# NUC

10.12.2024

Kemp, Yves
DESY HH

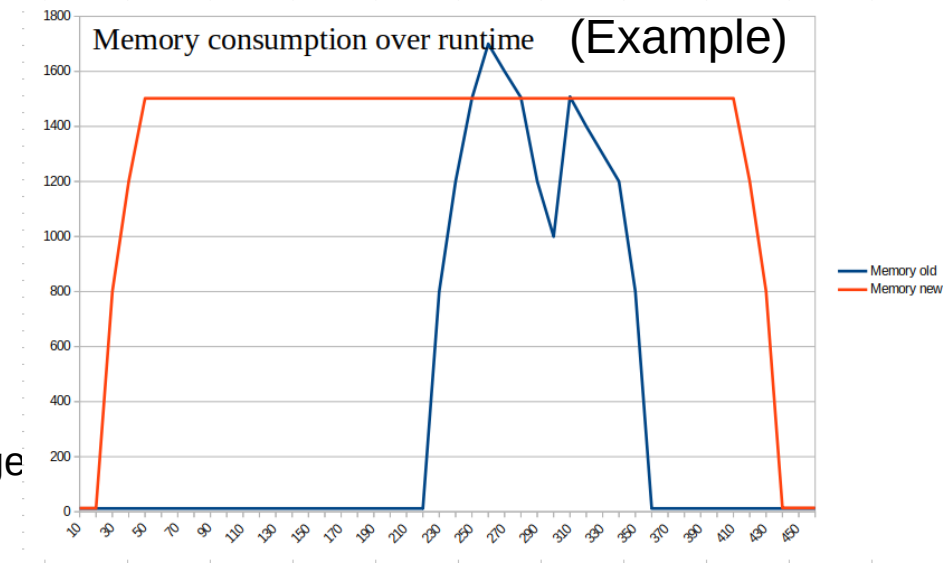
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

# NAF special incidences since last NUC

## A lot more problems than expected

- Jupyter Notebooks
  - Some problems with start of ‘small’ notebooks on faulty nodes were caused by 2 checks not working correctly after syntax change in Condor (fixed since last week)
  - The memory over commitment problems from earlier and locked worker nodes did not help either
  - The current Jupyterhub version has some bugs related to reconnecting to running notebooks and removing links to stopped notebooks, the sync of the internal routing scheme with the ‘reality’ is sometimes failing and locking users out
    - We have identified the problem and run different probes to narrow the problem down & limit the fallout, a bug report is in progress
- Memory over-commitment
  - Historical limit at 3 x reserved memory
  - Possible by horizontal filling, peak usage and delayed job starts
  - Since a while jobs have a different memory allocation behavior
    - Reserve everything early, free nothing till the end of job runtime
  - Swapping is not practical: large nodes with a lot of cores and relatively small disks – disk performance also not adequate for active memory usage



# NAF special incidences since last NUC

## A lot more problems than expected

- Massive bulk submission of high memory jobs did send a lot of worker into swap and inoperability resulting in loss of all jobs on the affected nodes
- New memory policy in place: 1.2 x reserved memory is the new limit (violating jobs go into hold with according hold reason)
- Memory accounting in Condor does not seem to be very exact but we are in contact with the developers and raised our concerns (again interaction with CgroupV2 seems to be part of the problem)
- Problem with CgroupV2 process tree handling of condor revealed at the same time, processes going stale, blocking slots on a large scale
- Update of Condor from feature version to last LTS version as according to changelogs and Linux with RHEL 9.4 → 9.5 upgrade fix at least partly the problems we see
- Currently the CVMFS client is locking whole workernodes under some conditions that are not yet fully clear to us (filed a bug report and keep investigating)
- In addition the recently updated YFS client is unstable and causes occasional kernel panics bug fixes came out last week and were installed and activated over the weekend

# NAF special incidences since last NUC

## Common development of usage behavior

- Usage patterns have changed a lot recently, people use ‘phd-homemade’ frameworks e.g. based on DASK on top of the NAF
- The house keeping of DASK itself is partly broken which leads to a lot of jobs not being removed, running into time limits instead
- In addition these frameworks get handed over to other users and provide default features like automatic retransmission
- We see partly success rates of these users below 10% (90% +x of lost jobs) because people do not request ressources correctly and only a minimal fraction of their jobs escapes accidently from being stopped for violating memory or time boundaries. Automatic retransmission hides the problem and just keeps resending the jobs until all made it
- Little consciousness on the user-side, no reaction to automated e-mails apparently
- Maybe think about setting limits based on past failure rates?
- Keeping condor happy must become one of the main goals in order to use it successfully

# NAF special incidences since last NUC

## Usage of the WGS

- Not only as a part of the sustainability workshop Q&A session we regularly receive complains through tickets about login server being ‘slow’
- Apart from rare individual problems like a full AFS home and similar issues the misusage of the WGS as a workhorse for interactive analysis is usually the problem
- In the past we did not feel authorized to kill individual processes of users and relied on the users educating each others which seems to work less or at least we receive more complains than in the past
- We have rolled out a new OOM killer that kills user processes rendering the hosts unresponsive
  - Details still to be worked out, e.g. user feedback/information is currently missing
- The foreseen procedure for higher individual non-batch workloads is to reserve a decent sized slot by using ‘condor_submit -i’ and then run everything in the slot.
- We circulated an e-mail with instructions and got positive feedback, same thing works with Jupyter notebooks with a little bit of tweaking
- Will send an additional e-mail with instructions and put those in a Q&A section that will be prominent in the documentation

# DUST Status

## Operational Issues

### Software Upgrades

- 2024-10-22..23: Software upgrade of existing storage block
  - No issues or user impact during upgrade
  - Enclosure firmware update failed
    → solved ~2 weeks later with the vendor
- 2024-10-17: Software upgrade of NFS servers
  - Upgrade to latest LTS GPFS and NFS server (Ganesha) release
    → Update worked without issues
  - Evening: Ganesha hanging, which triggered failovers, but failover was hanging as well?!
    → sent out trouble notice to naf-users

### Ganesha Upgrade Issues

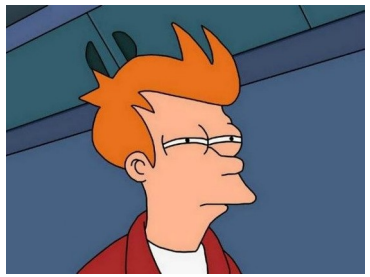
- Workaround: Restart all Ganesha servers
- No stable situation in the night, NAF dead in the morning due to DUST issues
  - NAF Jhub has hard operational dependency on DUST
    → very high impact
- Performed downgrade of Ganesha to older version
  → stable operation restored, opened support case with vendor
- Vendor answer: 1x known stability issue, 1x bug in LTS release (fixed in latest feature release)
  → applied efix on 2024-11-13, now stable operation

# DUST Status

## IDAF Migration

### IDAF

- Reminder: Simplify directory structure, merging smaller group to bigger group, new /data/dust mountpoint
  → see last NUC for details and docs.desy.de/idaf
- New storage block with 2.2 PiB integrated to DUST on 2024-11-25
- Maxwell migration finished on 2024-12-02
  - Suspicious silence from users, no tickets with questions or complaints
    → perhaps it's a smaller issue than we've envisioned?!



### Migration Plan

- Keep going with the migrations to restore stable structure ASAP
- Proposed plan to NUC:
  - H1, ZEUS, HERMES & FHLABS migrated
  - BELLE on 2024-12-17: medium size and little overlap with other groups
  - CMS: Christmas break for shadow copy, migration in 2nd or 3rd week of January
  - ATLAS, ILC, AXION experiments: end of January/early February
- If we stick to this plan: Migration finished by end of January or early February 2025
  → stable structure restored

# DUST Status

## BELLE Migration – Timeline & Procedure

- BELLE first NAF user group for migration on 2024-12-17, 8:00 AM to 4:00 PM
  → blueprint for all other groups
  - 1st shadow copy finished in ~36h, delta copy ~6h
    → ~8h of downtime required
- Limit job submission for BELLE users, implemented today by Condor Admins
  - Jobs with runtime ending **before** maintenance → scheduled as usual
  - Jobs with runtime ending **after** maintenance → will remain queued and run after maintenance
  - Goal: reduce badput to avoid killing many jobs on maintenance day, jobs with short runtime will run as usual
- Unmount of /nfs/dust/belle1+2 on WN and WGS
  → kill remaining processes on WGS nodes
- After downtime: Users have to adjust paths

# Miscellaneous

- PoF-IV review going on … some load on people, and potentially requests for user-views/statistics on short notice
- 7.1.2025: Work in the compute center and network:
  - Network connection between RZ1&RZ2 and the campus and internet down for ~.5h
  - NAF jobs will most likely continue to work
  - Grid jobs might be disturbed if external connectivity is used
  - Any interactive access will break (WGS, jupyter,...)