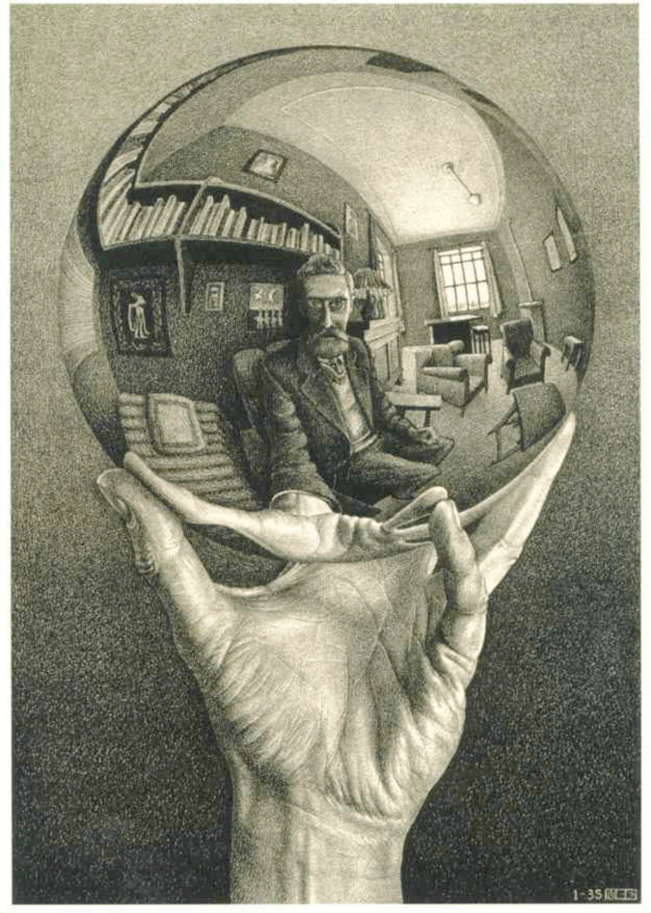
NAF Perspectives.



Andreas Haupt, Yves Kemp

NAF NUC F2F 11.11.2009
DESY Hamburg

How far? Timeline

- We have the time remaining of the TeraScale Alliance (2012)
- … and we have the time after 2012
- 2010 is very important: We should be able to react very fast to changing and increasing needs!
  - 2010 also hardware renewals
- In the beginning of LHC data taking, a flexible and multipurpose facility is most helpful
  - Available HW should similar to Grid during this period
  - This period might be until 2012, afterwards analysis can be better streamlined, and also better move to CTDR facilities (I.e. Grid) (Disclaimer: My personal view)
- Money: Have asked for additional funding
  - The NAF is well seen: good critics by DESY PRC, well perceived by German users community
  - **YOU (NUC)** should tell us, what you need!
- Manpower: Tight as always everywhere:-)

Some thoughts about the NAF concepts

- The NAF is complementary to the Grid
  - The NAF can never replace the Grid! Users MUST use the Grid to some extent for their work. It is important for all to know and live this!
  - The NAF [makes / should make] working with the Grid easier.
- The key part of the NAF is the central (dCache) storage
  - It is both Grid and NAF, and origin of all/most analysis.
- Optimal workflows with the “fast cluster file system” (aka Lustre) still not found (my personal opinion)
  - Different usage profiles between experiments
  - We see some usage pattern that does not fit the current technology (or maybe no filesystem based technology at all…)
  - Maybe there even is no “optimal” workflow
  - Nevertheless, a file system like this one seems to be needed
  - … next slide

Cluster file systems

- Difficult business: Quote from Amazon “Simple Storage Service”
  - *“Building highly scalable, reliable, fast, and inexpensive storage is difficult. Doing so in a way that makes it easy to use for any application anywhere is more difficult. “*
  - No silver bullet. (Amazon would claim something else here:-))
- Technology
  - Have opted for Lustre for different reasons
  - Lustre only partially has evolved as we expected / wanted / needed
  - We are currently looking into alternatives (mainly Hamburg site because of larger scope)
  - For the moment, we still use Lustre, no alternative chosen yet.
- Very long term (my personal view)
  - dCache will go the way of NFS 4.1 and pNFS (p=parallel). Works in the lab, but is not there in the NAF in 2010. Some dCache related issues also unclear to me at this point (open time, small files, …)
  - Nevertheless: NFS 4.1 and pNFS will change lots of things. But not soon:-)

Network technology

- We have opted for InfiniBand two years ago
  - Was the right choice at that point
- Now, 10Gbit Ethernet over copper is there, costs identical to IB (and going down)
  - We will slowly migrate infrastructure to 10Gbit Ethernet
  - IB-Infrastructure not lost: IB and 10 GbitE can be combined up to some point
  - Complete migration will take ~liftetime of a current server
- This helps for future planning:
  - One single network infrastructure allows for more flexibility in resource assignment and storage provisioning

CPU / server technology

- You all know: Multi/Many-core is not around the corner, it is there
- Particle physics is not really ready for that
  - Working groups in experiments / HEPIX / … only slowly starting
- Potential Problems with future purchases
  - Power/core is not increasing, #core/server will increase
  - If Memory/core stays identical (2-3 GB), Memory configuration difficult
  - Memory (RAM and cache) bandwidth?
  - Local hard drive is problematic: If we stick with 1job/core, local IO to disk is always random read/write. This is deadly for spindle disks, SSD could be solution???
- We are not alone with these problems, we are closely looking into alternatives, and also following what others do.

Operating system

- 2010 should be SL5, also on WGS.
  - Remember already current issues with SL4 and InfiniBand/Lustre
  - Remember October 2010, end-of-life of RHEL 4…
- SL6 probably not an issue for the NAF in 2010.
  - Also other distributions not requested / needed / wanted …

"Distributed NAF"

- This is in the proposal, and up to now, we have managed to have the NAF distributed over two sites
- Some concepts might be used elsewhere
- The NAF as a facility as it is now is difficult to envisage being more distributed
  - This would need more investigation
- My personal opinion: We already have the Grid as distributed facility.
  - Maybe "distributed NAF" something like "enhanced Grid"?