REDWOOD Track 3: Monitoring

Tania Korchuganova (Pitt)

02 Oct 2024 REDWOOD monthly meeting

Historic data access and extraction

Done:

- Organising access and extraction of historic distributed workload management (ATLAS PanDA) and distributed data management (ATLAS Rucio) systems metadata for analysing, solving optimization problems, and building models in Track 1&4 which includes:
  - Computational tasks (workflow): input/output datasets, time of submission/start/end etc
  - Computational jobs (workload): time of submission/start/end, success/failure, resource utilization metrics (walltime/CPU/IO), computing resource (Grid site/HPC/Cloud) etc
  - Input dataset size and replica locations
- Extracted a tiny sample of metadata (15 tasks, ~10k jobs) as an introduction to the available metrics
- Created initial scripts and a guide how to extract all needed information
- 6 months of metadata extracted by the Track 4 team:
  - ~2.4M tasks, ~0.9M input datasets, ~185M jobs, ~140 computing sites
  - Extracted metadata itself is a few hundreds of GB

Job error classification

It is not easy to understand the root cause from the error message of jobs, it can be site / task configuration / network / user code / PanDA component issue, etc.

Implement interface in the BigPanDA monitoring system that allows experts and operation people to categorize the job errors -> training dataset for a model and can be used in the future for automatic actions (e.g. do not take into account site related errors in retry mechanism in PanDA)

6 components which report error code & diagnostic message PanDA:

Pilot
Exe
Transformation
DDM
Task Buffer
Job Dispatcher

ddm, 200: expected output panda.1002122753.994485.lib._41491449.40242259257.log.tgz is missing in pilot JSON **pilot, 1152:** File transfer timed out during stage-out: panda:panda.1002122753.994485.lib._41491449.40242259257.log.tgz to UNI-FREIBURG_SCRATCHDISK, copy command timed out: TimeoutException: Timeout reached, timeout=748 seconds')]:failed to transfer files using copytools=['rucio']

exe, 2002: payload execution failed with 220 **pilot, 1305:** payload execution failed with 220 **trans, 220:** Proot: An exception occurred in the user analysis code

pilot, 1201: Job killed by signal: SIGTERM

pilot, 9000: Starting job was killed because queue went offline

pilot, 9000: Unknown error

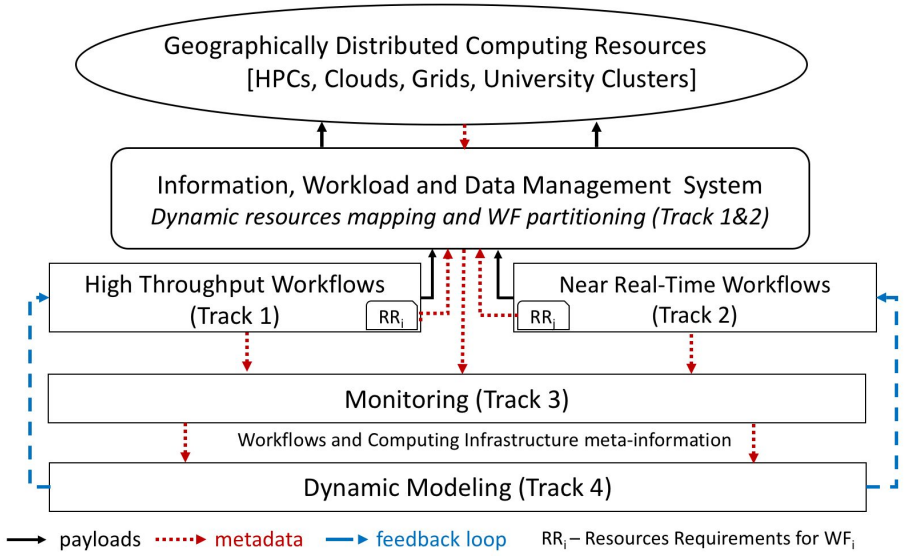
Plans

- Data placement and workload distribution monitoring
  - Currently the only available on a single task level - need to promote it to a global level
- Complex workflows monitoring:
  - Improve overall representation of the workflow progress and add more functionality for debugging problems
- Exploring how to consolidate the SimGrid output and the existing workload management system metrics in order to compare and validate
  - Will work together with Paul, Raees, Joe and Fred, who are working on creating on the system simulation with SimGrid

In the process of redesign of the BigPanDA monitoring system due to using outdated technology stack on front-end level and therefore nonoptimal implementation of the interactive and dynamically updating interfaces.

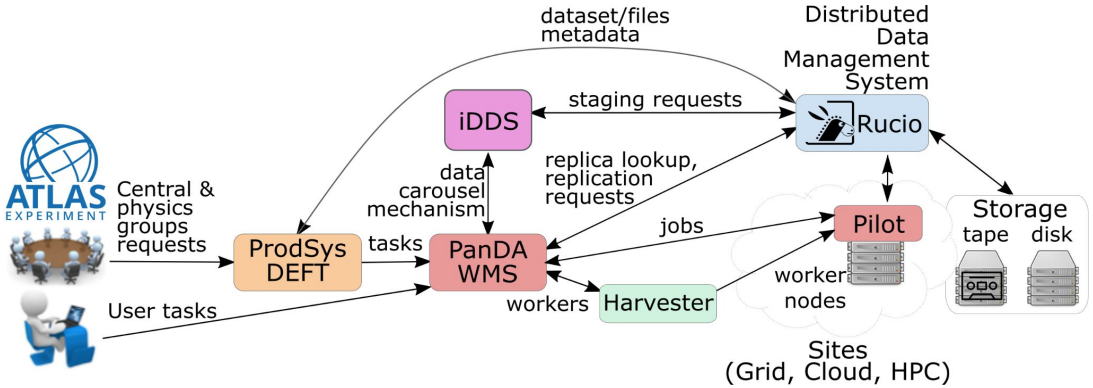
Back UP

Track 3: monitoring



"Developing a monitoring system to present coherent data placement and workload distribution to enable early failure detection"

Workflow Management System in ATLAS



- **DEFT**: Database Engine For Tasks
- **PanDA**: Production ANd Distributed Analysis System
- **Harvester**: resource-facing service between the PanDA and collection of pilots
- **Pilot**: the execution environment on a worker node
- **iDDS**: Intelligent Data Delivery System
- **Rucio**: Distributed Data Management System