News from NHR@FAU
cluster configurations, resource monitoring, case studies

HPC Café, 13 September 2022

HPC Services, NHR@FAU

16:00-16:30 coffee & cake
16:30 start of the presentation

High Performance Computing

cluster configurations

Major changes in September 2022

Updates in September maintenance - Meggie

- **Tier3 parallel computer Meggie**
  - OS upgrade from CentOS7 to **AlmaLinux8** (again a RHEL clone)
  - Batch system and queue configuration still is the same
  - $FASTTEMP is not (yet) available again due to hardware issues; a replacement part could not yet be purchased because of EOL of the hardware
  - **Recompilation of applications is likely to be required** as old libraries, etc. are no longer available
  - Configuration is still fine-tuned. Open a ticket for requests.
  - No “AccessKey” for legacy job performance data anymore.
    Meggie has been integrated into **ClusterCockpit**
    => https://monitoring.hpc.fau.de
  - Available software mostly in line with Fritz / Woody-NG

  Usage: parallel workload (single and multi-node)

Updates in September maintenance – Woody (I)

- **Tier3 throughput computer Woody(-NG) - hardware**
  - Changed hardware: all nodes now have 8 GB per core (7.75 available)
  - Two node types:
    - Thin (w1xxx) with fast quad-core desktop-like CPUs; already known from the past
    - Big (w2xxx) with current server CPUs (2x 16 cores per node); partially financed by ECAP
    - The w2xxx nodes (and the new login nodes) support AVX512 while the w1xxx don't.
  - New login nodes are (currently) called woody-ng.nhr.fau.de (will be changed to **woody.nhr.fau.de** after Sept. 18th)
  - The new woody login nodes cannot be used to submit jobs to TinyGPU/TinyFAT
  - If software is compiled for AVX512 instructions, the binary will not run on the w1xxx nodes; if software is not compiled for AVX512, it may loose 1.7x performance on w2xxx
    => it may be required to compile two versions and select the binary at runtime

Updates in September maintenance – Woody (II)

- **Tier3 throughput computer Woody(-NG) – software environment**
  - OS of Woody changed from Ubuntu 18.04 to **AlmaLinux8** (RHEL clone)
  - Batch system changed from PBS/torque to **Slurm**
  - Recompilation of applications is almost always required
  - Configuration is still fine-tuned. Open a ticket for requests.
  - Job performance data now available via **ClusterCockpit**
  - Available software often in line with Fritz / Meggie

Usage: single-core / single-node throughput workload (no multi-node)

Updates in September maintenance – Tiny*

- **Tier3 systems TinyFAT, TinyGPU/Slurm**
  - The OS is unchanged Ubuntu 20.04
  - The batch system is unchanged Slurm
  - The **login node** “woody.rrze.uni-Erlangen.de”/”woody3” has been updated from 18.04 to **Ubuntu 20.04** and, thus, now matches the Tiny*/Slurm compute nodes
  - The login node “woody.rrze” (woody3) will be renamed to **“tinyx.nhr.fau.de”**
  - **The new WoodyNG login nodes cannot be used to submit to Tiny***
  - No job performance data available at the moment.
    Tiny* will be integrated into ClusterCockpit later

  Usage:
  - GPU workload
  - large memory demands (but note: the AMD nodes in TinyFAT are all financed by 3 groups of the Physics department who have priority access)

Things currently broken …

- quota / shownicerquota.pl currently cannot display data for /home/woody ($WORK for most people)
  - We changed the underlying file system and ZFS does not support remote quota queries. We try to find a solution.

- *.rrze.uni-erlangen.de vs. *.nhr.fau.de
  - The HPC login and compute nodes are currently in two different domains.
  - Some (login) nodes have an alias for both, but many do not.
  - Expect more nodes to appear in *.nhr.fau.de

- **`openmpi/4.1.3-intel*`** was broken on Woody-NG and Meggie8 until today

Final shutdowns on Sept. 18th

- **Emmy** will be shutdown for ever after more than 9 years of operation
  - Save data from $FASTTMP of Emmy before Sept. 18th or the data is lost
  - Transition to Woody-NG or Meggie

- **Woody/PBS** (i.e the remaining w11xx nodes with only 2 GB/core) will be shutdown for ever
  - Transition to Woody-NG

- Remember the **name change of woody[3].rrze to tinyx.nhr.fau.de as frontend for TinyFAT/TinyGPU with Slurm**

- **TinyGPU/PBS** (GTX1080/1080Ti) nodes will be shutdown

Access to Fritz + Alex

- **Via NHR project application**
  - https://hpc.fau.de/systems-services/documentation-instructions/nhr-application-rules/
  - Normal projects (with or without “DFG” project); Application possible at any time
    - Alex: 8.000-60.000 A40-h per year / 4.000-40.000 A100-h per year
    - Fritz: 1-10 mio core-h
    - Application possible at any time
  - Large projects: ~3 time the compute time; cut-off deadline 1st of each quarter
- **Via free FAU basic service:** only limited resources
  - in particular for preparation an NHR application (thus, no NHR-test/porting projects for FAU people)
  - In exceptional cases for groups with demand below the NHR normal projects which (a) cannot be done on the Tier3 systems and (b) show good performance. Proof for both required. https://hpc.fau.de/tier3-access-to-{alex,fritz}/

NHR applications

- NHR applications are **per project not per user**
  - For project in the normal category, we distinguish between with/without previous evaluation by DFG, BMFB, etc.
  - https://hpc.fau.de/systems-services/documentation-instructions/nhr-application-rules/
- The **PI must hold a PhD** (professors without PhD are also ok; but no students or PhD students)
- Multiple running projects are possible but must address different topics
- NHR projects get **new HPC accounts** using the new HPC portal.
  The PI or its technical contact are responsible for maintaining the accounts
- **Access** to the HPC systems with these accounts from the HPC portal is **by SSH keys only**; login to the portal is with SSO/DFN-AAI (using your IdM account). There is never a password stored for these HPC accounts. Certain services may therefore be not available right now (e.g. Jupyterhub)

HPC portal

- There is no new schedule (“plan”) yet on the general transition to the HPC portal also for regular Tier3 HPC accounts.
- For the free basic HPC service continue using the well known paper forms for the time being.
  At latest 2023 will bring a paper less solution for everyone … ☺

Job monitoring with ClusterCockpit

- https://monitoring.nhr.fau.de/
  - IdM-based HPC accounts:
    - login with HPC account + HPC passwort
  - HPC accounts from the new HPC portal:
    - Login to the HPC portal
    - Click on your account
    - And follow “go to ClusterCockpit” (which is next to “Add SSH key”)
  - Manager will in the future see the data of all their group members; no ETA yet

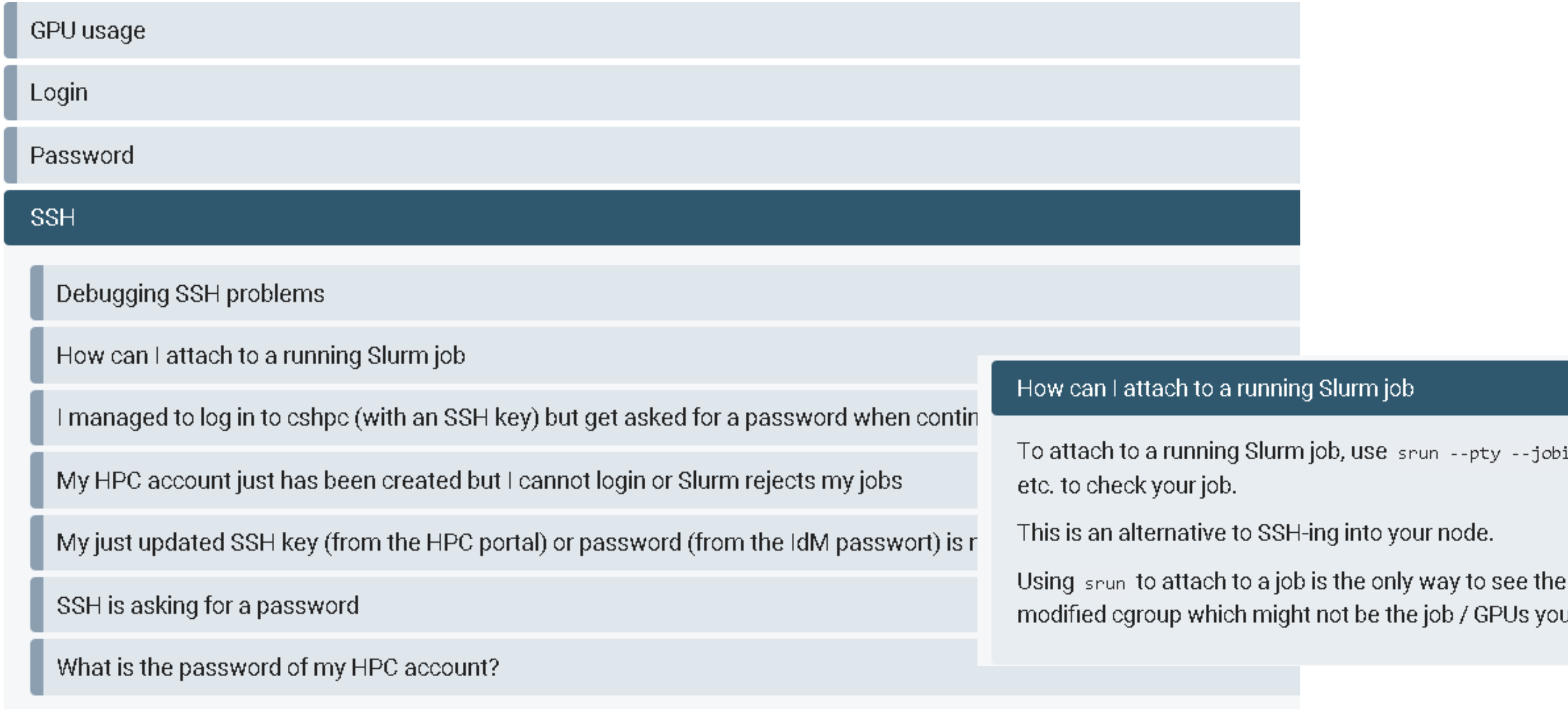
NEW

ClusterCockpit: System integration status

- **Fully integrated**:
  - Fritz
  - Meggie
  - Alex (with node sharing)
  - Woody (with node sharing, this includes woody-ng)
  - Emmy (for past jobs only, soon)
- We are planning to integrate all production NHR systems in the near future
- On all systems there are still **missing metrics** or **metric unit issues**! We are working on fixing all issues.
- Please report issues or feature requests to hpc-support@fau.de!

Additional documentation

- https://hpc.fau.de/faqs/
- E.g.



GPU usage

Login

Password

SSH

Debugging SSH problems

How can I attach to a running Slurm job

I managed to log in to cshpc (with an SSH key) but get asked for a password when contin

My HPC account just has been created but I cannot login or Slurm rejects my jobs

My just updated SSH key (from the HPC portal) or password (from the IdM passwort) is r

SSH is asking for a password

What is the password of my HPC account?

How can I attach to a running Slurm job

To attach to a running Slurm job, use `srun --pty --jobi` etc. to check your job.

This is an alternative to SSH-ing into your node.

Using `srun` to attach to a job is the only way to see the modified cgroup which might not be the job / GPUs you

New way to attach to running (GPU) Slurm jobs

- To **attach to a running Slurm job**, use **`srun --pty --jobid YOUR-JOBID bash`**. This will give you a shell on the first node of your job and you can run top, nvidia-smi, etc. to check your job.

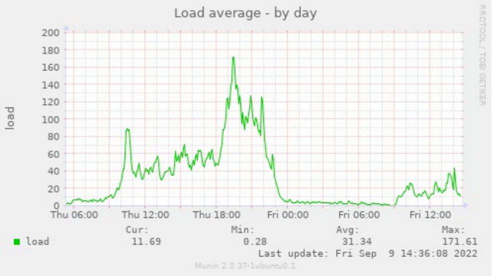
- This is an alternative to SSH-ing into your node.

- Using **`srun`** to attach to a job is the only way to see the correct GPU if you have multiple GPU jobs running on a single node as SSH will always get you into last modified *cgroup* which might not be the job / GPUs you are looking for.

General reminder: file system usage

- All filesystems available as /home/* are a shared resource!
- We still (again?) see many jobs (especially from ML) which do massive IO permanently. Opening/closing hundreds or even many thousand times per minute all the time is NO option for HPC cluster.

Load of one of the fileservers

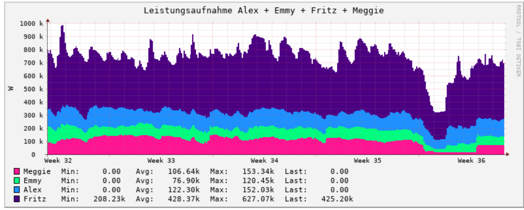


- There has been plenty of information in January’s HPC Café.
  Check these slides and ask us if you need help!
  And avoid to have hundred or even millions of (tiny) files. There must be better ways to handle data also for your domain.

  https://hpc.fau.de/teaching/hpc-cafe/
  Tuesday, January 18, 2022: Using File Systems Properly

HPC service and energy saving measures

- The HPC systems consume ~1 MW altogether => 8 GWh per year
- We don’t know if the University is aware of that. So far nobody talked to us.
- **HPC might be one of the services which might be asked first to shutdown if there is a real energy shortage in the coming months …**



- We’ll keep you informed if we get bad news.
- Our current plans for the worst case would be
  1) NHR projects are asked to burn as many compute cycles as soon as possible
  2) **We may give up turbo mode (or even throttle CPUs and GPUs more) – may come even without being forced due to rising energy costs**
  3) Reduce the number of nodes (in particular for the free Tier3 service as NHR is paying for operation); the head of the University would then have to decide whether “research” or “education” (bachelor/master thesis) should prioritized
  4) Complete shutdown of some/all systems; but our central systems will run longer than local HPC clusters in the institutes!

Final steps

- Do not forget to acknowledge NHR@FAU
  - Acknowledge support by NHR@FAU using our standard text: https://hpc.fau.de/systems-services/hpc-usage-reports/
  - Send us a copy of your paper to nhr-redaktion@lists.fau.de

- Link your publications in CRIS to the HPC infrastructure

- Next cut-off for NHR@FAU large scale project proposals is October 1st 2022

https://hpc.fau.de/systems-services/systems-documentation-instructions/nhr-application-rules/

http://tiny.cc/NHRFAU-Application

case studies

THANK YOU! – Questions?

NHR@FAU

https://hpc.fau.de