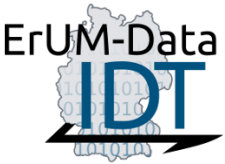
Opportunistic extension of a local compute cluster with NEMO resources for HEP workflows

Stefan Kroboth, Michael Böhler, Anton J. Gamel, Benjamin Rottler, Markus Schumacher

Physikalisches Institut, Albert-Ludwigs-Universität Freiburg

14th Annual Meeting of the Helmholtz Alliance
"Physics at the Terascale"
24.11.2021



ATLAS-BFG (Tier 2/3)

- Compute center for the ATLAS experiment
- ≈ 3200 cores (25 kHS06) / ≈ 2.5 PB storage (dCache)
- ATLAS production/analysis jobs
- Jobs by users of local HEP research groups
- Scheduler: SLURM

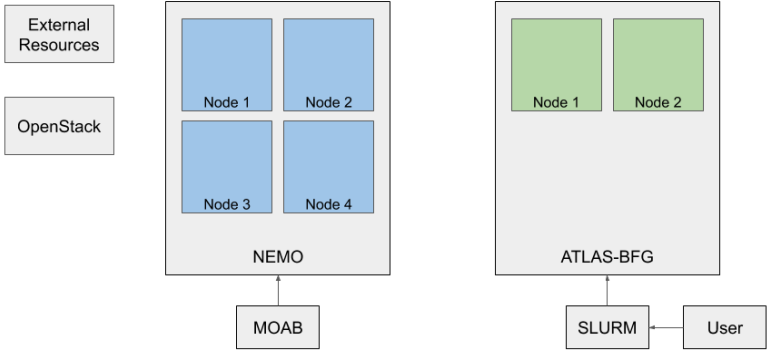
bwForCluster NEMO

- HPC cluster at Freiburg University
- ≈ 18000 cores (340 kHS06) / ≈ 800 TB storage (BeeGFS)
- Different software setup than ATLAS-BFG
- Scheduler: MOAB

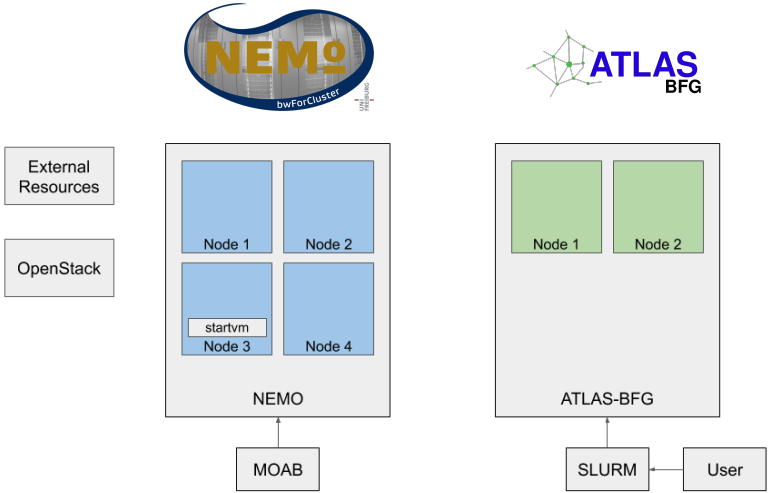
COBaID/TARDIS

- Opportunistically integrates resources from NEMO into ATLAS-BFG
- Based on demand and availability

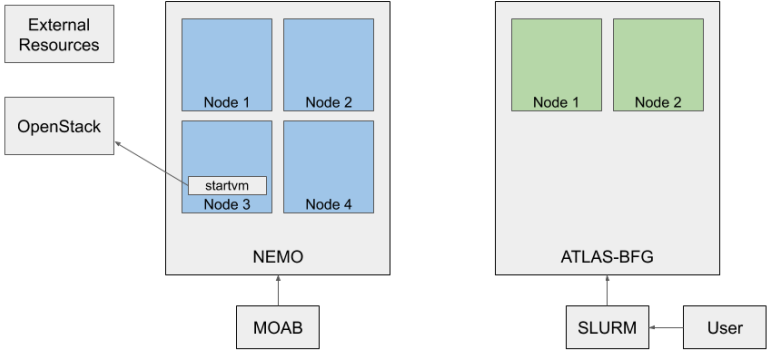
Resource integration



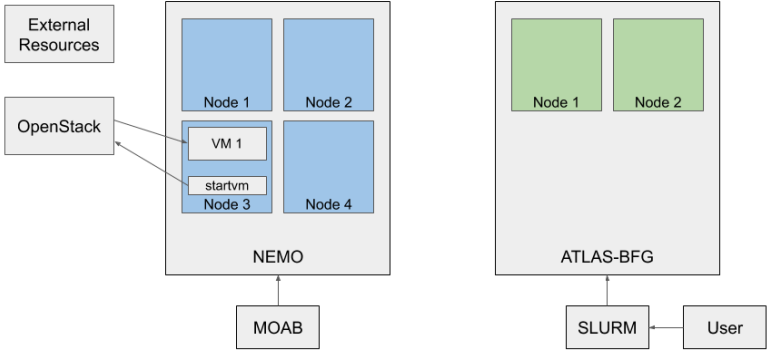
Resource integration



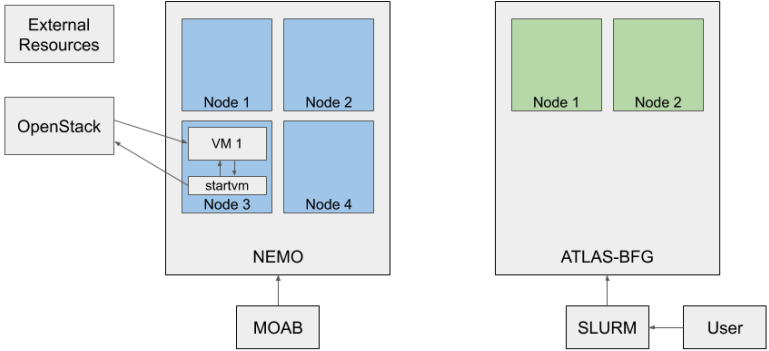
Resource integration



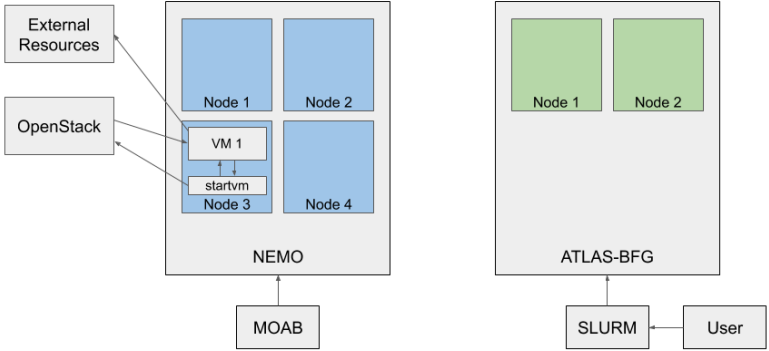
Resource integration



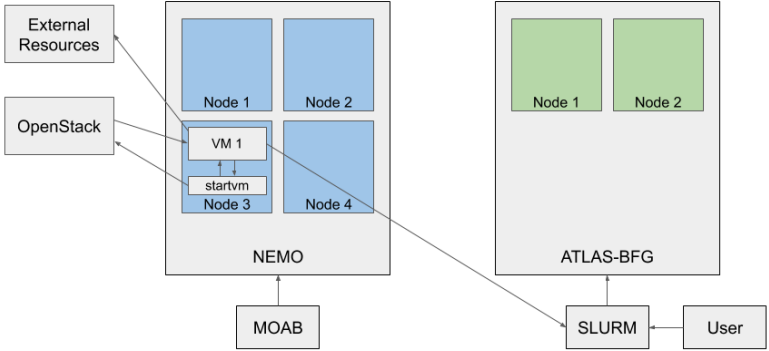
Resource integration

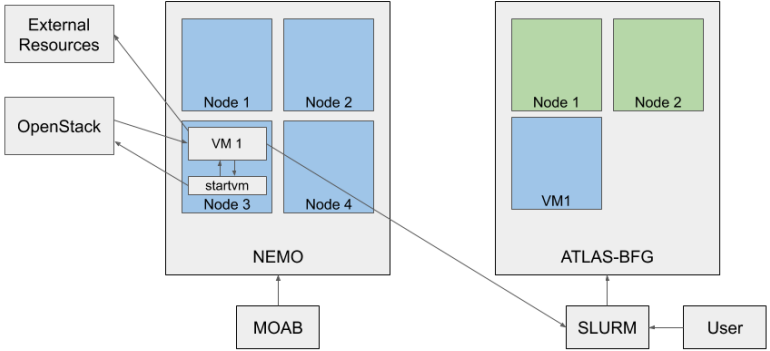


Resource integration

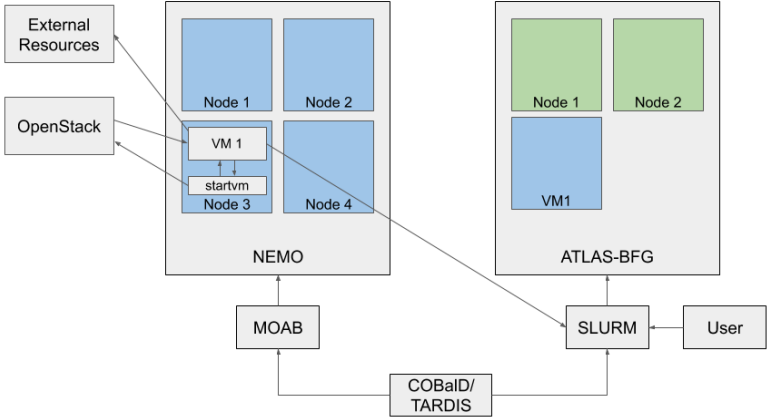


Resource integration




NEMO
bwForCluster
ATLAS
BFG
External Resources
OpenStack
Node 1
Node 2
VM 1
startvm
Node 3
Node 4
NEMO
Node 1
Node 2
VM1
ATLAS-BFG
MOAB
SLURM
User

Resource integration



COBalD

- COBalD is an Opportunistic Balancing Daemon
- Decides whether to request or release resources

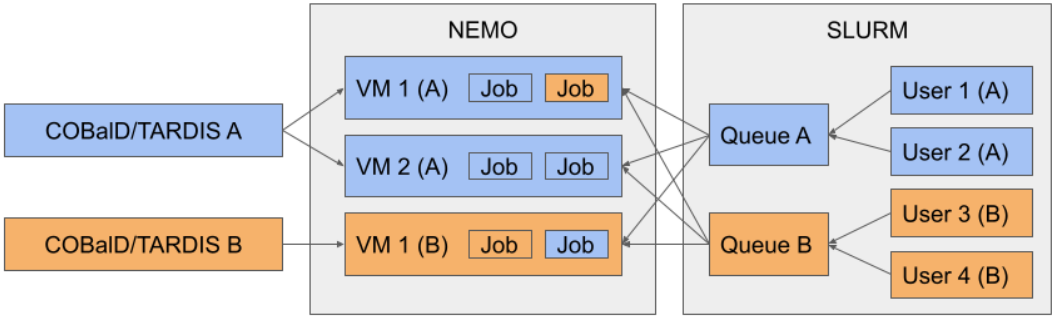
TARDIS

- Transparent Adaptive Resource Dynamic Integration System
- Interaction with the batch systems
- Resource integration

Main developers from group of Prof. Günter Quast (KIT)
`https://github.com/MatterMiners`

Local setup @ ATLAS-BFG

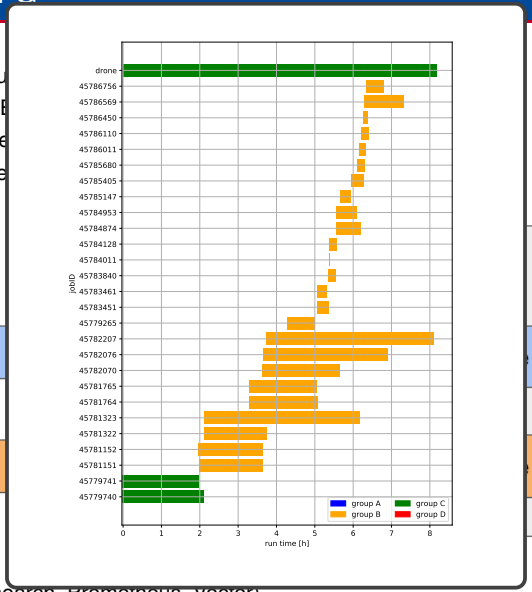
- Four local HEP research groups (A to D) with a share in NEMO
- Each served with its own COBalD/TARDIS instance
- Each has its own SLURM queue/partition
- Efficient use of resources due to sharing VMs across HEP groups



- Monitoring (Grafana, Elasticsearch, Prometheus, Vector)

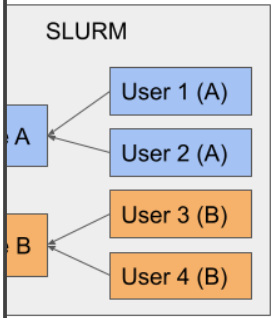
Local setup @ ATLAS-BFG

- Four local HEP research grou
- Each served with its own COB
- Each has its own SLURM que
- Efficient use of resources due



COBalD/TARDIS A

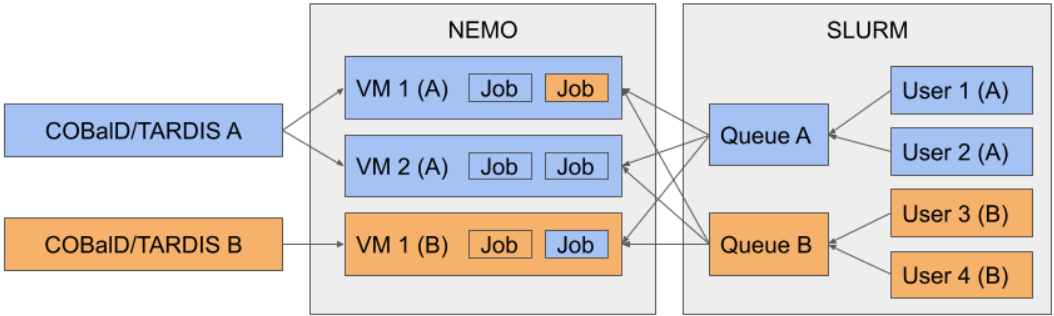
COBalD/TARDIS B



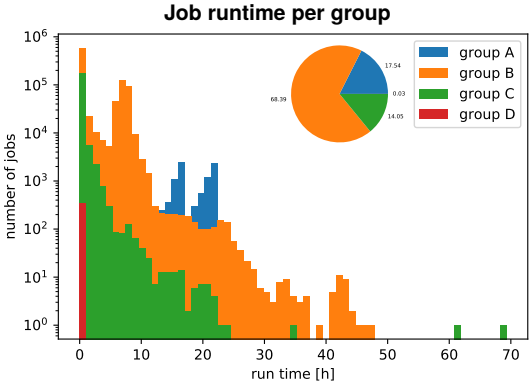
- Monitoring (Grafana, Elasticsearch, Prometheus, Vector)

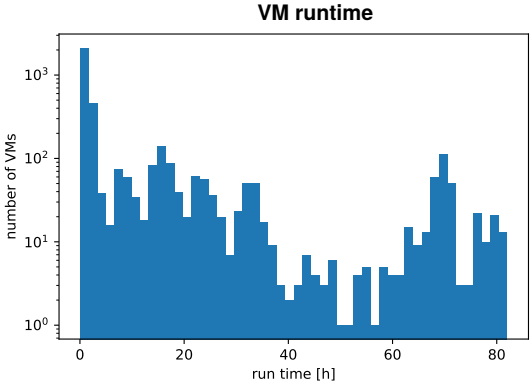
Local setup @ ATLAS-BFG

- Four local HEP research groups (A to D) with a share in NEMO
- Each served with its own COBalD/TARDIS instance
- Each has its own SLURM queue/partition
- Efficient use of resources due to sharing VMs across HEP groups



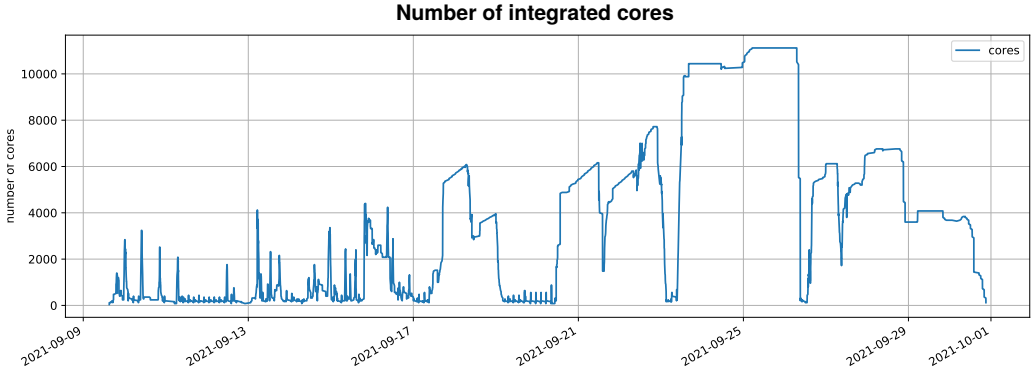
- Monitoring (Grafana, Elasticsearch, Prometheus, Vector)





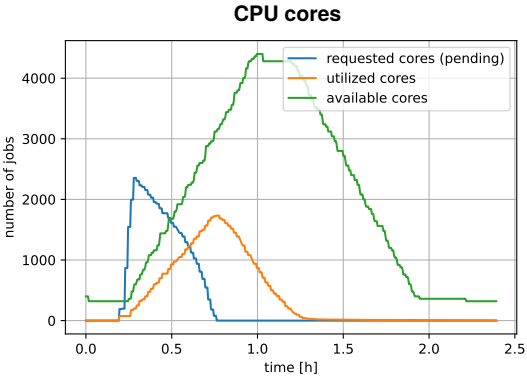
- Cluster usage varies heavily across HEP groups
- Long running and fully utilized VMs are preferable for efficiency

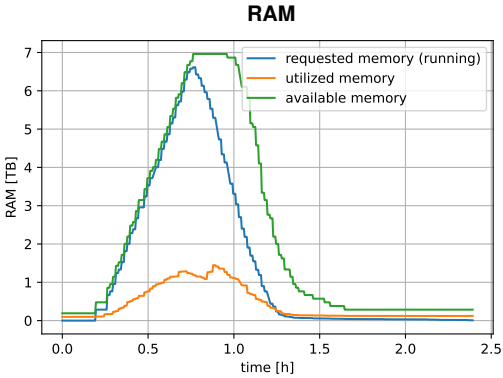
Integrated cores in September 2021



- Resources are only integrated when needed and released otherwise
- Efficient use of resources across cluster boundaries

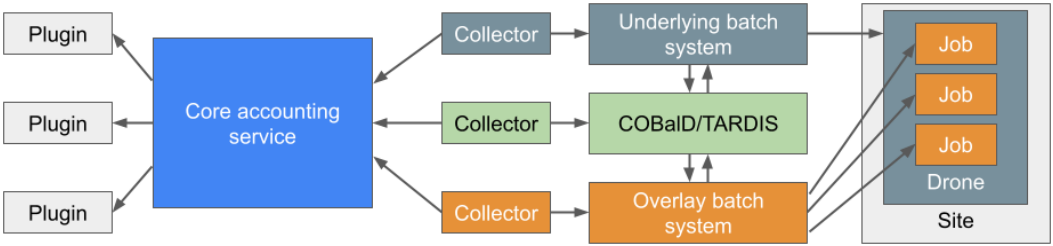
Dynamics





- Some CPU cores running idle because of RAM requirements of jobs
- Steepness of increase in available resources could be a measure for the responsiveness of the setup
  - Influenced by: COBalD/TARDIS settings, NEMO utilization, ...

Accounting



- Opportunistic resources are hidden behind other resources
- Track usage of all resources
- Can be used for
  - Accounting with an experiment such as ATLAS/CMS/…
  - Inter-site billing between sites sharing resources
  - Matching fairshare/priority to provided resources
- Currently in development
  - Core services and prototype of framework for collector and plugins ready
  - Prototype of collector plugin for COBalD/TARDIS ready
  - Next step: test fairshare/priority matching in production

Conclusion and outlook

Conclusion

- Integrated NEMO resources into ATLAS-BFG in an opportunistic manner transparent to the users
- Successfully in operation for more than two years
- Providing an additional $2 \cdot 10^6$ CPU hours since beginning of 2020

Outlook

- Integrate other resources such as Cloud providers (bwCloud)
- Accounting
  - Enable prototype in production environment: adjust group priority on ATLAS-BFG (fairshare) based on provided resources per group on NEMO
  - Develop further items necessary for a complete accounting framework