Oar, A Versatile Resource and Job Management System: Overview and Ecosystem

Olivier Richard

[May 2021]

1/ DATAMOVE Team, LIG, INRIA, Univ. Grenoble Alpes

https://oar.imag.fr

https://github.com/oar-team/
Overview: OAR

- **RJMS**: Resource and Job Management System (aka Batch Scheduler)
- **Main features**: Same than Slurm, PBSpro, Torque/MAUI, LSF, GridEngine, Flux Framework...
- **Suitable for**:
  - **Production**: middle size HPC centers, tested dedicated to experimentation (Grid5000, Iot-Lab)
  - **Research**: topology aware scheduler, dynamic jobs, energy, HPC - Big Data convergence...
Design Principles

- **High level software components**
  - relational database engine (PostgreSQL)
  - SSH, Taktuk Parallel launcher for low level management operation (job launching and control).

- **High Modularity**
  - Central automaton and modules (scheduler, submission handling, jobs executions...)

- **Simple to customize**
  - < 50K Line Of Code (LOC) (lower code complexity than other systems)

- Evaluated up to 80K (emulated) resources same completion time/resource utilization level than Slurm
Research Topics and Needs

Research Topics

- Hybridation
  - HPC-BigData Convergence: Mixing HPC and BigData Workloads (PhD Michael Mercier)
  - Elastic Computing / Cloud Bursting
- Feedback loops to prevent IO bottleneck (PhD Quentin Guilloteau)
- Workload Analysis (PhD Salah Zrigui)

Needs:

- Simulation of infrastructure: *Batsim*
  - Resource consumptions, application behaviour
    - (PhDs Millian Poquet, Adrien Faure, Clément Mommessin)
- Experimentation: *Nixos-compose* (TBA)
  - Reproducibility of distributed systems
  - Variation, Tranposition (container, VM, Grid’5000)
  - Use of Functional Package Manager (NIX)
HPC-BigData Convergence: Mixing HPC and BigData Workloads

- Idle HPC resources used for BigData workload
  - HPC jobs have priority
  - BigData Framework: Spark/Yarn, HDFS
  - Evaluating costs of starting/ceasing tasks (Spark/Yarn) and data transfers (HDFS)
Mixing HPC and BigData Workloads: OAR + Spark/Yarn
- **CiGri**: Lightweight Grid Middleware
- *Bag-of-tasks* applications, use besteffort OAR’s jobs
- CiGri’s jobs add I/O pressure, control the number of CiGri’s jobs to submit to avoid I/O congestion
- Apply Control Theory/Principle
Batsim: an Infrastructure Simulator

- Separation between Batsim’s core simulator and schedulers/orchestrators
  - Protocol JSON compatible (Flatbuffer based in next major release)
- Apps/Workloads resource consumptions (CPU, Network, Energy, IO)
- Based on SimGrid
- Difficulties: apps/workload modelisation, platform characterization
Research Topics and Needs

Research Topics

- Hybridation
  - HPC-BigData Convergence: Mixing HPC and BigData Workloads (PhD Michael Mercier)
  - Elastic Computing / Cloud Bursting
- Feedback loops to prevent I/O bottleneck (PhD Quentin Guilloteau)
- Workload Analysis (PhD Salah Zrigui)

Needs:

- Simulation of infrastructure: Batsim
  - Resource consumptions, application behaviour
    - (PhDs Millian Poquet, Adrien Faure, Clément Mommessin)
- Experimentation: Nixos-compose (TBA)
  - Reproducibility of distributed systems
  - Variation, Tranposition (container, VM, Grid’5000)
  - Use of Functional Package Manager (NIX)