PCOG Team Meeting
Fall 2014

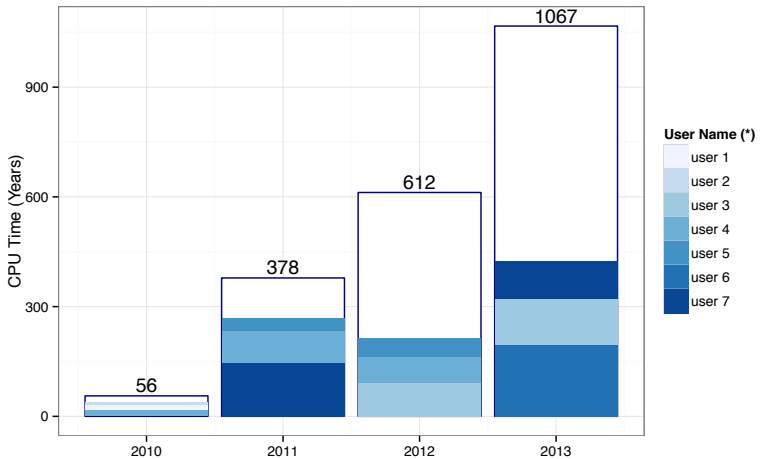
Joseph Emeras

SnT - CSC Research Unit,

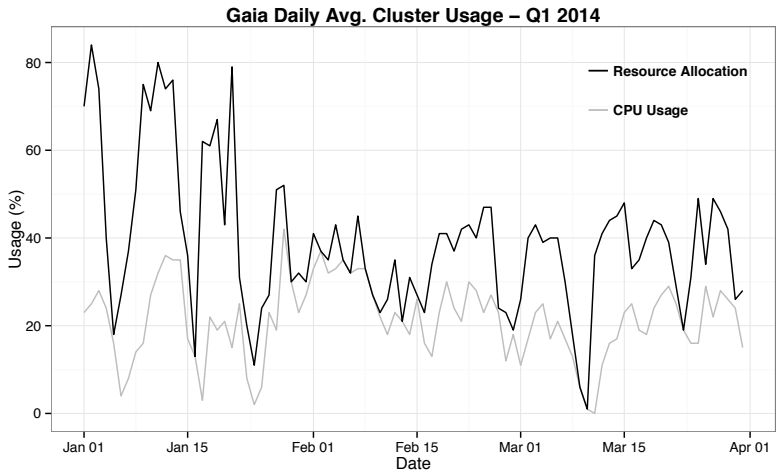
University of Luxembourg, Luxembourg

Summary

Observation of Gaia Cluster



Platform Utilization vs. Top Users Area.



Resource Allocation vs. CPU Usage.

- 151 nodes, 2000 cores, 200 users, 21.178 TFlops
- Period 2012 - 2013: 3M jobs, average throughput = 3 jobs/min.
- Few large users, many small users
- Varying CPU usage vs request

UNIVERSITÉ DU LUXEMBOURG

Do users have identifiable usage patterns?

Workflow

1. Data Collection (Colmet)
   - Data Aggregation and Discretization (R parallel - HDF5)
2. Unsupervised Learning – Characterization of jobs consumption
   - Data clustering: expert knowledge based
3. Supervised Learning (SVM - Multiclass - poly kernel - 10-fold X validation)
   - Performance Evaluation (10-fold X evaluation on Accuracy - AUC - Cohen's Kappa)

Data Collection

Colmet

- OAR Team product
- taskstats, cgroups, HDF5: lightweight and scalable
- collects info about jobs CPU, Mem, disk IO

Data

- 3 months trace, raw compressed size: 10GB
- 84 active users, 51859 jobs
- 6.05×10^9 metrics

Outcomes

Jobs Classification

User behavior tagging

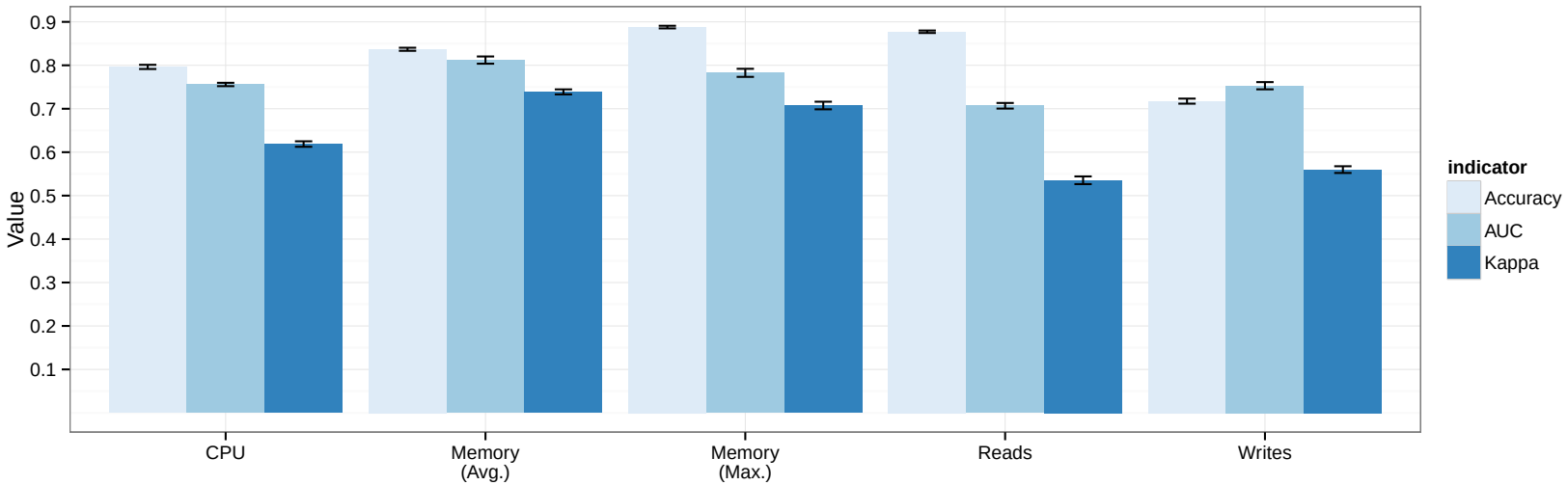
- 20 CPU intensive
- 26 Memory intensive
- 9 IO intensive
- 10 low resources
- 54 unclassified (mixed usage pattern)

Learning

- Jobs Consumption models
- CPU, Memory, disk IO
- learning based on jobs input data: user name, nb. of nodes, walltime, queue...

10 low usage users cost 5 years of CPU time.

Prediction Performance



Results

- information retrieval rate of 71% to 89%
- probability of having accurate prediction between 0.7 and 0.8 based uniquely on job input data

Summary

On-going work

- Collect AWS EC2 prices and history
- Map clusters nodes (GFLOPs) to EC2 instances (ECUs)
- Derive a cost model from this mapping