universität freiburg

# Accounting with AUDITOR

## WLCG Operations Coordination Meeting

Michael Böhler
October 3rd 2024

Bundesministerium
für Bildung
und Forschung

IDIUM

Physikalisches Institut
Albert-Ludwigs-
Universität Freiburg

# AUDITOR

???

- What is AUDITOR?
- Which technologies are used?
- Who is AUDITOR?
- What can we do with AUDITOR?
- What are existing use cases?

- AUDITOR stands for **A**ccco**u**nting **D**ata Handl**i**ng **T**oolbox For **O**pportunistic **R**esources

# AUDITOR

???

- What is AUDITOR?
- Which technologies are used?
- Who is AUDITOR?
- What can we do with AUDITOR?
- What are existing use cases?

**Internal Auditor (IA)**

*[in-ˈtər-nᵊl ˈȯ-də-tər]*

A trained professional employed by companies to provide independent and objective evaluations of financial and operational business activities, including corporate governance.

Investopedia

- AUDITOR stands for **A**cco**u**nting **D**ata Hand**l**ing **T**oolbox For **O**pportunistic **R**esources

# AUDITOR

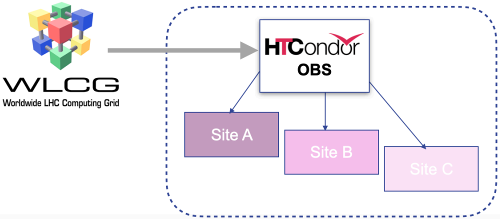
???

- What is AUDITOR?
- Which technologies are used?
- Who is AUDITOR?
- What can we do with AUDITOR?
- What are existing use cases?

- AUDITOR stands for **A**cco**u**nting **D**ata Handl**i**ng **T**oolbox For **O**pportunistic **R**esources

# Original Motivation

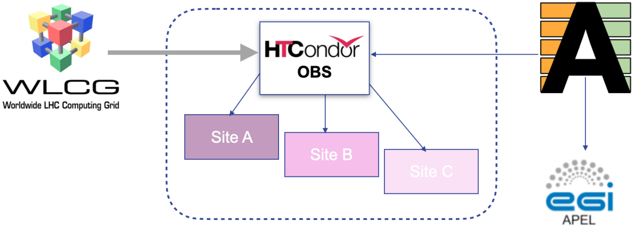
## Accounting opportunistic resources



- COBalD/TARDIS allows multiple resources to be clustered in an **O**verlay **B**atch **S**ystem
  - Sub clusters **cannot be accounted individually** with existing tools
  - Requires a dedicated mechanism for accounting
- Challenges
  - Vastly different infrastructures
  - Many potential use cases
- **AUDITOR** provides multi-purpose accounting ecosystem

# Original Motivation

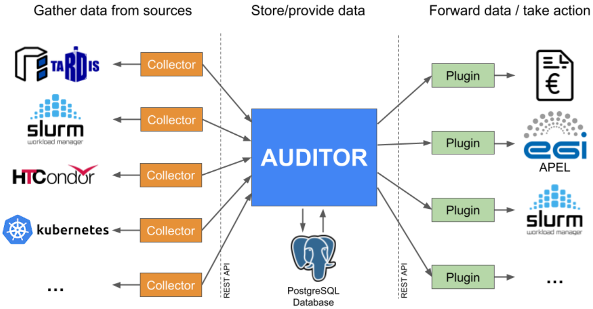
Accounting opportunistic resources



- COBalD/TARDIS allows multiple resources to be clustered in an **O**verlay **B**atch **S**ystem
  - Sub clusters **cannot be accounted individually** with existing tools
  - Requires a dedicated mechanism for accounting
- Challenges
  - Vastly different infrastructures
  - Many potential use cases
- **AUDITOR** provides multi-purpose accounting ecosystem

# AUDITOR

## Accounting Ecosystem



### Modular accounting ecosystem

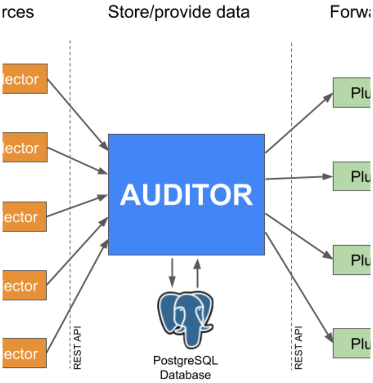
- **Collectors**
  - Accumulate data
- **Core component**
  - Accept data
  - Store data
  - Provide data
- **Plugins**
  - Take action based on stored data

### Documentation and code

https://github.com/ALU-Schumacher/AUDITOR

# Auditor

Core component

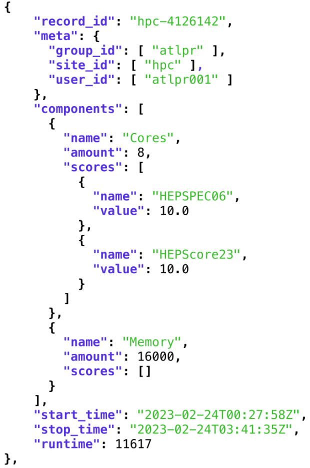


- Implemented in **Rust**
  - Access via REST interface
- Unit of accountable resources: **Record**
- Data stored in PostgreSQL
- Completely stateless
  - No dataloss
  - Suitable for high availability setups
- Provided as **RPM** or **Docker container**
- Client libraries in Rust and Python

# Record

Unit of accountable resources

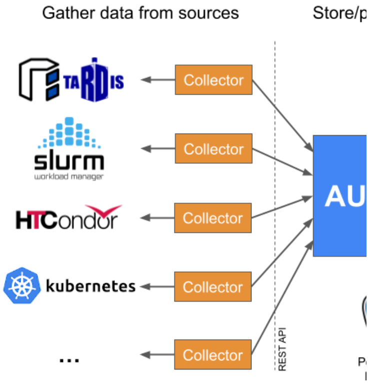
- `record_id`: uniquely identifies the record
- `meta`: multiple key value pairs of the form `String -> [String]`
- `components`: arbitrary number of resources that are to be accounted for (CPU, RAM, Disk, GPU, . . . )
  - `scores`: (multiple) accounting scores supported
- `start_time`, `end_time`: datetime in UTC
- `runtime`: calculated as `end_time - start_time`



```
{
    "record_id": "hpc-4126142",
    "meta": {
      "group_id": [ "atlpr" ],
      "site_id": [ "hpc" ],
      "user_id": [ "atlpr001" ]
    },
    "components": [
      {
        "name": "Cores",
        "amount": 8,
        "scores": [
          {
            "name": "HEPSPEC06",
            "value": 10.0
          },
          {
            "name": "HEPScore23",
            "value": 10.0
          }
        ]
      },
      {
        "name": "Memory",
        "amount": 16000,
        "scores": []
      }
    ],
    "start_time": "2023-02-24T00:27:58Z",
    "stop_time": "2023-02-24T03:41:35Z",
    "runtime": 11617
},
```

# Collectors

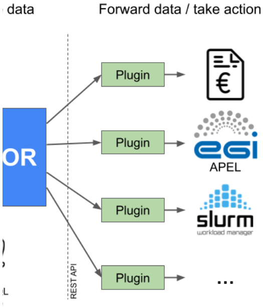
Accumulate data



- **TARDIS Collector**
  - Collect drone information
- **SLURM Collectors**
  - Collect information about SLURM jobs via SLURM CLI commands
- **HTCondor Collector** (developed @ KIT)
  - Equivalent of SLURM collector for HTCondor
- **Kubernetes Collector** (developed @ Uni Wuppertal)
  - Collects information from Kubernetes pods

# Plugins

Take action based on stored data

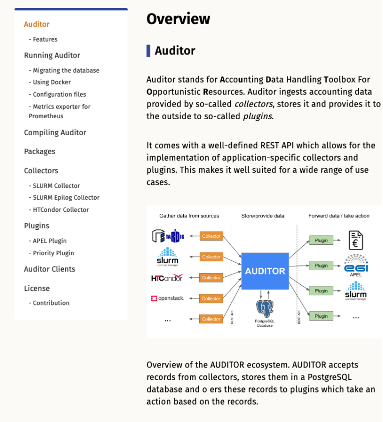


- **Priority plugin**
  - Compute priorities from a list of records
  - Update priorities on a batch cluster
- **APEL accounting plugin**
  - Properly accounts individual sites behind COBalD/TARDIS
  - Reports accounting data to the APEL accounting platform
- **Utilization report** (future project)
  - Analyse requested vs. consumed resources of a user
  - Send a weekly report with possible savings and $CO_2$ footprint

# AUDITOR

## Contributors and Documentation

- Extensive documentation



- Auditor
  - Features
- Running Auditor
  - Migrating the database
  - Using Docker
  - Configuration files
  - Metrics exporter for Prometheus
- Compiling Auditor
- Packages
- Collectors
  - SLURM Collector
  - SLURM Epilog Collector
  - HTCondor Collector
- Plugins
  - APEL Plugin
  - Priority Plugin
- Auditor Clients
- License
  - Contribution

### Overview

#### Auditor

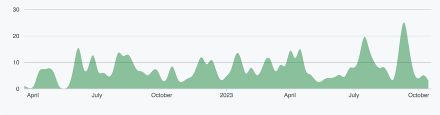
Auditor stands for **Accounting Data Handling Toolbox For Opportunistic Resources**. Auditor ingests accounting data provided by so-called *collectors*, stores it and provides it to the outside to so-called *plugins*.

It comes with a well-defined REST API which allows for the implementation of application-specific collectors and plugins. This makes it well suited for a wide range of use cases.

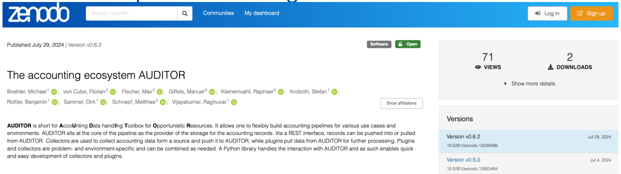
Overview of the AUDITOR ecosystem. AUDITOR accepts records from collectors, stores them in a PostgreSQL database and o ers these records to plugins which take an action based on the records.

https://doi.org/10.21203/rs.3.rs-4741479/v1

- 8 contributors
- from 3 universities
  - Freiburg (main developement), KIT, Uni Wuppertal
- 15 releases - latest v0.6.2
- Continuous improvements: Commits



- Also on https://zenodo.org/records/13239266



Published July 29, 2024 | Version v0.6.2

### The accounting ecosystem AUDITOR

Boehler, Michael[1]; von Cube, Florian[2]; Fischer, Max[2]; Giffels, Manuel[2]; Kleinemuehl, Raphael[3]; Kroboth, Stefan[1]; Rottler, Benjamin[1]; Sammel, Dirk[1]; Schnepf, Matthias[2]; Vijayakumar, Raghuvar[1]

**AUDITOR** is short for **A**cco**U**nting **D**ata hand**I**ing **T**oolbox for **O**pportunistic **R**esources. It allows one to flexibly build accounting pipelines for various use cases and environments. AUDITOR sits at the core of the pipeline as the provider of the storage for the accounting records. Via a REST interface, records can be pushed into or pulled from AUDITOR. *Collectors* are used to collect accounting data form a source and push it to AUDITOR, while *plugins* pull data from AUDITOR for further processing. Plugins and collectors are problem- and environment-specific and can be combined as needed. A Python library handles the interaction with AUDITOR and as such enables quick and easy development of collectors and plugins.

71 VIEWS

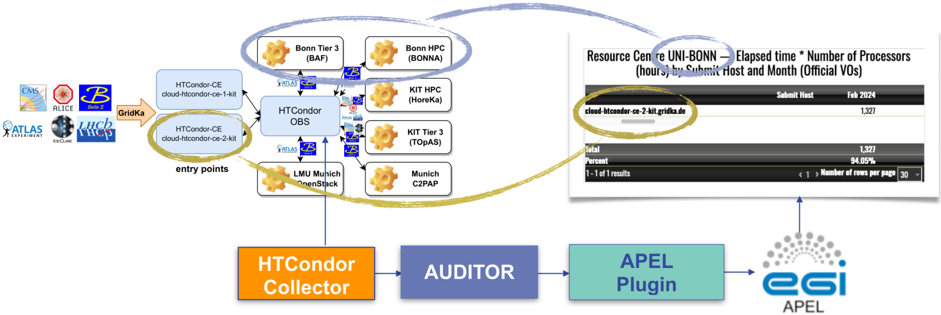
2 DOWNLOADS

Versions

| | |
|---|---|
| Version v0.6.2<br>10.5281/zenodo.13239266 | Jul 29, 2024 |
| Version v0.5.0<br>10.5281/zenodo.12653484 | Jul 4, 2024 |

# WLCG Accounting Use Case

First working prototype



Resource Centre UNI-BONN — Elapsed time * Number of Processors (hours) by Submit Host and Month (Official VOs)

| Submit Host | Feb 2024 |
|---|---|
| cloud-htcondor-ce-2-kit.gridka.de | 1,327 |
| Total | 1,327 |
| Percent | 94.05% |

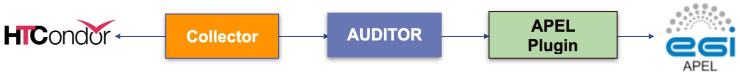
1 - 1 of 1 results ‹ 1 › Number of rows per page 30

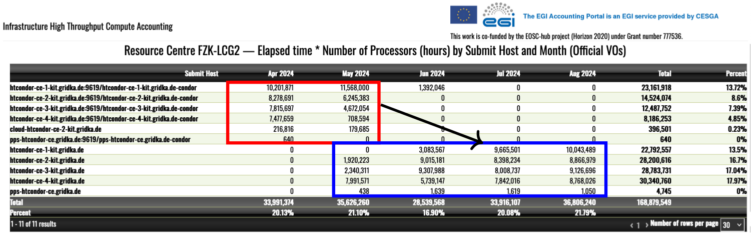
- Grid infrastructure hosted and maintained in Karlsruhe, resources provided by Bonn
- **AUDITOR** accounting pipeline allows to account for sub-clusters individually

# AUDITOR Accounting

In Production for German ATLAS T1

- KIT replaced accounting of the APEL client by AUDITOR pipeline in May 2024





The EGI Accounting Portal is an EGI service provided by CESGA

Infrastructure High Throughput Compute Accounting

This work is co-funded by the EOSC-hub project (Horizon 2020) under Grant number 777536.

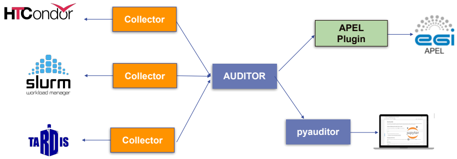
Resource Centre FZK-LCG2 — Elapsed time * Number of Processors (hours) by Submit Host and Month (Official VOs)

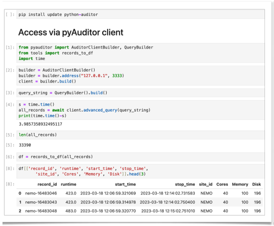
| Submit Host | Apr 2024 | May 2024 | Jun 2024 | Jul 2024 | Aug 2024 | Total | Percent |
|---|---|---|---|---|---|---|---|
| htcondor-ce-1-kit.gridka.de:9619/htcondor-ce-1-kit.gridka.de-condor | 10,201,871 | 11,568,000 | 1,392,046 | 0 | 0 | 23,161,918 | 13.72% |
| htcondor-ce-2-kit.gridka.de:9619/htcondor-ce-2-kit.gridka.de-condor | 8,278,691 | 6,245,383 | 0 | 0 | 0 | 14,524,074 | 8.6% |
| htcondor-ce-3-kit.gridka.de:9619/htcondor-ce-3-kit.gridka.de-condor | 7,815,697 | 4,672,054 | 0 | 0 | 0 | 12,487,752 | 7.39% |
| htcondor-ce-4-kit.gridka.de:9619/htcondor-ce-4-kit.gridka.de-condor | 7,477,659 | 708,594 | 0 | 0 | 0 | 8,186,253 | 4.85% |
| cloud-htcondor-ce-2-kit.gridka.de | 216,816 | 179,685 | 0 | 0 | 0 | 396,501 | 0.23% |
| pps-htcondor-ce.gridka.de:9619/pps-htcondor-ce.gridka.de-condor | 640 | 0 | 0 | 0 | 0 | 640 | 0% |
| htcondor-ce-1-kit.gridka.de | 0 | 0 | 3,083,567 | 9,665,501 | 10,043,489 | 22,792,557 | 13.5% |
| htcondor-ce-2-kit.gridka.de | 0 | 1,920,223 | 9,015,181 | 8,398,234 | 8,866,979 | 28,200,616 | 16.7% |
| htcondor-ce-3-kit.gridka.de | 0 | 2,340,311 | 9,307,988 | 8,008,737 | 9,126,696 | 28,783,731 | 17.04% |
| htcondor-ce-4-kit.gridka.de | 0 | 7,991,571 | 5,739,147 | 7,842,016 | 8,768,026 | 30,340,760 | 17.97% |
| pps-htcondor-ce.gridka.de | 0 | 438 | 1,639 | 1,619 | 1,050 | 4,745 | 0% |
| Total | 33,991,374 | 35,626,260 | 28,539,568 | 33,916,107 | 36,806,240 | 168,879,549 | |
| Percent | 20.13% | 21.10% | 16.90% | 20.08% | 21.79% | | |

1 - 11 of 11 results

- AUDITOR is able to provide the accounting of the DE-Tier 1
  - one of the largest WLCG Tier-1s

# Collecting Accounting Info with AUDITOR





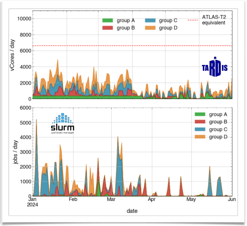
```
[ ]: pip install update python-auditor
```

### Access via pyAuditor client

```
[1]: from pyauditor import AuditorClientBuilder, QueryBuilder
     from tools import records_to_df
     import time

[2]: builder = AuditorClientBuilder()
     builder = builder.address("127.0.0.1", 3333)
     client = builder.build()

[3]: query_string = QueryBuilder().build()

[4]: s = time.time()
     all_records = await client.advanced_query(query_string)
     print(time.time()-s)

     3.9857358932495117

[5]: len(all_records)

[5]: 33390

[6]: df = records_to_df(all_records)

[8]: df[['record_id', 'runtime', 'start_time', 'stop_time',
         'site_id', 'Cores', 'Memory', 'Disk']].head(3)
```

[8]:

| | record_id | runtime | start_time | stop_time | site_id | Cores | Memory | Disk |
|---|---|---|---|---|---|---|---|---|
| 0 | nemo-16483046 | 423.0 | 2023-03-18 12:06:59.321069 | 2023-03-18 12:14:02.731583 | NEMO | 40 | 100 | 196 |
| 1 | nemo-16483043 | 423.0 | 2023-03-18 12:06:59.314978 | 2023-03-18 12:14:02.750400 | NEMO | 40 | 100 | 196 |
| 2 | nemo-16483048 | 483.0 | 2023-03-18 12:06:59.320770 | 2023-03-18 12:15:02.751010 | NEMO | 40 | 100 | 196 |



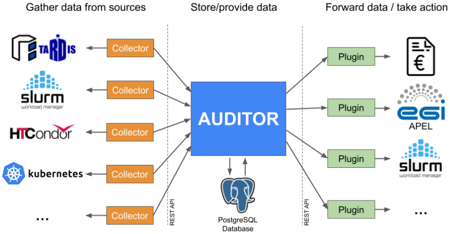
- ▶ accounting data can be collected in one or more AUDITOR instances from multiple sources
- ▶ APEL plugin can report for one or more queues
- ▶ pyauditor allows to integrate AUDITOR client into python env

# Conclusion

AUDITOR

- Provides an accounting ecosystem for various use cases
- Allows to account for different resources shared by one overlay batch system
- Allows to collect accounting data from multiple sources
- provision via containers independent of the OS

- Flexible structure of records and ecosystem allows to quickly adapt to future use cases



# References

AUDITOR

Website: https://alu-schumacher.github.io/AUDITOR/

GitHub: https://github.com/ALU-Schumacher/AUDITOR/

FIDIUM: https://fidium.erumdatahub.de

Email: auditor@physik.uni-freiburg.de

Michael Boehler

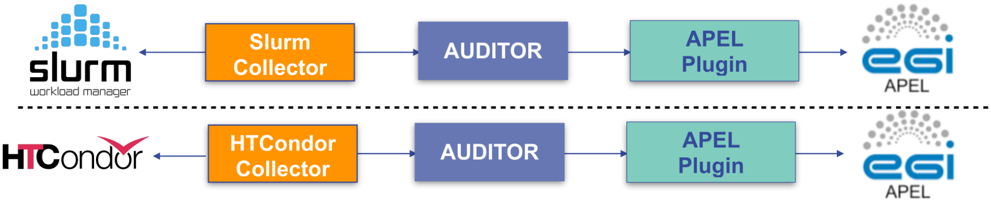
Albert-Ludwigs-Universität Freiburg

michael.boehler@physik.uni-freiburg.de

# Back-Up...

# WLCG Accounting Use Case

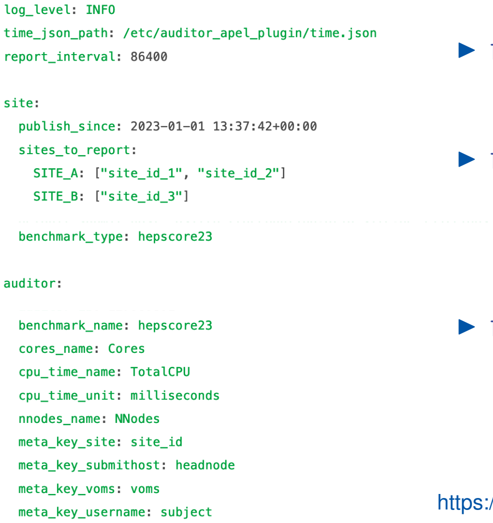
APEL Accounting with AUDITOR



- Collect accounting data from SLURM or HTCondor
- Store data as records in AUDITOR DB
- APEL plugin retrieves records from AUDITOR
  - creates APEL job summary from records
  - sends summary to defined APEL server
- Sites planing to use AUDITOR for accounting:
  - DESY-HH, Uni Wuppertal, ... ← ATLAS DE T1 (GridKa) moved reporting to AUDITOR

# APEL Plugin

## Configuration



```
log_level: INFO
time_json_path: /etc/auditor_apel_plugin/time.json
report_interval: 86400

site:
  publish_since: 2023-01-01 13:37:42+00:00
  sites_to_report:
    SITE_A: ["site_id_1", "site_id_2"]
    SITE_B: ["site_id_3"]

  benchmark_type: hepscore23

auditor:

  benchmark_name: hepscore23
  cores_name: Cores
  cpu_time_name: TotalCPU
  cpu_time_unit: milliseconds
  nnodes_name: NNodes
  meta_key_site: site_id
  meta_key_submithost: headnode
  meta_key_voms: voms
  meta_key_username: subject
```

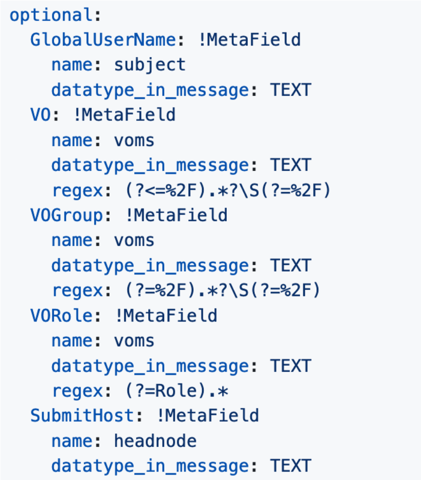
- `block 1`: configure serivce
  - file to store current state
  - time in seconds between reports
- `block 2`: configure site(s) to be reported
  - sites_to_report:
    keys: names of the sites in the GOCDB,
    values: corresponding site names in AUDITOR records
- `block 3`: configure metrics to be reported
  - meta_key_voms:
    key in meta field to be used as voms

https://alu-schumacher.github.io/AUDITOR//v0.5.0/#apel-plugin

universität freiburg

# APEL Plugin

Configuration in Release v0.6.2

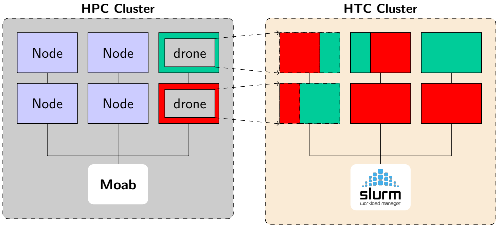


```
optional:
  GlobalUserName: !MetaField
    name: subject
    datatype_in_message: TEXT
  VO: !MetaField
    name: voms
    datatype_in_message: TEXT
    regex: (?<=%2F).*?\S(?=%2F)
  VOGroup: !MetaField
    name: voms
    datatype_in_message: TEXT
    regex: (?=%2F).*?\S(?=%2F)
  VORole: !MetaField
    name: voms
    datatype_in_message: TEXT
    regex: (?=Role).*
  SubmitHost: !MetaField
    name: headnode
    datatype_in_message: TEXT
```

- dynamic mapping of any MetaField via regex
  - this allows to report accounting data for different VOs submitted with **tokens**
- plugin configuration a bit more complicated, but much more flexible

# Priority Use Case

Adapting priorities on HTC cluster based on provided HPC resources

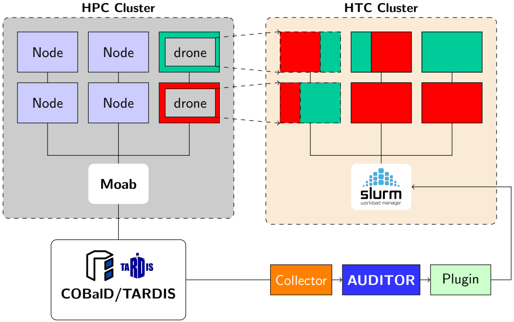


- HPC resources integrated with COBalD/TARDIS
- Several HEP groups provide HPC resources
- Resources shared among HEP groups
- How to guarantee fair share on HTC cluster?

- **TARDIS collector** retrieves info of provided resources on the NEMO cluster
- **AUDITOR** accounts for provided resources of individual groups [A and B]
- **Priority plugin** adjusts priorities on HTC cluster

# Priority Use Case

Adapting priorities on HTC cluster based on provided HPC resources

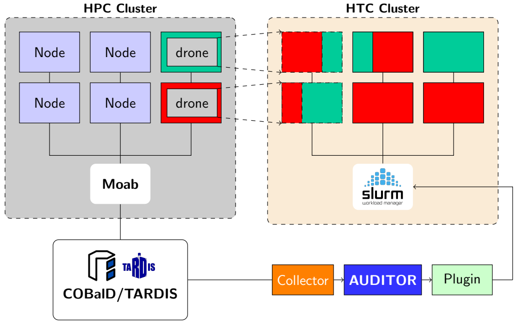


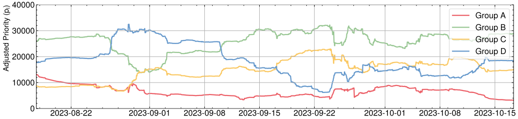
- HPC resources integrated with COBalD/TARDIS
- Several HEP groups provide HPC resources
- Resources shared among HEP groups
- How to guarantee fair share on HTC cluster?

- **TARDIS collector** retrieves info of provided resources on the NEMO cluster
- **AUDITOR** accounts for provided resources of individual groups [**A** and **B**]
- **Priority plugin** adjusts priorities on HTC cluster

# Priority Use Case

Adapting priorities on HTC cluster based on provided HPC resources



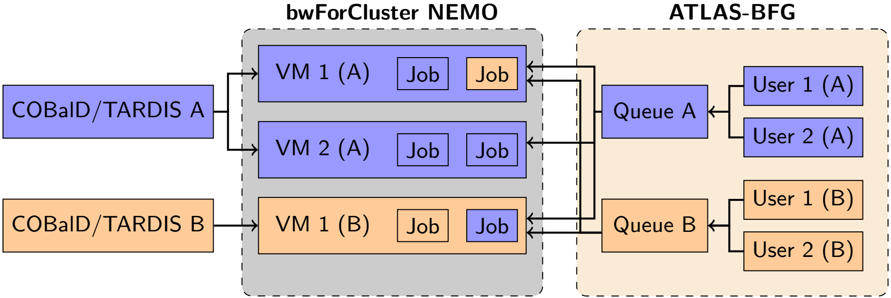


- HPC resources integrated with COBalD/TARDIS
- Several HEP groups provide HPC resources
- Resources shared among HEP groups
- How to guarantee fair share on HTC cluster?

- **TARDIS collector** retrieves info of provided resources on the NEMO cluster
- **AUDITOR** accounts for provided resources of individual groups [**A** and **B**]
- **Priority plugin** adjusts priorities on HTC cluster

# Priority Use Case

## Motivation



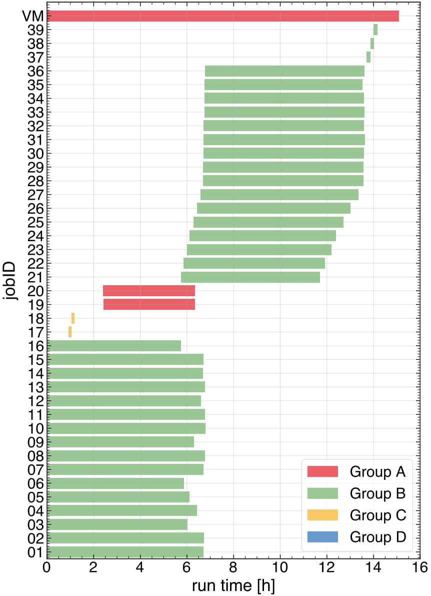
- Four local HEP research groups (A to D) with a share in NEMO
- Each served with its own COBaID/TARDIS instance
- Each has its own SLURM partition (job queue)
- Efficient use of resources due to sharing VMs across HEP groups

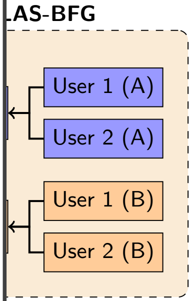
# Priority Use Case

Motivation

COBalD/TARDIS

COBalD/TARDIS





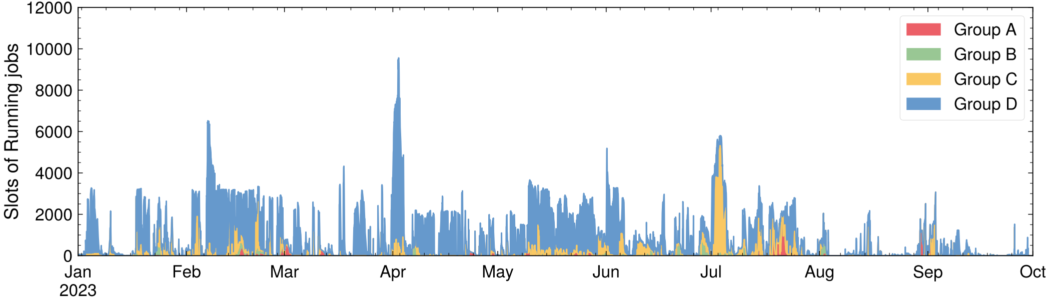
- Four local HEP rese
- Each served with its
- Each has its own SL
- Efficient use of resou

# Priority Use Case

Motivation

bwForCluster NEMO

ATLAS-BFG

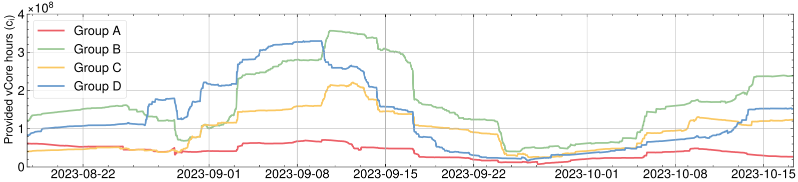


- Each has its own SLURM partition (job queue)
- Efficient use of resources due to sharing VMs across HEP groups

# Priority Use Case

Results from HEP groups @ University of Freiburg

- Provided resources of the four local HEP groups

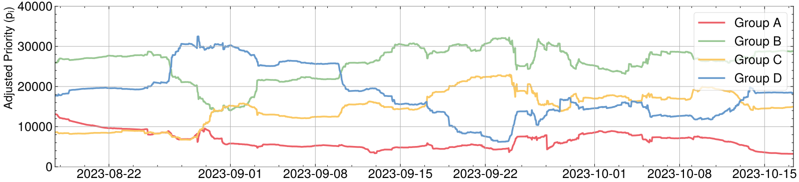


Integral of provided vCore hours over last 14 days for each group:

$$c_i = \int_{t_{\text{now}} - 14\,\text{d}}^{t_{\text{now}}} N_i(t)\mathrm{d}t$$

with $i \in A, B, C, D$

- Priority is adjusted according to the provided resources



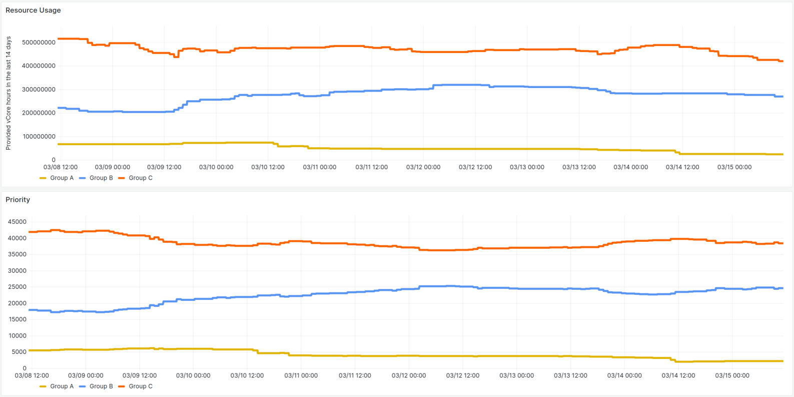
Priority $p_i$ is defined as:

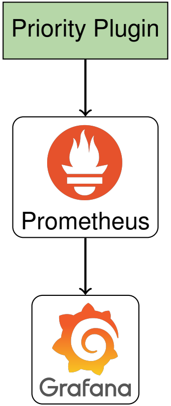
$$p_i = \frac{c_i}{\sum_j c_j}(p_{\text{max}} - p_{\text{min}}) + p_{\text{min}}$$

with $i, j \in A, B, C, D$;
$p_{\text{min}} = 1$; $p_{\text{max}} = 65535$

# Real-time monitoring

e.g. Priority Use Case, but also AUDITOR stats





- Resource usage and priority can be made available for Prometheus
- Real-time monitoring of priority adjustments (with e.g. Grafana)