The German National Analysis Facility

What it is and how to use it efficiently

Andreas Haupt, Stephan Wiesand, Yves Kemp
GridKa School 2010
Karlsruhe, 8th September 2010

HELMHOLTZ GEMEINSCHAFT

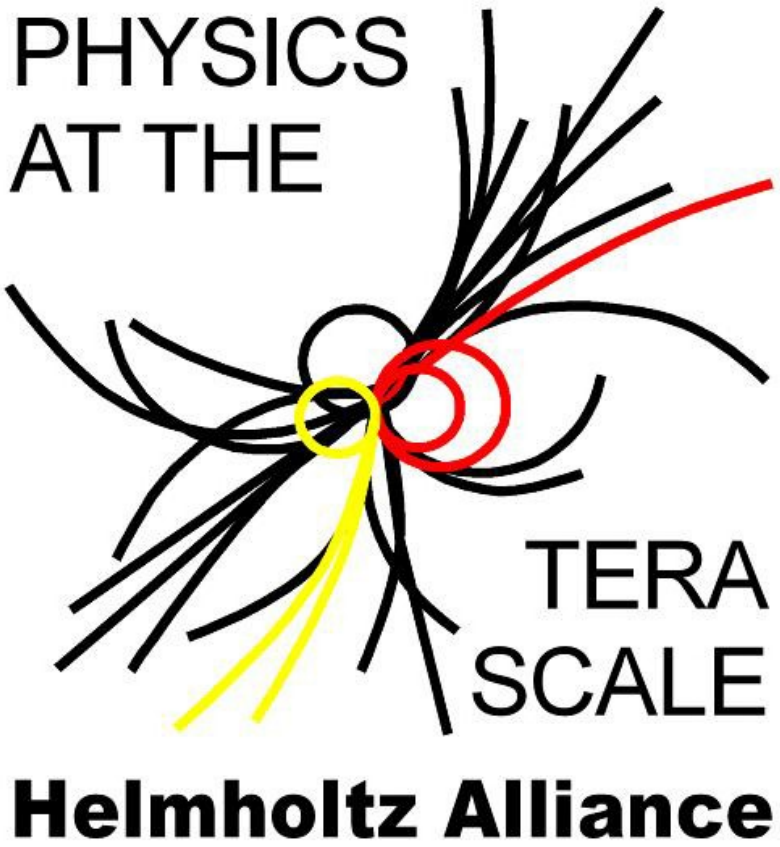
DESY

Outline

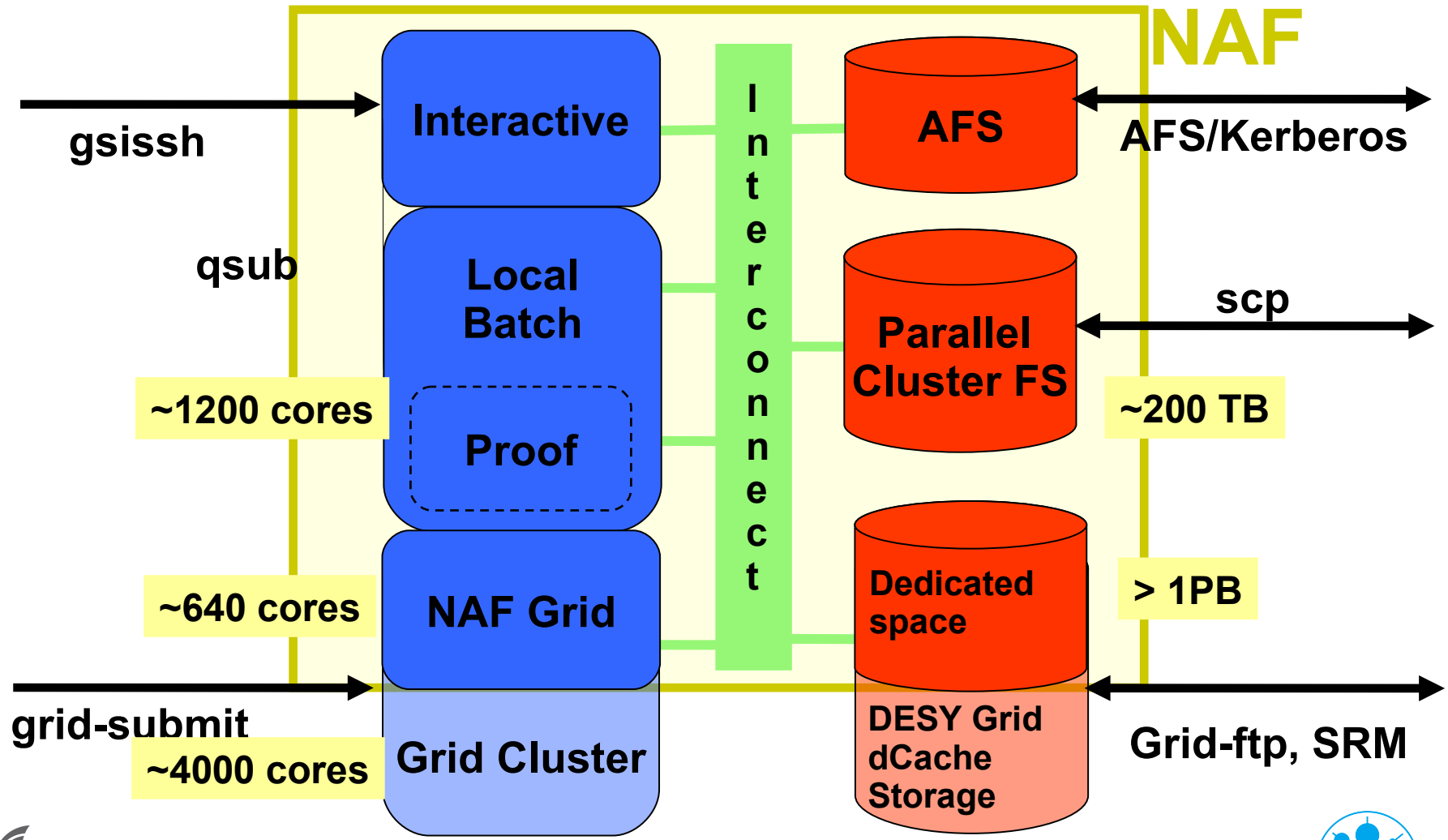
- NAF? What's that?
- The NAF blueprint
- How does the login work – a short introduction
- The NAF interactive work group servers
- The NAF batch system
- The NAF storage systems
- NAF support – how to get help ...
- Some tips how to use the resources best

What's the NAF

- a general purpose and flexible analysis platform for the German LHC experiments (Atlas, CMS, LHCb) & ILC
  - locality of the analysis data is a key feature
- provide interactive access to large scale computing resources coupled to the data
- close contact to users
  - general technical support by NAF administrators, experiment internal support
  - NUC (NAF user committee)
- Distributed over the DESY sites Hamburg & Zeuthen



The NAF blueprint

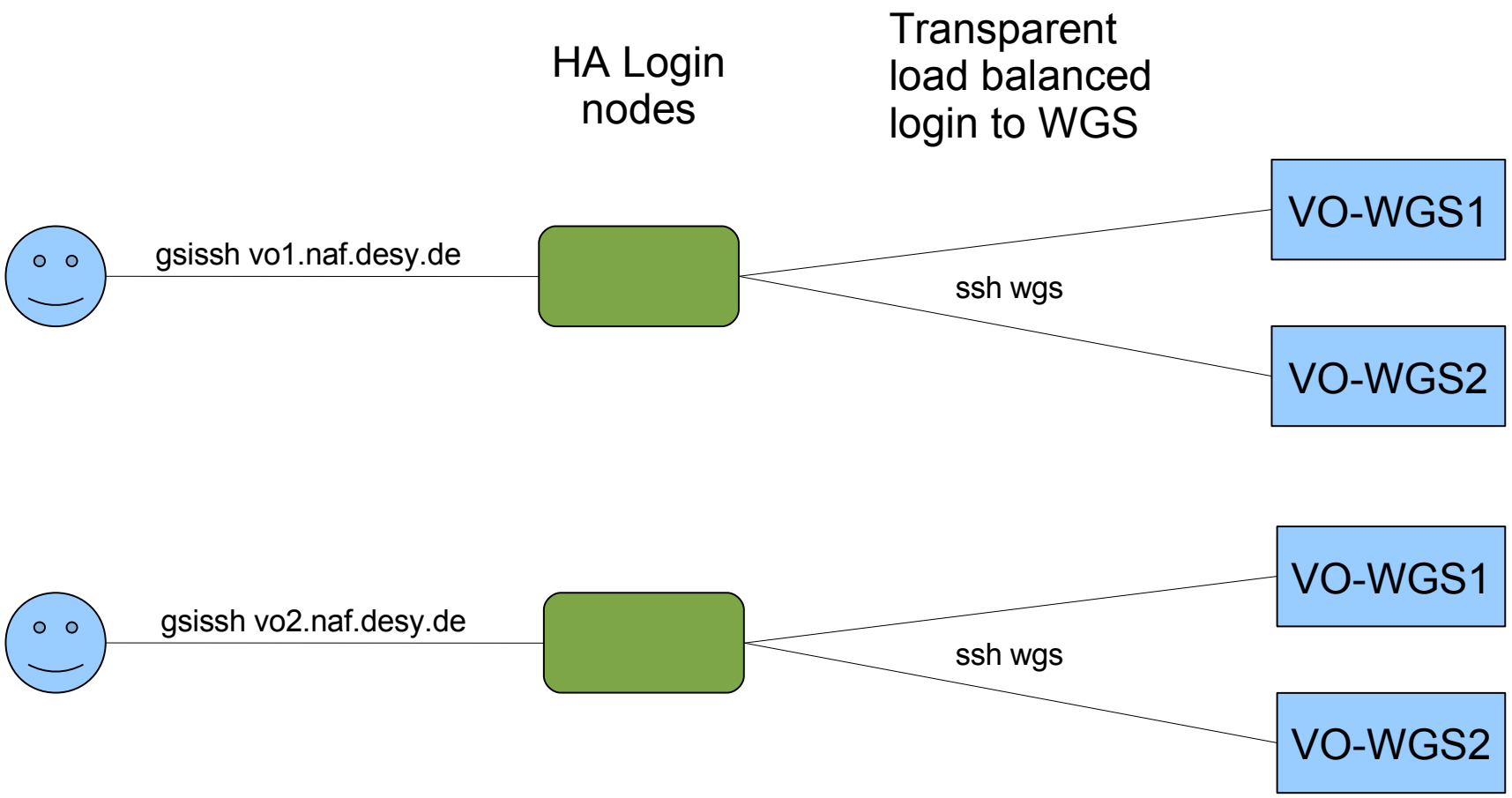


How does the login work?

- Login based on the same technology that is used for authentication in the “grid world”
  - X509 globus proxy certificates
- gsissh
  - An extended ssh client that allows authentication with globus proxy certificates
- Automatic generation of NAF Kerberos5 ticket / AFS token out of proxy certificate at login time
  - Transparent SSO access to e.g. AFS and other NAF services

```
[oreade38] ~ % voms-proxy-init -rfc
Your identity: /O=GermanGrid/OU=DESY/CN=Andreas Haupt
Enter GRID pass phrase for this identity:
Creating proxy ..................................... Done
Your proxy is valid until: Mon Aug 23 22:52:18 2010
[oreade38] ~ % gsissh login.naf.desy.de
Last login: Tue Aug 17 15:17:09 2010 from oreade38.ifh.de
Last login: Fri Jul 30 14:52:02 2010 from tcsh6-vm5.naf.desy.de
[ahaupt@tcx032]~%
```

Transparent login process to work group servers (WGS)



Main login problems

- You are asked for a password
  - You don't have a valid globus proxy (same holds true if you get a message like this:)

```
[blade84] ~ % gsissh login.naf.desy.de
The authenticity of host 'login.naf.desy.de (141.34.229.43)' can't be established.
RSA key fingerprint is 9e:5a:a5:c2:c6:7e:1e:6a:e3:d9:4c:72:62:99:d7:3b.
Are you sure you want to continue connecting (yes/no)?
```

- You get an error message saying “no RFC compatible proxy”
  - Globus proxy not RFC compatible
  - Check “**voms-proxy-info**” - “type” must be of kind “**RFC compliant proxy**”

Main login problems (2)

- You get an error message stating that all experiment work group servers are temporary unavailable
  - Shouldn't actually happen but does in case of e.g. major problems
  - Report it (but probably a monitoring service already noticed it)
- High load on the WGS (is the load balancing corrupt?)
  - All other servers might be even busier
  - The load balancing takes some time to react, this might only be temporary
  - In rare cases, the load balancing does not work correctly. This usually is only a symptom of other, more severe problems. If the problem persists for a longer period (30 mins), please inform naf-helpdesk@desy.de

The interactive workgroup servers

- Access to all NAF storage systems
- Software setup identical to farm nodes
- Meant for developing and testing software, handling the produced data
- See which other WGS are available: wgs-info
- Autoproxy, automatic token renewal

The NAF batch system

- GridEngine 6.2u5 
  - Open source version of SGE (now called Oracle Grid Engine ...)
  - Unfortunately no clear future for this “free” version right now ...
- More than 1200 cpu cores
- Features included in the NAF setup:
  - Automatic AFS token / Kerberos5 ticket provisioning
  - Automatic VOMS proxy generation and renewal, if configured by user
- You can request an interactive slot on a batch worker node
  - qrsh
  - You need to request resources as you would do it with qsub
  - In case the farm is full, you might want to use the switch “**-now n**“
  - http://naf.desy.de/general_naf_docu/working_with_the_local_batch_system/interactive_batch_usage/

NAF batch system: requesting resources

- Gridengine is designed to choose the best node / queue with respect to the resources you request
  - You don't specify the queue as in PBS/Torque
  - Just say what your job needs via stacking the different resources...
    - qsub -l h_vmem=2G,h_cpu=05:00:00 my_job
- Available resources:
  - http://naf.desy.de/general_naf_docu/working_with_the_local_batch_system/requesting_resources/
  - h_cpu: CPU time limit (e.g. 7000 -> 7000 seconds, 05:30:00 -> 5 hours and 30 minutes)
  - h_vmem: virtual memory limit (e.g. 750M, 1.5G)
  - site: specify the NAF location the job should run on (e.g. due to “close” data)
    - Only available: **hh** or **zn**

The NAF batch system: parallel jobs

- Typical use case in HEP: one process runs on one cpu core
  - This also reflects the standard batch system configuration: one job reserves one slot
- Different ways exist to parallelise jobs
  - PROOF, OpenMP, MPI
- There are different so called “parallel environments” configured in the NAF batch system
  - Can be requested with qsub / qrsh switch “**-pe <pe name> <number of slots>**”
  - Handle different use cases:
    - proof: request proof slots on different worker nodes
    - multicore: request a number of slots on a single node -> e.g. for multithreaded jobs
    - mpi: run mpi jobs distributed over several worker nodes

Batch system best practices

- Typically experiments already have job submission frameworks (Ganga, CRAB, ...) that should do things right ... but:
- Use array jobs in case you need to run lots of similar tasks
  - e.g. qsub parameter "**-t 1-100**" submits your job 100 times
  - Faster and easier for you, reduces load on batch system
  - The environment variable $SGE_TASK_ID holds the task number inside the job
- Optimize your job throughput
  - Only request resources you really need (especially h_vmem and h_cpu)!
  - In case you are using "large" (i.e. very high h_vmem) or parallel jobs, request job reservation
    - "**-R y**" qsub / qrsh parameter
- Read the documentation ... ;-)
  - http://naf.desy.de/general_naf_docu/working_with_the_local_batch_system/best_practises/

NAF batch system troubleshooting

- Your job doesn't start
  - Maybe you requested resources that are not available
    - Use qsub parameter "**-w e**" to let the batch system reject such jobs
    - We could generally switch it on but in case of some minor transparent maintenance this sometimes rejected valid jobs ...
  - The farm is full ...
    - Check the queue status with "**qstat -g c**"
    - But even if there are free slots in some queues it doesn't mean a job can start there – other limits (e.g. shortage in host memory) might apply
- You can see jobs STDOUT/STDERR only after the job has finished
  - That only happens in case those files are placed in AFS – Lustre shouldn't show that behaviour
- Some of your jobs die / have a non-zero exit status:
  - Use the monitoring at: https://www-zeuthen.desy.de/dv-bin/batchssl/stat/naf/jobs//

NAF addons (1)

- ini
  - http://naf.desy.de/general_naf_docu/naf_features/setup_environments/
  - Prepares environment for special purposes (e.g. set up a special ROOT version)
  - Just type “ini” to get an overview of all available targets

NAF addons (2)

- Get an AFS token on your pc / notebook
  - Use **/afs/naf.desy.de/products/scripts/naf_token <account>**
  - Needs to have "**grid-proxy-init**" in your PATH (e.g. a sourced gLite-UI)
  - Only works on Linux clients currently
- Automatic VOMS-proxy generation and renewal:
  - http://naf.desy.de/general_naf_docu/naf_features/autoproxy/
  - includes German group extension (/atlas/de, cms/dcms, ilc/de, ...)

```
[ahaupt@tcx032]~% ini autoproxy
autoproxy scripts now in PATH variable and X509_USER_PROXY set
Initializing Module autoproxy...
[ahaupt@tcx032]~% ap_gen.sh
NOW: Creating and uploading a proxy valid for 30 days to myproxy server
Your identity: /O=GermanGrid/OU=DESY/CN=Andreas Haupt
Enter GRID pass phrase for this identity:
Creating proxy ..................................... Done
Proxy Verify OK
Your proxy is valid until: Wed Oct  6 13:58:13 2010
A proxy valid for 720 hours (30.0 days) for user ahaupt now exists on tcsh2-vm5.naf.desy.de.
[ahaupt@tcx032]~% touch .globus/.autoproxy
```

The NAF storage systems

- AFS
  - Holds home directories and experiment software
  - Accessible worldwide under the common path /afs/naf.desy.de/
- dCache
  - Holds main experiment data
  - Accessible worldwide via several grid tools
- Lustre
  - Main scratch area for large analysis data
  - Only available on interactive NAF nodes

AFS

Volume Location Database
cluster at application level

- volume based
  - namespace is constructed from embedded mount points
  - R/O replication, asynchronous
  - transparent migration
  - volume quotas (2 TB max)

Fileservers

- metadata:
  - volume location data: small amount, low transaction rate
    - no scalability problems (at our size)
  - per file metadata resides on the fileserver, within the volume
    - scales ok

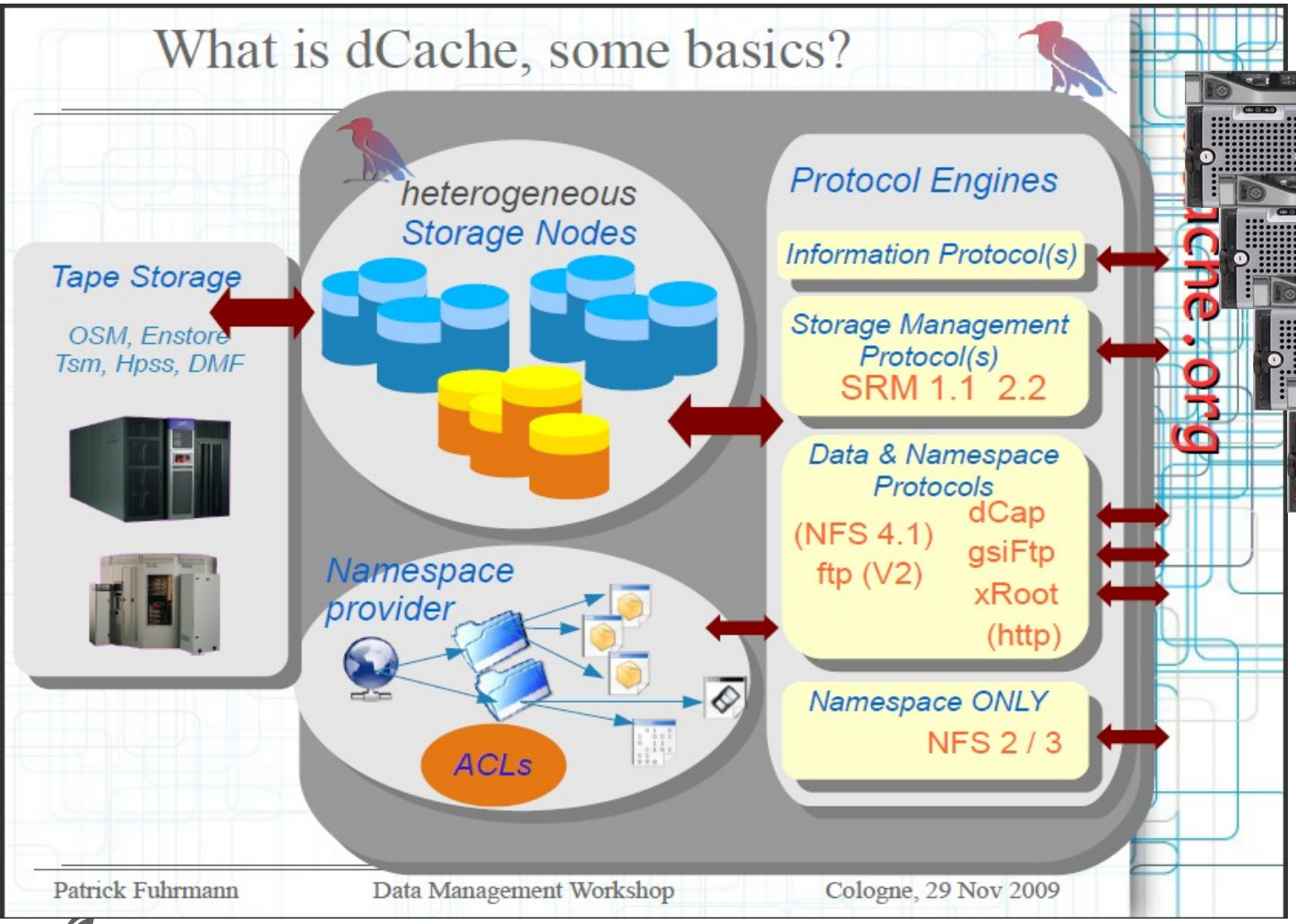
AFS

- Home directory volume with backup
  - Initial quota 1GB typically
  - Holds your code ...
- AFS scratch volume (~/scratch) can be much larger but without backup
- Token for the NAF AFS cell from your notebook / desktop:
  - /afs/naf.desy.de/products/scripts/naf_token <account>
- Structure your data in volumes
  - Your experiment admins will create them for you

AFS pros and cons

- PROs:
  - reasonably secure
  - group space administration delegated to group admins (afs_admin)
  - backup selectable per volume (matching quota)
    - separate group quotas for space with/without backup
    - files from backup can be retrieved by users
  - usable ACLs (per directory), working the same way on each client
  - metadata transaction capacity scales with number of fileservers
- CONs:
  - AFS token required for authenticated access (might expire)
  - client relatively slow
    - persistent client side cache helps in some cases, hurts in others
    - has much improved in recent years, more improvements soon
  - volumes are confined to their fileserver partition
    - data is not distributed over fileservers automatically

dCache – the overview



Head node(s)

Pool nodes

- Replication
- Migration
- HSM optional
- Grid storage element (SRM)

dCache in the NAF

- http://naf.desy.de/general_naf_docu/naf_storage/working_with_dcache/
- Dctools examples:

```
[ahaupt@tcx032]~% ini dctools
dc-Tools now in PATH variable
Initializing Module dctools...
[ahaupt@tcx032]~% dcls -l /pnfs/ifh.de/data/lhcb/user/ahaupt
-r-------- lhcbsgm    lhcbsgm    1073741824 Aug 16 16:25 test.1g.1
[ahaupt@tcx032]~% dcget /pnfs/ifh.de/data/lhcb/user/ahaupt/test.1g.1 /tmp/test.1g
[ahaupt@tcx032]~% ll /tmp/test.1g
-rw-r--r-- 1 ahaupt support 1073741824 Sep  6 14:04 /tmp/test.1g
```

- Other commands:
  - dcmkdir (creates directory)
  - dcrmdir (removes an empty directory)
  - dcput (puts a file onto dCache)
  - dcdel (deletes a file in dCache)

dCache Pros and Cons

- PROs:
  - most versatile
  - many different access options
    - local access via dcap, gsidcap
  - access from anywhere via gsiftp, srm
    - all NAF dCache storage is grid-enabled
  - in future, will add WebDAV, pNFS (NFS 4.1)
  - very good aggregate performance
- CONs:
  - no immediate POSIX access
    - pNFS will remedy this, but may take a while
  - files cannot be modified, only deleted and rewritten (won't change in future)
    - But in HEP “write once – read often” typical use case
  - modest single client performance, no Infiniband support
  - Head Node is equivalent to Lustre MDS
    - single point of failure, limits scalability
  - dCache is not suitable for small files!

Lustre

Metadata server

- looks like a single POSIX filesystem to the client
- files are distributed round robin across OSTs when created
  - automatically
- single files can even be striped across OSTs (not advisable for common usage)

Object storage servers

unclear future after Oracle's SUN-acquisition

Lustre Pros and Cons

- PROs:
  - high & scalable data performance, large filesystems
    - without hassle for users
  - fast client
    - single client easily saturates a GbE connection
    - uses the operating system cache
  - supports modern, fast interconnects (Infiniband)
    - have seen 500 MB/s for a single client-server connection
  - multihomed servers & clients possible
- CONs:
  - metadata for each and every file resides on a single MDS
    - aggregate lookup/open/create performance limited by single server
    - can be a real problem if many clients rapidly access different files
  - a small file (say,1 kB) takes up as much space on the MDS as on the OSS
    - and accessing it probably causes more work on the MDS
  - not suitable for (many) small files
  - storing large amounts of data in small files is always a bad idea
    - but on Lustre, it's particularly bad (performance worse than AFS not unlikely)

NAF support – how to get help ...

- NAF has a shared support model
  - Experiments provide a first contact point via mailing list:
    - naf-<vo>-support@desy.de
  - NAF operators provide a ticket system
    - naf-helpdesk@desy.de
- Regular NUC meetings
  - On every second Wednesday in a month
  - http://naf.desy.de/nuc
  - Raise problems that disturb the work
  - Contact your experiment representatives!

NAF best practices

- Don't overload directories
  - Use a subdirectory structure
  - 1000 files per directory should be enough
- Avoid building software in network file systems
  - Compile in /tmp
  - Install into AFS
- Avoid the use of X11 applications on any NAF system
  - Use the applications on your desktop / notebook
  - Access the input files via AFS

NAF best practices (2)

- Get your data to your home institute ...
  - Most experiments already have a user friendly data distribution system (e.g. dq2 for Atlas) – use them!
  - Register your data there and use the builtin data replication mechanisms (normally the replication is the done between the dCache instances at the several sites)
  - scp'ing the data from NAF-Lustre to your home is usually rather slow - avoid it if you can
- In case you are unsure how to do things best or observe problems – contact us!
  - naf@desy.de

That's it folks!

General NAF documentation and news:
http://naf.desy.de/