Karlsruhe Institute of Technology

Dynamic Transparent Integration and Management of Opportunistic Resources with COBaID/TARDIS

ErUM Data Cloud Workshop, Karlsruhe 27.06.2019

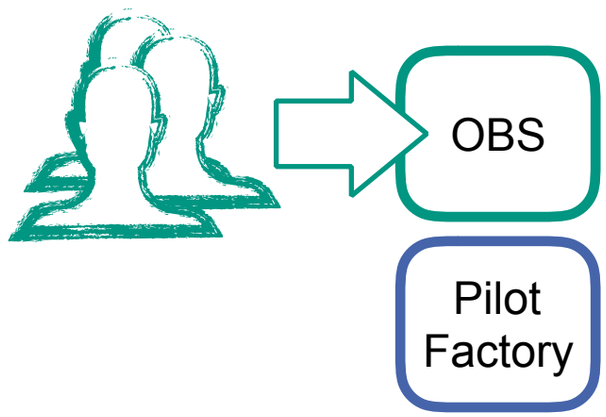
Manuel Giffels, René Caspart Max Fischer, Eileen Kühn, Matthias Schnepf, Florian v. Cube

Institute for Particle Physics (ETP) & Steinbuch Centre for Computing (SCC)

Opportunistic Resources and their Challenges

- Very different to the traditional HEP environment
  - ➡ Virtualisation and containerisation techniques
- Temporary availability of opportunistic resources
  - ➡ Dynamic integration and workflow management
- Varying demand for opportunistic resources
  - ➡ Dynamic integration and workflow management
- Availability and suitableness for given tasks is not predictable
  - ➡ Dynamic reaction on availability and utilisation
- Each opportunistic resource is different (very heterogenous system)
  - ➡ Hard to manage for users and computing operations of experiments
  - ➡ Transparent integration of resources needed

Transparent Integration: Overlay Batch System (OBS)



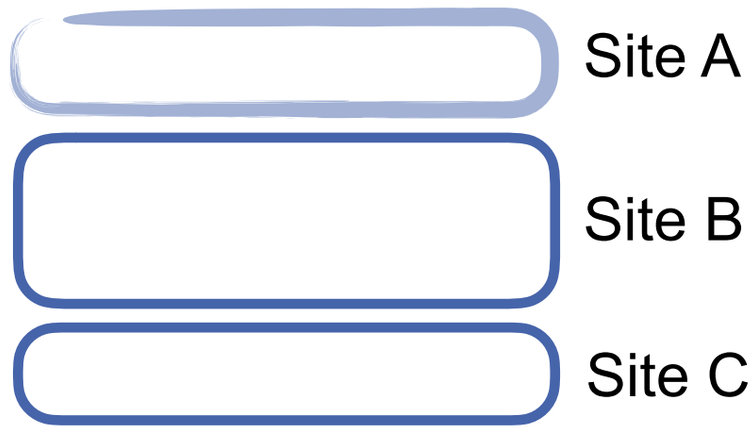
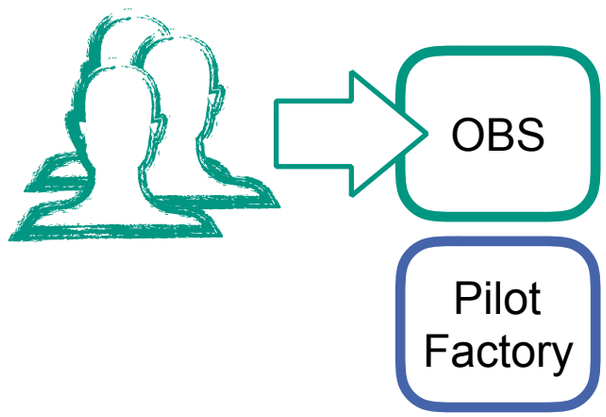


Illustration by Max Fischer (KIT)

Transparent Integration: Overlay Batch System (OBS)

Or how the Grid is used today:



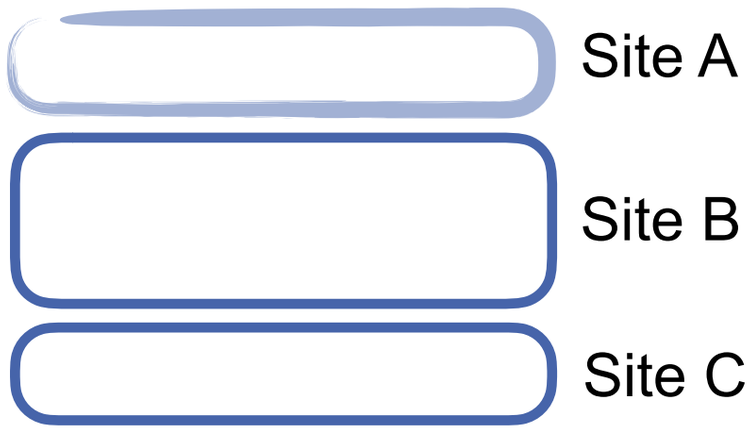


Illustration by Max Fischer (KIT)

Transparent Integration: Overlay Batch System (OBS)

Karlsruhe Institute of Technology

<u>Or how the Grid is used today:</u>

- Pilot factory submits placeholder jobs (pilots) to different sites

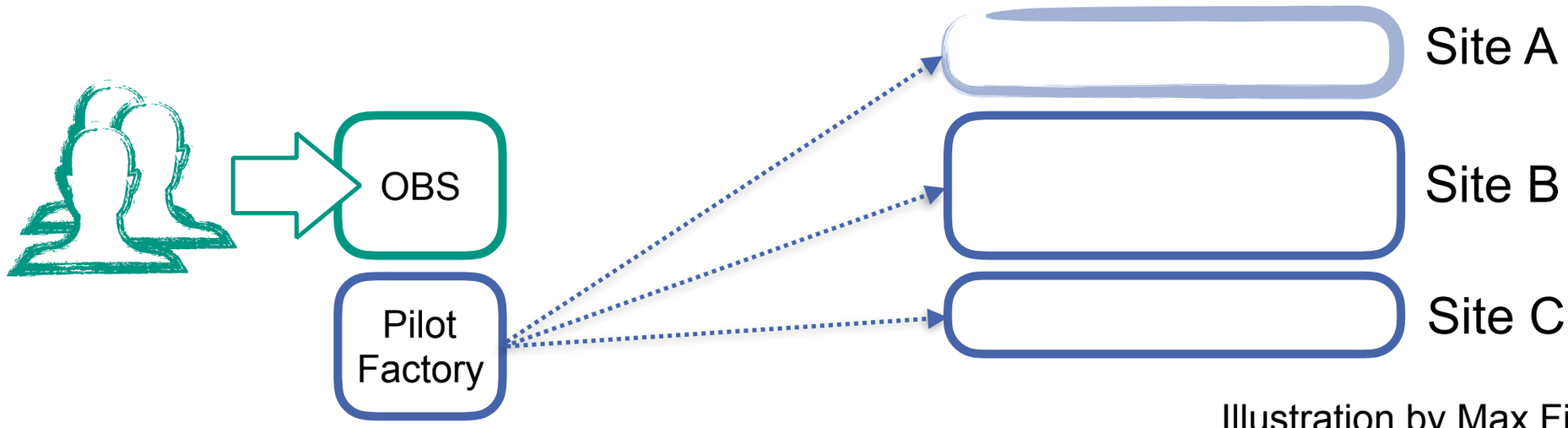
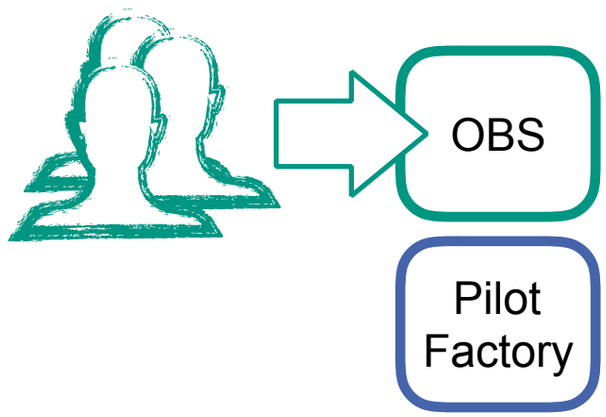


Illustration by Max Fischer (KIT)

Transparent Integration: Overlay Batch System (OBS)

Or how the Grid is used today:

- Pilot factory submits placeholder jobs (pilots) to different sites
- Pilot allocates resources at the site



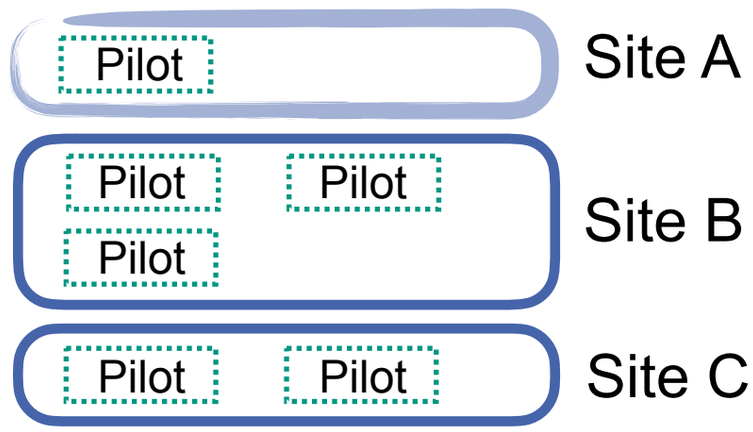


Illustration by Max Fischer (KIT)

Transparent Integration: Overlay Batch System (OBS)

Karlsruhe Institute of Technology

<u>Or how the Grid is used today:</u>

- Pilot factory submits placeholder jobs (pilots) to different sites
- Pilot allocates resources at the site
- Resources are integrated into the OBS
- Workload is pulled from the OBS

➡ Users interact only with one single-point-of-entry the OBS

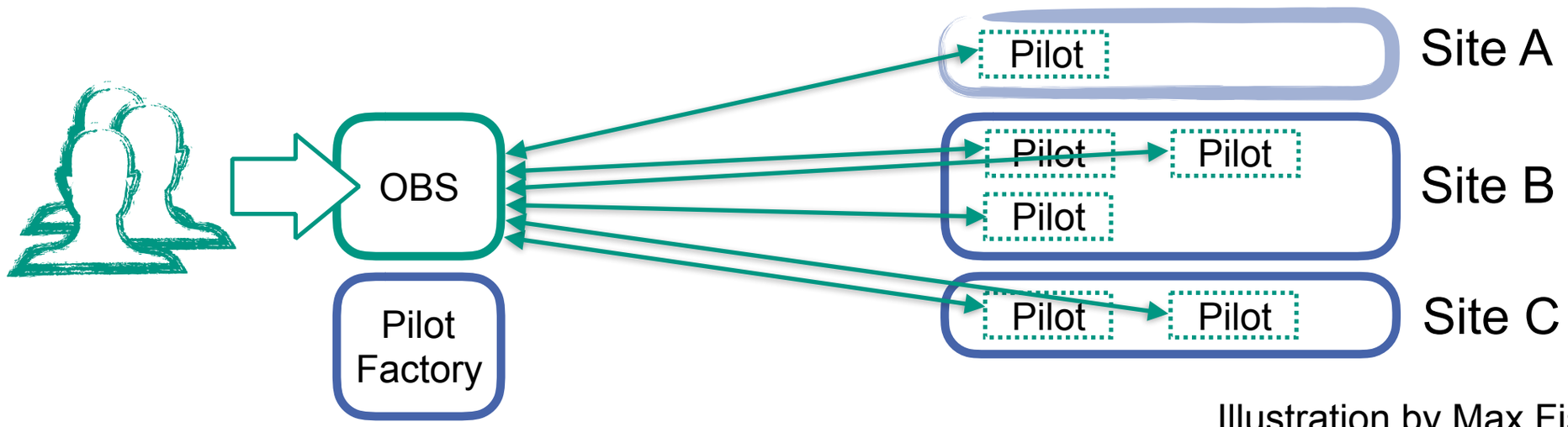
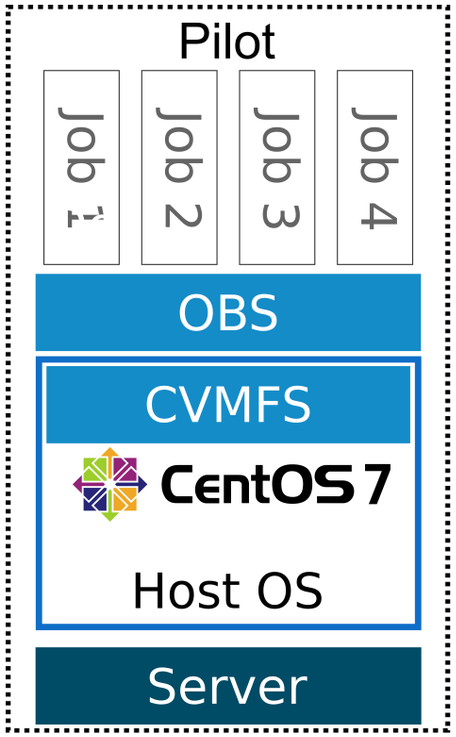
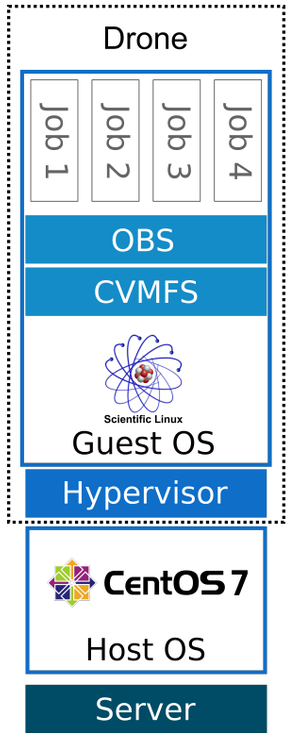


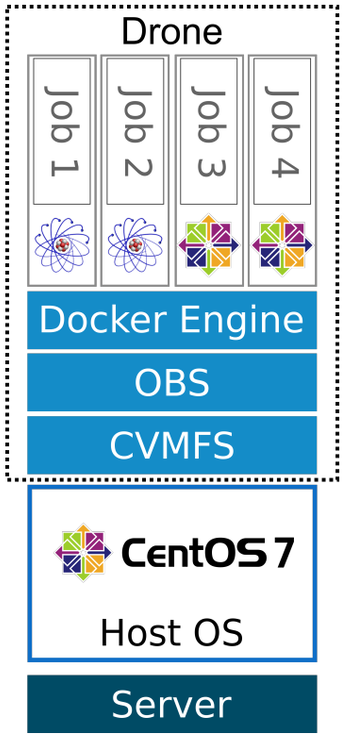
Illustration by Max Fischer (KIT)

Drone Concept

- Pilot concept
  - Allocate and integrate resources (via OBS)
  - Usually just a batch system daemon plus wrapper scripts
- Drone is a generalised pilot
  - Allocate and integrate resources (via OBS same as a pilot)
  - Provide dedicated environment (OS/Software) via virtualisation or containerisation techniques







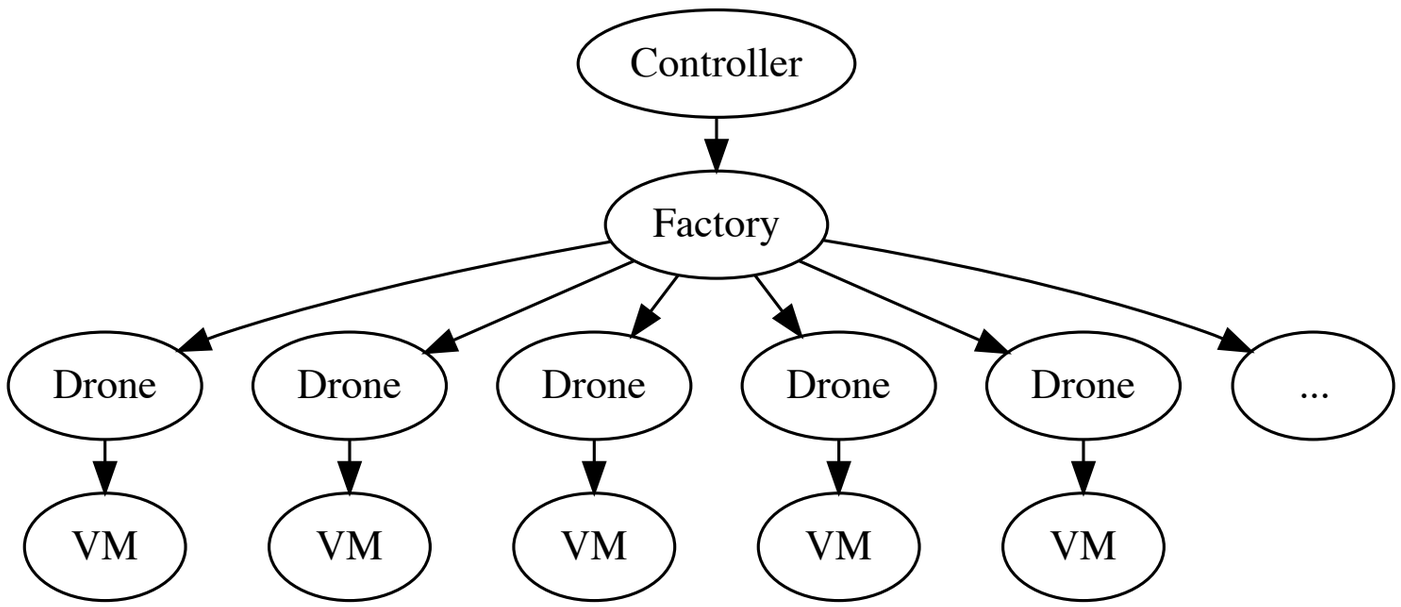
Dynamic Resource Integration with TARDIS

Transparent Adaptive Resource Dynamic Integration System (TARDIS):

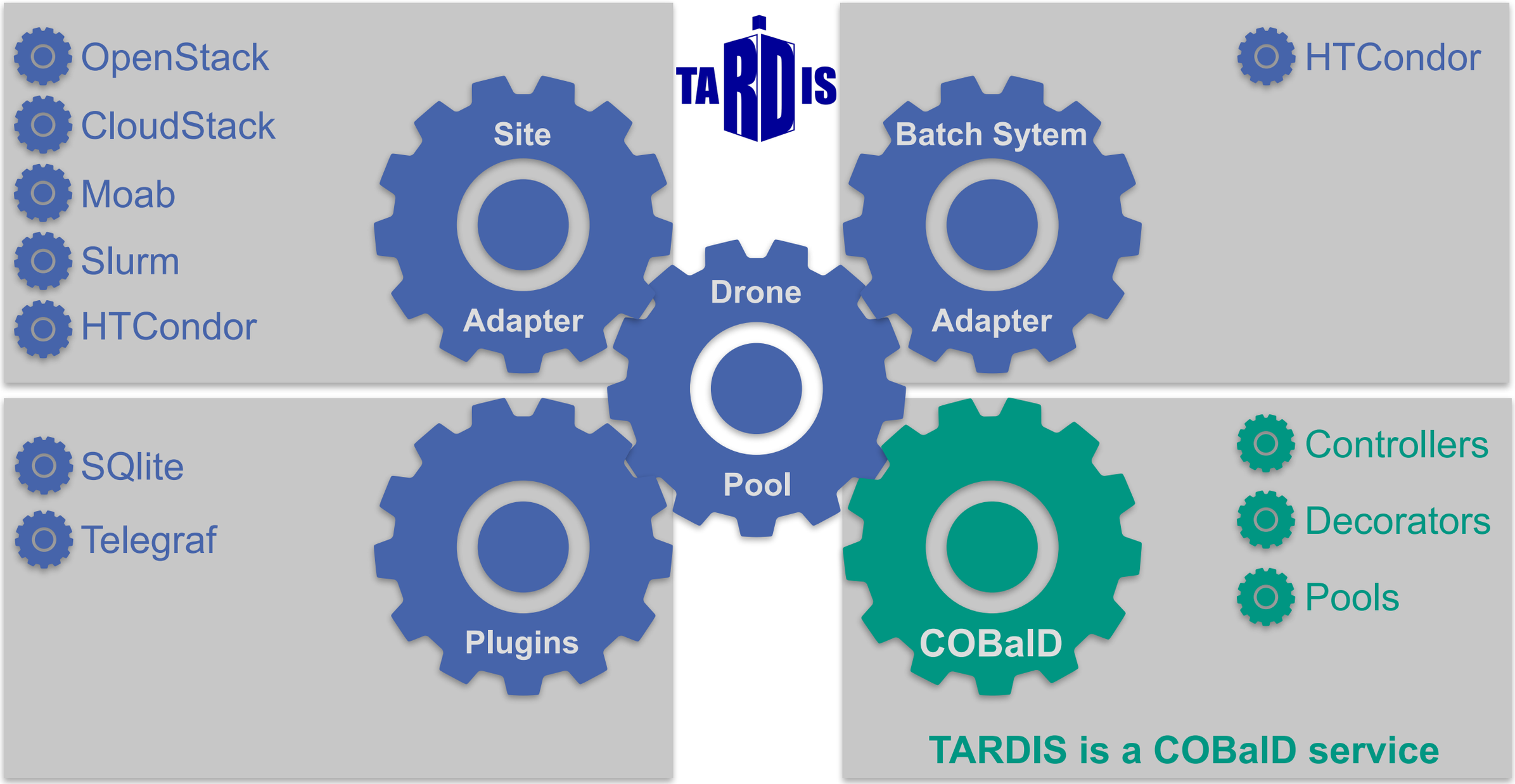
- Provides backends to various cloud providers and batch systems
- Allows dynamic orchestration of pre-built VMs and containers
- Asynchronous multi-agent COBalD service (ensures scalability)

Structure:

- A resource is represented by a Drone implementing the pool interface
- Drones are aggregated in dynamically sized composite pools
- Factory pool acquires and releases resources as desired by the controller

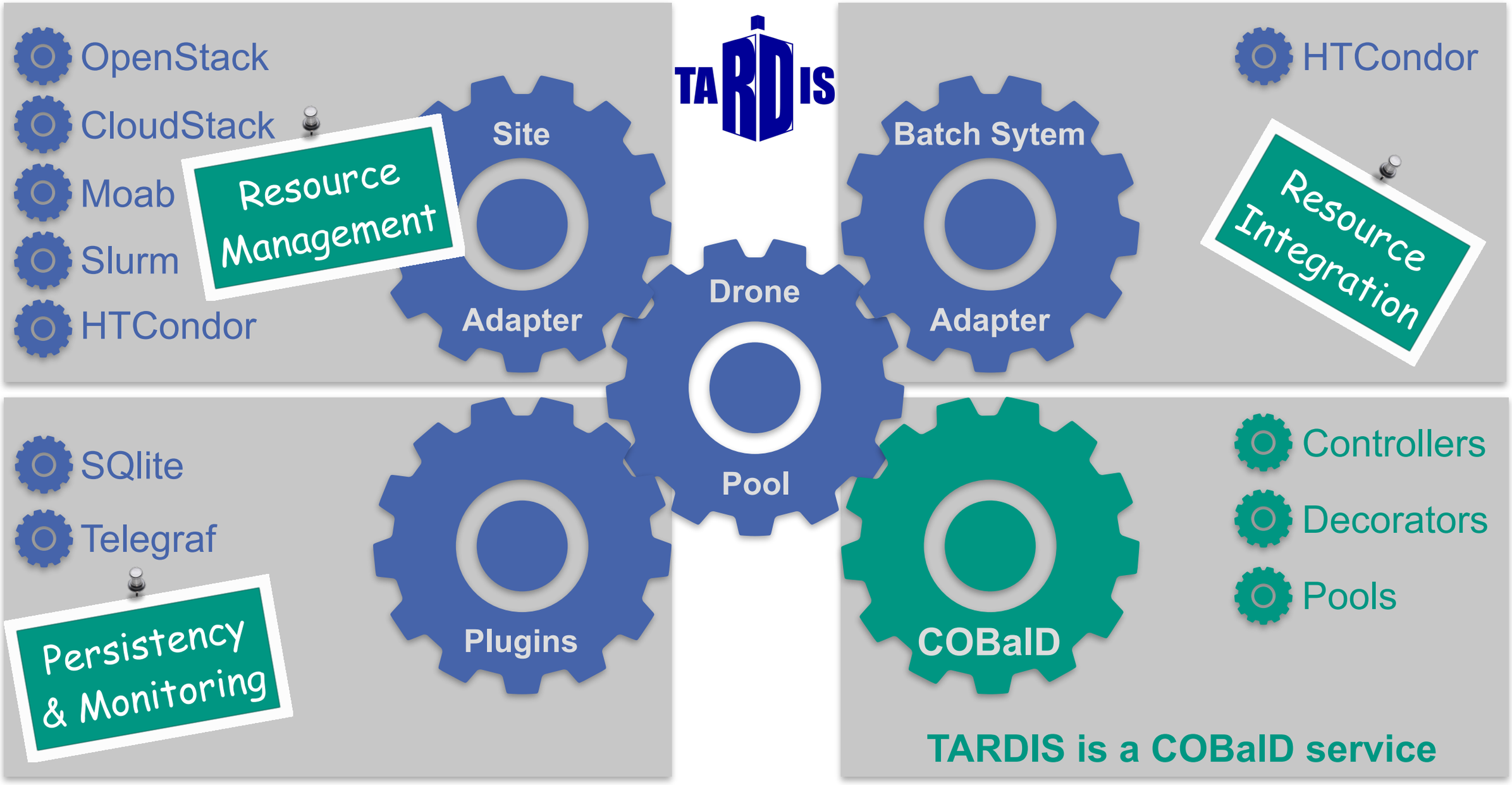


Components of COBalD-TARDIS



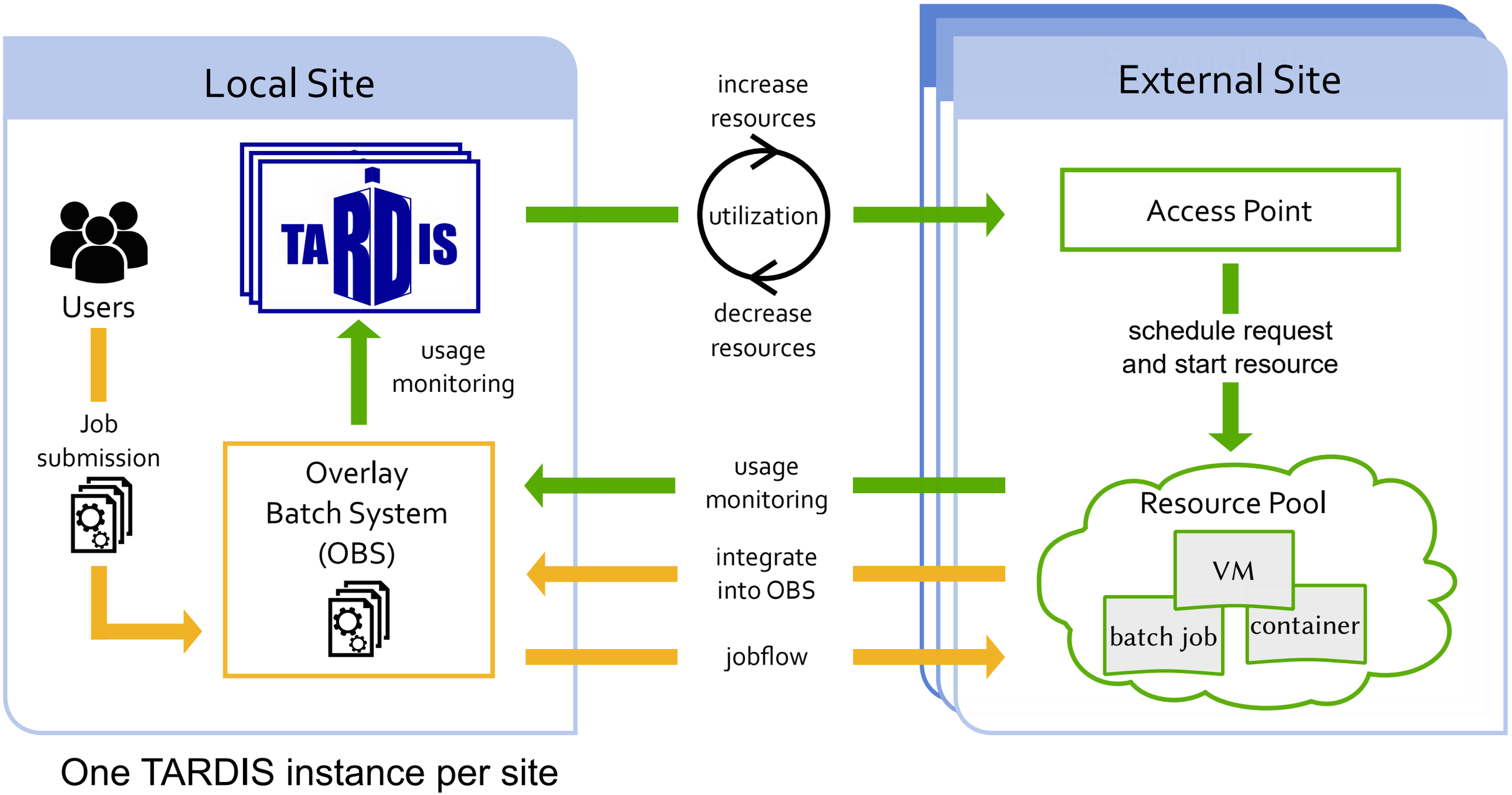
Well defined abstract base class interfaces to all components available!

Components of COBalD-TARDIS



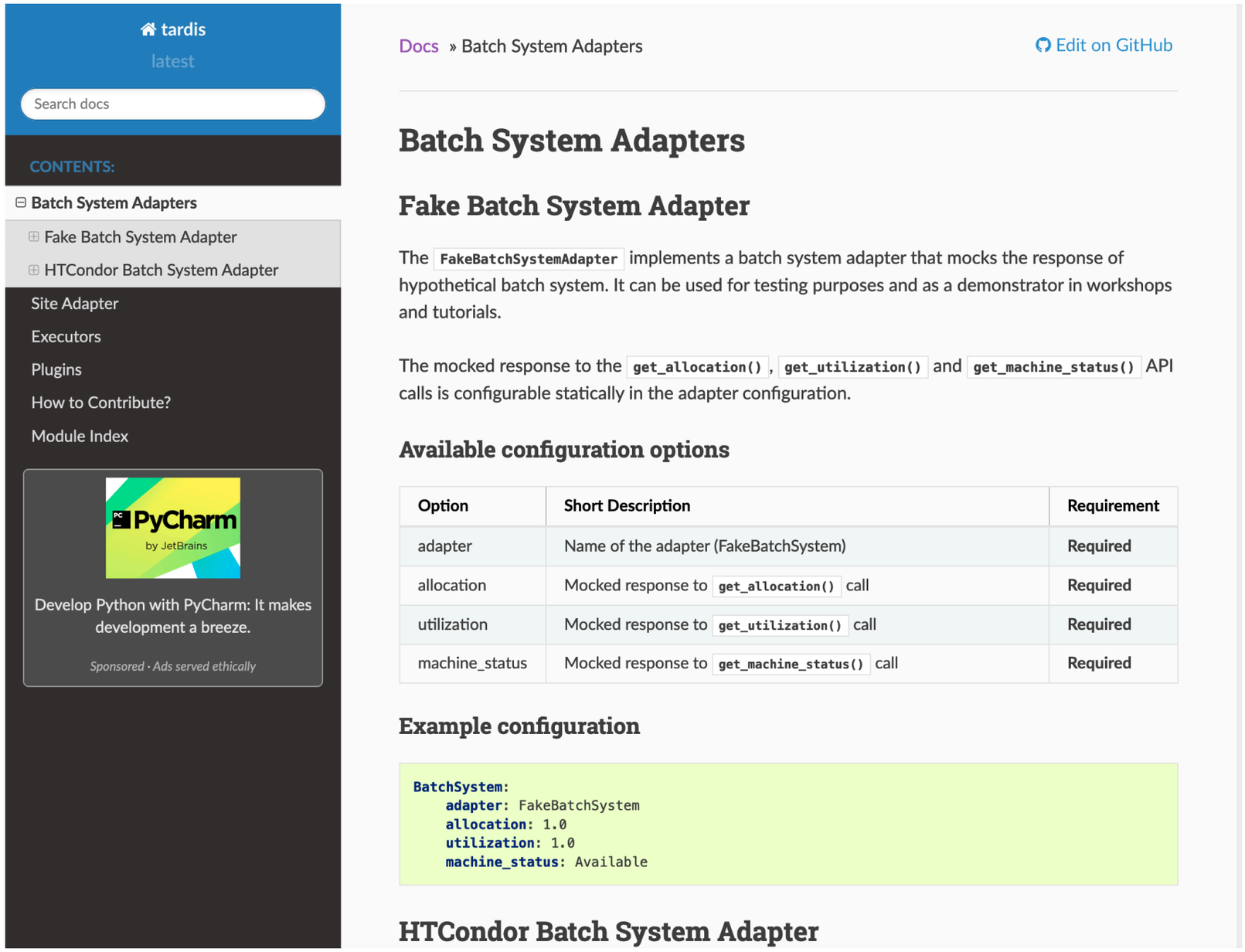
Well defined abstract base class interfaces to all components available!

The Entire Picture



One TARDIS instance per site

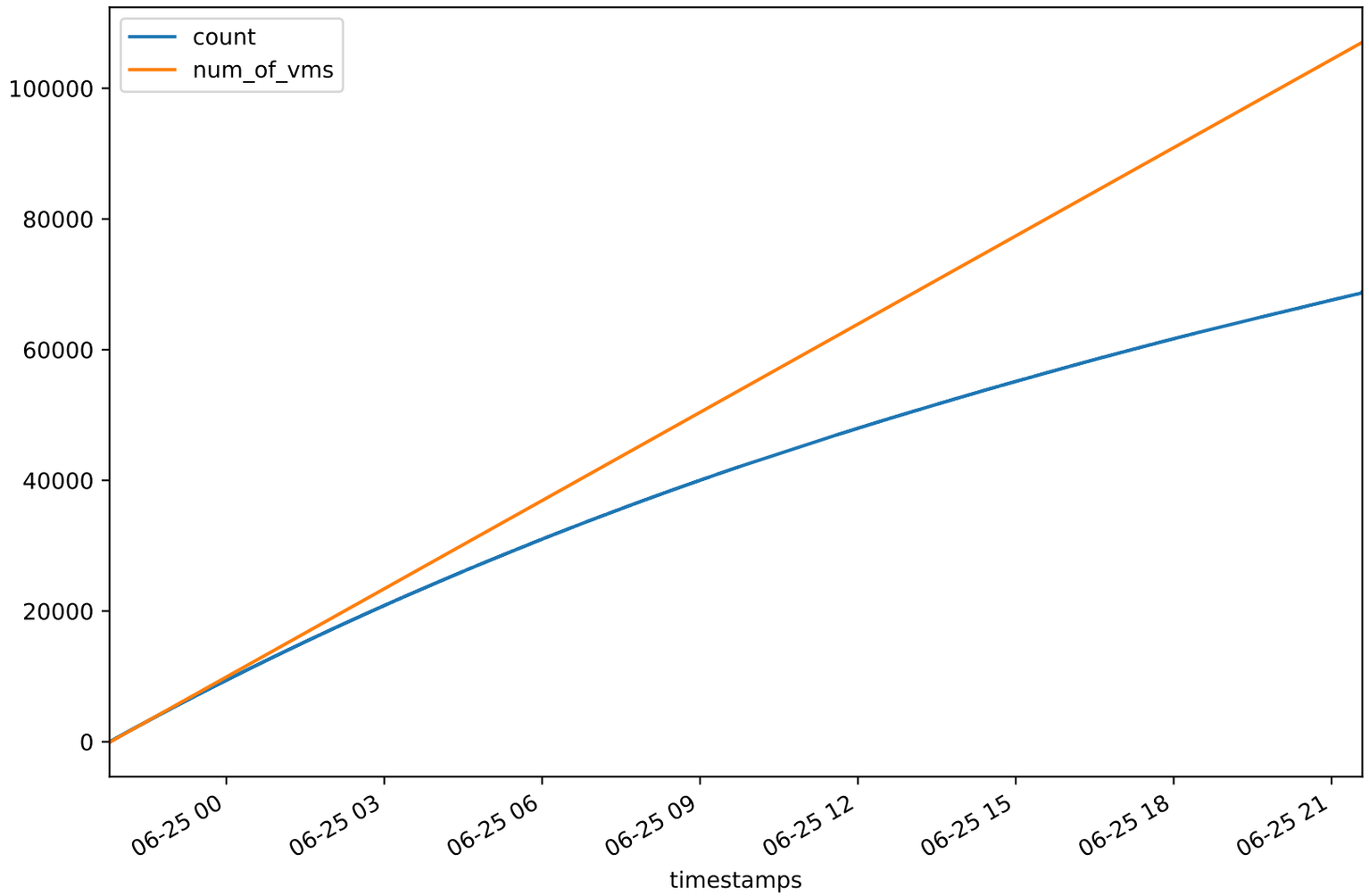
First Documentation Available



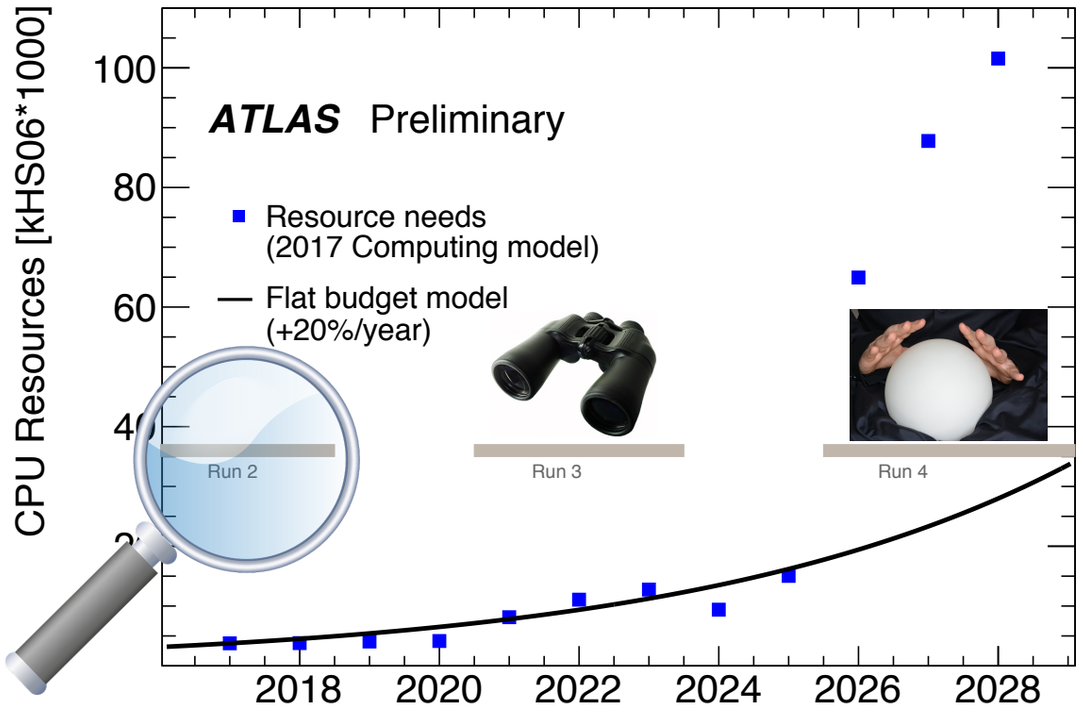
https://cobald-tardis.readthedocs.io/en/latest/index.html

BACKUP

Scalability



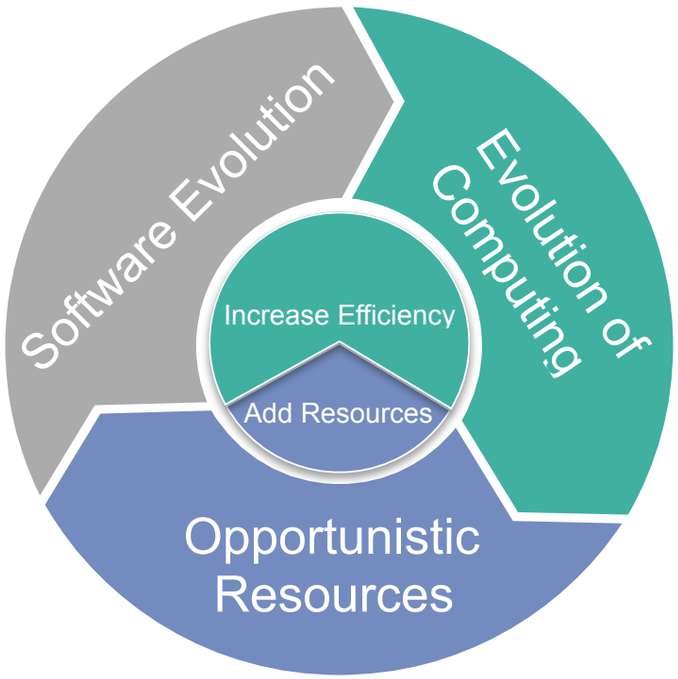
HEP Computing Challenge Ahead



Alves et al., A Roadmap for HEP Software and Computing R&D for the 2020s, *HSF-CWP-2017-001 (*2017)

- ATLAS/CMS CPU resource estimates
- Assuming flat budget and 20% technology advance per year
- CPU shortfall between needs and technology gains about factor 4 in 2027
- Situation for disk storage is even worse (Shortfall is about factor 7 in 2027)

➡ Critical conditions for HL-LHC (Run 4)



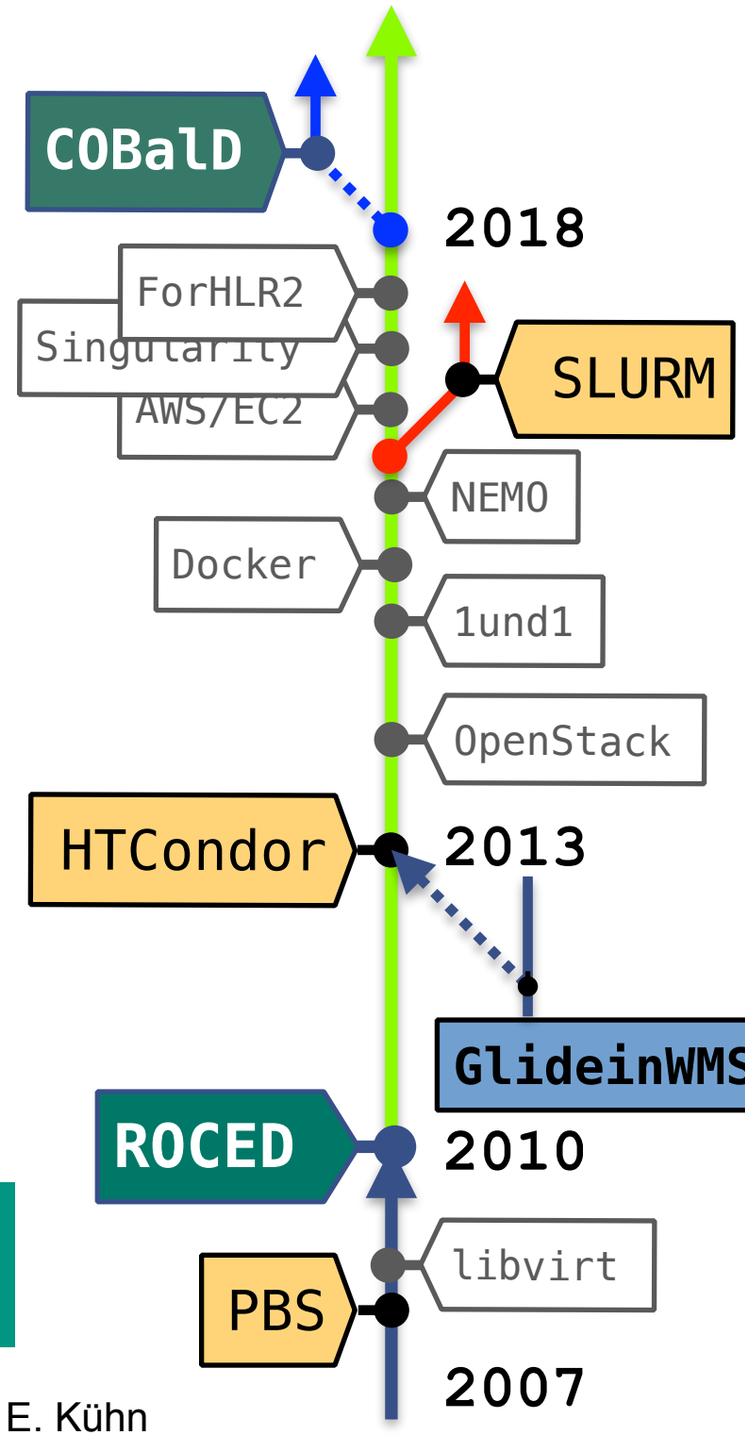
<u>New interesting research topics:</u>

- Exploitation of modern technologies
- Improvement of algorithms and utilization of ML
- Dynamic integration of opportunistic resources (HPC, cloud, volunteer computing)
- Data lakes and data caching technologies

Opportunistic Resources @ KIT (2007-present)

- Longterm experience with opportunistic resources, virtualization and containers
- Software development (Cloud Resource Manager) ViBatch, ROCED (➠COBalD/TARDIS)[1]
- Opportunistic Resources
  - Institute resources
    - Desktop cluster (Docker)
    - GridKa School Virtual Infrastructure (OpenStack)
  - HPC Cluster
    - ~~IC1@Uni Karlsruhe (ViBatch)~~
    - ForHLR II @ KIT (Singularity)
    - bwFORcluster NEMO @ Uni Freiburg (OpenStack)
  - Commercial cloud providers
    - AWS, 1&1 Cloud Services, OTC, ExoScale

All dynamically and transparently integrated in a single HTCondor instance using ROCED and CoBalD/TARDIS

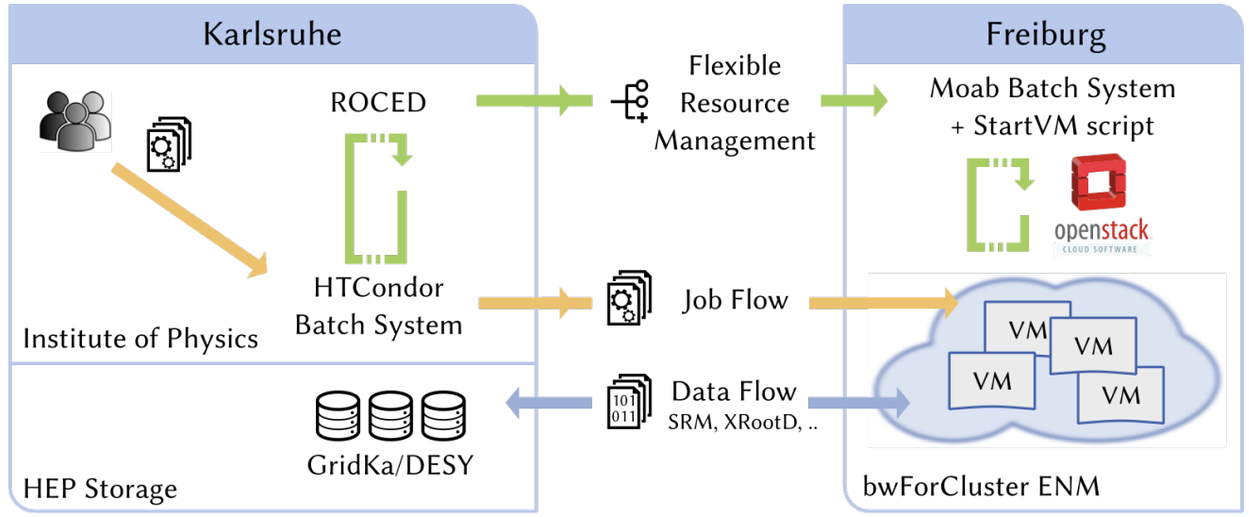


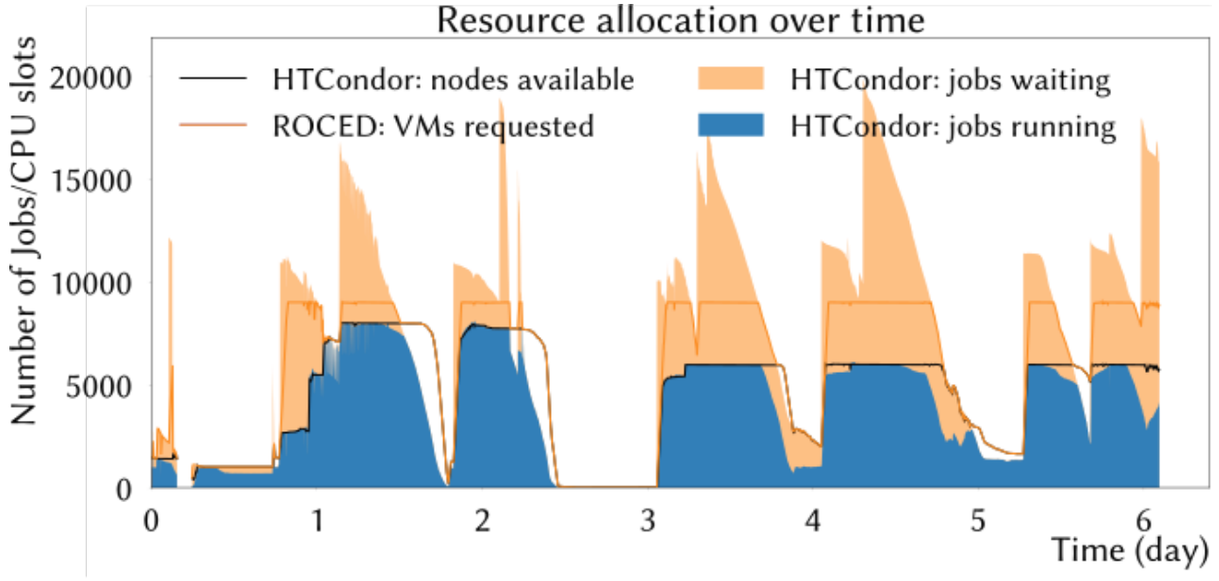
[1] COBalD - the Opportunistic Balancing Daemon (http://cobald.readthedocs.io/) by M. Fischer & E. Kühn
TARDIS - **T**ransparent **A**daptive **R**esource **D**ynamic **I**ntegration **S**ystem (https://github.com/giffels/tardis)

Illustration by Max Fischer (KIT)

Success Story - Opportunistic “Tier 1“ for a Day

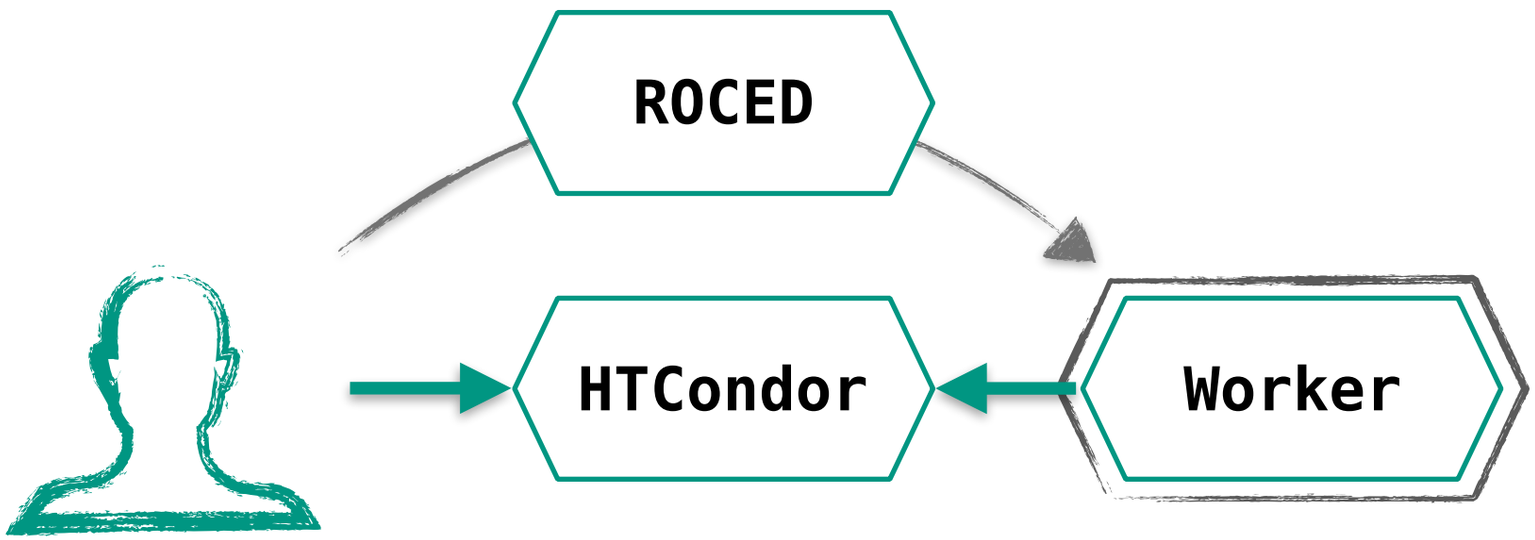
- Dynamically shared HPC Centre at Freiburg (three diverse communities)
- Virtualization is key component to:
  - Allow dynamic resource partitioning
  - Meet OS & software requirements
- ROCED cloud scheduler developed at KIT
  - On-demand resource provisioning
  - Transparent resource integration
- Suitable for CPU-intense workflows
- I/O-intense workflows need data caches and fast networks (more later!)





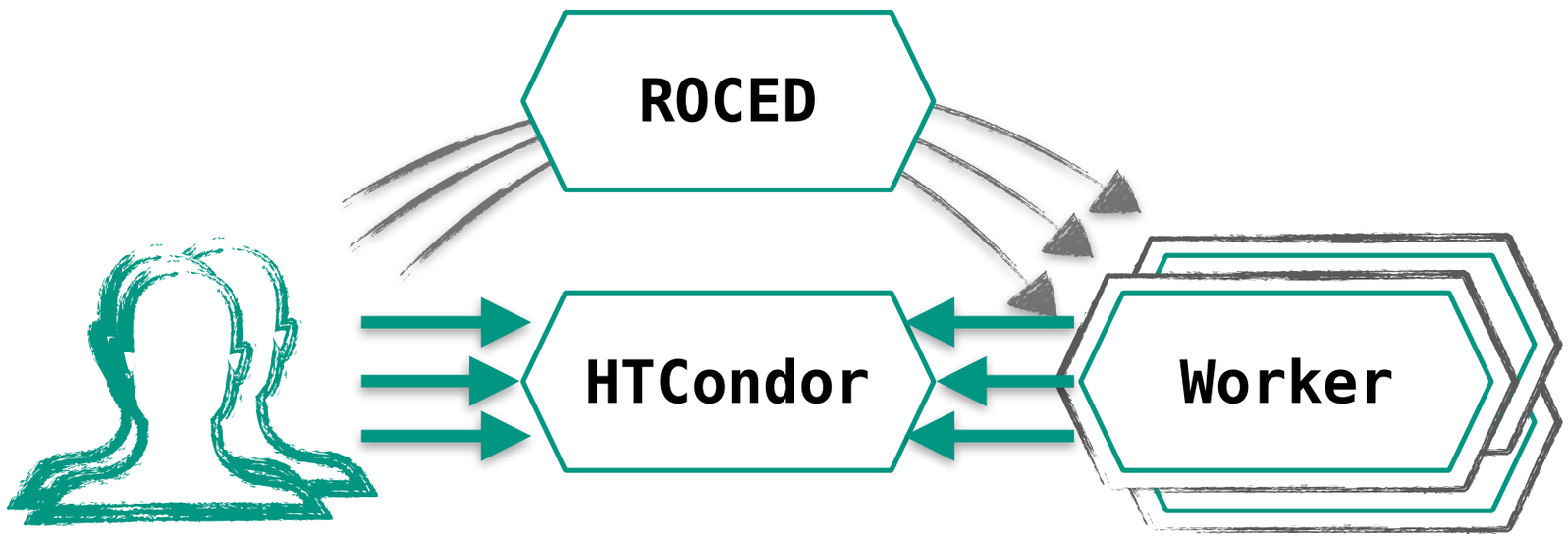
Very good experience, however HEP was involved in the project from the early beginning

ROCED Resumé



Slide by Max Fischer (KIT)

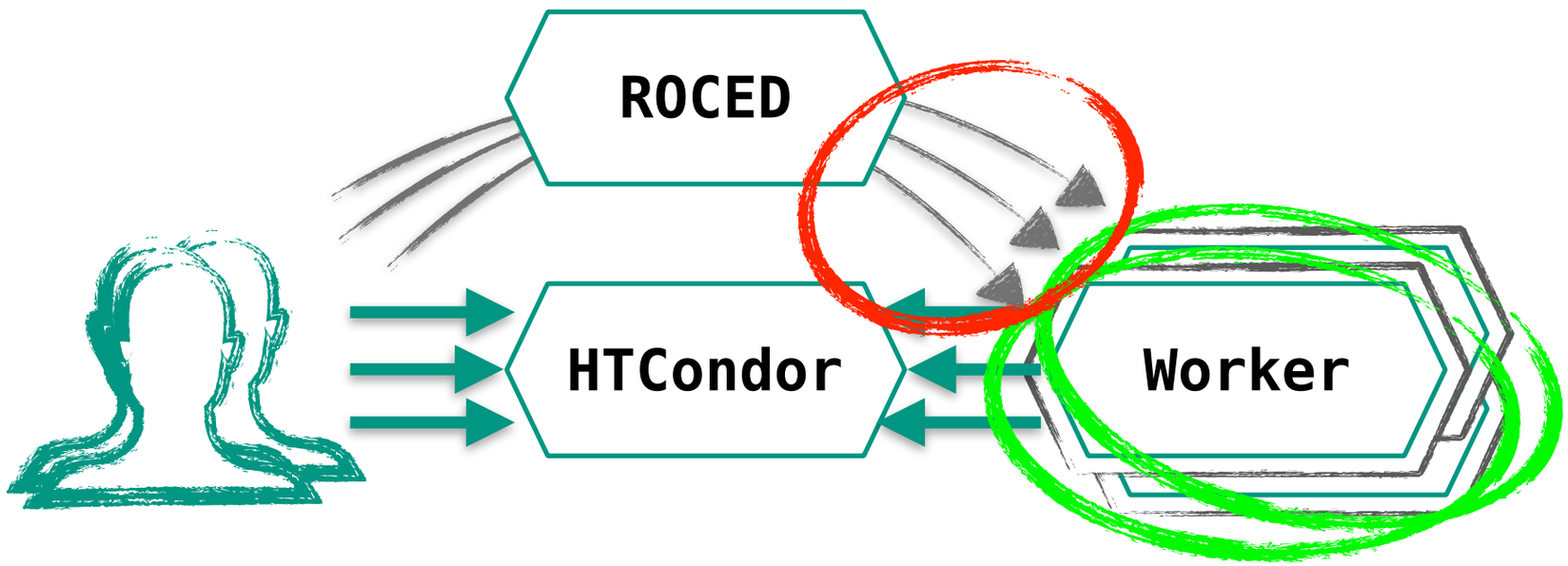
ROCED Resumé



Slide by Max Fischer (KIT)

ROCED Resumé

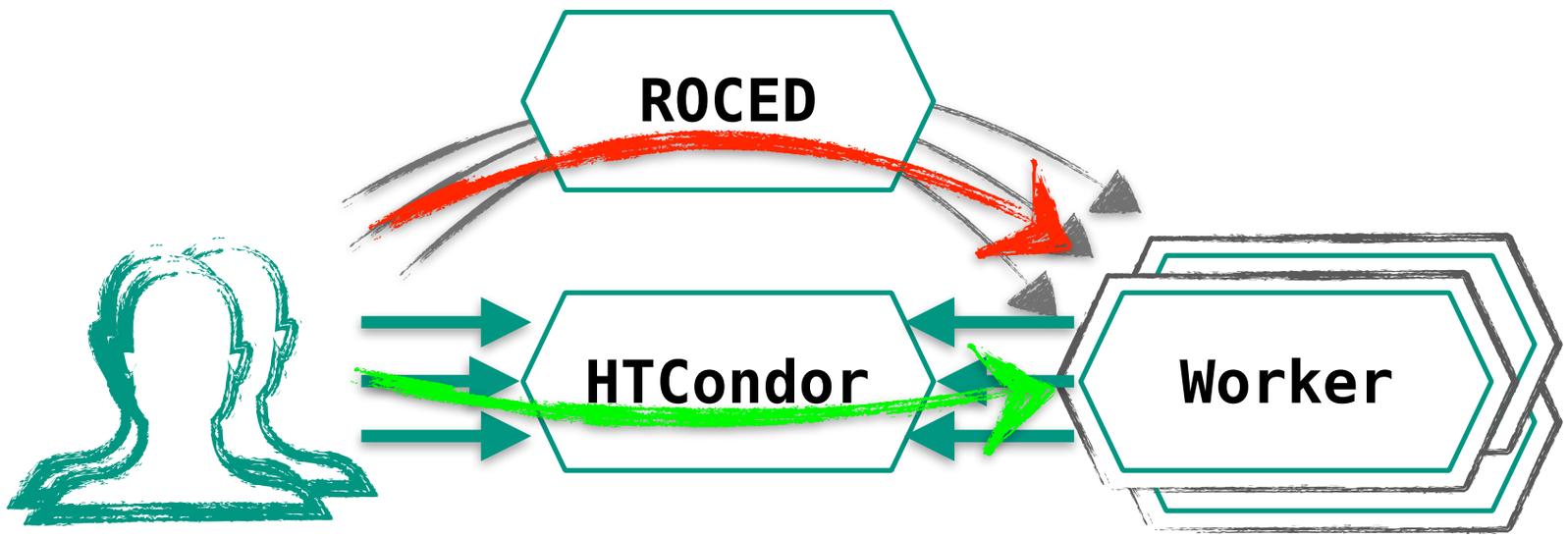
- Dynamic resources matching user demand
  - Trivial to support new providers for many users
  - Difficult to manage several providers for many users



Slide by Max Fischer (KIT)

ROCED Resumé

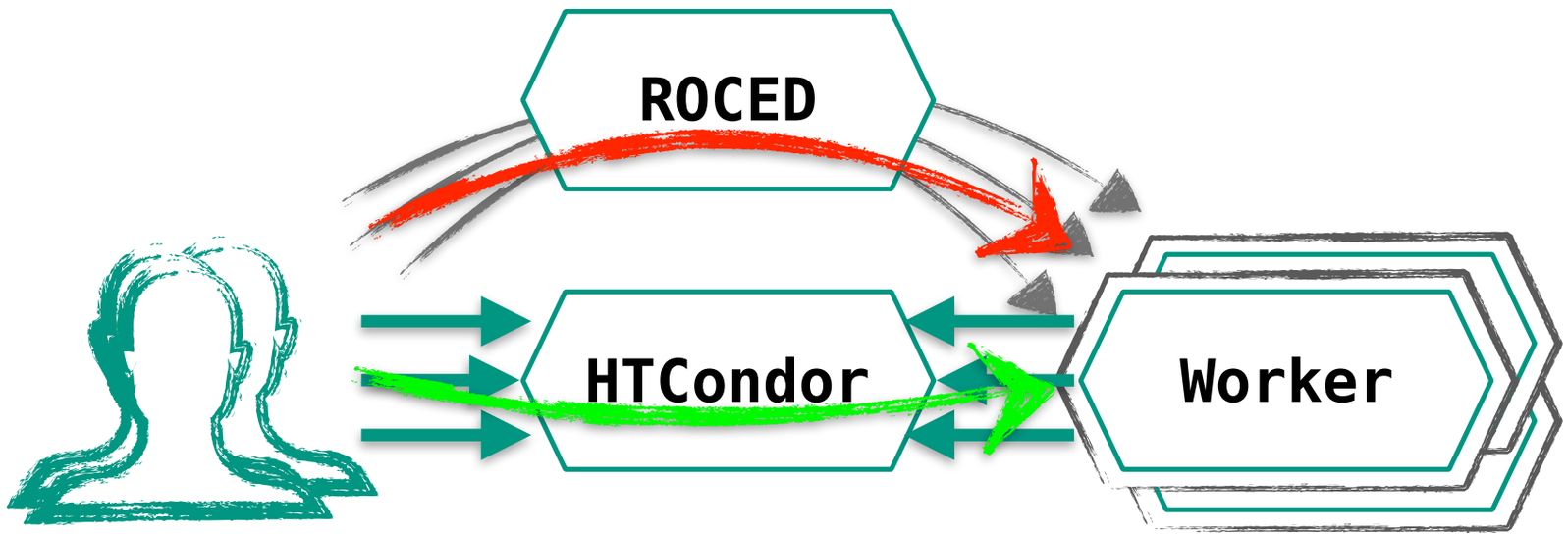
- Dynamic resources matching user demand
  - Trivial to support new providers for many users
  - Difficult to manage several providers for many users
- Resource aggregation in overlay batch system
  - Unreliable to predict resources required for jobs
  - Efficient to integrate resources, then match jobs



Slide by Max Fischer (KIT)

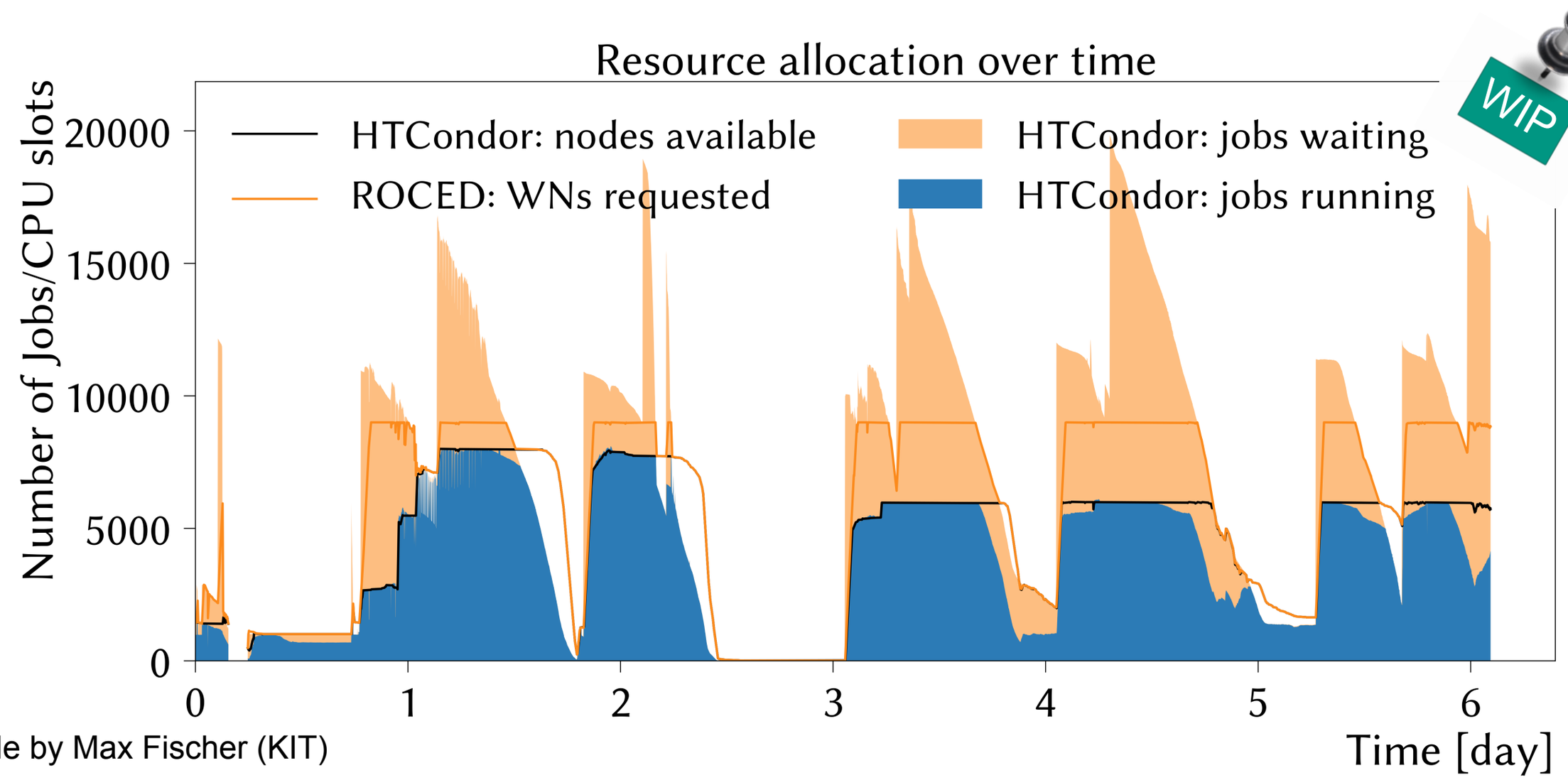
ROCED Resumé

- Dynamic resources matching user demand
  - Trivial to support new providers for many users
  - Difficult to manage several providers for many users
- Resource aggregation in overlay batch system
  - Unreliable to predict resources required for jobs
  - Efficient to integrate resources, then match jobs
- Yet it really works!



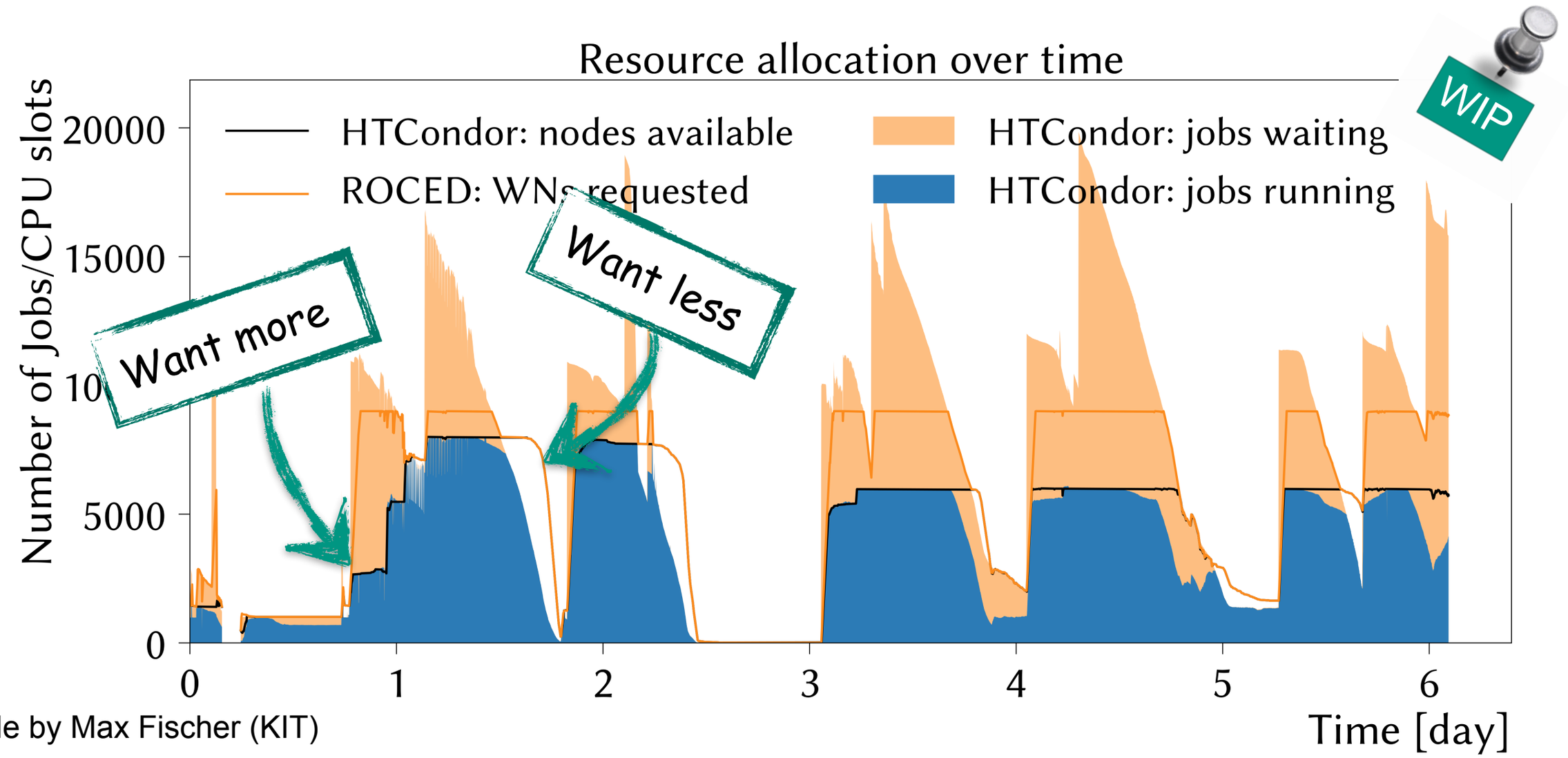
Slide by Max Fischer (KIT)

Pragmatic View to Resource Management



Slide by Max Fischer (KIT)

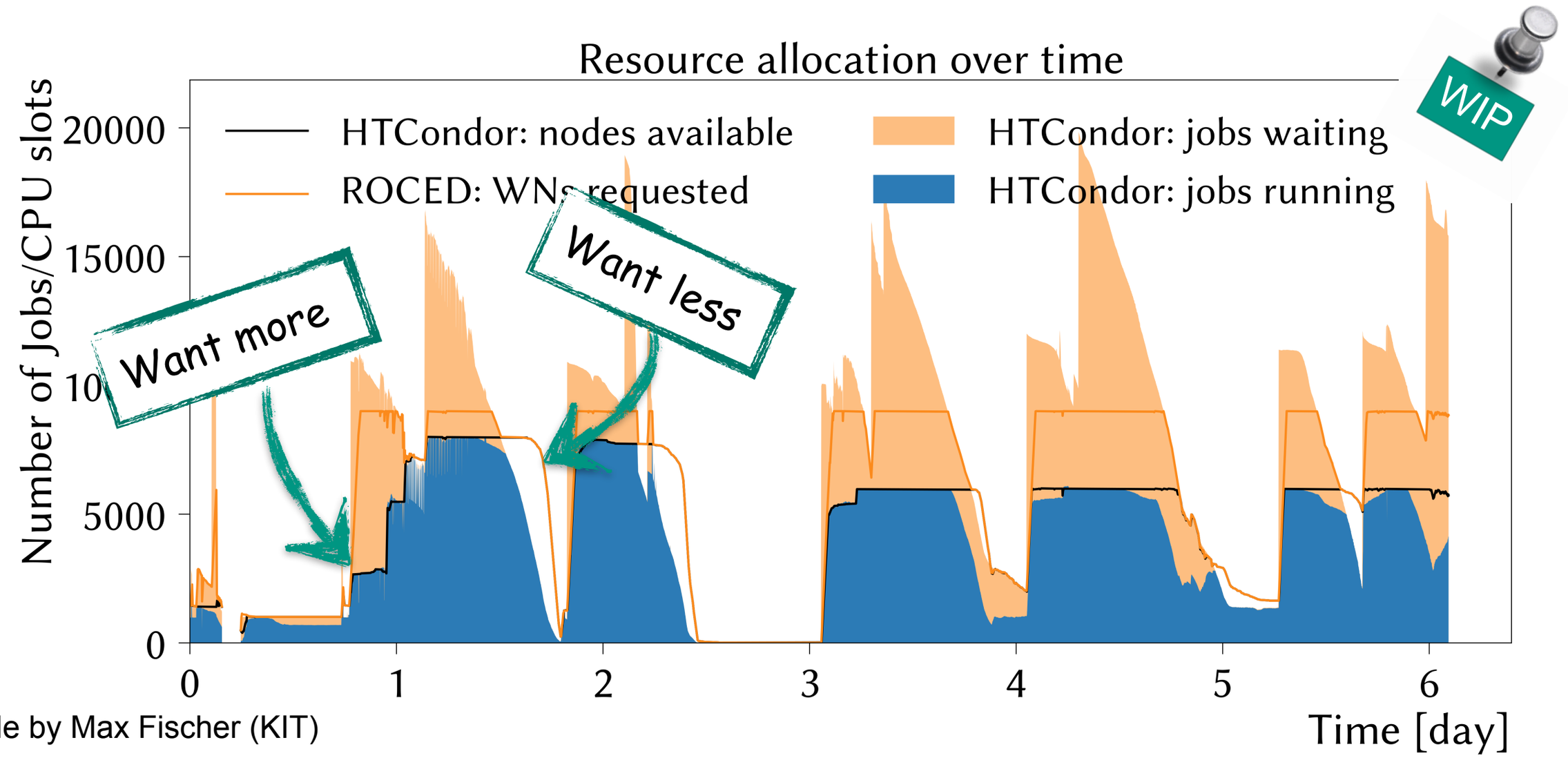
Pragmatic View to Resource Management



Slide by Max Fischer (KIT)

Pragmatic View to Resource Management

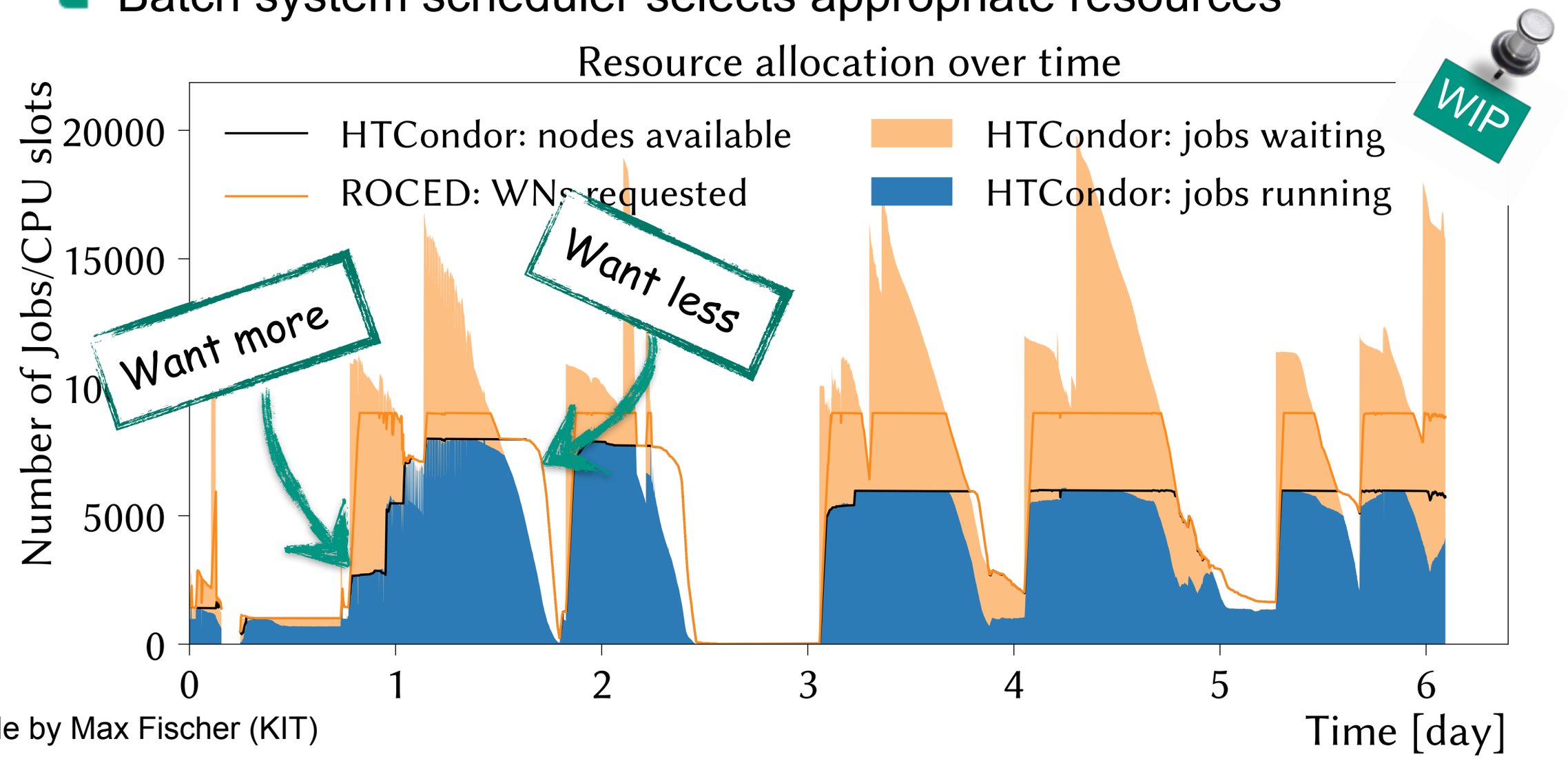
- New development for scalability and maintainability in HNSciCloud



Slide by Max Fischer (KIT)

Pragmatic View to Resource Management

- New development for scalability and maintainability in HNSciCloud
- Simple logic: more used resources, less unused resources
  - COBalD only watches, creates and disables resources
  - Batch system scheduler selects appropriate resources



Slide by Max Fischer (KIT)

Innovative Digital Technologies for Exploring Universe and Matter

- Joint proposal by HEP, Physics of Hadrons and Nuclei, Astroparticle Physics
- Covered Topics:
  - Development of technologies to utilize heterogeneous computing resources (Integration of Opportunistic Resources, Caching Technologies, Workflow Management)
  - Application and testing of those technologies in heterogenous computing resources
  - Deep Learning - Achieving knowledge through profound data-driven methods (Hardware-related Data Processing, Object Reconstruction, Simulation, Quality of Network Predictions)
  - Event reconstruction: Cost- and energy efficient utilization of computing resources
    (Alternative Algorithms and Architectures like GPUs)
- Funded in the scope of Digital Agenda programme (BMBF)

Proposal of:

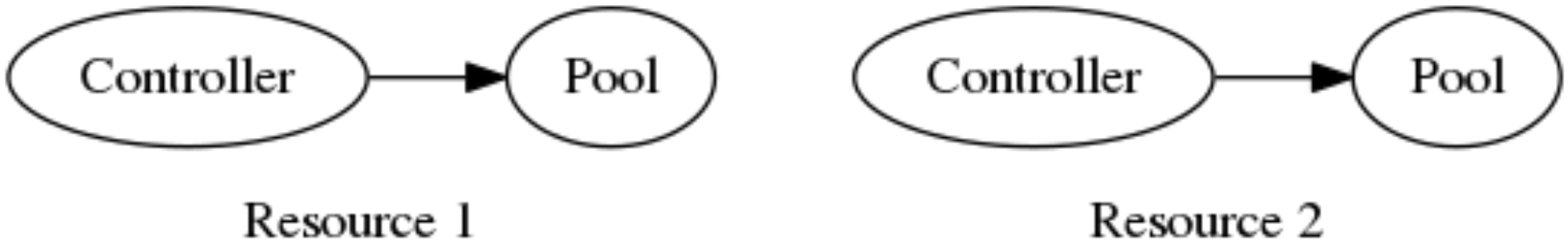
Associated Partners:

Resources

- COBalD: http://cobald.readthedocs.io/
- ROCED: https://github.com/roced-scheduler/ROCED
- TARDIS: http://cobald-tardis.readthedocs.io/
- COBalD Simulation: https://git.scc.kit.edu/fq8360/cobalt_sim
- COBalD Demo: https://github.com/MaineKuehn/cobald_demo

COBalD Resource Pool Model

COBalD Resource Pool Model



COBalD Resource Pool Model