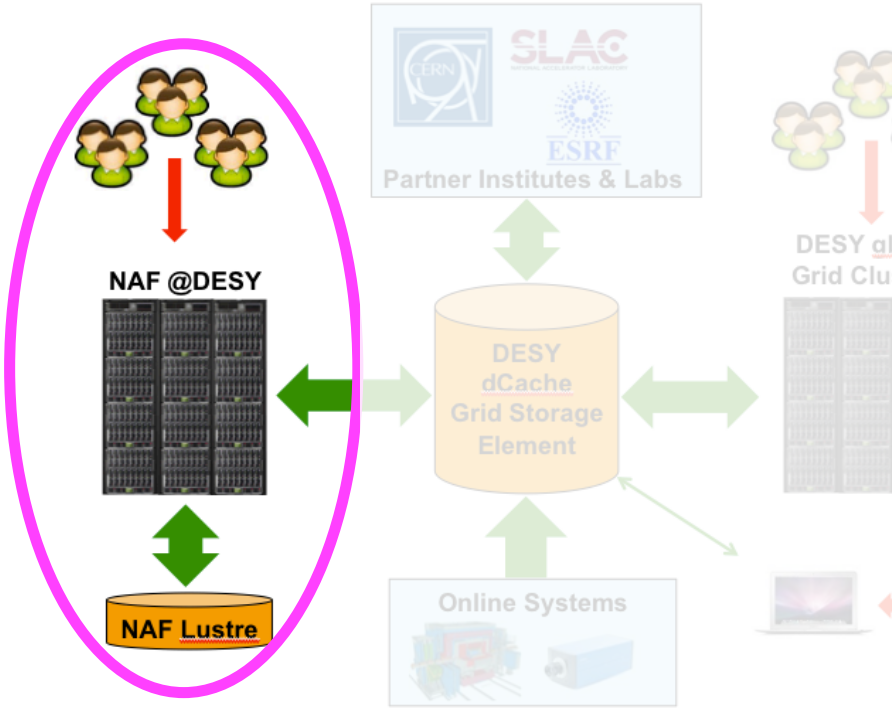
DESY Grid Services

- File Catalogue
  - File location (which SE?) and metadata about files
- VOMS Server
  - Virtual Organization: Digital equivalent of a group of people
  - Manages **authorization** within Virtual Organizations
- Workload Management
  - Send jobs to different Grid cluster
- Some words about **authentication**
  - Grid certificates: Based on X509 (implementation of a Public Key Infrastructure)
  - Issued by CA: One for each country (DFN and GridKa in Germany)
  - DESY serves as Registration Authority (RA) for members of DESY and Uni-HH, other institutes in D have their own or should create one with DFN or GridKa
- **Authentication and Authorization** always needs administrative procedures in the background! Independent of technical implementation



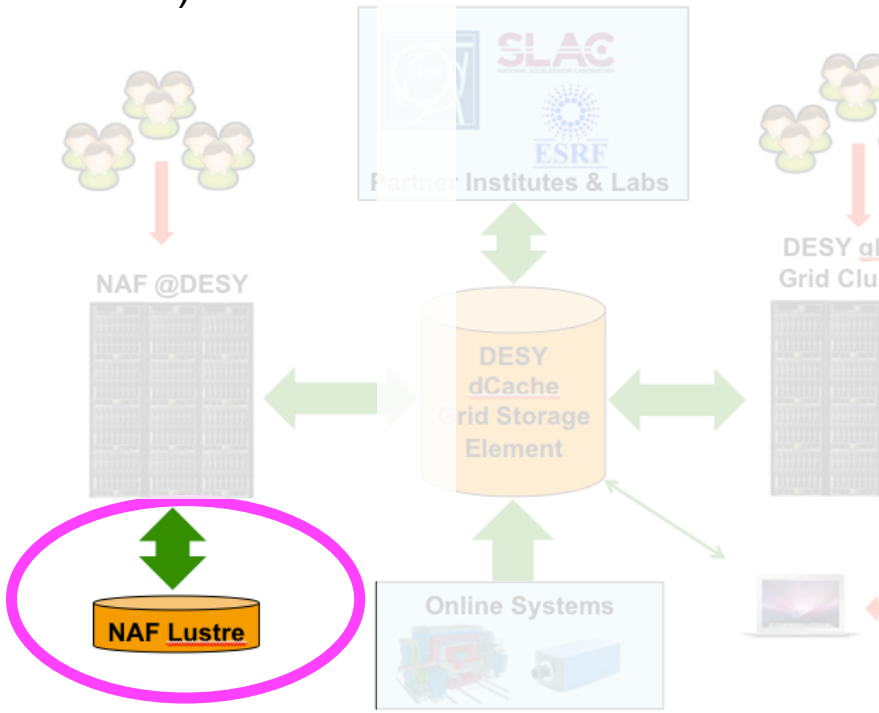
NAF: National Analysis Facility

- Designed as fast and versatile Analysis Facility for members of the German LHC community (~400 registered users)
- Complements the DESY Grid infrastructure
  - Interactive type work, small scale testing, high-throughput analysis
- dCache SE is the heart of the NAF data
- Lustre space:
  - Easier access than dCache
  - Caching of often used files
  - Pool space for user created small files
- (and AFS for $HOME, omitting this one now)
- About 1500 CPU cores, two DESY sites



Lustre: Setup and Experience

- Lustre: Connection to WNs over Infiniband
  - Faster than Gbit Ethernet (true end 2007, today 10 Gbit available)
  - Remote Direct RAM Access advantages over TCP/IP
- File Server: Mixture, all based on SATA disks
- ~120 TB currently



- Road-map not clear anymore (Oracle…)
- Lustre as a product difficult to handle
  - Instabilities on server and clients
  - Maintenance very invasive into operation
- User (mis-)use it for small files:
  - Meta-Data performance bad, seriously limits performance
- Basically, user want one single large file store
  - Lustre and dCache are not orthogonal enough

Backup Slides

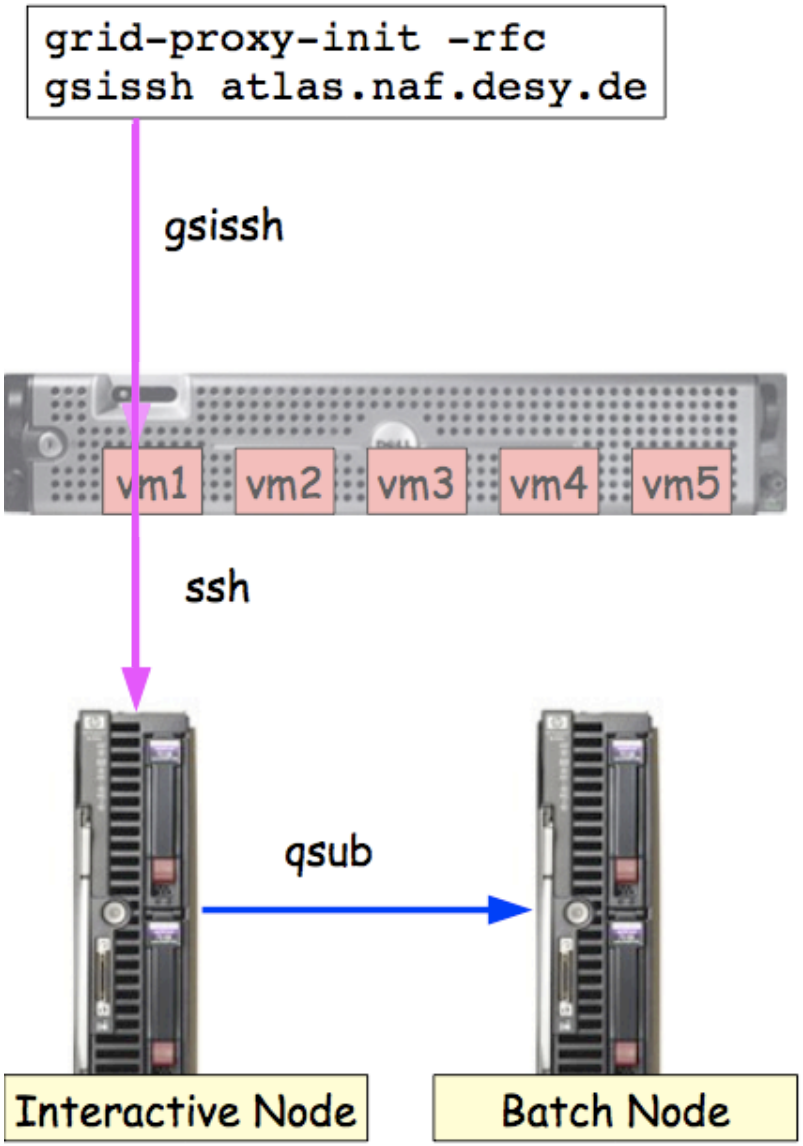
Example of Data-Grid Users

- **ILDG:** International Lattice Data Grid
- Use computing power outside the Grid, at Supercomputers (Juelich, APE,...)
  - Huge computing power to produce data, that will later be analyzed elsewhere
- Store and manage data on the Grid
- Built own semantic meta-data catalogue on top of Grid LFC
- Fine-grained access possible via VOMS
- VO hosted by DESY, DESY provides storage capacity

- http://ildg.sasr.edu.au/Plone

Access to Workgroup servers and AFS in the NAF

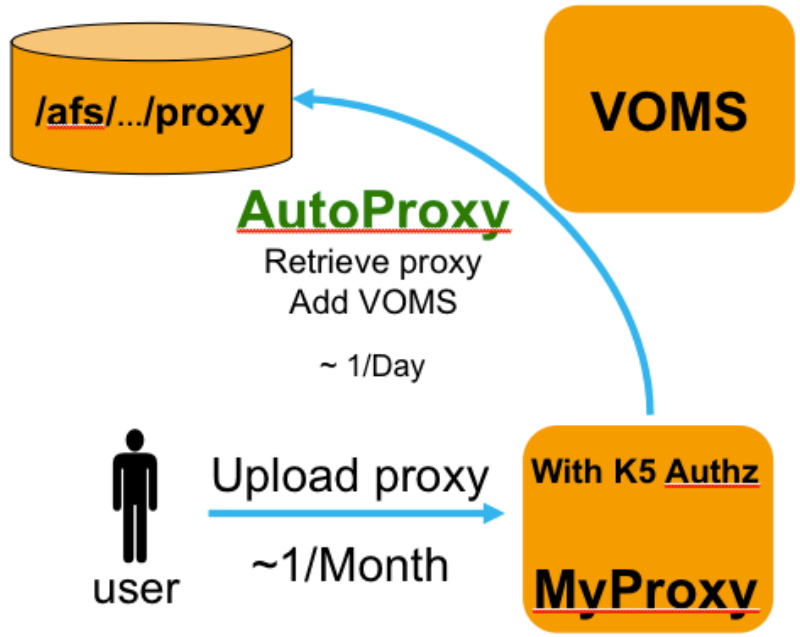
- *Idea:* Everyone doing LHC and ILC analysis has a Grid certificate
  - Can we use them to log into the interactive NAF?
- Krb5 ticket & AFS Token generated from proxy certificate
  - Login node: failover&load-balanced
  - Each login node only serves one VO
- Login node automatically redirects to Interactive Node
  - gsissh transparent to user
- Possibility to get AFS token also outside of NAF
  - Remote access to AFS via proxy
- Using Heimdahl Kerberos implementation!



Authentication in the NAF

- Grid **Authentication and Authorization** (AA)
  - VOMS-like structure implemented in registry
  - MyProxy used to renew X509+VOMS automatically
  - Global AA (X509+VOMS) and local AA (K5+GroupID) integrate well with the help of Registry
- AA needs admins and clear authorization schemes: Independent of implementation

X509 Proxy ⇔ AFS/K5 Integration



- Password-less login via X509
- Internally: K5 (hidden from user)
- Access to Grid via X509:
- AutoProxy: Always provide users with a valid VOMS proxy.

- „Single-Sign-On“ to all needed analysis resources facilitates users‘ work.