NUC

05.09.2024

Kemp, Yves
DESY HH

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

NAF special incidences (upgrades) since last NUC

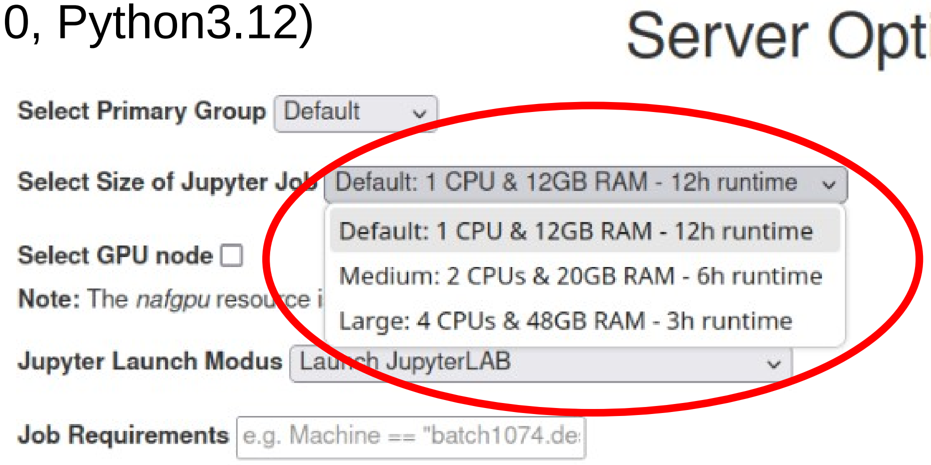
Upgrade to EL9 of all instances

- Upgrade to EL9 (RedHat Enterprise Linux) as planned and announced due to end-of-life of EL7
  - Phase 1: Provision of EL9 WGS for all VOs submitting on 2 EL9 worker in existing pool
  - Phase 2: Redirect EL9 WGS into new EL9 pool
  - Phase 3: Migrate ressources from the old pool into the EL9 pool
  - Migration completed 16-07-2024
- Migration for early EL9 user theoretically interruption free !
  - In reality there was a short gap of ~3h where both pools were not accessible due to a misconfiguration
- Lessons learned
  - Migrating to EL9 much more demanding than it would have been to EL8. EL9 surprisingly for us seemed like bleeding edge technology for batch systems (e.g. late-materialization, CGroupsV2 etc)
  - We underestimated the time we needed to clean up the old config in puppet and roll out a production type EL9 version of the pool
  - Always calculate some spare time – the very last minor upgrade of Condor 2 days before final shutdown of the old pool corrupted the Kerberos token handling of the pool and caused 3 days of grief to fix it

NAF Software

Next generation JUPYTER notebooks

- JHUB & notebooks upgraded (JupyterHub version 5.0.0, Python3.12)
- New notebook classes:
  - Default: 1 CPU / 12 GB RAM / 12h runtime
  - Medium: 2 CPUs / 20 GB RAM / 6h runtime
  - Large: 4 CPUs / 48 GB RAM / 3h runtime



Server Options

Select Primary Group: Default

Select Size of Jupyter Job: Default: 1 CPU & 12GB RAM - 12h runtime

- Default: 1 CPU & 12GB RAM - 12h runtime
- Medium: 2 CPUs & 20GB RAM - 6h runtime
- Large: 4 CPUs & 48GB RAM - 3h runtime

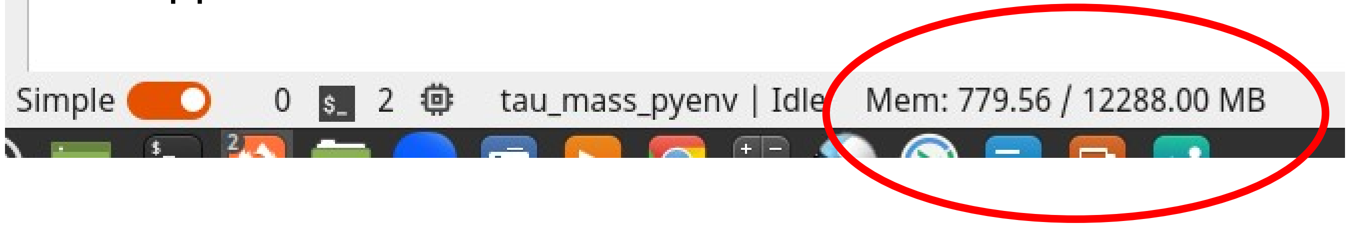
Select GPU node □

Note: The *nafgpu* resource i

Jupyter Launch Modus: Launch JupyterLAB

Job Requirements: e.g. Machine == "batch1074.de

- Default notebooks run on all pool nodes (similar setup to old pool)
- Medium & large notebooks run on 2 dedicated servers
- Feedback about new sizing and user experience appreciated
- RAM taxometer now in place



Simple 0 $_ 2 tau_mass_pyenv | Idle Mem: 779.56 / 12288.00 MB

- Suggestion: Have a ‘show-us-your-notebook’ session later this year in order to connect notebook users over VO/batchsystem borders and discuss further experiences and needs

NAF Storage (1)

dCache

- Experiments ATLAS and CMS have deprecated SRM for file access
- SRM was stopped for ATLAS and is no longer available
- SRM for CMS kept available until the Update tot 10.2 (next golden release in early ‘25)
- SRM for Belle II and ILC available until deprecated by experiments
- Update to RHEL9 → BDII no longer available and therefore a port must be given:
  - srm://dcache-se-cms.desy.de:8443/pnfs/desy.de/cms
  - srm://dcache-se-desy.desy.de:8443/pnfs/desy.de/belle
- Better yet: use the WebDAV endpoints
  - davs://dcache-atlas-webdav.desy.de:2880/pnfs/desy.de/atlas
  - davs://dcache-cms-webdav.desy.de:2880/pnfs/desy.de/cms
  - davs://dcache-desy-webdav.desy.de:2880/pnfs/desy.de/belle

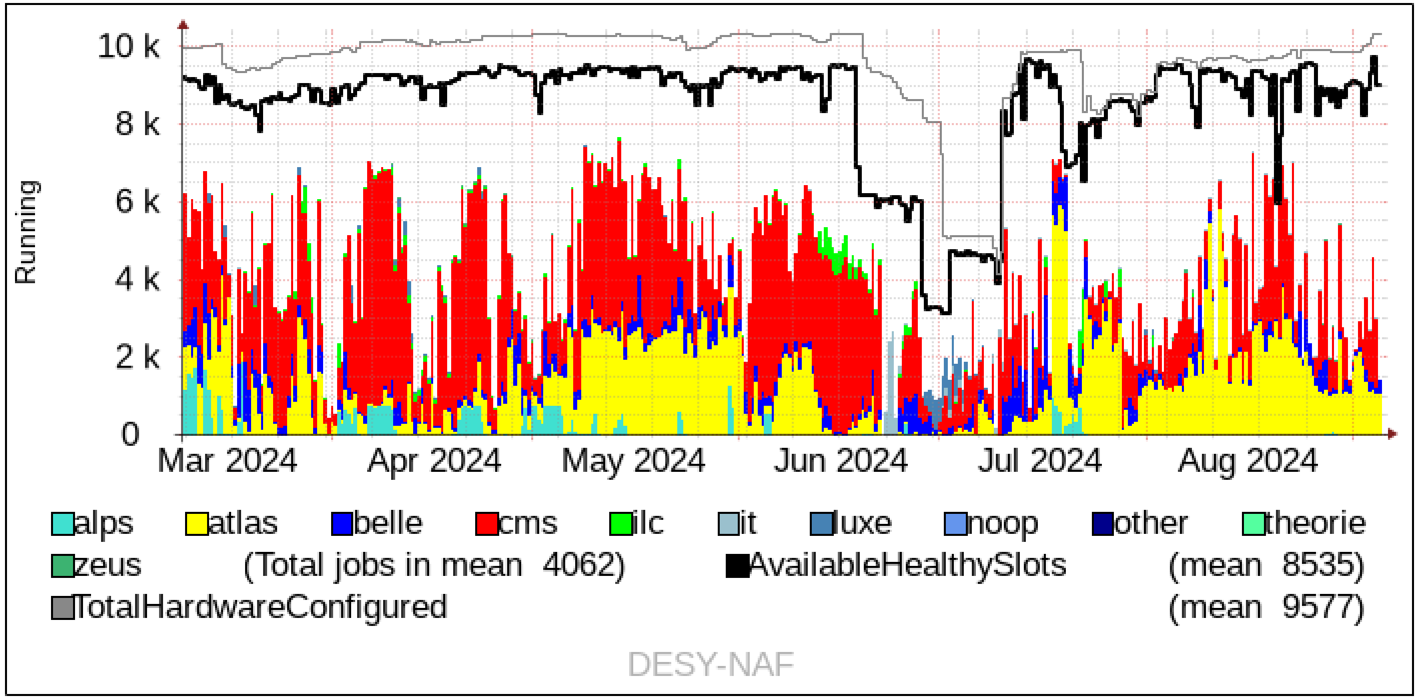
NAF Storage (2)

DUST

- Extension of Storage delivered, integration in September 2024
- Software upgrade of current DUST storage block and NFS servers
  → preparation for integration of new block
- As usual:
  - Upgrades are concurrent, no downtime required, “at risk”
  - Less bandwidth available for I/O operations
  - Short hangs during NFS failovers (<= 90s), applications will just stall
- Exact date/time TBD
  → will be announced through the usual support channels to users

Batch occupancy

- NAF occupancy quite low in the past 6 month … and decreasing after EL9 migration
-



Upcoming PRC

- Next PRC is 5/6 November 2024
- Usually, we have an combined NUC+PRC preparation meeting before
- Will propose a data first half of October

Upcoming DUST changes and new Login Concept Ideas for NAF

IDAF: Getting NAF & Maxwell closer

Name Surname
City, Date

Current User/Project Storage

Different Storage for NAF & Maxwell

NAF (HTC cluster)

- DUST as fast scratch & project space
- Quota per user & group, neither backup nor snapshots*
- No self-service: Registry Resource
- Very granular directory structure, possibility for multiple user directories
- Access via NFSv4 from NAF WGS and worker nodes
- Based on GPFS, connected to Maxwell InfiniBand fabric for internal communication

Maxwell (HPC-like cluster)

- BeeGFS as fast scratch & project space
- Neither quotas nor backup or snapshots
- Self-service: mk-beegfs
- Performance issues for some workloads and administrative issues (removal/adding of servers)
- Replace BeeGFS with DUST?
  - Unify scratch & project space between NAF & Maxwell
    → One more step towards IDAF :
    Interdisciplinary **D**ata and **A**nalysis **F**acility
  - Fun fact: DUST is already mounted on Maxwell...

DUST Extension

Subheading, optional

BeeGFS & DUST

- BeeGFS size: 1.6 PiB, need >= 2.0 PiB
- DUST size: 3.1 PiB, 2.0 PiB used
  → not enough space
- DUST Extension: ~2.0 PiB extension of DUST ordered, delivery September 2024
- But how to implement this?
  - New & dedicated filesystem for Maxwell?
    → 👎
- To get closer to IDAF:
  Extend current DUST and implement
  unified access from NAF and Maxwell

Placeholder

- Next slides for unified DUST on NAF & Maxwell

Current DUST Setup

Subheading, optional

Issues with current setup

- Very granular directory structure:
  ```
  /nfs/dust/GROUP/user/ACCOUNT
  /nfs/dust/GROUP/group/PROJECT
  ```

  -
    ```
    /nfs/dust/ilc/user/sdietric
    /nfs/dust/atlas/user/sdietric
    /nfs/dust/atlas/group/zeed
    ```
- Works well, for a limited number of groups…
  - Recent new groups:
    Axion (ALPS II, MADMAX, IAXO), LUXE, M-division, IT
  - Group == Registry Namespace
- Even worse on Maxwell: >= 50 groups

Naming Paths is hard

- Current directory scheme does not scale well
  - Duplicate user directories due to **GROUP**
  - High administrative overhead
  - Results in **too** granular quota management
- Mountpoint encodes a protocol
  - On Maxwell: /**gpfs**/dust/
  - On NAF: /**nfs**/dust/
- To unify access and reduce admin overhead, a restructure is necessary

Proposed Plan

Subheading, optional

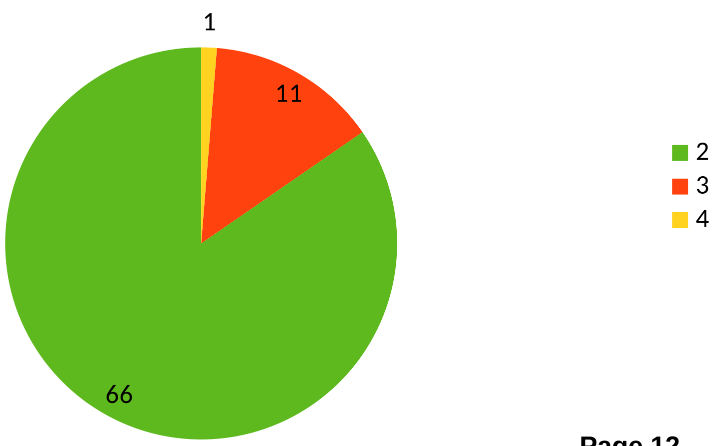
Simplified directory structure

- Protocol independent mountpoint `/data`
  - `/gpfs/dust | /nfs/dust`
    `→ /data/dust`
- Removal of **GROUP** in the user paths
  - User Directories
    `/nfs/dust/GROUP/user/ACCOUNT`
    `→ /data/dust/user/ACCOUNT`
  - Project Directories
    `/nfs/dust/GROUP/group/PROJECT`
    `→ /data/dust/group/GROUP/PROJECT`
- Result:
  single user directory & less admin overhead

Migration & Issues

- New directory structure requires data migration
  - How to merge users with multiple directories?
  - Access to user folder from multiple groups with UNIX mode bits?
- Migration proposal:
  Migration per-group, minimal downtime for final delta copy

# Users with multiple directories

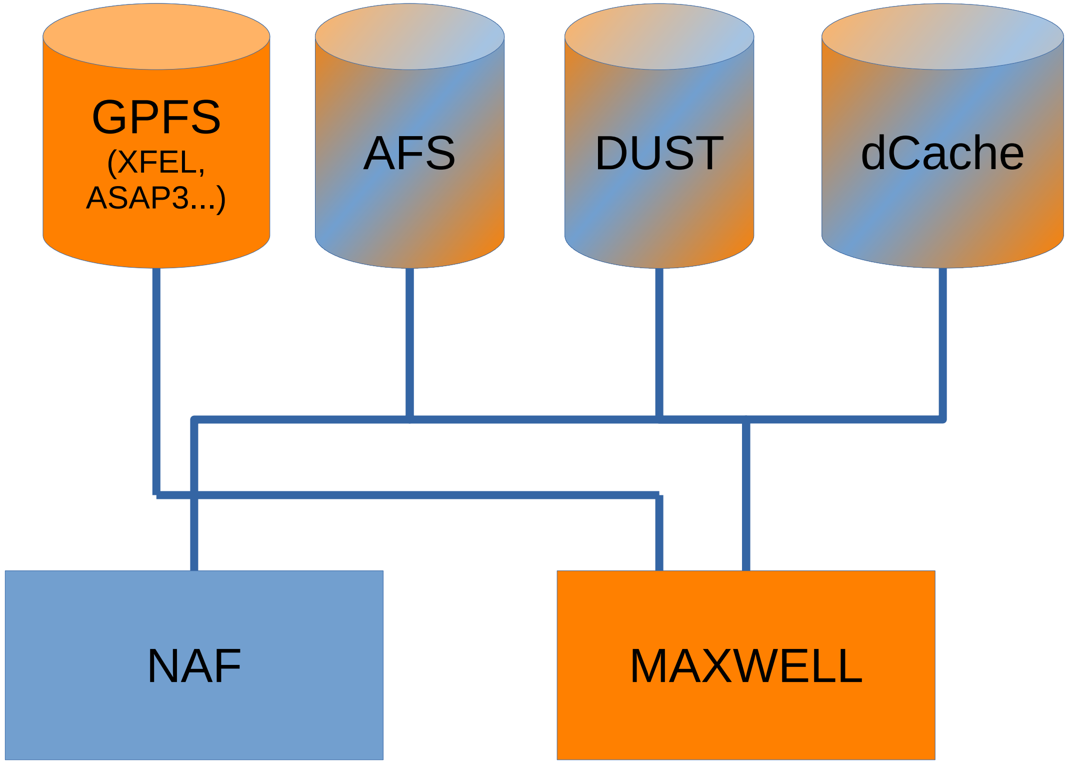


Result

Unified access to user/project space

Too long; didn't read

- Unified access to the same project space between NAF and Maxwell
  - New path: /data/dust/user & /data/dust/group
- Other filesystems, like /pnfs, AFS, CVMFS, NetApp NFS are *not* (yet?) affected by this change
  - Mountpoints are already mostly identical between NAF and Maxwell
- Single user directory needs some consideration for sharing data between different groups
- Reduced admin overhead results into lower entry burden for new users/groups

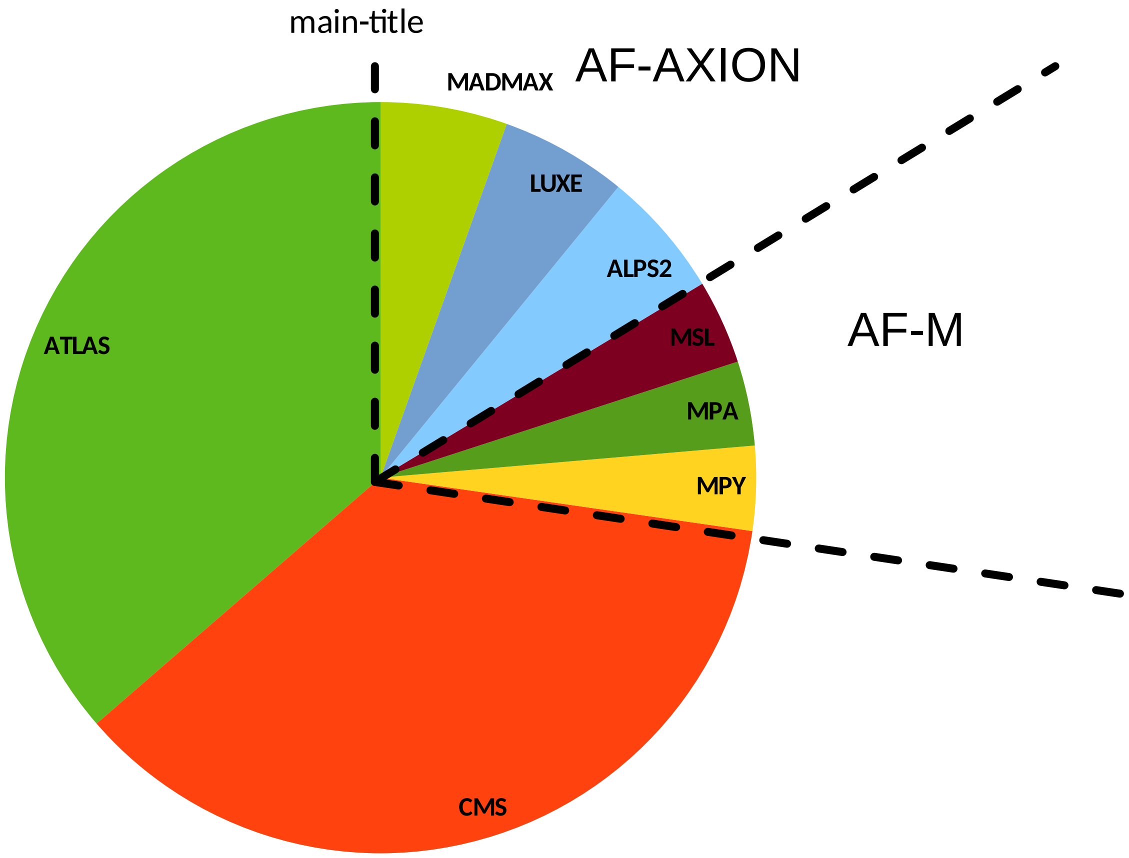


Quota Management

Reduce fragmentation by creating bigger groups

Simply Quota Management as well

- “Virtual” namespaces for groups of common interest
  - Reduces quota management overhead
  - No need to shuffle around maximum quota values
  - Flexibility: fragmentation still possible!
- Changes for current groups
  - Big groups (ATLAS, CMS): No changes
  - Smaller groups (Axion, M-Divison, Belle1/2):
    Group into bigger “virtual” namespaces
    → virtual namespace == RGY namespace
  - Very small groups:
    Introduction of catch-all resource
- Quota Management Tool: Amfora

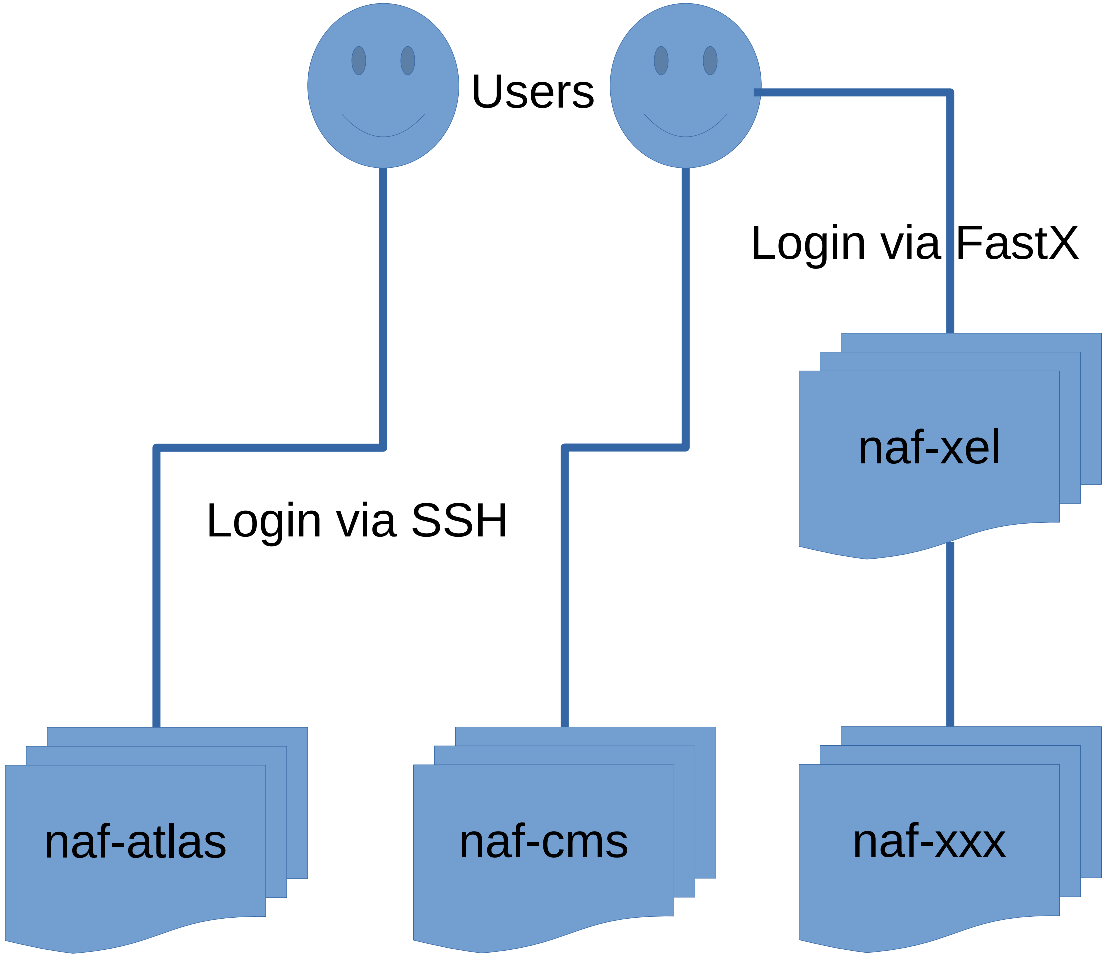


New Login Concept for NAF

Reduce fragmentation & easier graphical access

Current Login Concept

- Each group has its own WGS:
  - naf-GROUP.desy.de
    → naf-atlas.desy.de, naf-cms.desy.de, naf-alps.desy.de etc.
- Access tightly controlled via Registry resources
  - ATLAS users can not login on CMS nodes
- Primary group membership fakery
  - Primary UNIX group of users are changed to project group
    → ATLAS → af-atlas
    → CMS → af-cms
- High entry burden: Wanna test NAF? Yeah, we need to create a new WGS first...

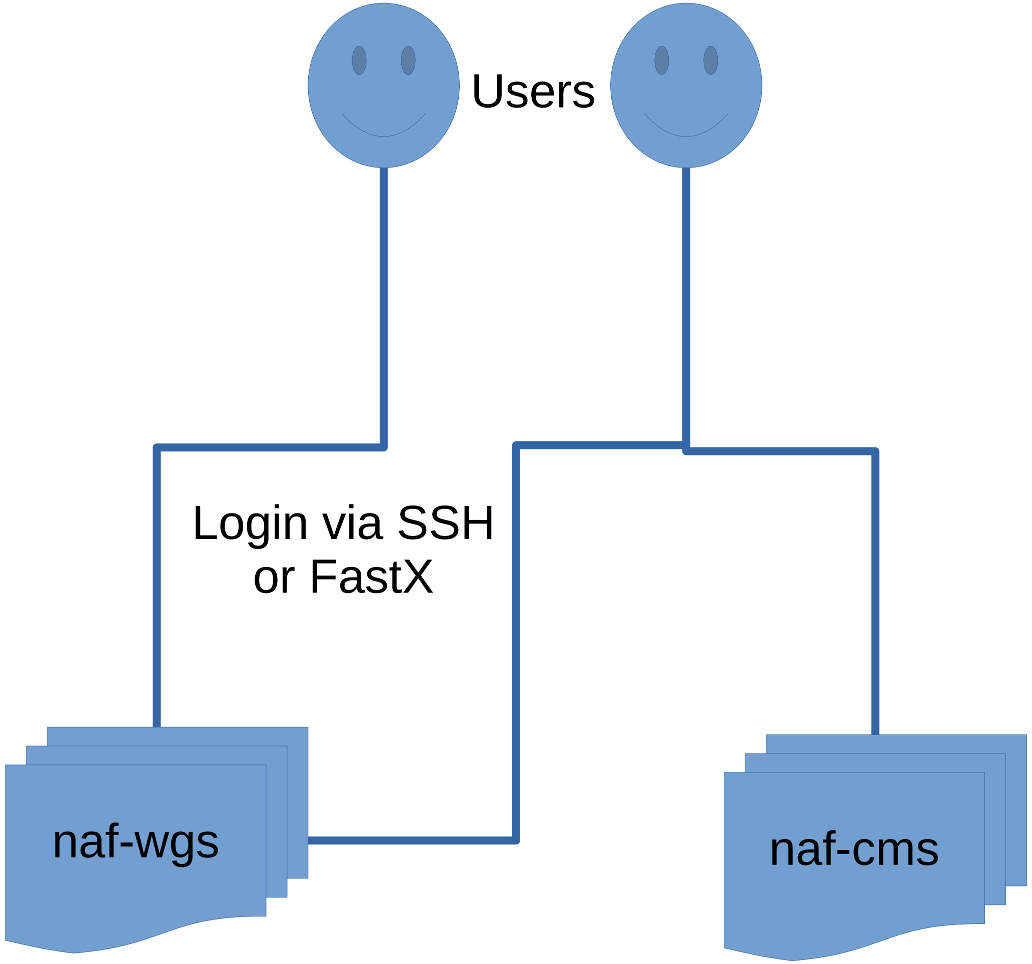


New Login Concept for NAF

Reduce fragmentation & easier graphical access

Current Login Concept

- Remove WGS per group concept
  → shared login nodes across all groups
  - Big groups can always buy dedicated HW
- Similar concept to Maxwell Display Nodes
  - Login either via SSH or directly via FastX
    → easier graphical access
- Drop primary group fakery
  - Primary group as defined in RGY
  - For DUST group space: No big deal, due to ACLs
  - For DUST user directories: sharing data across multiple groups might be harder



New Login Concept for NAF

Reduce fragmentation & easier graphical access

Access via Resources?

- TBD: How to grant access to naf-wgs or group specific wgs?
  - Old model: granular access for known NAF groups
  - Very granular: additional resources
  - Less granular: allow every batch users

/etc/security/access.conf		
	naf-cms: @af-cms	
	naf-atlas: @af-atlas	
naf-wgs: @batch-users → allow every batch user	naf-wgs: @af-axion @af-m @af-it → granular access, allow known NAF group	naf-wgs: @mpy-users @mpa-users → very granular access

Discussion @ IT:

- Same-WGS-for-all: Works well for Maxwell:
  - WGS-per-group simply would not work: each proposal would be its own group
  - Sharing data between proposals not foreseen, people use other means
- WGS-per-group: Works well for the larger NAF groups
  - Because there are (better: were) a small, static number of larger groups
  - Tedious for smaller groups
  - Sharing data between groups is technically possible via user directories
- Same-WGS-for-all @ NAF:
  - Would work for people only in one group, not sharing/accessing other groups data
  - People offering shared data might need (complicated?) tooling to set access rights correctly
- → Our take is: Do not change the WGS-per-group at the moment … but open for discussion