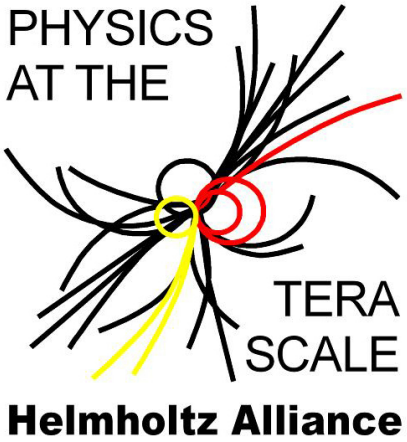
(Re-)Designing the National Analysis Facility: NAF 2.0



Yves Kemp, DESY IT
DV seminar
DESY, 11.2.2013

HELMHOLTZ
ASSOCIATION

DESY

Some history: The National Analysis Facility (NAF)

- The Helmholtz Alliance “Physics at the TeraScale” started Mid 2007
- One of its pillars was “Research topic Grid Computing”
- One of its projects was the “National Analysis Facility”
  - … complement the German Grid resources
  - … interactive and analysis resources
  - … to be used by members of German LHC (and ILC) institutes
  - … starting at the two DESY locations
- Started (according to my calendar) on 4.6.2007 with a meeting with CMS users
- LHC datataking: 10-19.9.2008 – since end 2009

PHYSICS AT THE TERA SCALE
Helmholtz Alliance

DESY

Some more history: Grid and Computing

- 2001: European DataGrid project launched / dCache.org project officially launched
- 2004: Foster/Kesselmann: The Grid: Blueprint for a new computing infrastructure
- August 2004: DESY operates Grid infrastructure
- 2005: LHC Computing TDR
- 2006: 1&1 hosts ~25.000 server in Karlsruhe CC
- 2006: Intel Dual-Core Systems are “state-of-the-art”, two-socket-systems
- End 2010: LHC Computing Grid: ~200k CPU core, 150 PB data
- End 2011: Amazon S3 ~550 PB data
- 2012: Amazon EC2 largest CC (Virginia) ~5k Racks . 150k-330k Server
- 2012: Facebook initial public offer: 104.000.000.000 USD

The Mobile Revolution

Mobile Computing in 2007

Mobile Computing in 2013

... Is this of relevance for NAF people?
… Yes, it is: Influences people and technology

Getting back to the NAF

- How did the NAF came to life?

- DESY IT (Hamburg) and DV (Zeuthen) created a NAF project group
- We asked the experiments for requests
  - Received by ATLAS, CMS and LHCB
  - Can be found at http://naf.desy.de/nuc/
- We had several discussion workshops, both internally and with experiments

- … let's have a look at what people asked for – five years ago

Starting with Atlas & CMS

HAMBURG • ZEUTHEN

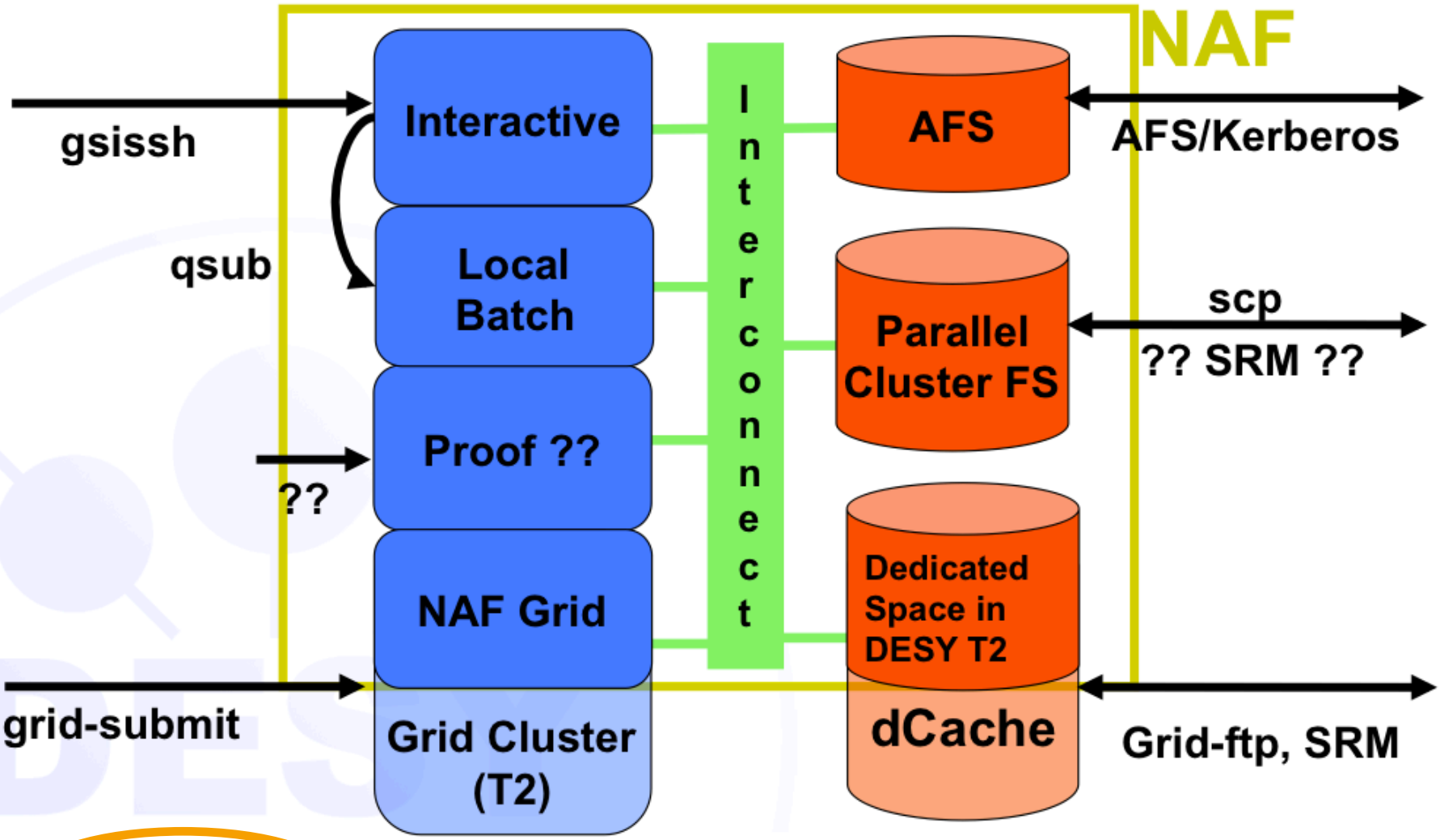
Reading through the requirements: Some points:

- **Interactive login: What for, How?**
  - **Code development, Experiment SW and tools, Grid-UI**
  - **Code testing, working on small data samples**
  - **Work-group-server**
  - **Uniform access, gsissh ("Single-Sign-On")**
  - **Central registry**
- **Personal/group storage**
  - **AFS home directories (and access to other AFS cells)**
- **High-capacity /High-bandwidth storage**
  - **Grid & local (with backup)**
  - **Grid-part: Enlargement of the T2 part**

Atlas & CMS cont.

- **Batch-like resources:**
  - Local access: short queue, for testing purpose
  - Fast response wanted
  - Large part (only) available via Grid-mechanisms
    - Guarantee fast response, dedicated (fair-)share (VOMS mechanisms)
    - For private/regional MC production
- **Hosted Data:**
  - AODs (Full set in case for Atlas, maybe trade some for ESD?)
  - TAG database
  - User/Group data
- **Additional services**
  - PROOF farm, with connection to high-bandwidth storage
- **Flexible setup**

First sketch of the infrastructure



What part of the NAF concept is not in the previous sketch?

- Requirement: **Minimal dependency** of DESY infrastructure
  - Not sure whether NAF would be limited to DESY – actually expansion plan to other sites
  - National instrument, not a DESY one – make this clear in the infrastructure
  - Many new, non-DESY based user with no connection to DESY
  - New scale – not sure if feasible in DESY infrastructure in 2007

- Lead to: **Rather independent NAF** infrastructure
  - Independent user registration, based on X.509 Grid certificates
  - Independent software stack (independent installation and configuration system)
  - Independent support channels
  - ...
  - **Not independent:** in DESY CC, using the same network, the same people, ...

The support model: A split one

- The provider know their infrastructure
  - They will help users with problems there
- Who is “they”?
  - General helpdesk as a first-level-support – entry point for normal users
  - Different expert groups in the second-level – not directly accessible to normal users
- The experiments know about their software and their internal workflows
  - They will help users with problems there
- Who is “they”? (… my view as an outsider to the experiments …)
  - German VO support group – experts as a first-level-support
  - Global VO support channels in the second-level

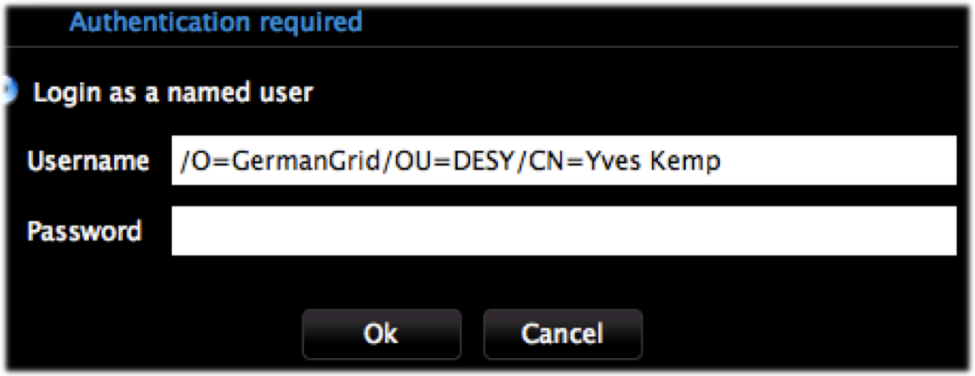
NAF 1.0: How did it go?

- Fast setup, many users, many successful analyses. General setup OK.
- Identified some weak points though:
  - Missing integration into “normal” DESY proved to be manpower intensive and decoupled NAF from advances in “normal” DESY infrastructure
  - A further spread to other sites in Germany did not happen
  - We never benefited from the two-site setup – actually, we suffered from it in terms of reliability and performance (latency!)

- Needs have evolved since 2007
  - More graphical tools needed
  - More software, e.g. also commercial one
  - General mobility ask for better remote capabilities
- **The NAF needs a fundamental redesign to continue its success story**

NAF 1.0: Login using Certificates

- In 2007: Integrate NAF with Grid
  - It was clear that Grid certificates must play a role
  - Grid certificates chosen as the only authentication method – no password
  - Lead to gsissh – worked well in the NAF 1.0 context
- Since 2007: Adoption of Grid certificates (or X.509 in general) has not increased
  - Gsissh not integrated into normal OpenSSH distribution
  - New communities not using X.509 (Photon)
  - No interest from commercial vendors to equip their tools with X.509
- Need for graphical login methods:
  - NX or the like
  - Works well – if you have username+password ☺

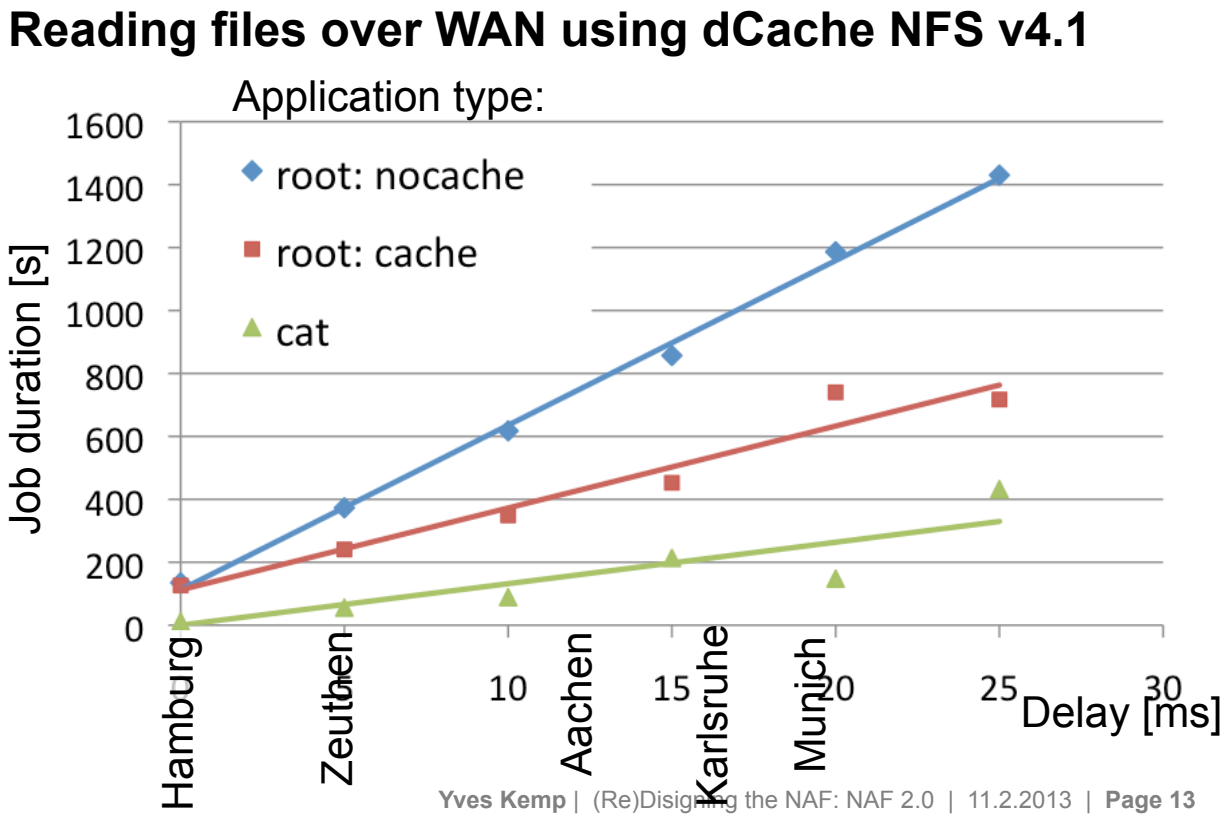


NAF 1.0: Detailed look at networking

- The two-site setup had as an effect that most disturbances at on site had repercussions on the other site and the NAF as a whole
  - Only a very small number of services were deployed in a redundant way
- Investigation on latency effects on dCache NFS 4.1 mounts for ROOT file access: Lab emulation of latency

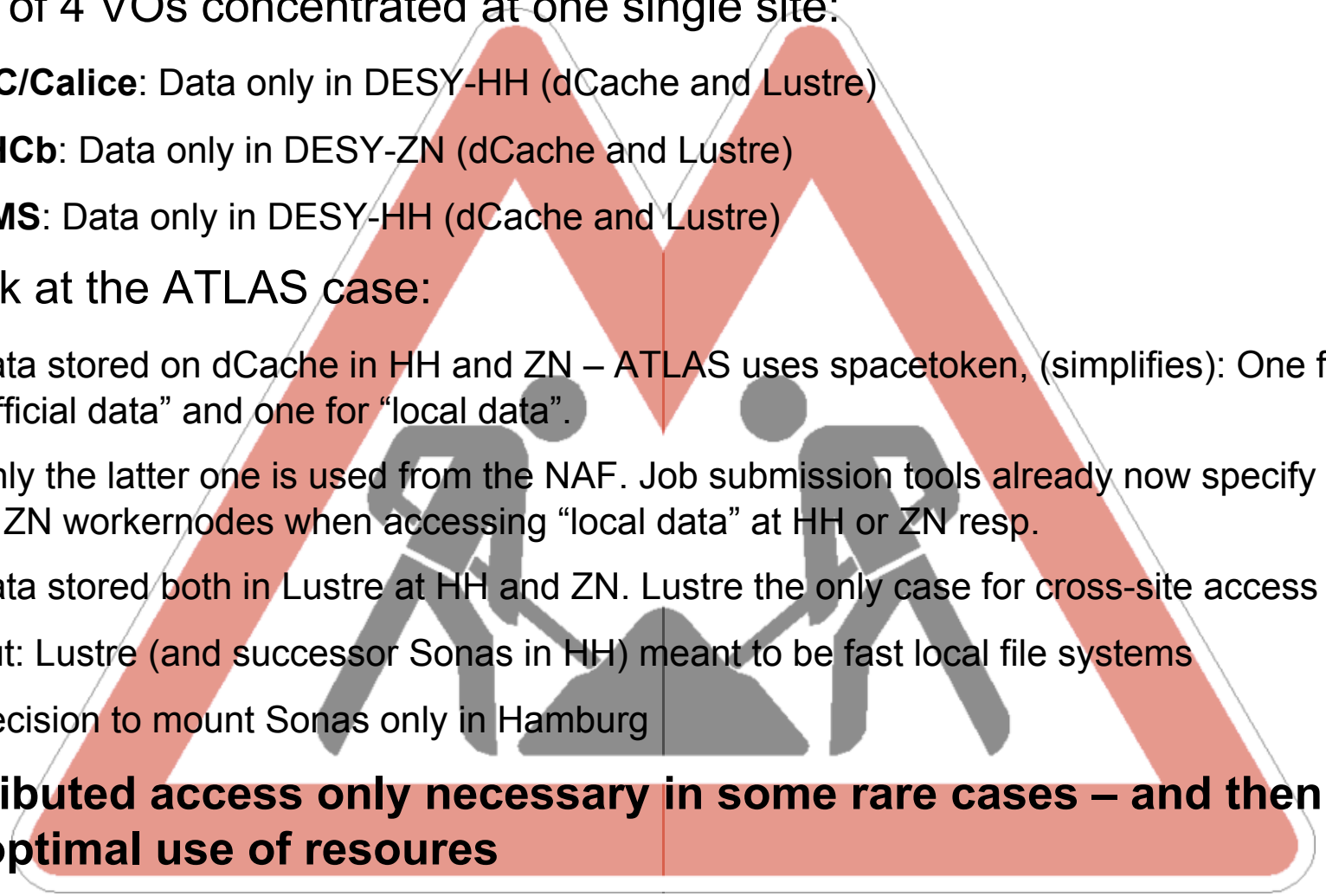
Ping times:
HH-WN -> HH-dCache
<0.2 ms

HH-WN -> ZN-dCache
~ 5.5. ms

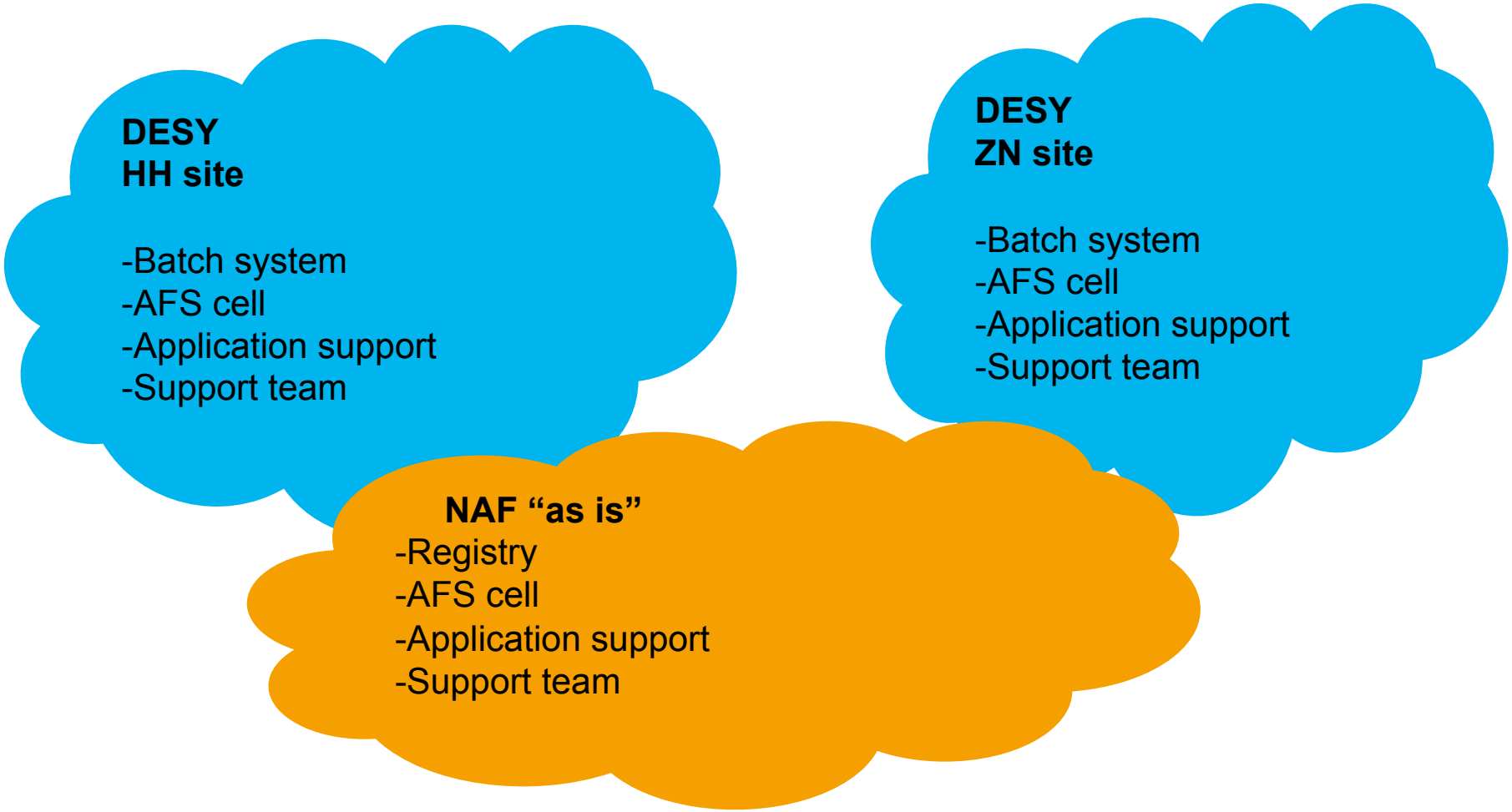


NAF 1.0: Is distributed access really needed?

- 3 out of 4 VOs concentrated at one single site:
  - **ILC/Calice**: Data only in DESY-HH (dCache and Lustre)
  - **LHCb**: Data only in DESY-ZN (dCache and Lustre)
  - **CMS**: Data only in DESY-HH (dCache and Lustre)
- A look at the ATLAS case:
  - Data stored on dCache in HH and ZN – ATLAS uses spacetoken, (simplifies): One for “official data” and one for “local data”.
  - Only the latter one is used from the NAF. Job submission tools already now specify HH or ZN workernodes when accessing “local data” at HH or ZN resp.
  - Data stored both in Lustre at HH and ZN. Lustre the only case for cross-site access
  - But: Lustre (and successor Sonas in HH) meant to be fast local file systems
  - Decision to mount Sonas only in Hamburg
- **Distributed access only necessary in some rare cases – and then it is not optimal use of resoures**

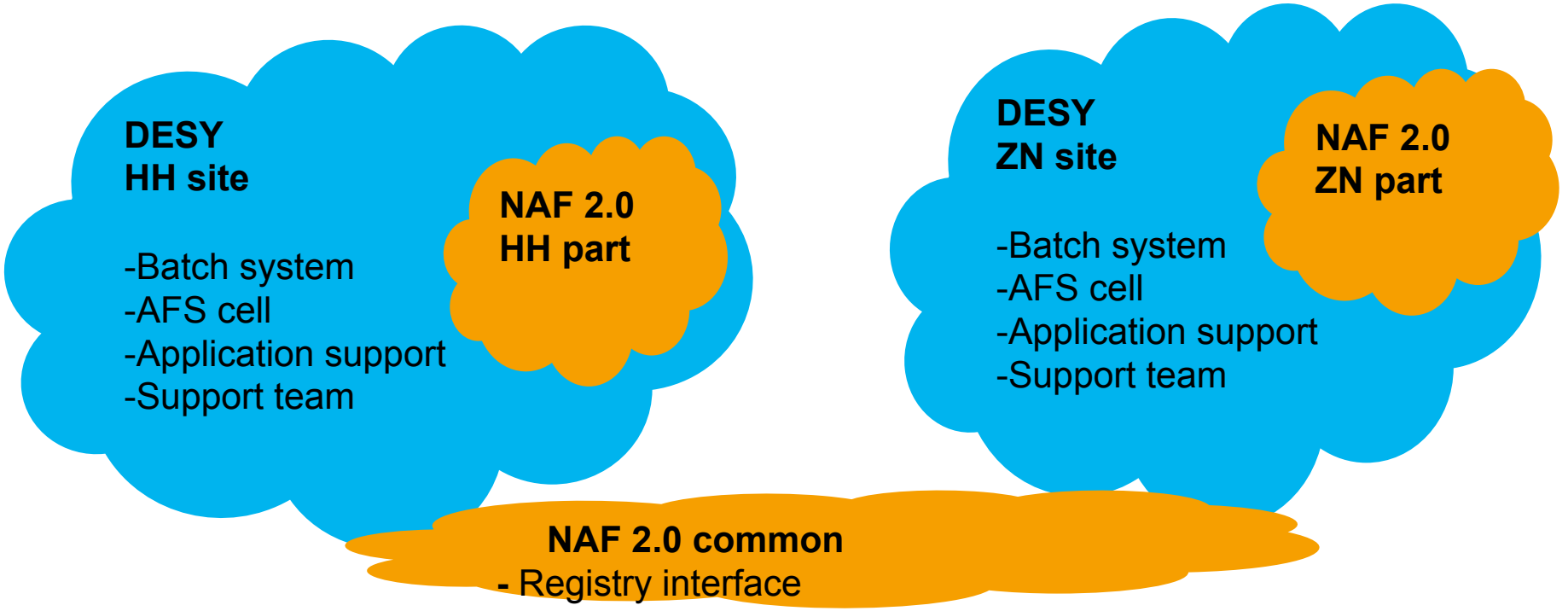


NAF 2.0: How? A very broad picture (oversimplified)



(in reality the two DESY sites are not as separated as shown here)

NAF 2.0: How? A very broad picture (oversimplified)



(in reality the two DESY sites are not as separated as shown here)

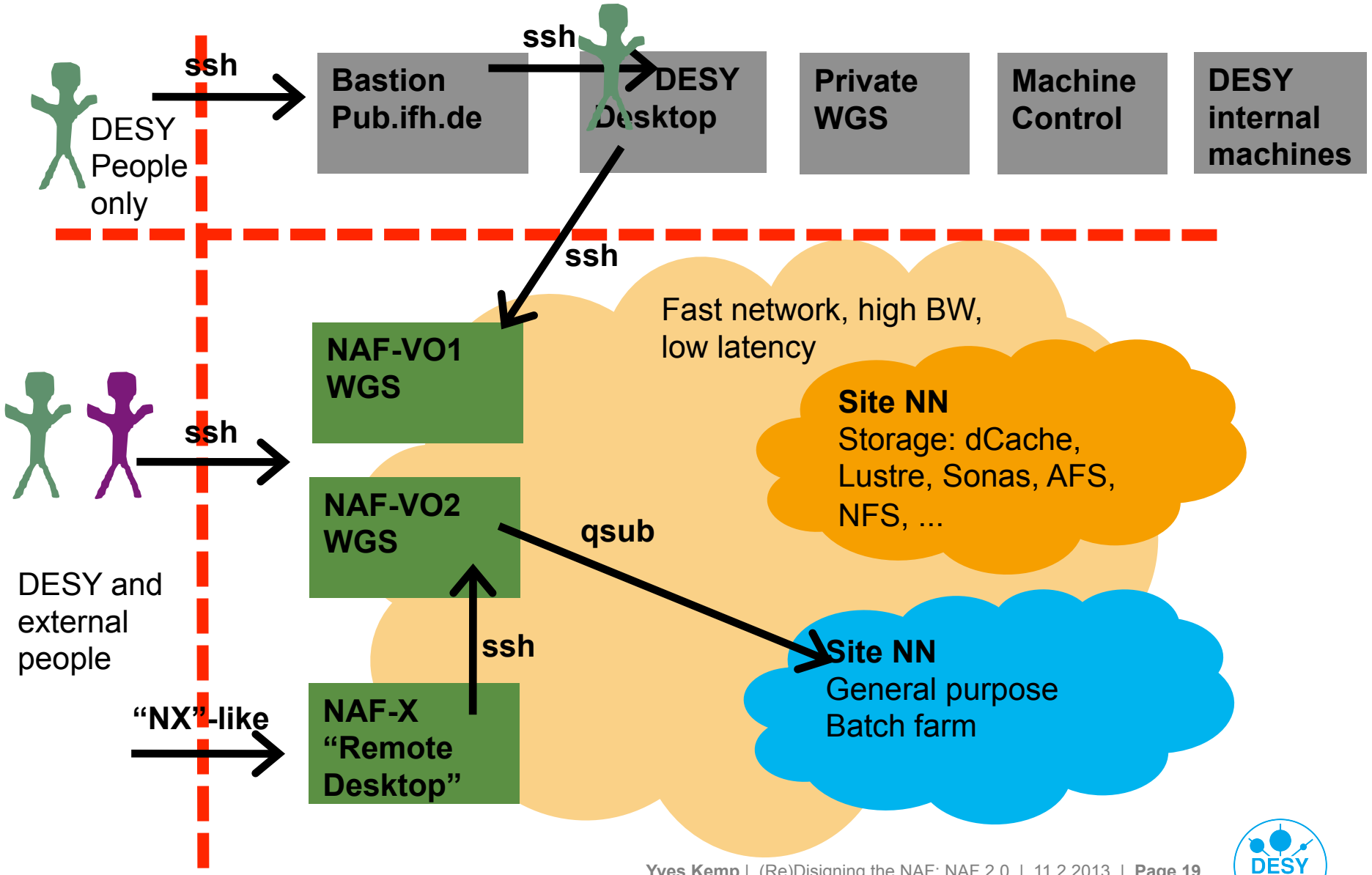
Future of the NAF: NAF 2.0 - What? _ 1

- Everyone will get a “normal” DESY account
- Will have access to a restricted set of “normal” DESY resources
  - The NAF 2.0 resources
  - Including several WGS
  - Including large batch system
  - Including $LargeFileStore (e.g. Sonas)
  - Including dCache access
  - …
- Technical details:
  - Closer integration into respective site (HH or ZN)
  - No data should go over the WAN
  - This is enforced in case of $LargeFileStore
  - Plain ssh+Password login – gsissh planned for later

Future of the NAF: NAF 2.0 - What? _ 2

- Support: Better integrated into site support, so more people know the infrastructure
- New developments
  - Ability to use DESY maintained software products
  - Graphical login (“NX”-like, using StarNet X-Win32/LIVE technology)
  - GridFTP access to Sonas planned
  - Support for new communities: BELLE
  - …
- The ATLAS case:
  - ATLAS is distributed over two sites, would lose a homogeneously looking system.
  - Decision: Expand resources at HH-site to offer full NAF capabilities at a single site (Storage, dCache & CPU)
  - Expansion will start in early 2013
  - Role of ZN-site will probably evolve to resource and support provider for local ATLAS group

NAF 2.0: Broad overview of one site



Current status in Hamburg:

Workgroupserver:

- ATLAS, CMS, ILC, BELLE – HeraFitter project to join
- Some SL5, others SL6 – depending on whishes of the VO (all 64bit)
- Accessible from outside – IT managed

Batch facility:

- Hamburg site general purpose batch farm BIRD is used
- Some new nodes purchased in NAF 2.0 context, some out-of-warranty machines from Grid cluster used for transition period
- Mixture of SL5 (~1500 cores) and SL6 (~200 cores)

Storage:

- dCache setup and access will stay untouched
- AFS: Using the DESY cell /afs/desy.de
- Lustre & Sonas: see later

Current status in Zeuthen

- A reminder: Decision: Expand resources at HH-site to offer full NAF capabilities at a single site (Storage, dCache & CPU)
  - The NAF 2.0 concept will however be used for local groups (Zeuthen and Humboldt e.g.)
- Workgroupserver exists for local ATLAS group
- 192 CPU cores in local Zeuthen batch farm
- 120 TB Lustre space
- 550 TB in ATLAS local group disk

  This space will probably not be expanded in future. The HH local group disk space token will serve as the main NAF 2.0 dCache space

Status: Getting an account

- Prerequisite for getting an account: Being known at DESY as a person
- We need to enter you in PIP system: PersonenInformationsPool
  - Name, Firstname, Affilitation, Date&Place of birth, … and this needs to be accurate, unique and somewhat certified
- DESY people (or Uni-HH, Humboldt Uni and other befriended institutes)

  - You are already registered in PIP !
  - Your normal DESY account (the one you use for bastion.desy.de) just needs the resource “batch” in the registry – and there you go
  - Already some (DESY-based) brave test users – Thank you!
- External people

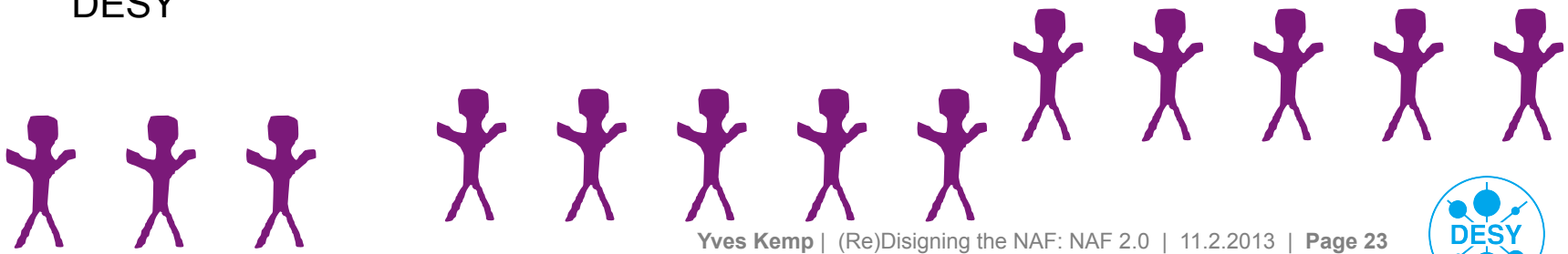
  - You need to get registered in PIP
  - Currently setting up a registration where you can enter all needed information – certified by your Grid certificate that you have in your browser
  - The PIP entry and the account creation is then done “automagically”
  - We expect the first sketch of the system latest end of February

… a side remark on the PIP issue

- This underlines the new role of DESY

- More and more people use only some DESY resources for a restricted time
  - HERA times: Lots of external people – but rather long and intense usage of resources
  - Photon science: Very mobile community, usage of many light-sources

- Secure, authenticated and authorized (and easy) access to computing important!
  - Cannot expect people to fill in a paper form and hand it in in person somewhere at DESY



When / how should users migrate?

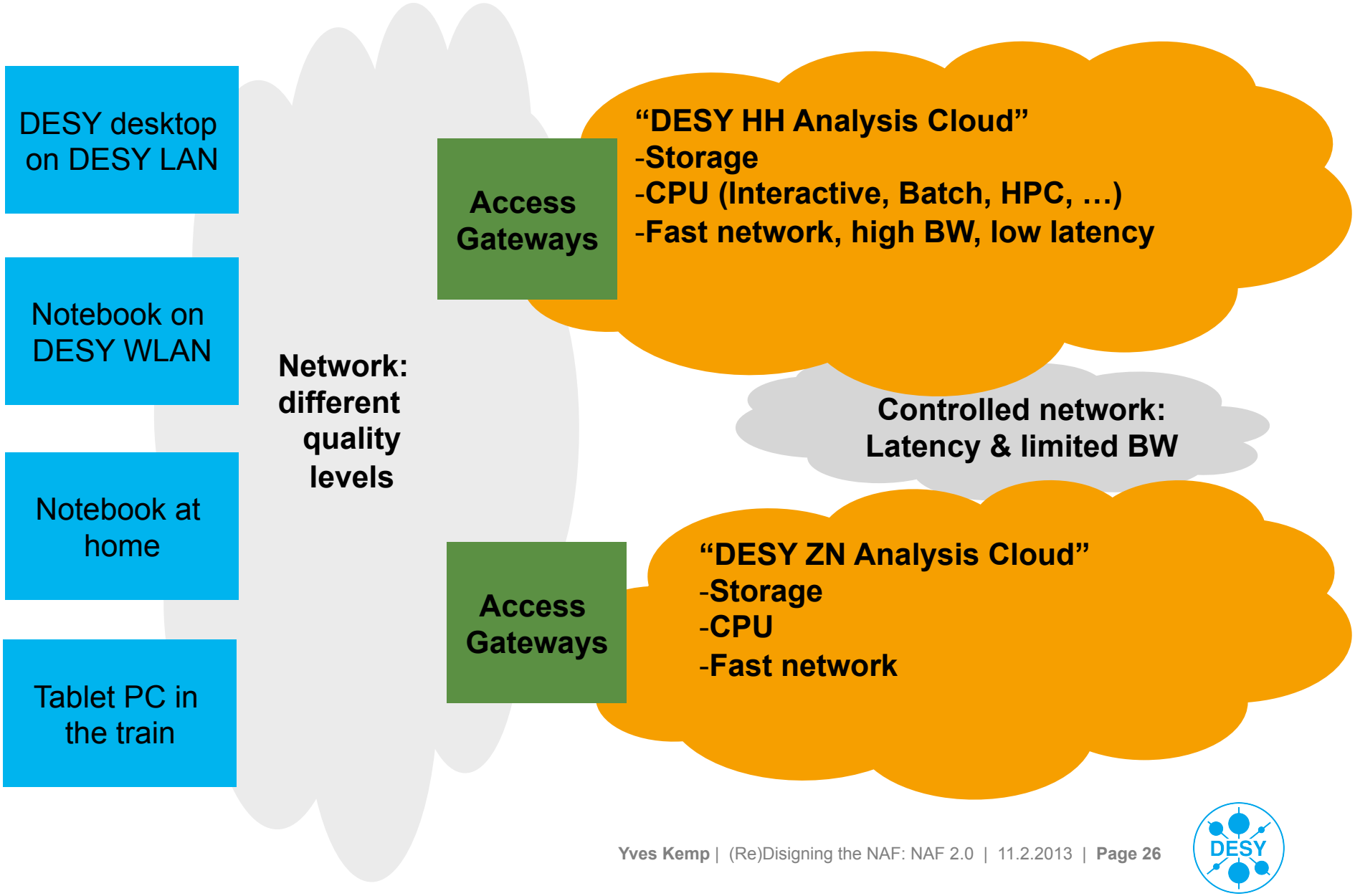
- NAF 2.0 is not yet fully operational
- Most parts can however already now being used
  - Depending on exact workflows, everything might be covered by NAF 2.0 already now
- If you are brave, you can/should try out now – contact your VO representative!

- We expect a larger migration to NAF 2.0 after the winter conferences
- There will be a coexistence NAF 1.0 and NAF 2.0
  - Details discussed with NAF Users Committee
- **… but our plan is to shutdown NAF 1.0 in 2013 !**

Lustre / Sonas and migration of data

- Lustre in HH: … will fade away with NAF 1.0 (or even before)
- Sonas in HH: Successor of Lustre
  - Parts are mounted in NAF 1.0
  - Other parts are mounted in NAF 2.0
  - … cannot mount the same part in both NAF: UID clashes (NFS v3 mount and security)
- Migration:
  - Sonas organized in “filesets”, e.g. each user has an own fileset
  - Can migrate one fileset at a time
  - … but it is either NAF 1.0 or NAF 2.0
  - … a problem for e.g. group filesets – these will probably need to be copied and provided twice
  - Lustre migration: e.g. copy (scp) Lustre@NAF1.0 to Sonas@NAF2.0

... No IT talk nowadays without talking about “The Cloud”



Summary and Outlook

- NAF 1.0 was (and is) a success
  - But the world has changed – need a redesign to continue success story
- NAF 2.0 is partially there right now
  - … if you are brave, you can/should try it now!
  - Plan to migrate to NAF 2.0 in 2013 – and shutdown NAF 1.0
  - Well on track – missing parts being worked on right now!
- NAF 2.0 a blueprint for communities beyond HEP? **YES!**