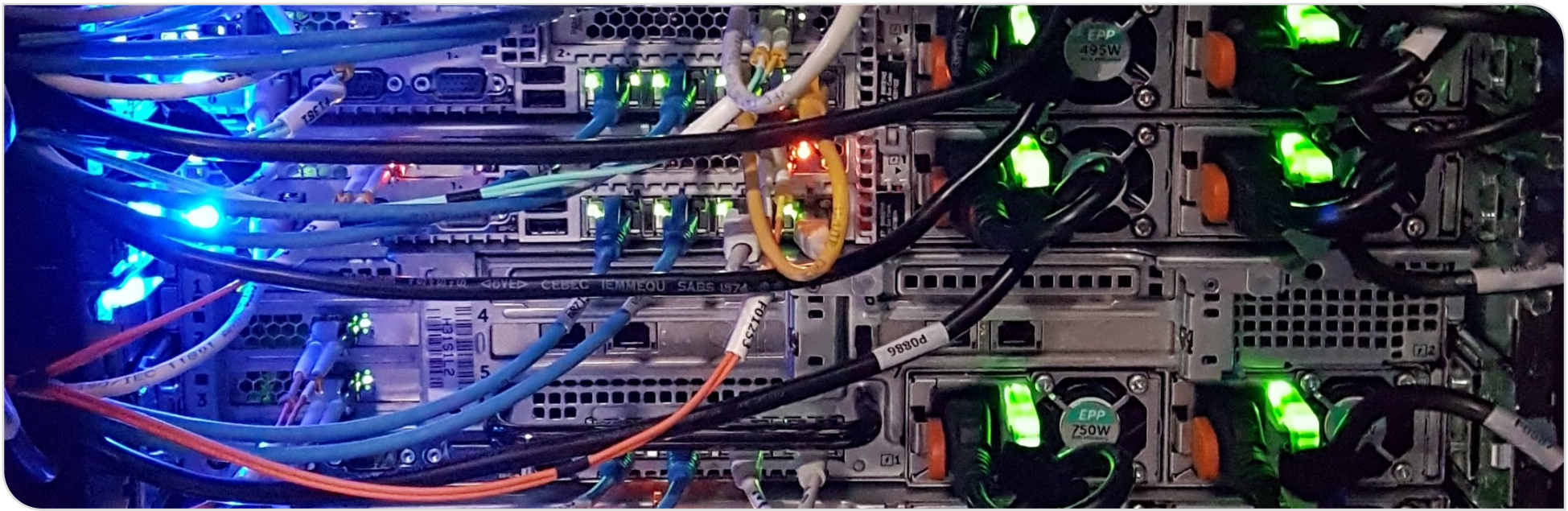
Karlsruhe Institute of Technology

Institute of Experimental Particle Physics

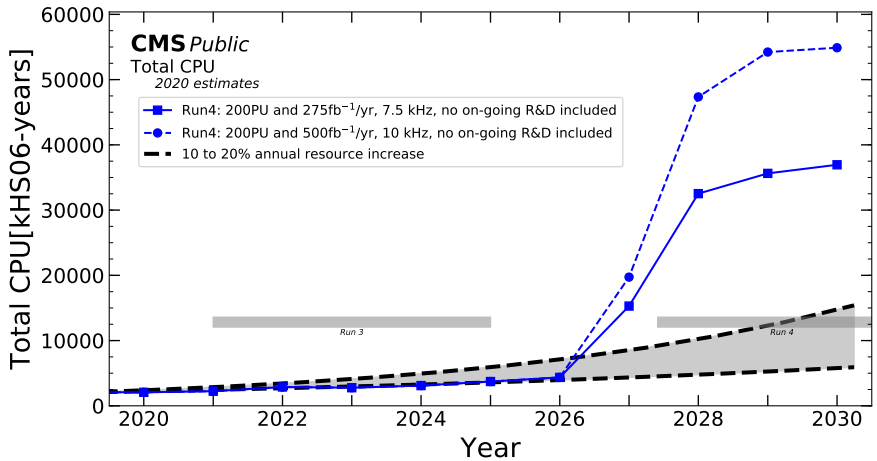
Opportunistic Cloud Computing for German HEP

R. F. von Cube, G. Quast, R. Caspart, M. Fischer, M. Giffels, C. Heidecker, R. Hofsaess, M. Horzela, E. Kuehn, M. J. Schnepf | ErUM Data IDT Collaboration Meeting, May 11, 2021



The HEP Computing Challenge I

- Clear challenge
  - Expect exploding demand for computing infrastructure
- Proposed solutions
  - Software improvements
  - **Integration of additional non-HEP resources**
  - Optimization of existing workflows



CMSOfflineComputingResults

The HEP Computing Challenge II

Access to multiple, heterogeneous resources

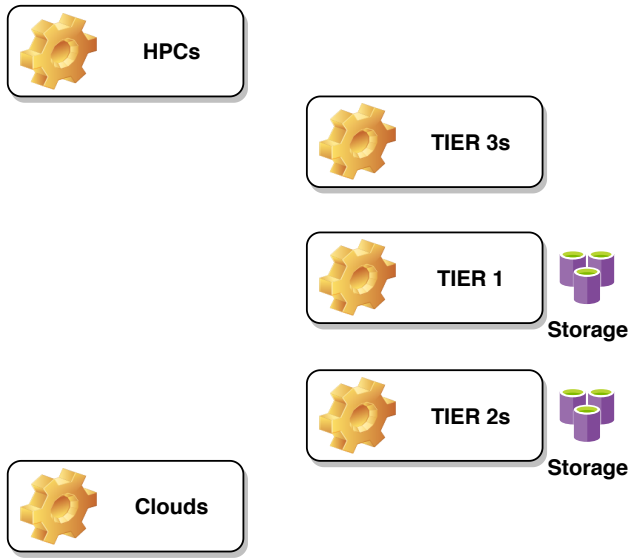
Difficult for experiments and users:

- Multiple identities, “submission types”
- Assessment which resource is available and suitable
- Experiments can’t negotiate with each resource provider

Computing in German (Astro-)Particle Physics

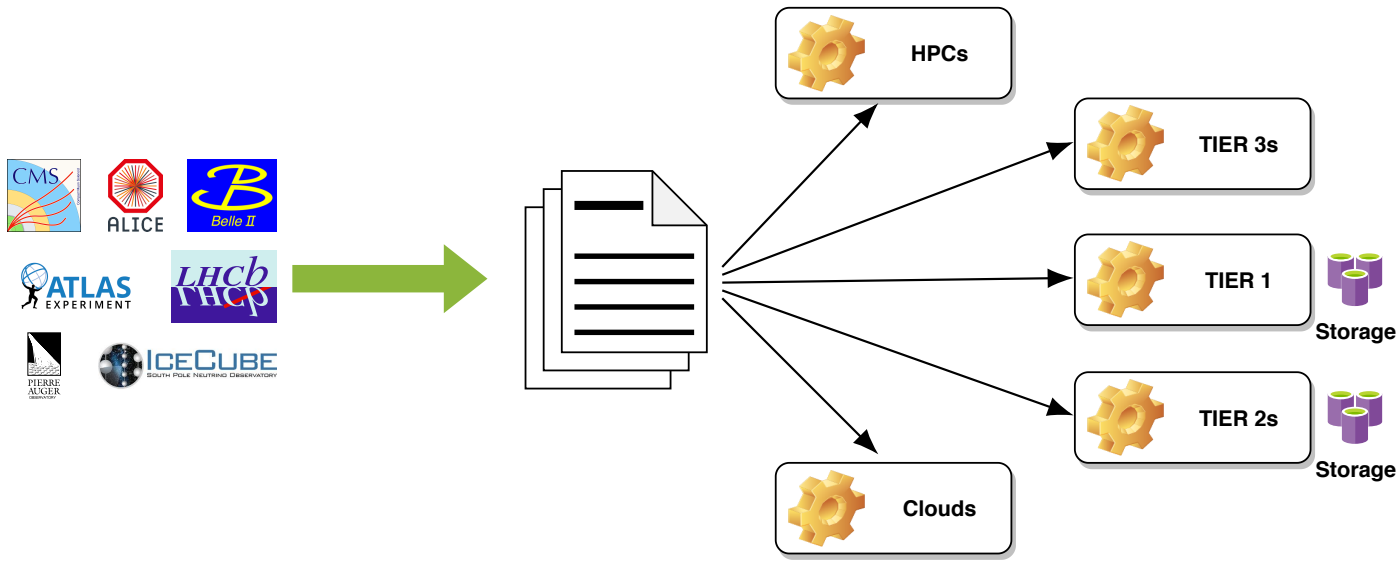
- Multiple experiments

Computing in German (Astro-)Particle Physics



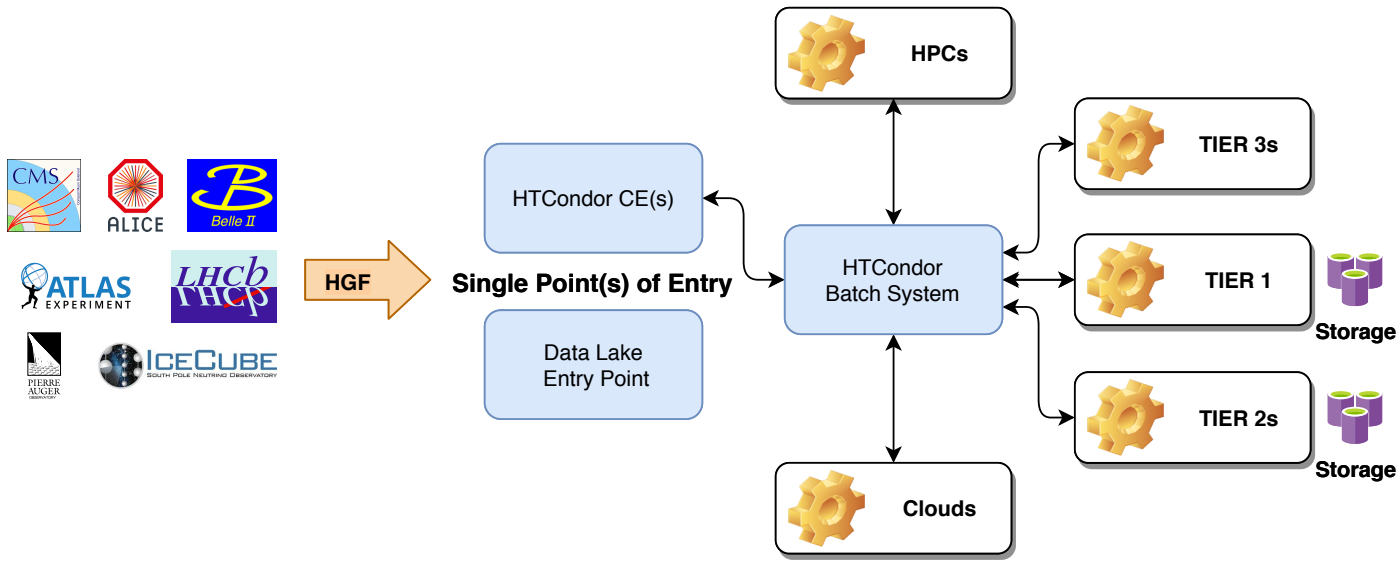
- Multiple experiments on heterogeneous resources

Computing in German (Astro-)Particle Physics



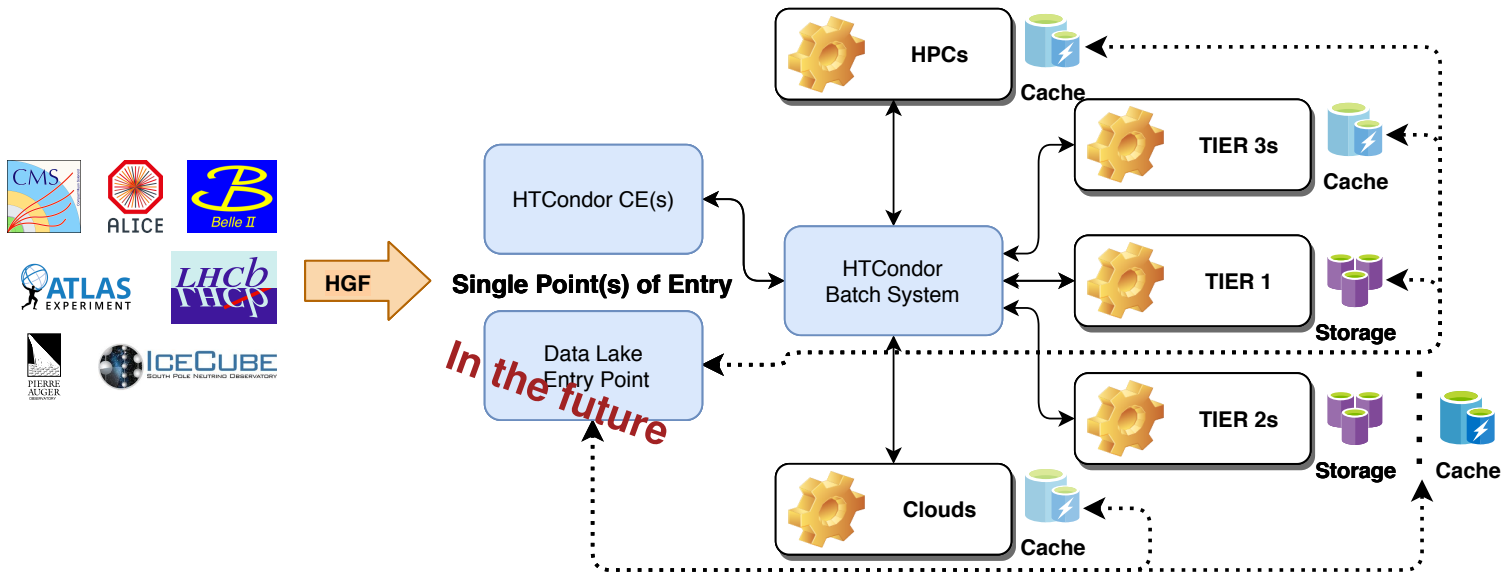
- Multiple experiments on heterogeneous resources

Computing in German (Astro-)Particle Physics



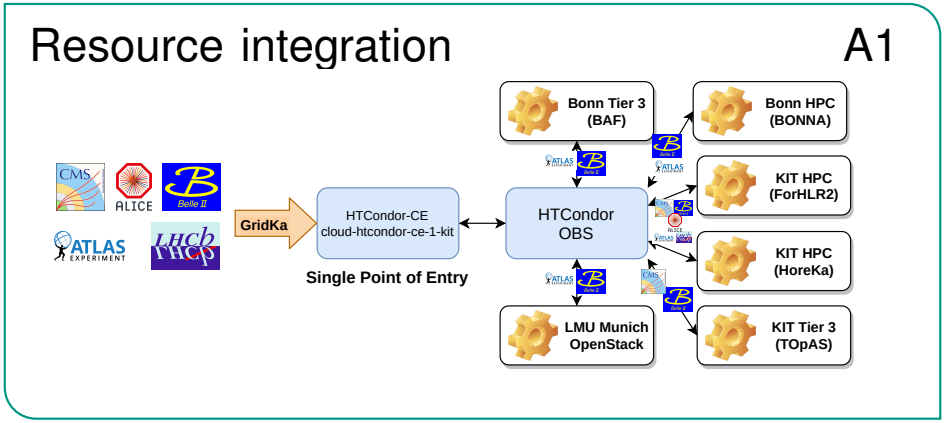
- Multiple experiments on heterogeneous resources
- Opportunistic, transparent resource integration and lightweight site operation

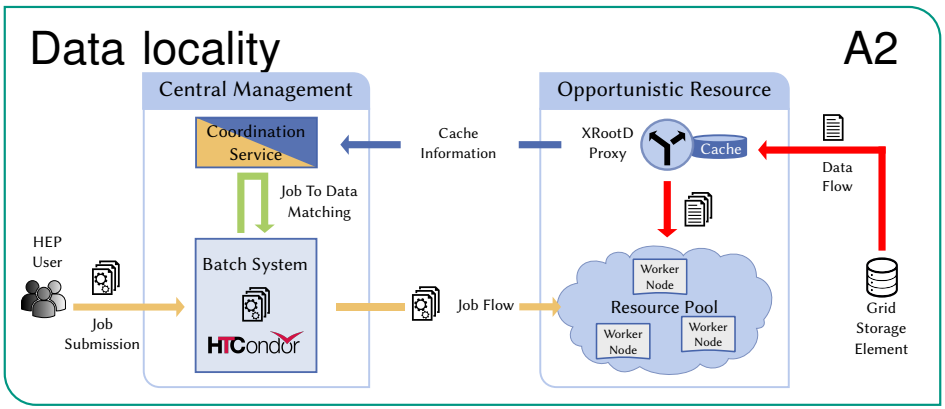
Computing in German (Astro-)Particle Physics

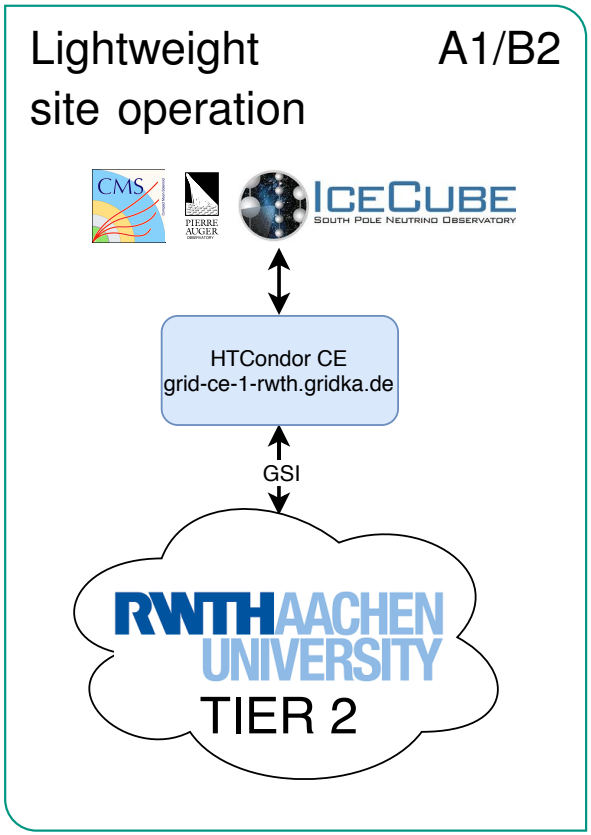


- Multiple experiments on heterogeneous resources
- Opportunistic, transparent resource integration and lightweight site operation
- Data locality for efficient usage

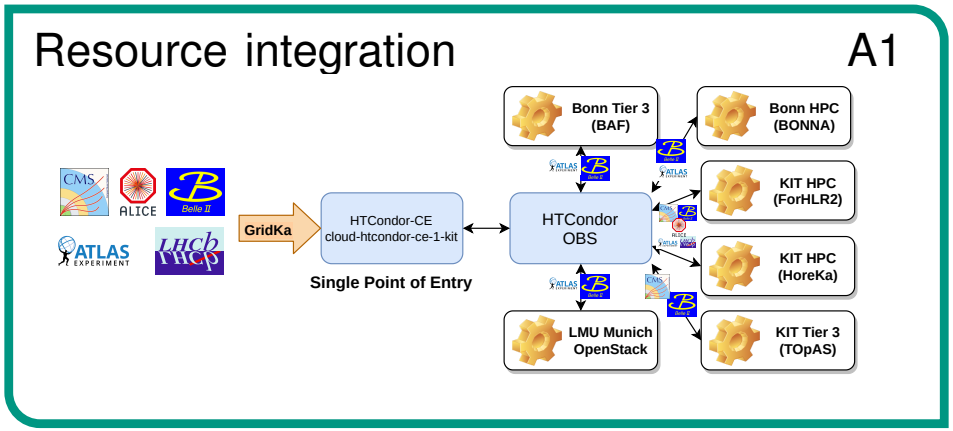
Areas of Development at KIT

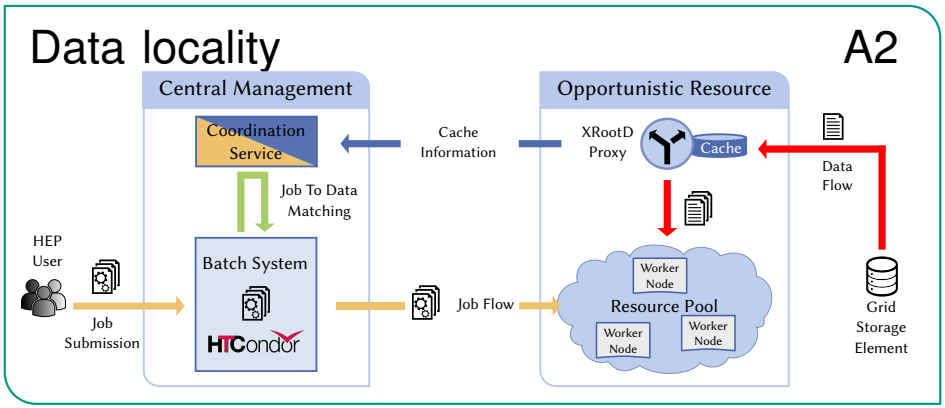


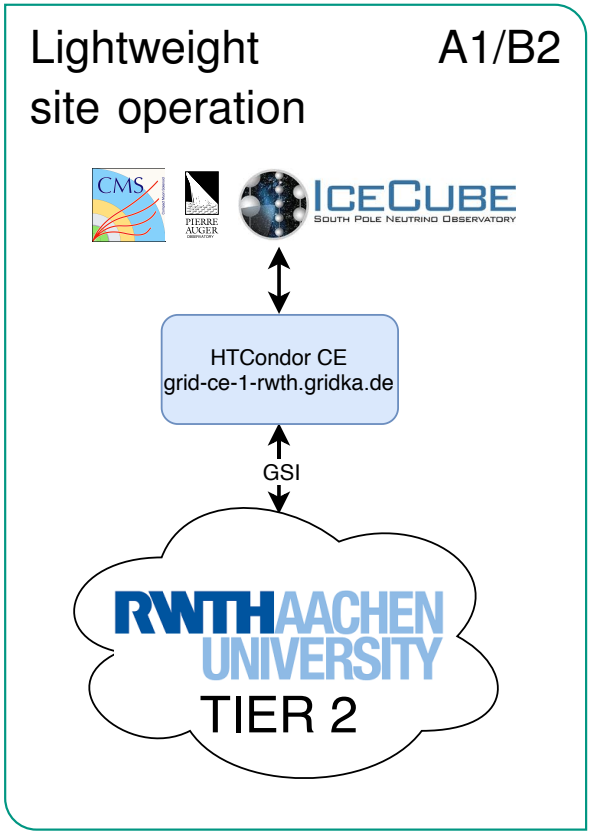




Areas of Development at KIT







COBALD/TARDIS Resource Manager

Karlsruhe Institute of Technology

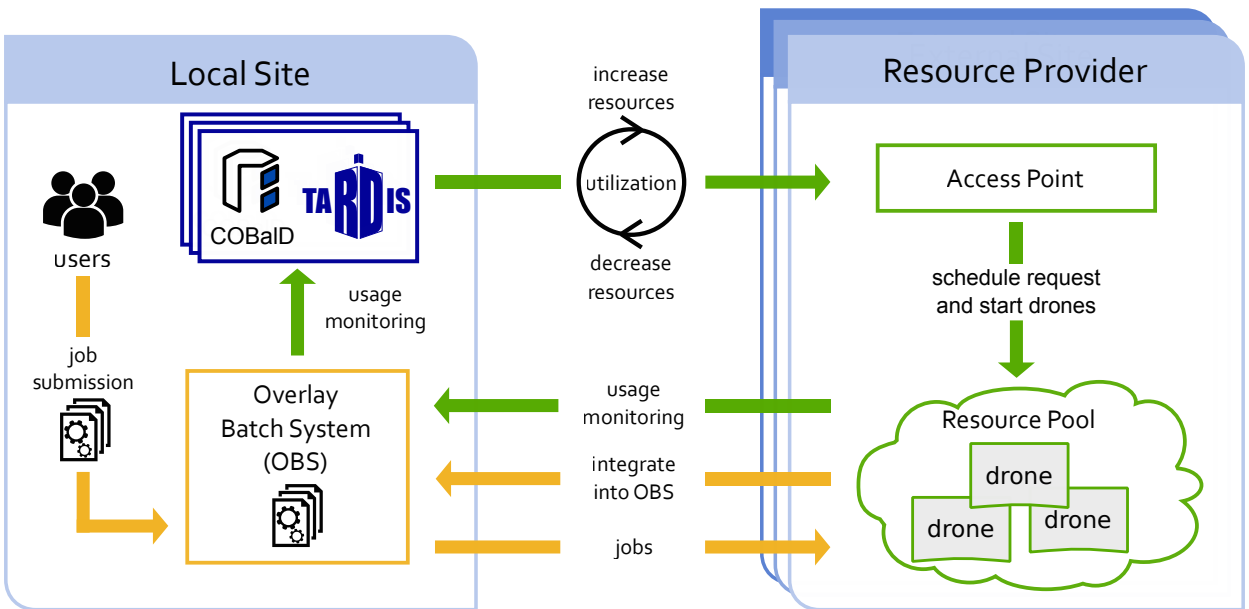
Dynamic integration through single point of entry

- Provide resources transparent to experiments and users
- React on resource demand

- Integration of resources into common overlay batch system (OBS)
  - Experiments and users authenticate only against single point of entry using standard grid technology
- Assessment of resource fit to current job mix through metrics allocation and utilization
  - COBALD/TARDIS decision based on feedback not prediction
  - Increase well-used resources, release badly used
- Allocation of resources through generalized pilots, so-called drones
  - Drone can be any executable, container, or virtual machine, depending on resource

Resource Integration

- allocate opportunistic resources: Cloud, HPC, university clusters
- provide opportunistic resources as extension to community-specific computing resources
- integrate resources transparently and based on current demand



Ressource Assessment with TARDIS and COBALD

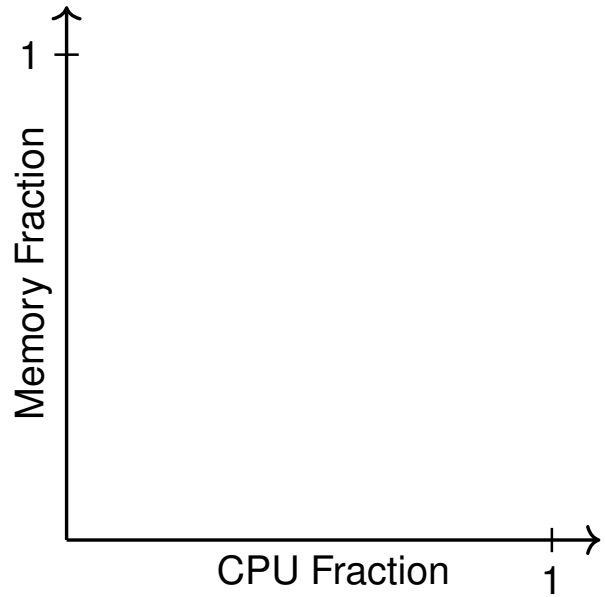
Karlsruhe Institute of Technology

TARDIS

Dynamically provisions and integrates resources into overlay batch system.

COBALD

Assesses the suitability of resources to the current job mix with metrics *allocation* and *utilization*.



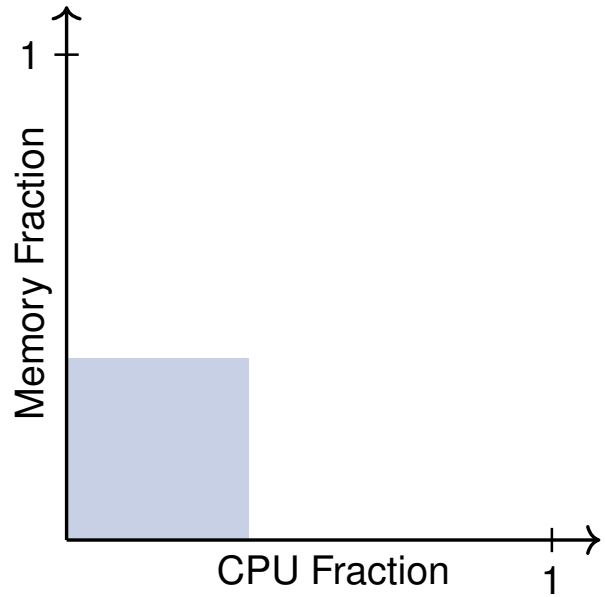
Ressource Assessment with TARDIS and COBALD

TARDIS

Dynamically provisions and integrates resources into overlay batch system.

COBALD

Assesses the suitability of resources to the current job mix with metrics *allocation* and *utilization*.



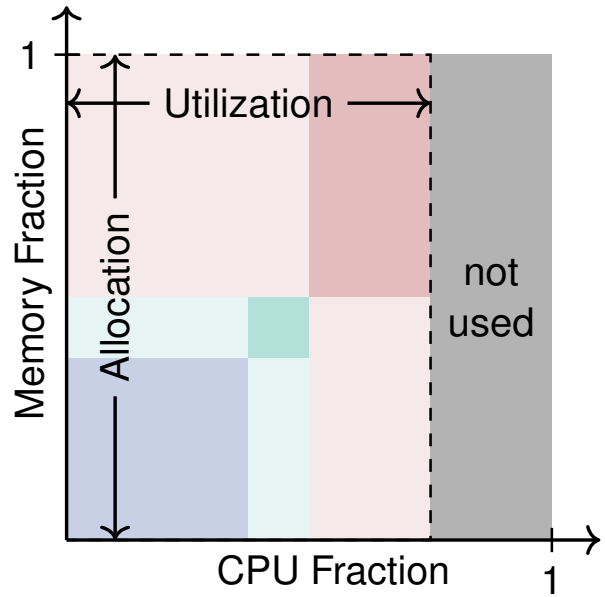
Ressource Assessment with TARDIS and COBALD

TARDIS

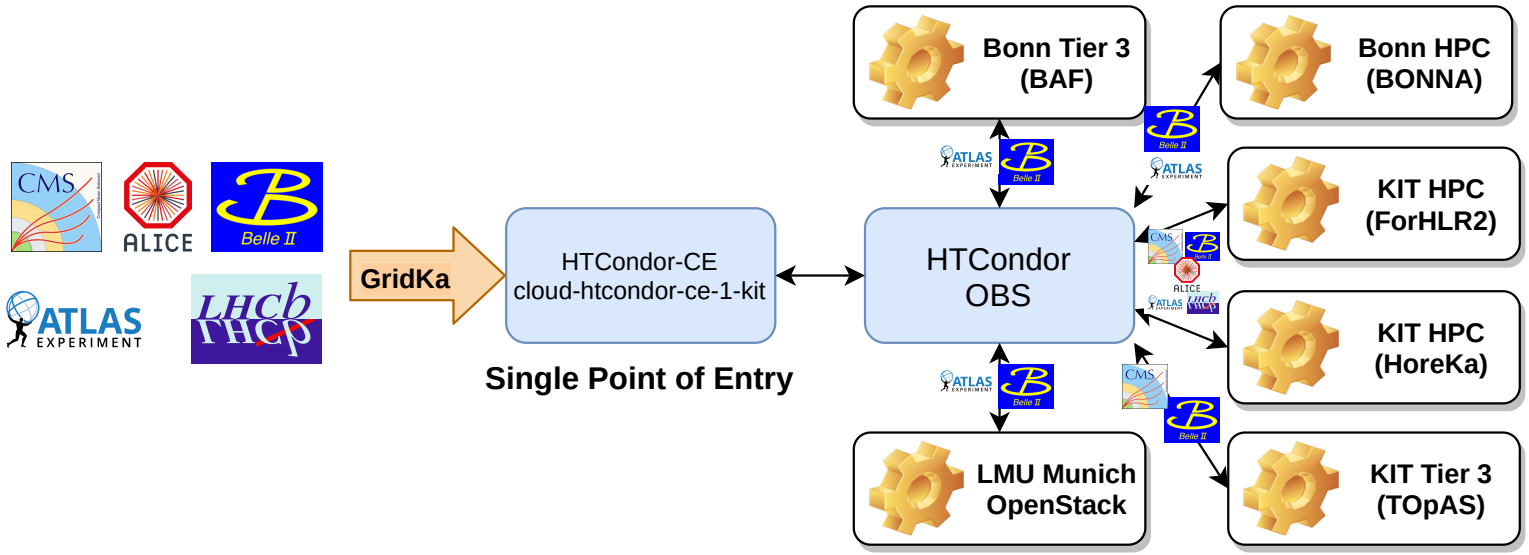
Dynamically provisions and integrates resources into overlay batch system.

COBALD

Assesses the suitability of resources to the current job mix with metrics *allocation* and *utilization*.



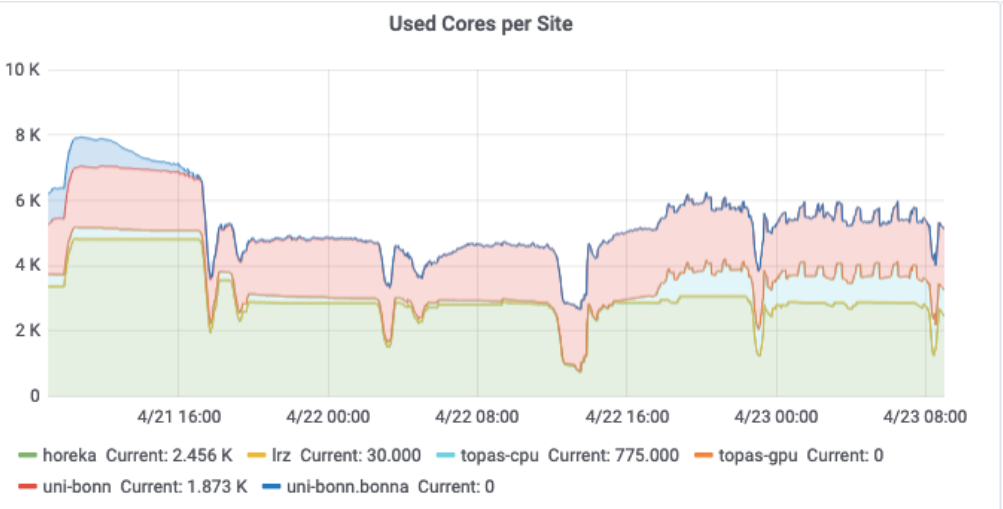
Resource Integration and Lightweight Site Operation

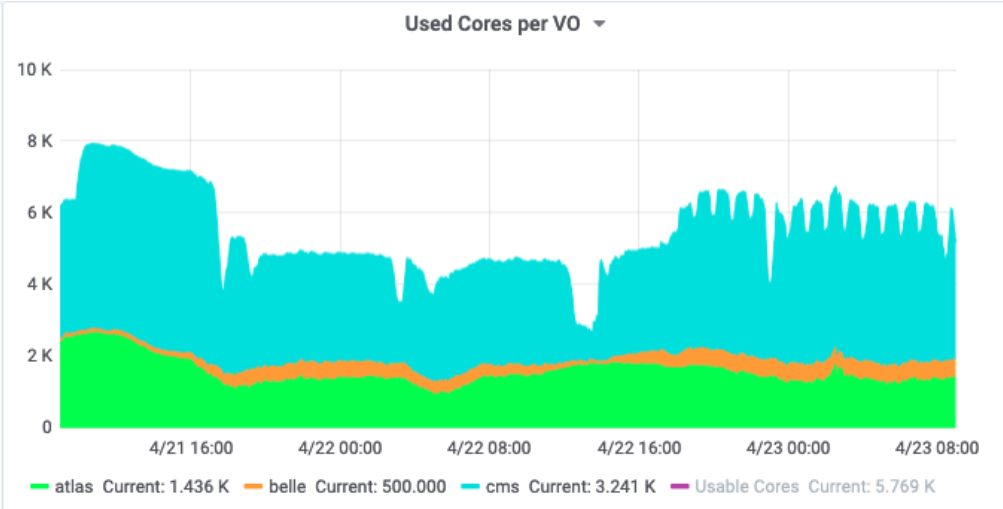


- built prototype of federated infrastructure with U Bonn, LMU, and KIT
- serves multiple VOs / Experiments

→ **wide-range and beneficial collaborations across communities**

Resource Integration and Lightweight Site Operation

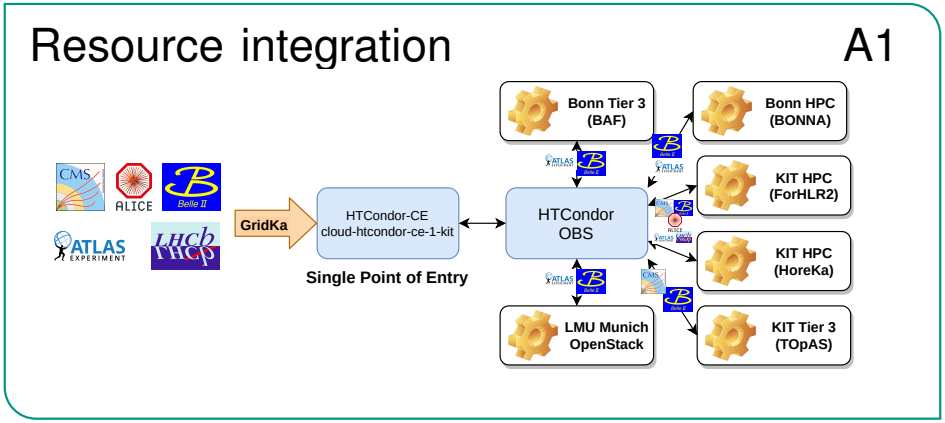


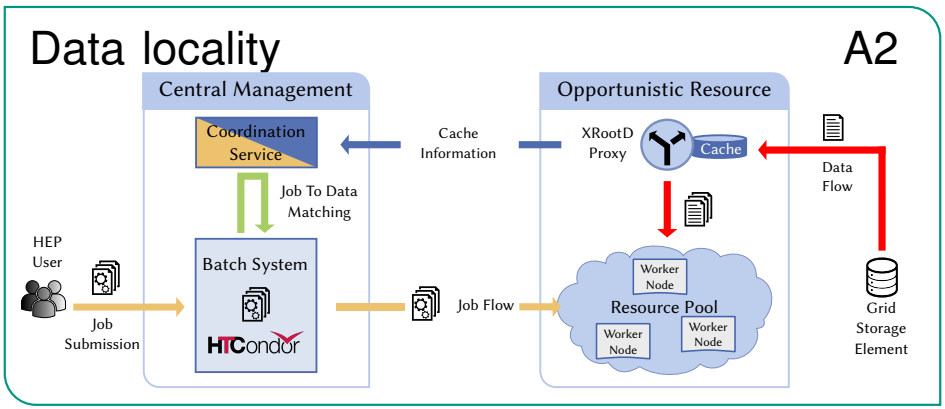


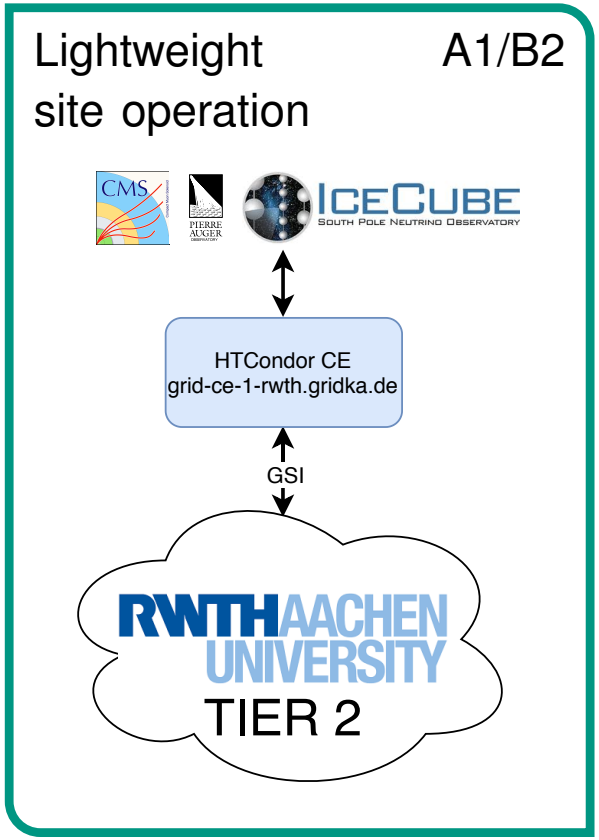
- built prototype of federated infrastructure with U Bonn, LMU, and KIT
- serves multiple VOs / Experiments

→ **wide-range and beneficial collaborations across communities**

Areas of Development at KIT







Lightweight Tier 2 Operations

- Compute elements (CEs) are entry points for users and experiments to grid sites
- Necessary to be operated for each site
- GridKa operates multiple CEs for the Karlsruhe Tier 1
- Deployment automated with puppet modules

Operate CE as a service for other sites

- Overhead is minimal to run additional CE
  - Only minor changes in “standard” HTCondor-CE configuration
- “Remote CE” implemented for the Aachen Tier 2 site $\Rightarrow$ Aachen doesn’t need to operate CE anymore

Interested Tier 2 sites may contact us

First Integration in a CMS Tier 2 WLCG Site

Karlsruhe Institute of Technology

- Aachen physics department operates standard WLCG tier 2 site
  - $\sim$ 5100 cores pledged to CMS plus storage and grid services
- Aachen researchers have access to university HPC cluster CLAIX
- Resources integrated into WLCG tier 2 site
  - Bash script submitted to CLAIX' SLURM workload manager sets up and starts HTCONDOR in unprivileged user account
  - Jobs are started in SINGULARITY containers, providing WLCG environment
  - Usage completely transparent to experiments and users
- CLAIX dynamically made available through COBALD/TARDIS

Integration of CLAIX: Challenges and Solutions

Karlsruhe Institute of Technology

Networking

- HTCondor daemons communicate using the “Condor Connection Broker”

Singularity

- Support for nested Singularity containers for GlideInWMS-pilots
  - Activation of user namespaces, usage of sandbox-image
- Singularity bind mounts unset in container

CVMFS

- Config `CMS_LOCAL_SITE` is dangling symbolic link on host
- Local SITECONF is provided through bind mount within container

Integration of CLAIX: Challenges and Solutions

Karlsruhe Institute of Technology

Networking

- HTCondor daemons communicate using the “Condor Connection Broker”

Singularity

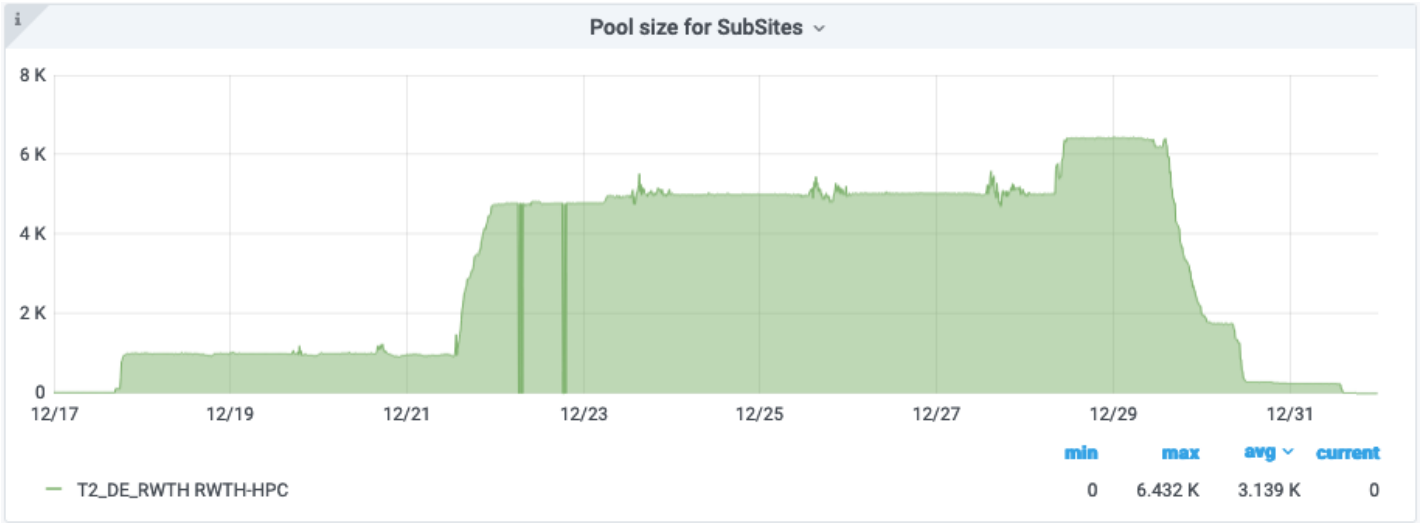
- Support for nested Singularity containers for GlideInWMS-pilots
- Activation of user namespaces, usage of sandbox-image
- Singularity bind mounts unset in container

CVMFS

- Config `CMS_LOCAL_SITE` is dangling symbolic link on host
- Local SITECONF is provided through bind mount within container

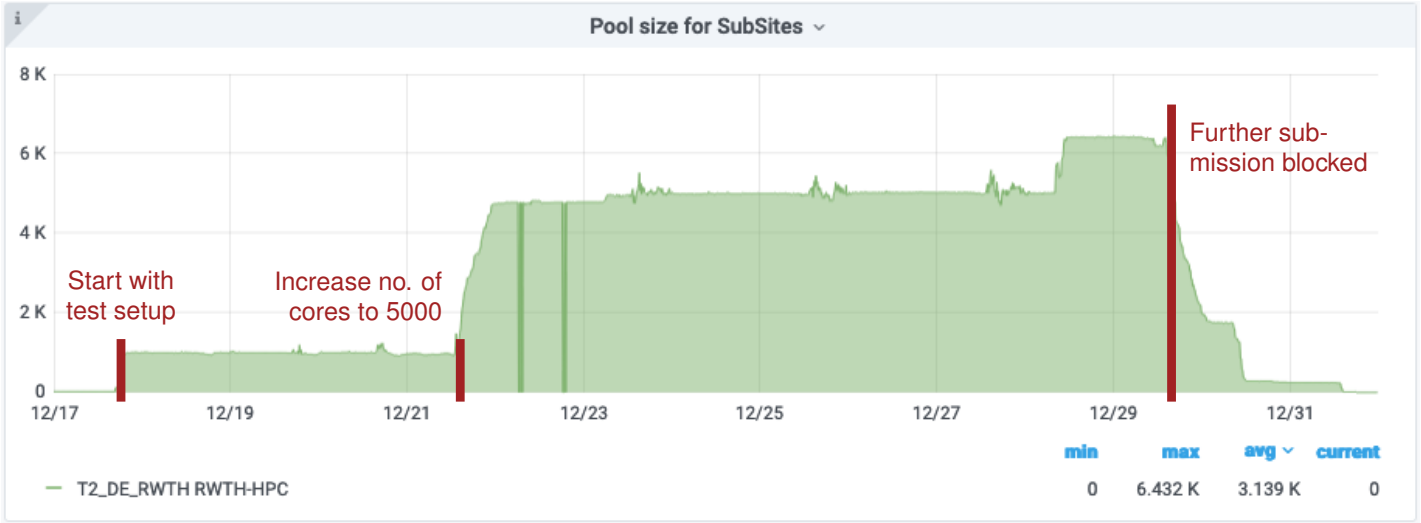
To overcome some challenges, close communication with CLAIX was essential, however, very productive!

Aachen: Scaling up



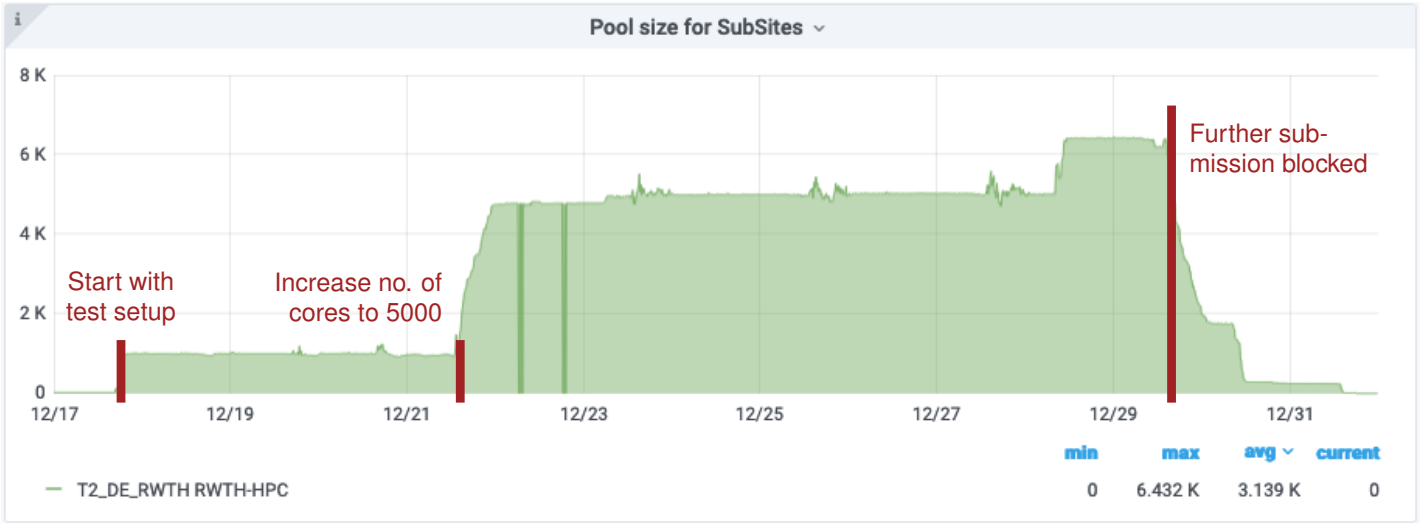
- Smooth start of test setup after close communication with CLAIX and Aachen grid team
- Virtually doubled number of cores, available to experiment, for a week

Aachen: Scaling up



- Smooth start of test setup after close communication with CLAIX and Aachen grid team
- Virtually doubled number of cores, available to experiment, for a week
- Submission blocked by HPC center after using $\sim$ 7 times monthly quota

Aachen: Scaling up



Thanks to Aachen grid team and CLAIX service personnel!

- computing project of 7.5 Mio core-h for the next year just got approved

Summary

- COBALD/TARDIS allows for transparent usage of opportunistic resources
- Resources are integrated dynamically into one overlay batch system
- Aachen tier 2 CMS pledge was doubled for a week

Join us on on

- GitHub: github.com/MatterMiners
- Gitter: gitter.im/MatterMiners/community
- Twitter: twitter.com/matterminers