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

Olivier Richard [1]

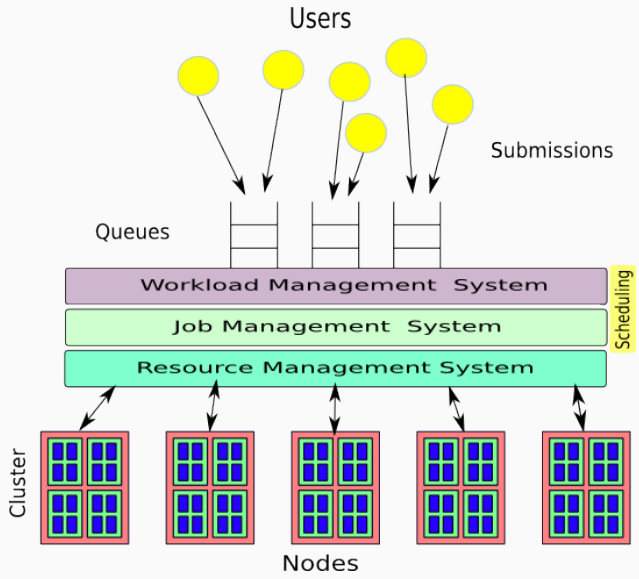
[May 2021]

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

https://oar.imag.fr

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

Overview: OAR



- RJMS: Resource and Job Mangement 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)
  - scripting language (OAR2: Perl, OAR3: Python) (execution engine, core modules, scheduler).
  - SSH, Taktuk Parallel launcher for low level managment operation (job launching and control)
- High Modularity
  - Central automaton and modules (scheduler, submission handling, jobs executions. . .)
- *Simple* to cutomize
  - $<$ 50K Line Of Code (LOC) (lower code complexity than other systems)
- Evaluated upto 80K (emulated) resources same completion time/resource utilization level than Slurm

Research Topics and Needs

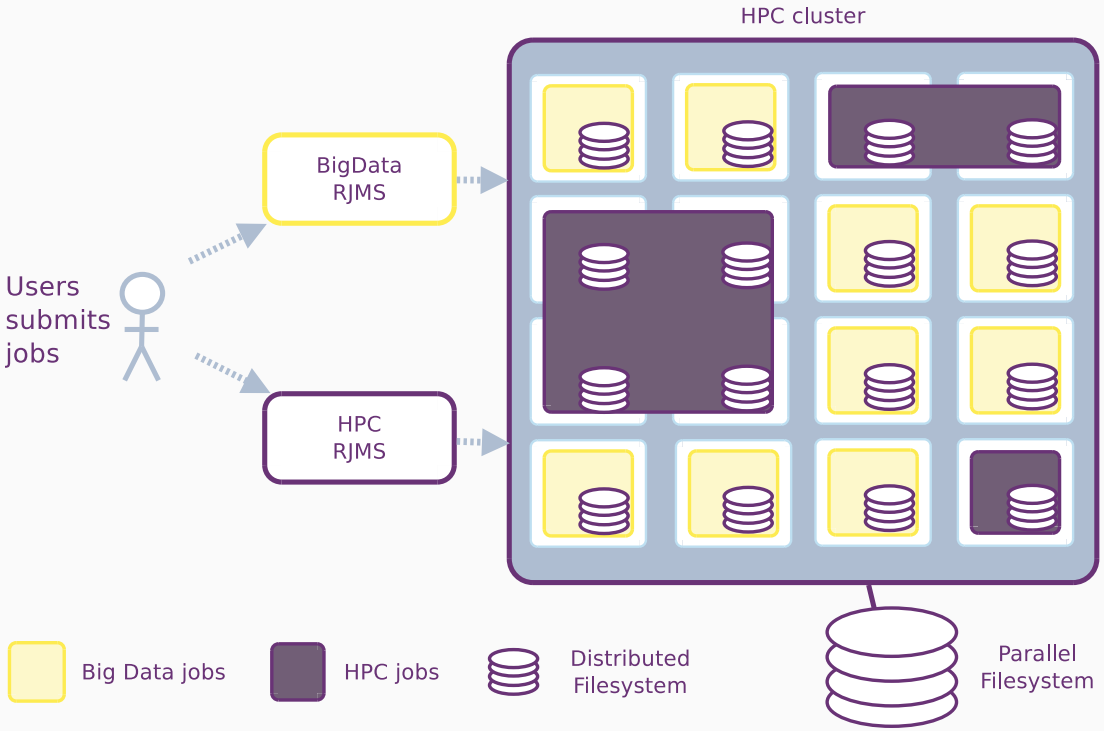
Research Topics

- Hybridation
  - HPC-BigData Convergence: Mixing HPC and BigData Workloads (PhD Michael Mercier)
  - Elastic Computing / Cloud Bursting
- Feedback loops to prevent IO bootleneck (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 distribituted systems
  - Variation, Tranposition (container, VM, Grid'5000)
  - Use of Functional Package Manager (NIX)

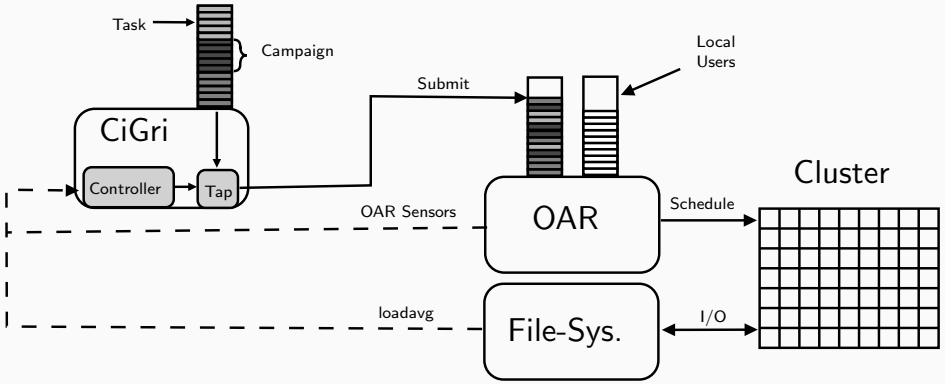
HPC-BigData Convergence: Mixing HPC and BigData Workloads

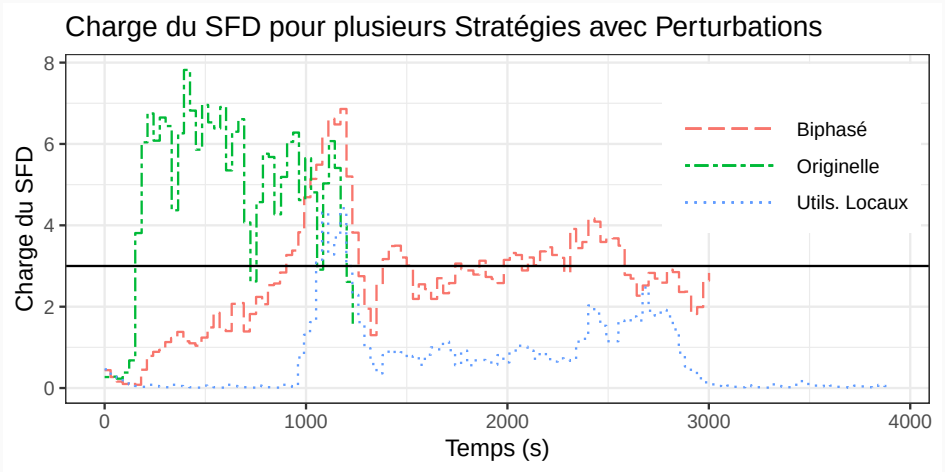


- Idle HPC resources used for BigData workload
  - HPC jobs have priority
  - BigData Framework: Spark/Yarn, HDFS
  - Evaluating costs of starting/stopping tasks (Spark/Yarn) and data transferts (HDFS)

Mixing HPC and BigData Workloads: OAR + Spark/Yarn

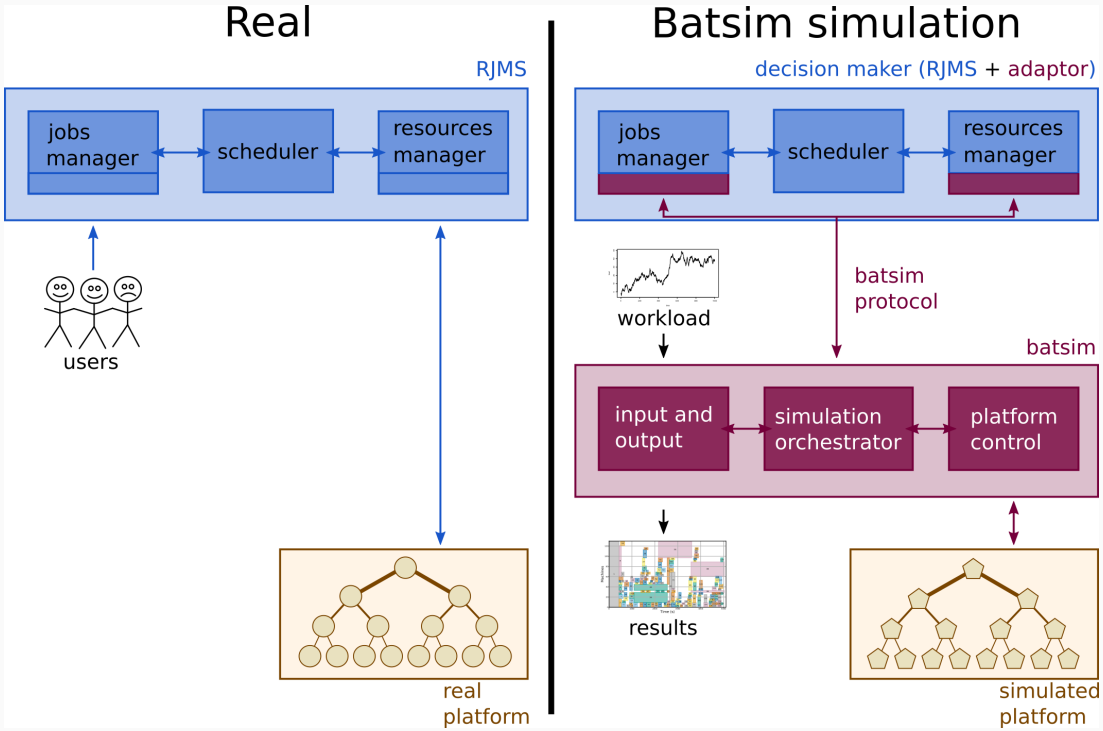
IO Congestion Avoidance with Feedback Loop





- 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 Insfrastructure Simulator



- Separation between Batsim's core simulator and schedulers/orchestrators
  - Protocole 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 bootleneck (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 distribituted systems
  - Variation, Tranposition (container, VM, Grid'5000)
  - Use of Functional Package Manager (NIX)