NAF Status

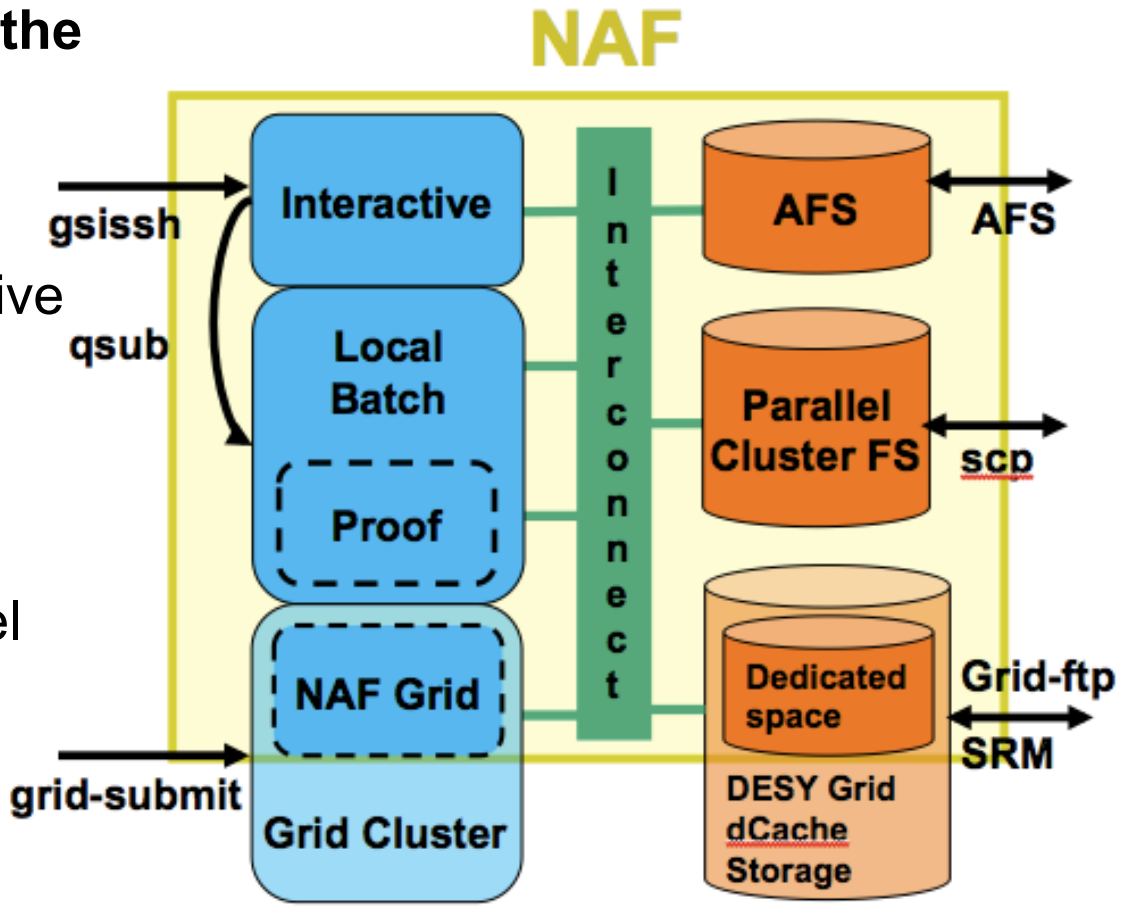
Y. Kemp for NAF admin team
H. Stadie for NUC
4th annual Alliance Workshop
Dresden, 2.12.2010

HELMHOLTZ ASSOCIATION

DESY

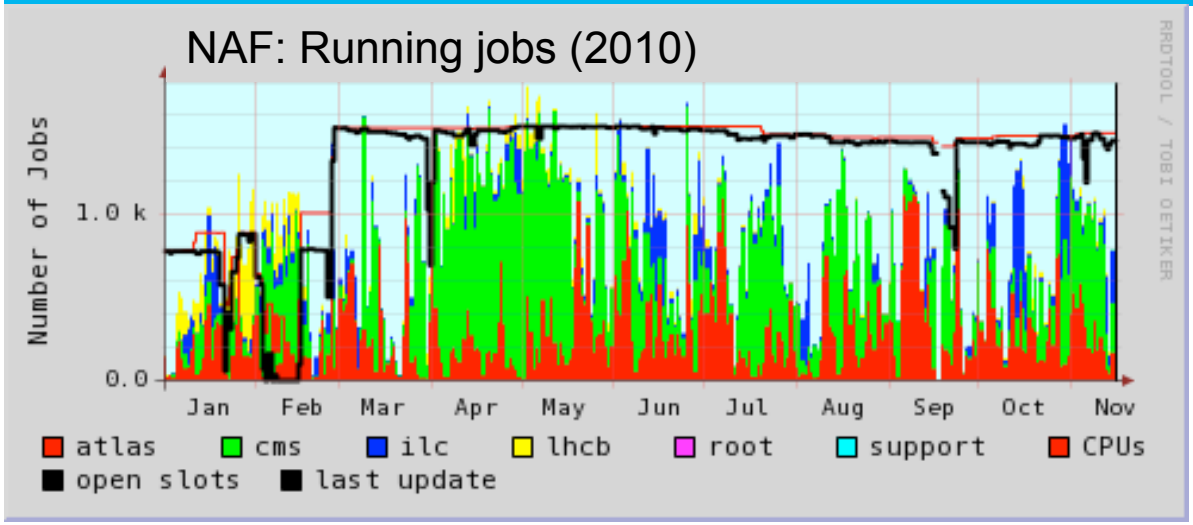
NAF introduction in one minute

- **Access to experiment data on the Grid Storage Element**
- CPU cycles for analysis: interactive and local batch
  - Complement the Grid resources
  - New techniques like PROOF
- Additional storage: Lustre parallel file system
- Home directories on AFS, accessible from everywhere



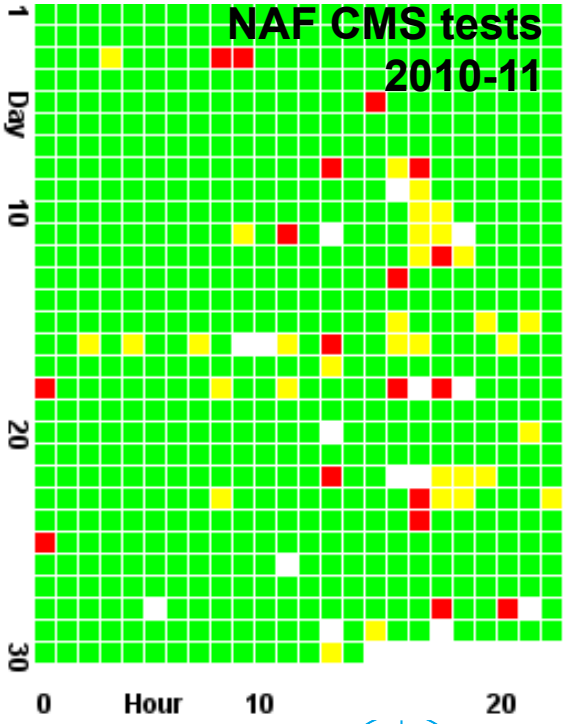
http://naf.desy.de/

NAF well used



- Very “peaky” behavior: Try to keep overall utilization below 75% and peaks below 90%: Will add hardware in 2011 (starting now).

- Availability and reliability
  - One of the most important aspects for users – and admins
  - Availability and reliability is ~97.5%, similar to the DESY Grid … but this does not tell the whole story: The 2.5% failures affect you and your work much more than in the Grid!
  - We want and need to get better! ... Have a look at the following slides.



Major problems in the past months:

- ”Data on dCache not available, slow transfers”
  - Some problems with dCache file server availability ➔ Under investigation / solved
  - User code sometimes causing denial-of-service: e.g. not closing files after reading them will keep them open for the duration of the job. Only a certain number of files can be kept open at the same time ➔ Other jobs cannot open files
  - Slow data transfer: Can have many different causes. It is known that older ROOT files are written in a bad way for reading them efficiently. Sometimes a file server is also overloaded ➔ ROOT versions to be changed by experiments & Improvements on dCache side constantly done
- “Lustre not working properly”
  - Lustre does not like small files: Keep your code / SVN / output files outside of Lustre! We provide you AFS-Scratch volumes for such purposes!
  - Other users might do harmful operations and affect your speed or even accessibility
  - To increase stability, Lustre data in HH is going via TCP/IP instead of InfiniBand
  - In general: Lustre future is unclear: ORACLE: DESY looking into alternatives
  - … but we recognize that there is a need for “easy access” file store

AFS problems in the past months

- "AFS hangs, Login impossible, shell is frozen, jobs die, …"
- We had severe troubles with NAF AFS cell in the past months
- Investigation very difficult and painful, even asking developers for help
- Patched AFS kernel module: Solved some problems
- It turned out that major problem is due to interference between SGE and AFS.
  - Similar jobs (e.g. one user submitting many jobs): All STDOUT and STDERR end up in files in the same directory
  - These files are created at time of job start ➔ If cluster is rather empty, can be several hundreds of jobs: Files are created and read simultaneously in the same directory
  - Fileserver ensures consistency of client cache through callbacks
  - A storm of callbacks between AFS server and AFS clients will basically paralyze the fileserver and the clients, when jobs read in the directory with the .e/.o files
- We think we finally have solutions / workarounds!

Solutions to AFS problem: What NAF can/will do

- Limit number of jobs / user: Ad hoc and drastic measure
- Throttle start of jobs: Implemented today
- Possible long-term solution
  - Change STDOUT/STDERR files with prologue and epilogue methods
  - Write into separate directories
- … and we now have a simple recipe for you to help us by defusing your jobs: See next slide

Solutions to AFS problem: What YOU can do

Change the submission command like this:

```
qsub -j y -o /dev/null <other requirements> <your jobscript>
```

Have as the very firsts lines in your job script something like:

```
exec > "$TMPDIR"/std.out 2>"$TMPDIR"/std.err
```

(this will store the files locally on the WN)

($TMPDIR is unique during job execution, you can of course add $JOB_ID, $SGE_TASK_ID … to the filename)

… and at the very end of your job script, copy these files over to some location on AFS, preferably in a subdirectory

… any maintainers of CRAB / GANGA / … here? Can you implement this for all users?

… we prepare a web page, and inform all users soon

Reminder of the NAF support channels

- Got a problem with your experiment setup?

→ naf-[atlas,cms,ilc,lhcb]-support@desy.de

- Got a problem with the NAF fabric (or are not sure where problem resides)?

→ naf-helpdesk@desy.de

  - Experiment supporters: You know the different system experts and you can use them directly

- If you think your job causes a problem:

→ **We need you to contact us and help us making the NAF better!**