Karlsruhe Institute of Technology

Institut für Experimentelle Teilchenphysik

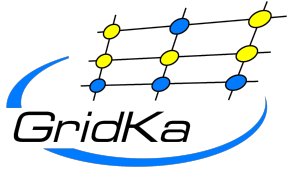
Scientific Computing Center

Institute Computing Infrastructure with Dynamic Extension

Matthias J. Schnepf for the HEP Computing Group at KIT | 16. April 2024

HEP Computing at KIT

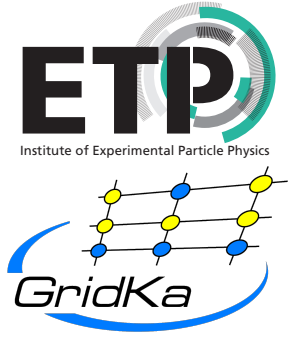
- Physics institute (ETP)
  - about 40 users
  - local Physics group for mainly Belle II and CMS as well as small groups
  - small computing infrastructure
- German Tier1 GridKa
  - supports several VOs: ALICE, ATLAS; Belle, CMS, LHCb, DARWIN, BaBar, IceCube, Piere Auger
  - large computing infrastructure



HEP Computing at KIT

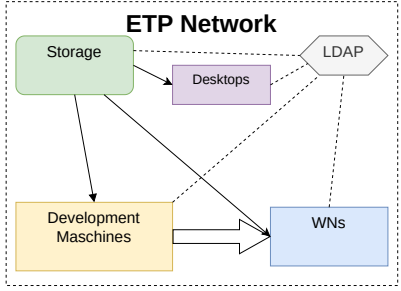
- Physics institute (ETP)
  - about 40 users
  - local Physics group for mainly Belle II and CMS as well as small groups
  - small computing infrastructure
- German Tier1 GridKa
  - supports several VOs: ALICE, ATLAS; Belle, CMS, LHCb, DARWIN, BaBar, IceCube, Piere Auger
  - large computing infrastructure

⇒ ETP can be used to test and develop new technologies



ETP computing infrastructure

- LDAP user management
- desktop PCs (Ubuntu)
- storage systems (RHEL8)
- big development machines RHEL8 with CC7 container
- HTCondor batch system with some WNs (Ubuntu)
  - about a dozen WNs
  - temporary included desktop PCs
- LDAP and storage only accessible within the ETP network

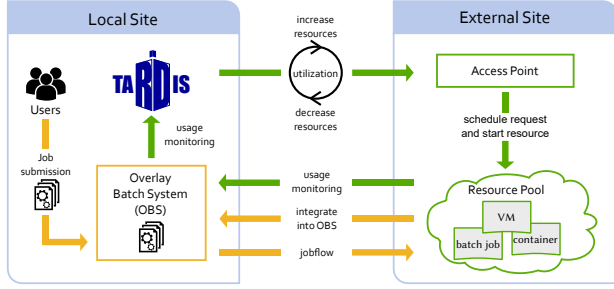


NEMO Cluster

- one of several cluster for different scientific communities
- NEMO: Neuroscience, Elementary Particle Physics, Microsystems Engineering and Materials Science

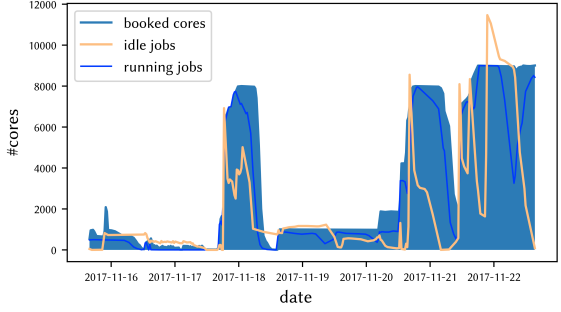
NEMO Cluster

- one of several cluster for different scientific communities
- NEMO: Neuroscience, Elementary Particle Physics, Microsystems Engineering and Materials Science
- started ETP VMs on NEMO WNs via MOAB
- resource manager COBalD/TARDIS via pilot like drones



NEMO Cluster

- one of several cluster for different scientific communities
- NEMO: Neuroscience, Elementary Particle Physics, Microsystems Engineering and Materials Science
- started ETP VMs on NEMO WNs via MOAB
- resource manager COBalD/TARDIS via pilot like drones
- got dynamically up to 9000 CPU cores for ETP
- successor cluster will come this year without VMs ⇒ using apptainer



Throughput Optimized Analysis System (TOpAS)

- End-User Cluster at GridKa
- 100 Gbit s^{-1} network
- 11 WN with 1 PB hard drive CEPHFS storage system for XRootD caching
- 7 GPU nodes (8x V100, 24x V100s, 24x A100)
- ETP jobs get send to TOpAS via HTCondor flocking
- backfilling with jobs from GridKa via COBalD/TARDIS and preemption

Software Distribution

- container
  - same software environment on all nodes
  - usable any image from docker hub
  - default docker image for vanilla jobs mimic software environment on the development machines (mostly used)
  - job based transition from SLC6 → CC7 → CentOS Stream 8 → RHEL-like 9 (soon)
  - docker preferred to support all images from docker hub
  - switching for some resources to apptainer and selected image distribution via CVMFS

Software Distribution

- container
  - same software environment on all nodes
  - usable any image from docker hub
  - default docker image for vanilla jobs mimic software environment on the development machines (mostly used)
  - job based transition from SLC6 → CC7 → CentOS Stream 8 → RHEL-like 9 (soon)
  - docker preferred to support all images from docker hub
  - switching for some resources to apptainer and selected image distribution via CVMFS
- analysis software
  - accessible via mount on all batch system resources
  - ETP CVMFS server
  - sandbox via HTCondor (small) or Grid storage

Grid Storage

- extra space at GridKa for local CMS (2.8 PB) and Belle II (250 TB) group
- GridKa dCache is accessible from all ETP resources including NEMO and TOpAS
- Users can read/write data and software from/to GridKa dCache
- authentication VO based and VO groups via certs
- GridKa dCache is more performant than ETP storage

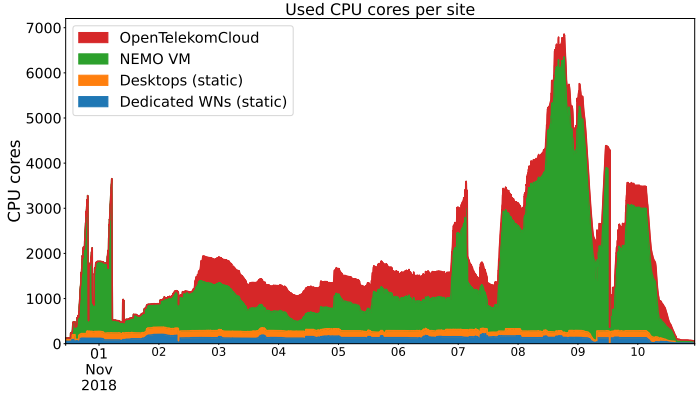
Job Submission

- default image for vanilla universe job *mschnepf/slc7-condocker*
- *+RemoteJob = True* required to run jobs outside of ETP WNs
- *+RequestWalltime* necessary to schedule to time limited resources (HPC)

```
Universe = docker
docker_image = mschnepf/slc7-condocker
+RemoteJob = True
+RequestWalltime = 3600

request_GPUs = 1

transfer_input_files = mnist_training.py
Executable = ./executable.sh
RequestCPUs = 1
RequestMemory = 8000
request_disk = 5000000
accounting_group = belle

x509userproxy            = /tmp/x509up_u12089
Output = out.txt
Error = err.txt
Log = log.txt


Queue
```

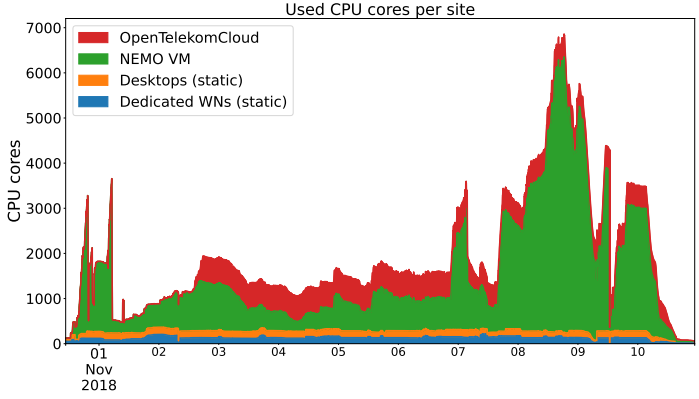
Other Computing Resource

- resources temporarily available to ETP
- Flocking to DARWIN development machine
- COBalD/TARDIS pilot like resources
  - Open Telekom Cloud during the Helix Nebula Science Cloud
  - VM for training at KIT
  - KIT HPC cluster



Other Computing Resource

- resources temporarily available to ETP
- Flocking to DARWIN development machine
- COBaID/TARDIS pilot like resources
  - Open Telekom Cloud during the Helix Nebula Science Cloud
  - VM for training at KIT
  - KIT HPC cluster
- transparent integrated into ETP batch system



User Experience

- most user just use local resources because it is simple
- power users are happy about the extra resources
  - code adjustments necessary
  - submission tools help
- simple selection of software environment via container

powered by DALL-E 3

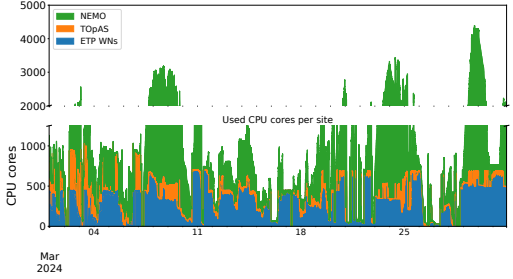
User Experience

- most user just use local resources because it is simple
- power users are happy about the extra resources
  - code adjustments necessary
  - submission tools help
- simple selection of software environment via container
- candies are helpful to talk to users and ask for update/changes

powered by DALL-E 3

Summary

- ETP has a basic computing cluster
- additional resources can dynamically and transparently integrated via COBalD/TARDIS and HTCondor flocking
- controlled software environment via container
- power users need to invest some time
- ETP is a good test field for GridKa



Backup