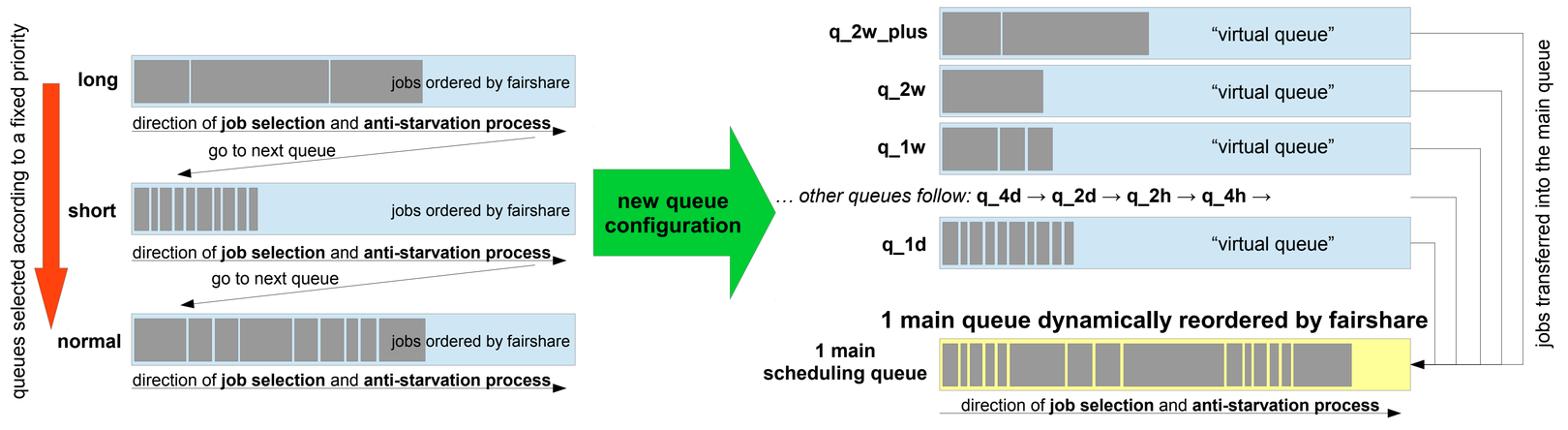
The Performance of the Czech National Grid Infrastructure after Major Reconfiguration of Job Scheduling System

Dalibor Klusáček and Šimon Tóth
Faculty of Informatics, Masaryk University, Brno, Czech Republic
xklusac@fi.muni.cz, toth@fi.muni.cz

1. Introduction

This work describes the outcomes of a **large reconfiguration of the job scheduling system** in the Czech National Grid MetaCentrum which was done in 2014. With the significant growth of MetaCentrum (1,500 CPU cores in 2009 vs. 10,000 CPU cores in 2014) we had to revise our scheduling approaches to better reflect the increased size of the system and the growing heterogeneity of hardware resources and users' workloads.

2. Queue Reconfiguration [2]

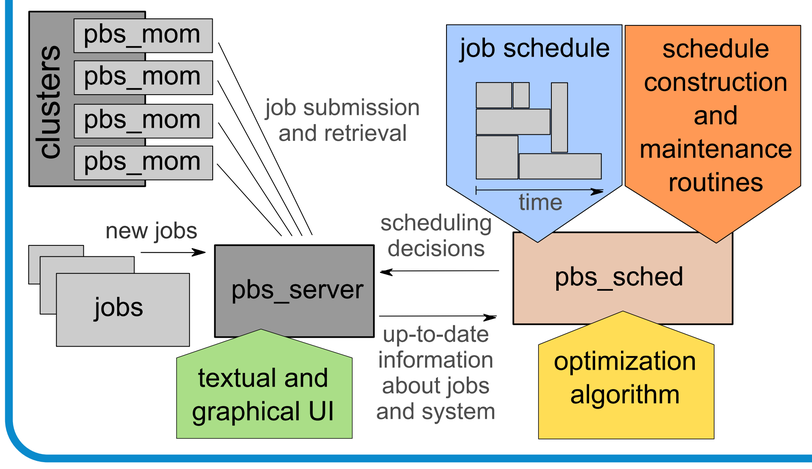


3. Complex Fair-sharing

Main features of the applied solution [3]:

- **an extended Proc. Equiv. (PE) metric**
- **a multi-resource aware mechanism**
- **fair regardless the heterogeneity of jobs and machines**
- **reflects various speeds of machines**
- **a job penalty (i.e., a user priority) is not scheduler-dependent**

4. Plan-based Scheduler with Optimization

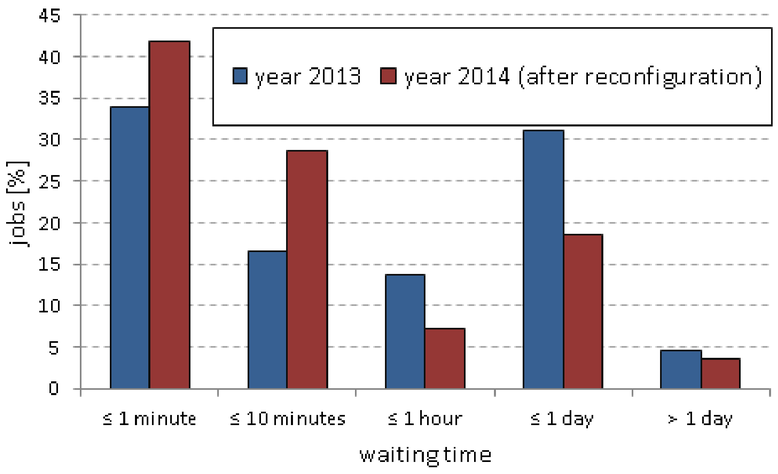


Newly developed TORQUE scheduler [1]:

- **Complete job schedule** data structures
- **Schedule construction** using backfill-like algorithm
- **Maintenance routines** adjust the schedule in time upon dynamic events such as (early) job completions
- **Schedule optimization** metaheuristic

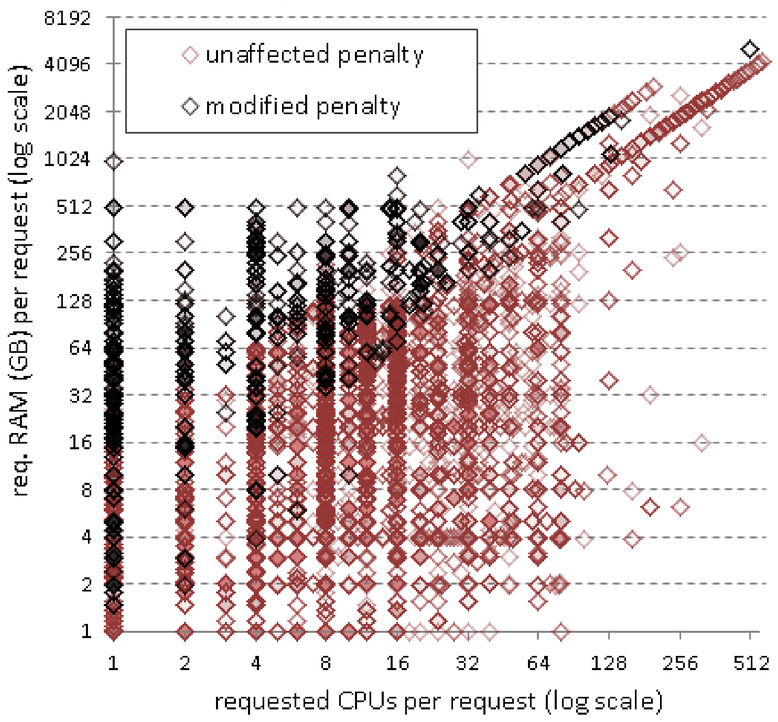
5. The Results of Complex Reconfiguration

Queue Reconfiguration



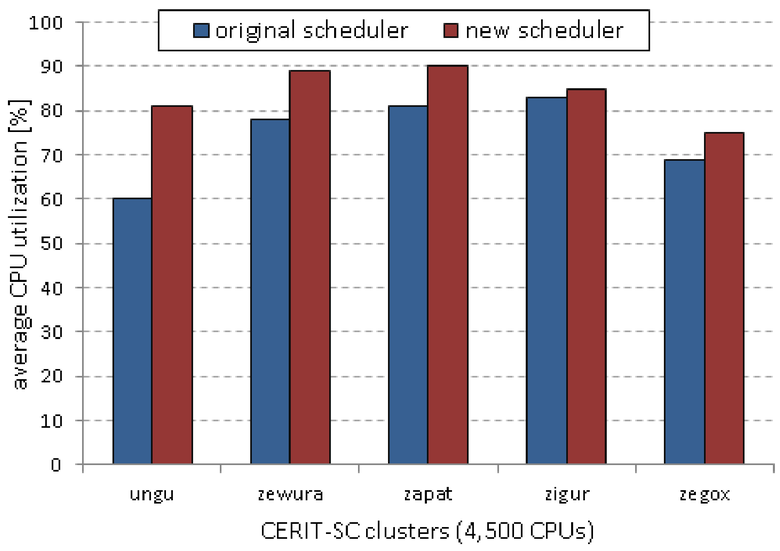
- **Average wait time decreased**
  5.6 hours → 3.9 hours
- **Nearly twice as many jobs processed**
  1.16 milions → 2.12 milions
- **Higher CPU time utilization**
  utilized CPU hours increased by 23%
- **Improved system utilization**
  system utilization increased by 9.1%

Complex Fair-Sharing



- **RAM-heavy jobs affected**
  average wait time increased significantly (3.9 hours → 16.7 hours)

Plan-based Scheduler



- **Improved utilization**
  efficient backfilling with planning
- **Planning and predictability**
  advanced job-to-machine mapping
- **Problem detection and avoidance**
  scheduling plan allows for advanced problem detection

6. Conclusion and Future Work

So far, **the reconfiguration, new fair-sharing solution and the new scheduler used in CERIT-SC seem to work as expected:**

- increased throughput and utilization
- significantly reduced job wait times
- improved fairness
- higher penalties for RAM-heavy jobs

In the future, we will further consider:

- complex evaluation of the performance of the new plan-based scheduler
- a development of a heuristic to dynamically adjust the amount of resources assigned to different queues in the system

Acknowledgments

We kindly acknowledge the gracious support of the Grant Agency of the Czech Republic provided under the grant No. P202/12/0306. We also acknowledge the tight cooperation with MetaCentrum including the access to the MetaCentrum's computing infrastructure and historic workload-related data.

References

[1] D. Klusáček, V. Chlumský, and H. Rudová. Optimizing user oriented job scheduling within TORQUE. In *SuperComputing - the 25th International Conference for High Performance Computing, Networking, Storage and Analysis (SC'13)*, 2013, (poster).

[2] D. Klusáček and Š. Tóth. On interactions among scheduling policies: Finding efficient queue setup using high-resolution simulations. In *Euro-Par*, 2014.

[3] D. Klusáček and H. Rudová. Multi-resource aware fairsharing for heterogeneous systems. In *Job Scheduling Strategies for Parallel Processing*, 2014.