Karlsruhe Institute of Technology

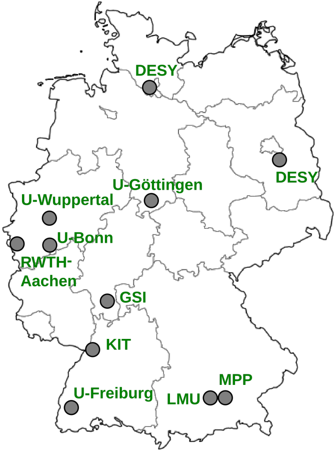
ErUM-Data IDT

Dynamic Integration of Opportunistic Computing Resources for HEP in Germany

Matthias J. Schnepf on behalf of the KIT HEP-Computing team | 14. February 2022

Situation in Germany

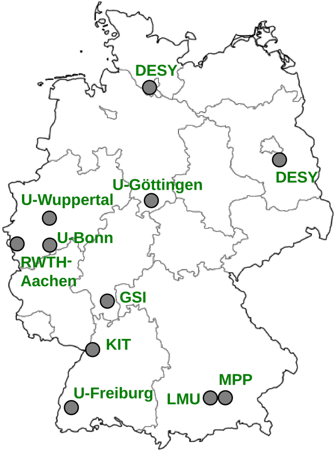
- federated states
  - individual funding
  - resources distributed through different states
  - limited access to resources (state bounded)
- challenges for HEP Computing
  - resource demand for HEP is increasing
    - HL-LHC
    - Belle II
    - various astroparticle physics experiments
  - complex infrastructure due to several resource provides



based on A.Streit
https://indico.desy.de/event/32315/contributions/114438/

Situation in Germany

- federated states
  - individual funding
  - resources distributed through different states
  - limited access to resources (state bounded)
- challenges for HEP Computing
  - resource demand for HEP is increasing
    - HL-LHC
    - Belle II
    - various astroparticle physics experiments
  - complex infrastructure due to several resource provides
- opportunities
  - various resource providers
  - interesting topics for students
  - one goal: efficient usage of available resources



based on A.Streit
https://indico.desy.de/event/32315/contributions/114438/

Computing Resources in Germany

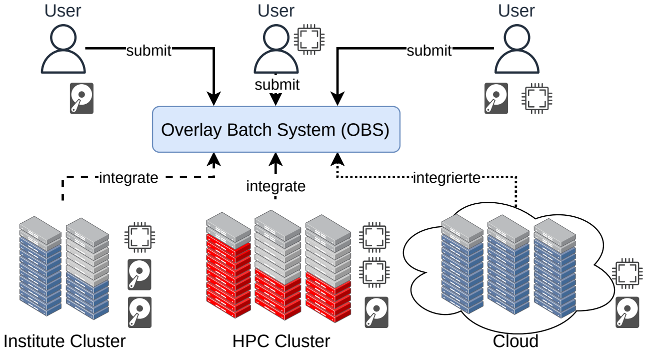
- HEP dedicated Grid sites
- opportunistic resources
  - resources not dedicated for HEP (Grid) computing
    - software environment
    - access methods
    - resource provision: VM, container, pilot job

Computing Resources in Germany

- HEP dedicated Grid sites
- opportunistic resources
  - resources not dedicated for HEP (Grid) computing
    - software environment
    - access methods
    - resource provision: VM, container, pilot job
  - cloud provider
    - good availability to cover peak loads
    - usually high costs
  - HPC cluster
    - availability depends on operation mode: backfilling / share
    - increase cluster utilization by the usage of unused resources due to multi-node scheduling
  - institute cluster
    - usually backfilling
    - increase cluster utilization by using free resources

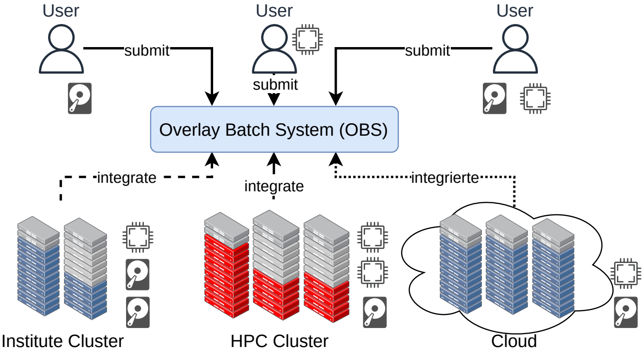
Integration of Resources

- Overlay Batch System as single point of entry



Integration of Resources

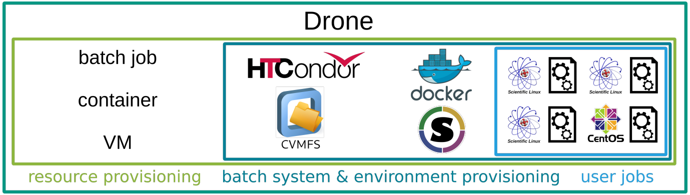
- Overlay Batch System as single point of entry



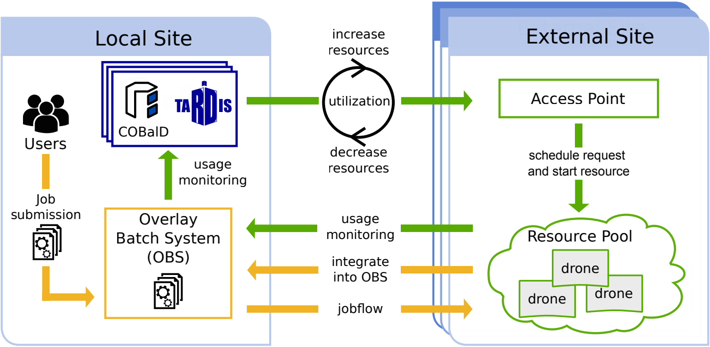
- How to integrate resources from different kinds of providers?
- How many resources of which type are needed at which provider?

Generalized Pilot Concept

- pilot concept
  - placeholder job allocates resources
  - worker node instance of an Overlay Batch System (OBS) starts payload jobs inside the pilot job
  - requires software environment
- generalized pilot concept ⇒ drone concept
  - resource allocation as
    - batch job
    - virtual machine
    - container
  - provides full Grid software environment
  - drone/pilot/job can run inside a drone

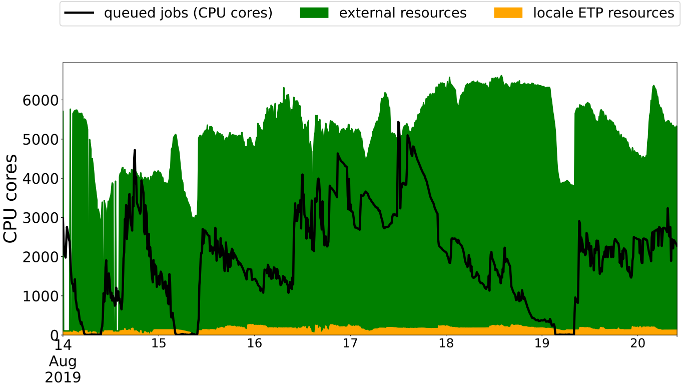


Resource Management: COBalD & TARDIS



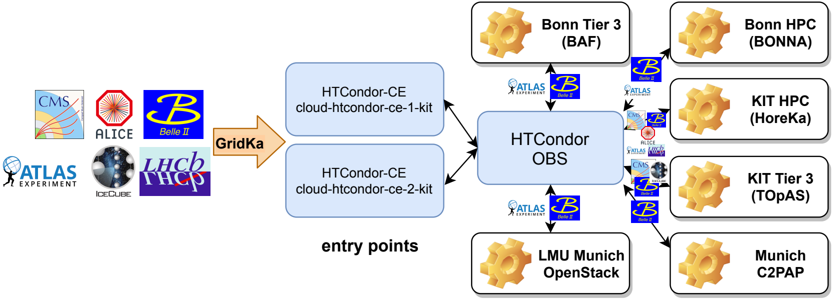
- initiated and coordinated by KIT
- load balancing daemon COBalD (COBalD - the Opportunistic Balancing Daemon)
- life cycle management TARDIS (Transparent Adaptive Resource Dynamic Integration System)

Dynamic Resources for End-User at ETP



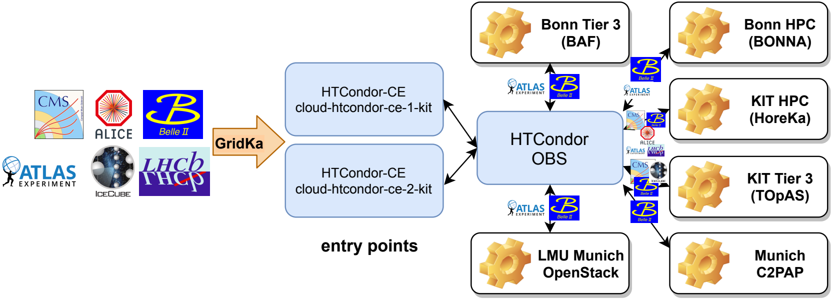
- dynamic resources into ETP batch system
- HPC cluster NEMO at Freiburg is main resource provider for ETP

German Federated Computing Infrastructure



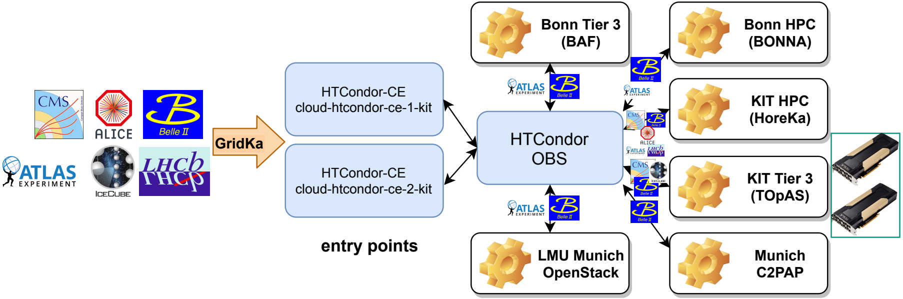
- sites are connected to prototype setup at GridKa
- transparent provisioning of resources to the majority of HEP experiments, see monitoring
- integration of further resources in the future - fully transparent and experiment independent

German Federated Computing Infrastructure



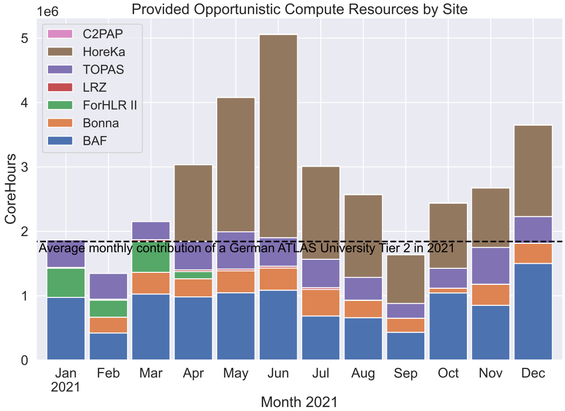
- sites are connected to prototype setup at GridKa
- transparent provisioning of resources to the majority of HEP experiments, see monitoring
- integration of further resources in the future - fully transparent and experiment independent
- further development in optimization, accounting and managing of multiple COBalD/TARDIS instances

German Federated Computing Infrastructure



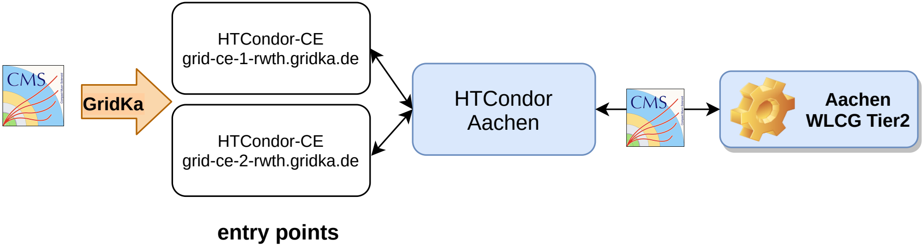
- sites are connected to prototype setup at GridKa
- transparent provisioning of resources to the majority of HEP experiments, see monitoring (with GPUs)
- integration of further resources in the future - fully transparent and experiment independent
- further development in optimization, accounting and managing of multiple COBalD/TARDIS instances
- development of workflows with GPU support in the Grid

Provided Resources via GridKa Setup in 2021



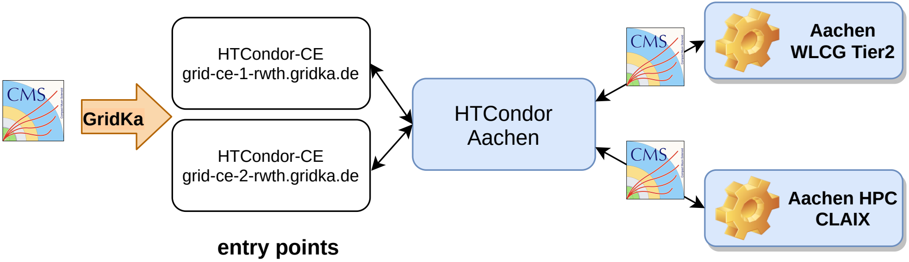
- up to more than 5 mil. CPU h provided in one month
- more than the average monthly contribution of a German ATLAS University Tier 2 in 2021 are provided

Lightweight Tier2 Site: Aachen



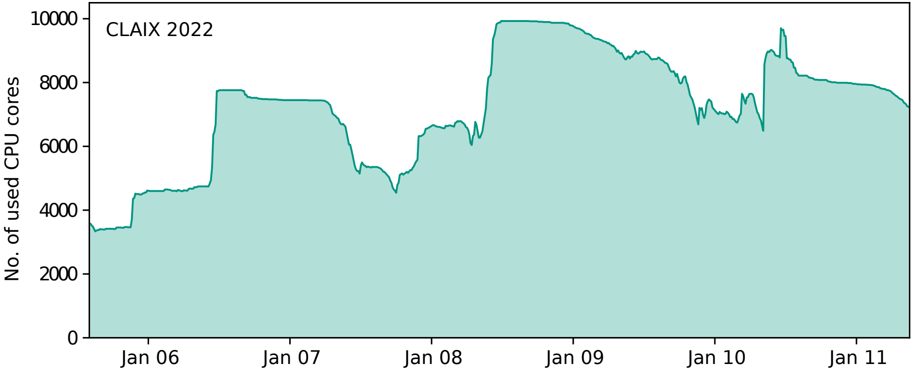
- GridKa operates extra CEs for Aachen (possible for other sites)

Lightweight Tier2 Site: Aachen



- GridKa operates extra CEs for Aachen (possible for other sites)
- Aachen runs lightweight resource manager COBalD/TARDIS and their HTCondor instance
- COBalD/TARDIS integrates resources from Aachen HPC cluster CLAIX
- resource grant of 10 mil. CPU h on CLAIX for HEP per year

Provided Resources from CLAIX to CMS



- up to 10.000 additional CPU cores to CMS provided by the Aachen HPC Cluster CLAIX
- 2900 CPU cores are pledged to the CMS experiment provided by the Aachen Tier 2 cluster

COBalD/TARDIS Community

- Workshops organized by KIT
  - introductions in COBalD/TARDIS
  - hands-on sessions
  - help by site configuration and integration
- development from our partners
  - support of other batch systems and providers
  - monitoring plug-ins
  - several other optimizations
- special thanks to our partners for their contribution
  - Uni. of Bonn
  - Uni. of Freiburg

Summary and Outlook

- provisioning of opportunistic computing resources to end-users and the Grid
  - drone concept to integrate resources and provide needed software environment
  - lightweight resource management COBalD/TARDIS
- lightweight Tier 2 center
  - test case Aachen
  - CEs operated by GridKa
  - extend their computing resources with HPC resources managed by COBalD/TARDIS
- future work
  - improvement and optimization for COBalD/TARDIS
  - accounting
  - provide support for our current and future partner

- COBalD/TARDIS used by

UNI FREIBURG

RWTH AACHEN UNIVERSITY

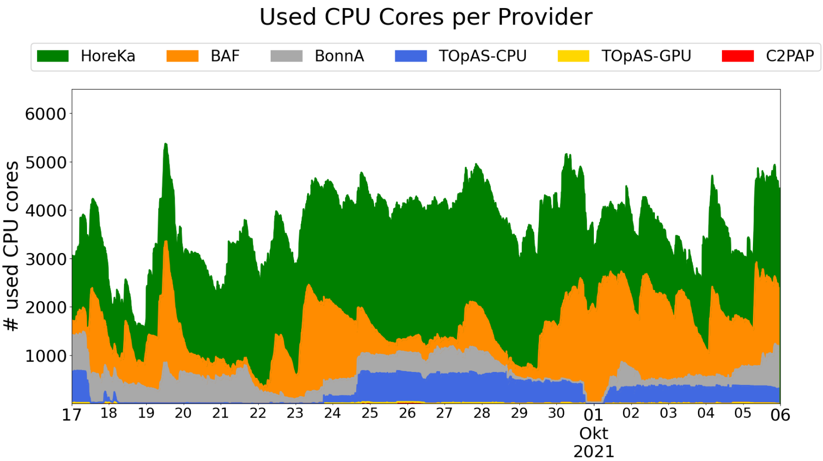
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

UNIVERSITÄT BONN

KIT Karlsruher Institut für Technologie

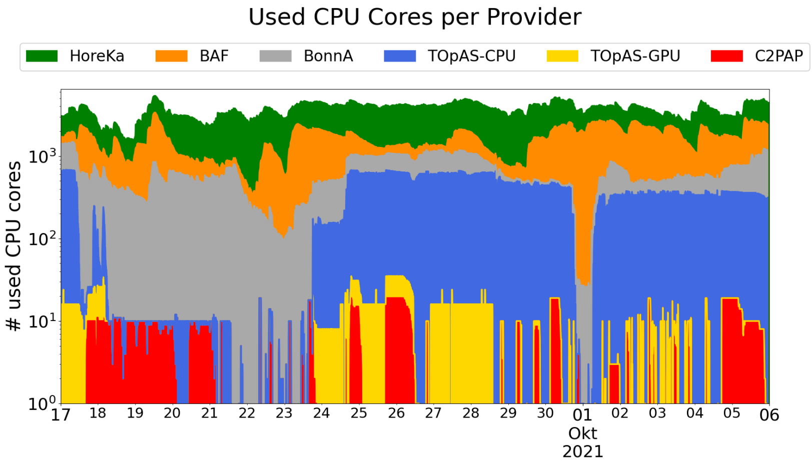
Backup

Provided Resources



- up to 17000 CPU cores provided by

Provided Resources



- up to 17000 CPU cores provided by

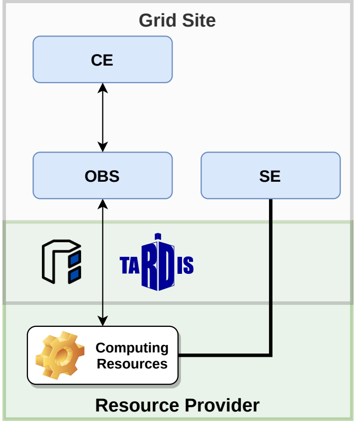
What We Provide

- COBalD & TARDIS
  - https://github.com/MatterMiners/cobald
  - https://github.com/MatterMiners/tardis
- help to setup OBS or integrate site
  - hands on sessions (integration of C2PAP cluster Munich within 4h)
- puppet module
  - https://github.com/unibonn/puppet-cobald
- wlcg-wn container
  - https://hub.docker.com/r/matterminers/wlcg-wn
  - https://github.com/MatterMiners/container-stacks/blob/main/wlcg-wn

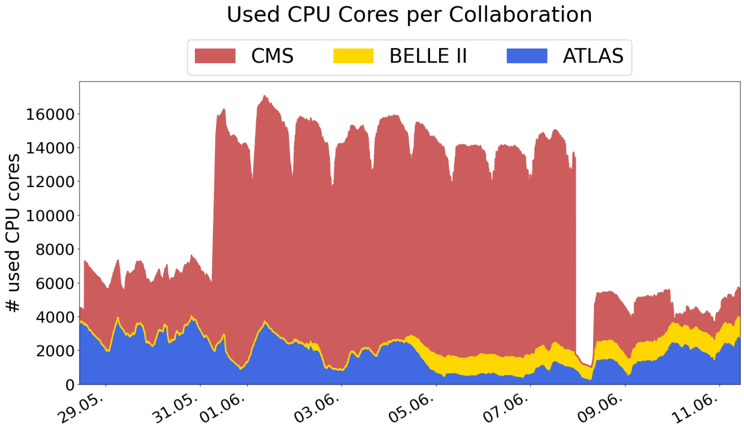
```
pip install cobald-tardis
```

Minimal Setup

- Grid Site
  - standard Grid site services
    - CE
    - OBS for resources
  - provide performant SE and outgoing network
- computing resource provider
  - accessible via HTCondor, Slurm, OpenStack, ...
  - virtualization or container with enables userspace
- COBalD/TARDIS instance
  - lightweight - multiple instances fit on one VM
  - needs just python and resource access
  - instances can be run by Grid site, resource provider, and third party



Provided Resources

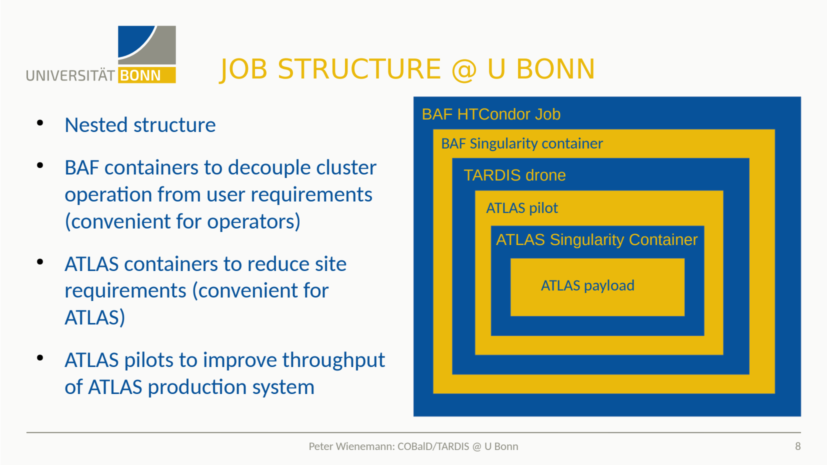


- used by several collaborations
- up to 17.400 CPU cores integrated

Supported Providers

- adapter to interact with provider
- providers
    - HTCondor
    - Moab
    - Slurm
    - CloudStack
    - OpenStack
    - Kubernetes
- further developments are welcome

Pilot inside a Drone



UNIVERSITÄT BONN

JOB STRUCTURE @ U BONN

- Nested structure
- BAF containers to decouple cluster operation from user requirements (convenient for operators)
- ATLAS containers to reduce site requirements (convenient for ATLAS)
- ATLAS pilots to improve throughput of ATLAS production system

BAF HTCondor Job

BAF Singularity container

TARDIS drone

ATLAS pilot

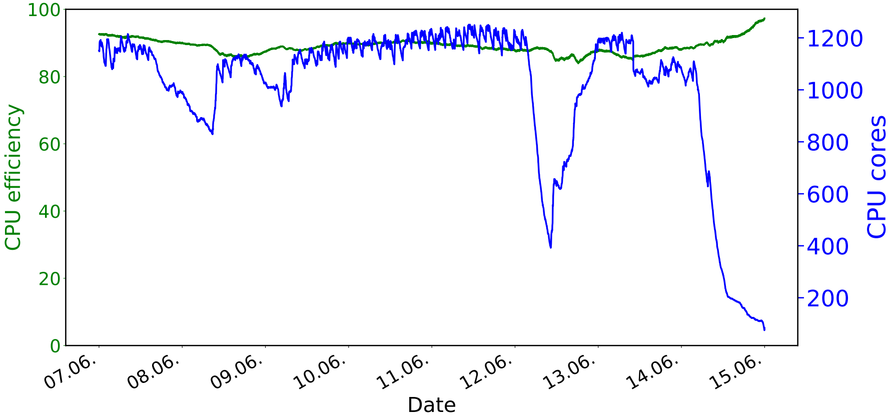
ATLAS Singularity Container

ATLAS payload

Peter Wienemann: COBalD/TARDIS @ U Bonn 8

Talk: Opportunistic Resource Mangement with COBalD/TARDIS at U Bonn from Peter Wienemann at the IDT-UM Meeting 30. Sep. 2019: https://indico.physik.uni-muenchen.de/event/22/

Used CPU cores and efficiency for Belle II



HTCondor Submit file for GPUs at GridKa

```
executable      = test.sh

universe        = grid
grid_resource   = condor cloud-htcondor-ce-2-kit.gridka.de cloud-htcondor-ce-2-kit.gridka.de:9619


request_cpus    = 8
arguments       = foo
request_gpus    = 1
request_memory  = 14000

should_transfer_files   = YES
when_to_transfer_output = ON_EXIT
x509userproxy           = /tmp/x509up_USERID

queue 1
```