ErUM-Data IDT

ERUM DATA IDT COLLABORATION MEETING 2020

STATUS AND PLANS – AREA A

Kilian Schwarz & Manuel Giffels

TOPICS OF AREA A

Development of technologies to enable utilization of heterogeneous computing resources

AP 1) Werkzeuge zur Einbindung • Scheduling von Cloud - Jobs • Container - Technologien • Checkpointing • Zugang zu Experiment-Datenbanken	**AP 2) Effiziente Nutzung** • Steigerung der Effizienz von daten-intensiven Anwendungen auf hetero-genen Ressourcen mittels „on the fly“ Datencaches
AP 3) Workflow Steuerung • Identifikation und Steuerung • In - Pilot Job Monitoring • Accounting • Optimierung durch data - mining	

PLANNED CONTRIBUTIONS TO AREA A

Aachen (friends):

- T2_DE_RWTH can be used for dynamic resource management tests [A1]

Bonn (associated partner):

- Dynamic resource management of T3 resources using (COBalD/TARDIS) [A1]
- Site in a box concept [A1] (no funding so far)

DESY (associated partner):

- Smart caching in WLCG data federation using dCache [A2]

Frankfurt/GSI:

- Singularity Containers to include HPC resources into Grid computing (e.g. ALICE T2@GSI) [A1]
- Developments of XRootD based coordinated distributed caching solutions [A2]

Freiburg:

- Dynamic resource management developments (COBalD/TARDIS) [A1]
- Contribute to the coordinated distributed caching solutions [A2]
- Development of monitoring, accounting tools and benchmarks [A3]

Karlsruhe:

- Development of the opportunistic resource manager COBalD/TARDIS [A1]
- Workflow management in heterogenous environments [A1]
- Development of a coordinated distributed caching solution using XRootD [A2]

München:

- Job log files analysis by using ML (anomaly detection) [A3]
- Development of XRootD based disk caching (XCache) [A2]

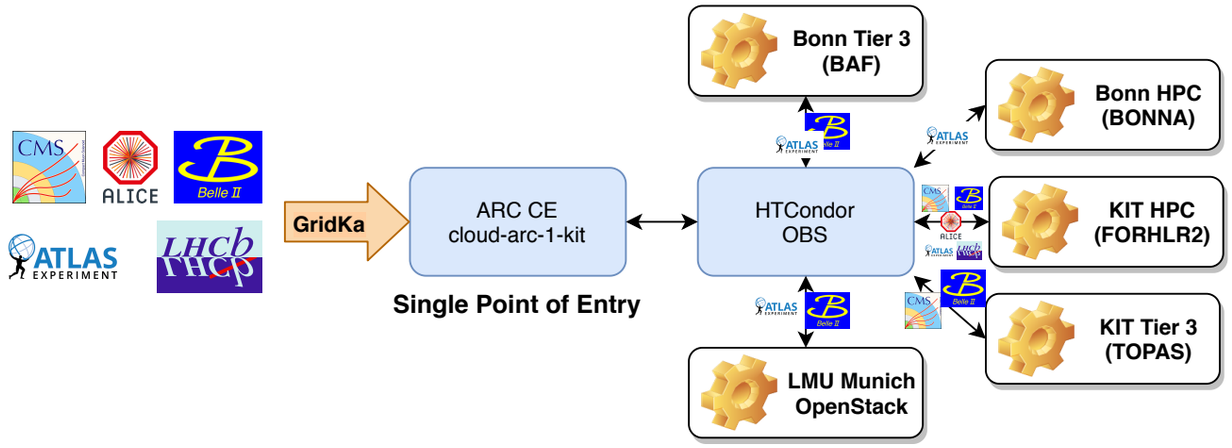
Wuppertal:

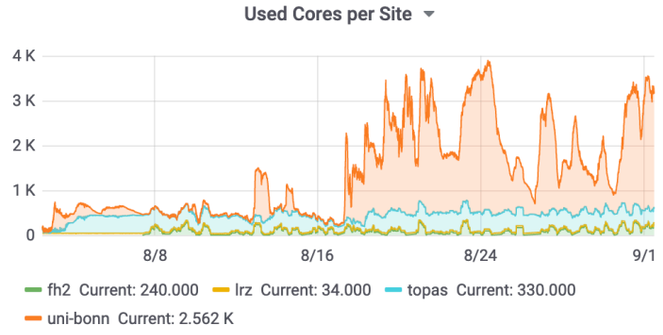
- Containerization of user jobs and services (VOMS, DB access, monitoring) focussing on lightweight solutions [A1, A3]

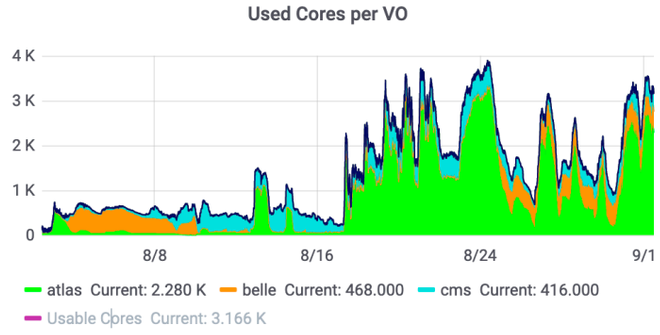
STATUS AND PLANS

STATUS & PLANS KARLSRUHE (A1)

Dynamic and Transparent Integration of Opportunistic Resources

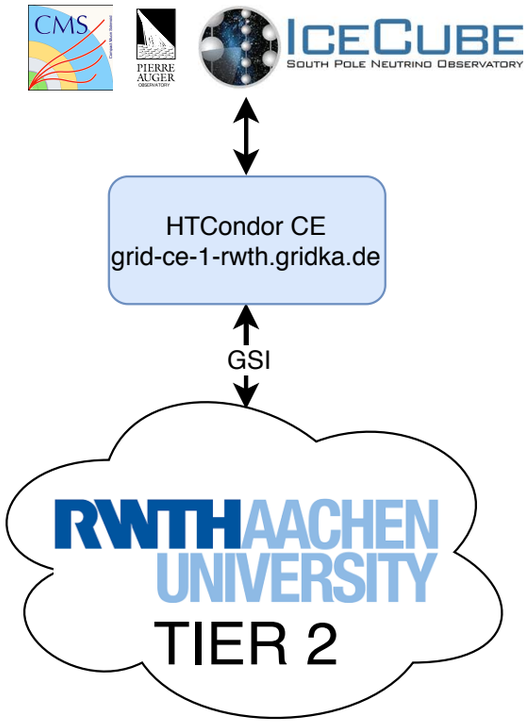






- Built a prototype of a federated infrastructure (smooth operation since more than a year)
  - Dynamic, on demand provisioning and transparent integration of heterogeneous resources with COBaID/TARDIS
  - Single point of entry to plethora of resources
  - Opportunistic utilization of clouds, HPCs and T3s
- Ready to add more parties from B2 (let us know)

Lightweight Site Operations



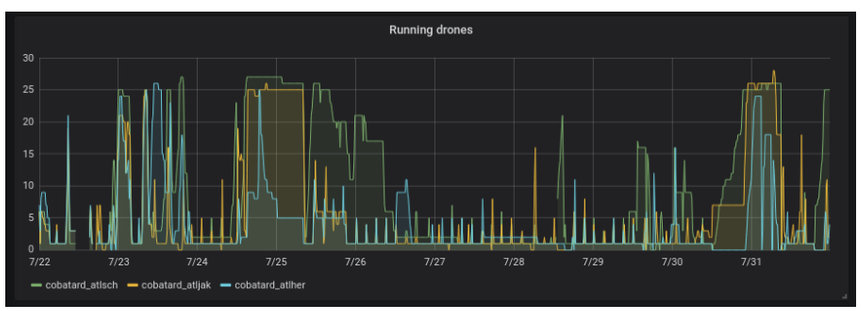
- Remote CEs at larger Grid sites allow for lightweight T2/T3 operations
- Reduces effort to contribute compute power
- Proof-of-concept: RWTH-CE@GridKa
- In stable operation since this summer

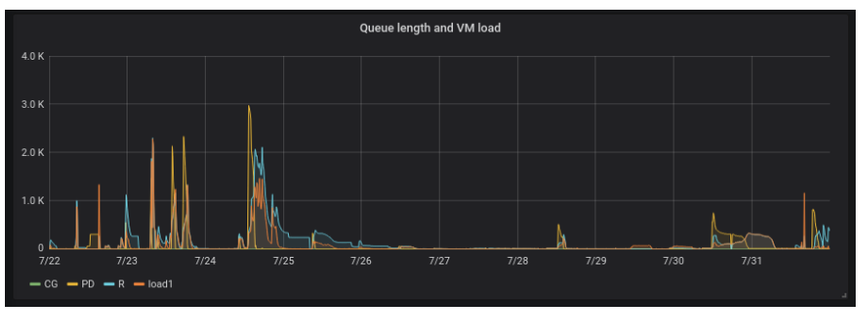
Status report Freiburg

Opportunistic Resources - scheduling cloud-jobs

- COBaID/TARDIS batch system adapter for Slurm working and in production at Uni Freiburg
- Extended the monitoring capabilities of COBaID/TARDIS:
  - Prometheus Monitoring Plugin: Number of booting/running/deleted drones reported to Prometheus
  - Elasticsearch Monitoring Plugin: Every state change of a drone is pushed to Elasticsearch
  - Allows for continuous monitoring and retrospective analysis of performance and problems
- Next steps
  - Autonomous and continuous health monitoring and problem handling based on gathered data
  - Performance analysis of entire setup and parameter tuning

Thanks to Manuel Giffels, Max Fischer, Matthias Schnepf and Eileen Kuehn!





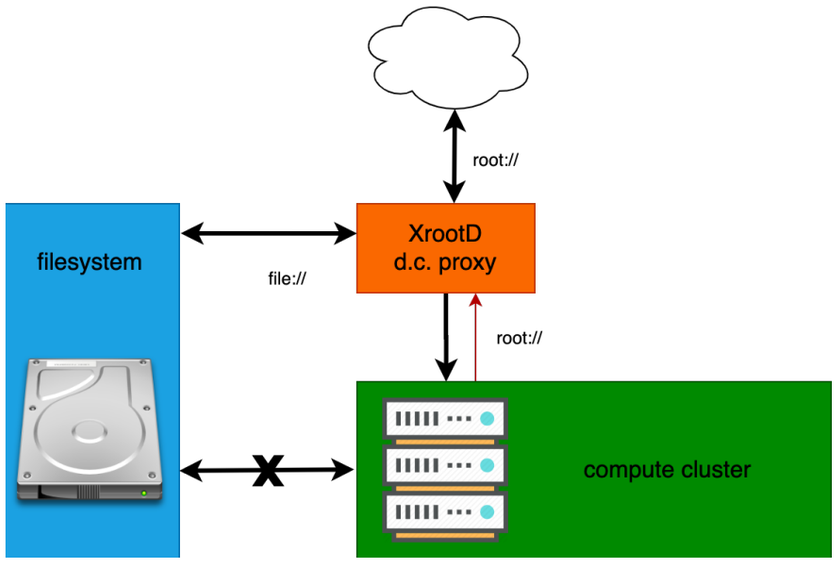
STATUS & PLANS WUPPERTAL

- **Our work with standalone containers was seminal to the containerization of production workflows**. ATLAS is developing an automatic pipeline for producing standalone images to run detector simulation at HPCs with restricted network connectivity. They plan to extend it to MC reconstruction
- ATLAS and CMS are currently attempting to converge on a similar data model for their detector conditions databases. This opens the possibility for developing common tools for handling conditions
- **Log analysis framework for generic user jobs**. Drawing on the example provided by the ATLAS job validation and reporting software, we are developing a log analysis framework for deployment in containers
- **The core components consists of:**
  - A message processing unit (structuring of log lines by matching regular expressions)
  - A storage unit (an optional search-engine index)
  - A software layer for configuration and report generation
- The framework can be easily deployed in pre-existing containers where payload execution generates log files
- From a review of open-source projects, *Fluentd* was chosen as the data-collecting component, and *elasticsearch* as the storage indexing component
- As a first use case, a prototype tuned to analyze ATLAS reconstruction job logs was developed. Data is structured according to: *service, level* and *message* fields
- The immediate goal is to submit a container to the Grid and generate a JSON formatted job report from the payload's log

STATUS FRANKFURT/GSI

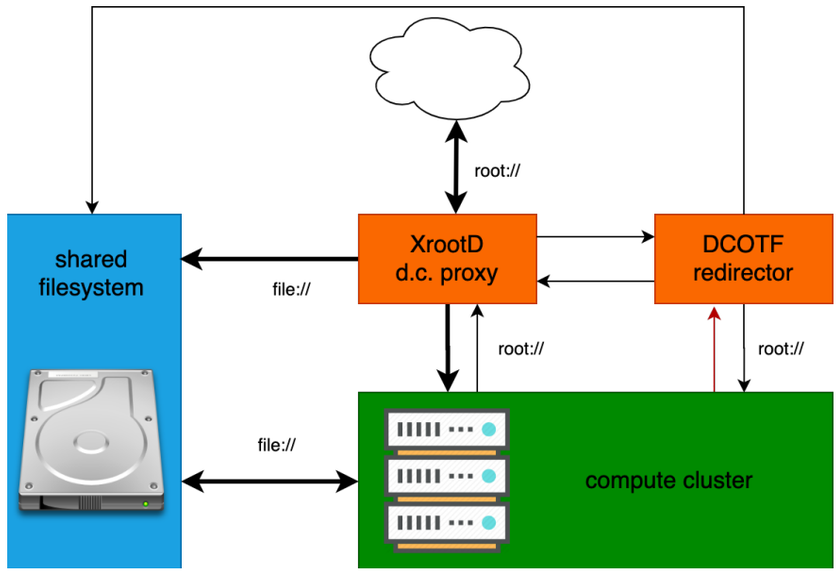
Evaluation of Different Approaches:

XCache



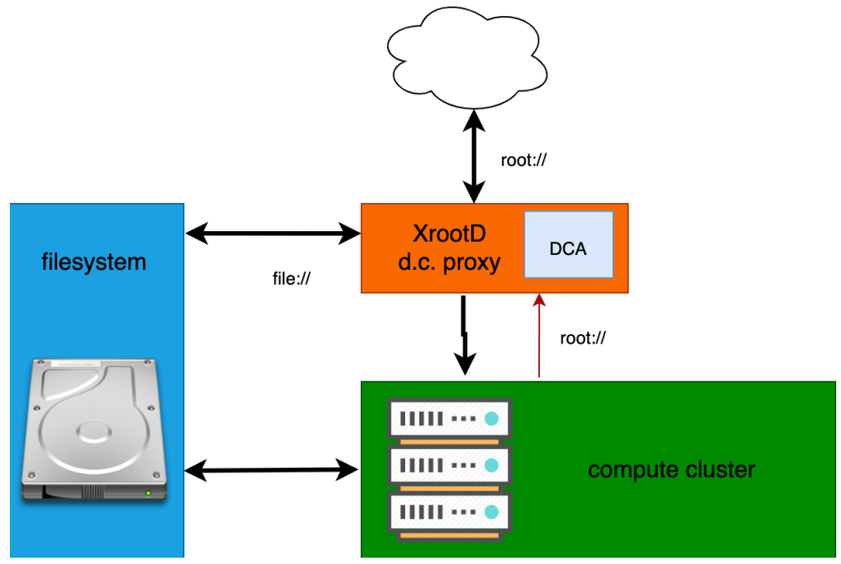
- Traditional setup

Disk-caching on the fly (DCOTF)



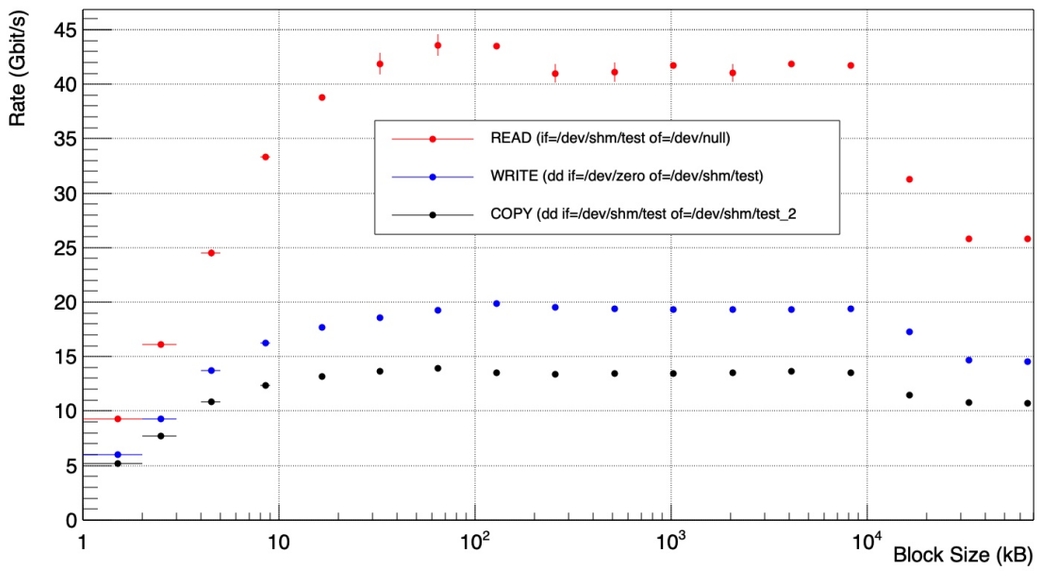
- Additional DCOTF redirector
- plus local redirect plugin

XCache direct cache access (DCA)



- No special redirector required
- Add access record (avoid purging)

RAM Disk Benchmarks



Platform	Copy Rate (Gbit/s)
DCOTF (v4.12.x)	16.16
XCache (v4.12.x)	13.22
XCache dca (v4.12.x)	16.18
XCache (v5.0-rc)	13.22
XCache dca (v5.0-rc)	13.29

50x 30GB files

All Frankfurt/GSI XRootD developments are also upstream contributions to XRootD!

PLANS FRANKFURT/GSI

- Setting up a new server to be added to the Goethe-HLR in Frankfurt with root privileges
  - 120Gbit/s private link to GSI (to be reestablished)
  - 100Gbit/s link to the cluster filesystem (unshared by other users)
  - Few TBs of NVMe SSD as a local filesystem in case the cluster filesystem is slow. (The cluster filesystem is always shared)
- Improvement and debugging of XrootD configuration files with the feedback from performance tests
- Development on the local redirect plugin to make use of XCache features, i.e.:
  - Caching blocks of files instead of complete files -not implemsented in the local redirect plugin
  - Status of data in cache (parsing .cinfo files) -not implemented in the local redirect plugin
- Submission of local redirect plugin to XrootD core package after debugging and aforementioned improvements
- Investigation of performance differences of DCOTF vs. XCache vs. XCache Direct Cache Access

Status report Freiburg

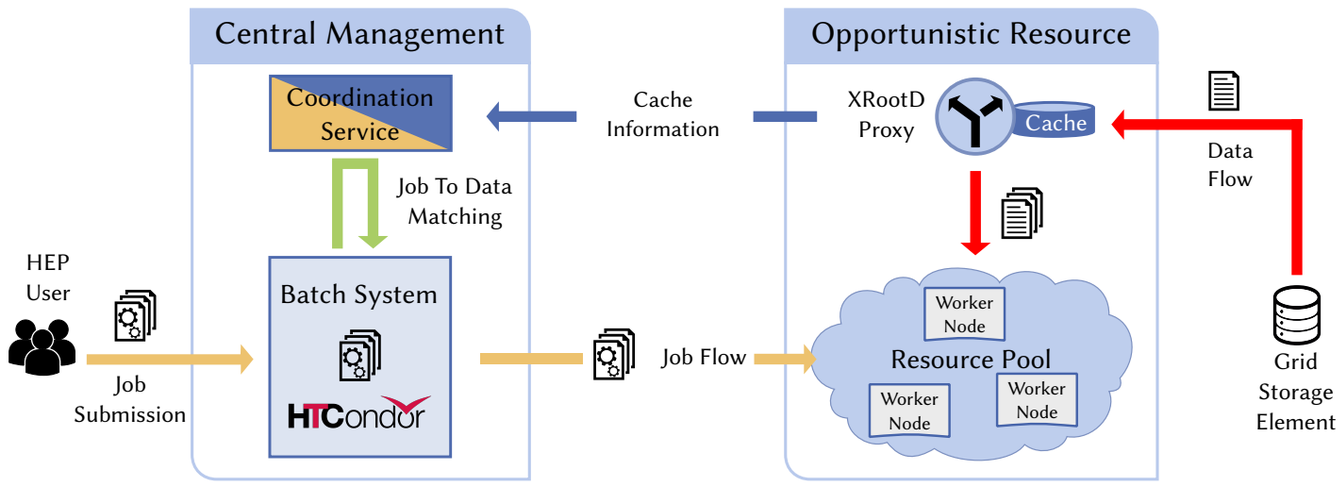
Opportunistic Resources - XRootD-Forward-Proxy Testing

- successfully tested sandbox provided by Dr. Kilian Schwarz et. al. (GSI) link: GitLab
  - nice test setup to get familiar with the infrastructure - works out of the box
- developed automatic provisioning of all needed components in Freiburg QA-environment (client, data manager, and forward proxy)
  - allows testing in environment close to production setup
  - could help to spot issue concerning file permission of cached files (currently: work in progress) - Thanks to Serhat Atay for his support

Next Steps:

- provide functional tests for needed components
- benchmark caching vs direct access, and caching one different file systems

STATUS & PLANS KARLSRUHE (A2)



- Prototyp implementation of the NaviX available
  - Utilizing commonly used tools XRootD and HTCondor
  - GSI plugins have been incorporated
  - Coordinates placement of data on distributed caches
  - Match jobs to computing resources taking into account data locality
  - Achievement: Increased CPU efficiency and data throughput
- Determining robust caching and coordination strategies turned out to be very complex
- Study different solutions with extensive simulations first
- Suitable simulator is in development at KIT
- Refine existing prototype to scalable caching solution

STATUS & PLANS LMU MUNICH

Xcache at LMU Munich

- Last year: Extensive tests in ATLAS production environment
  → read from nearby MPPMU storage via Xcache
  → provided valuable experience with and tests of the system
- Started to look into applications for analysis workflows
  - Test of "Virtual placement" system at ATLAS that directs jobs for the same datasets to the same Xcache sites
    → Can improve hit rates significantly
  - Tests currently paused since bug fixes for ROOT (TChain) take time to propagate into analysis releases
    (starting to test with workarounds)
  - Starting to test Xcache in the context of columnar data analysis
    → much higher I/O rates than for classical workflows
    → could be significantly improved with columnar storage formats
    (parquet, ROOT with large baskets)

Disk access pattern with default ROOT files:

Disk access pattern for columnar storage (parquet):

MORE DETAILS IN TOMORROWS PARALLEL SESSION

09:00 → 17:00 **Parallel session A+B**

Zoom link

09:00 **Log analysis and reporting framework for user jobs in containers** ¶ 20m

Speakers: Marcelo Vogel (Bergische Universität Wuppertal), Marcelo Vogel (Bergische Universität Wuppertal)

09:20 **Performance monitoring of opportunistic resources at ATLAS-BFG** 20m

Speaker: Stefan Kroboth (Albert Ludwigs Universität Freiburg)

09:40 **Testing technology and performance of XRootD-Forward-Proxy** 20m

Speaker: Dirk Sammel (University of Freiburg)

10:00 **Xcache for analysis workflows** 20m

Speaker: Nikolai Hartmann

10:20 **Opportunistic Computing and Lightweight Grid Operations** 20m

Speaker: Ralf Florian von Cube (KIT)

10:40 **Disk-Caching-On-The-Fly vs. XCache** 20m

Speaker: Serhat Atay (Goethe University Frankfurt)

SUMMARY

WP1 (Tools to integrate opportunistic resources):

- Built a prototype of a federated compute infrastructure
- Smooth operation since more than a year
- Software ready to add more interested parties from (B2)

WP2 (Efficient Utilization of Resources):

- Caching prototypes have been set-up using different approaches
- Developments and evaluations are ongoing

WP3 (Steering of Workflows/Monitoring):

- Prometheus, ES and Telegraf monitoring plugins for COBalD/TARDIS are available
- Prototype of log analysis framework is available
  - More work to be done on experiment overarching aspects

QUESTIONS?