Measuring the immesurable

analyzing the impact of different scheduling algorithms

Mgr. Šimon Tóth

Faculty of Informatics @ Masaryk University

April 9, 2013

Czech National Grid

- available for non-commercial research

Czech National Grid

- available for non-commercial research
- interesting hardware

Czech National Grid

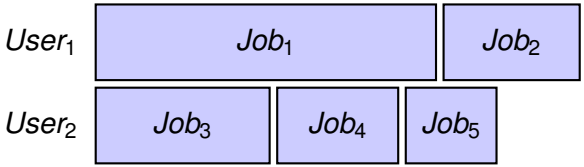
- available for non-commercial research
- interesting hardware
  - GPU clusters
  - clusters with infiniband
  - machines with up to 1TB RAM

Czech National Grid

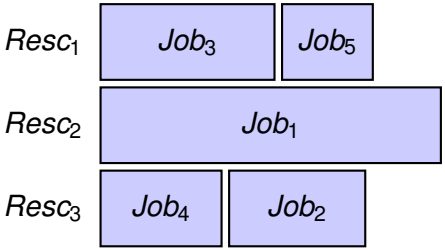
- available for non-commercial research
- interesting hardware
  - GPU clusters
  - clusters with infiniband
  - machines with up to 1TB RAM
- a lot of very expensive and useful software

`http://metavo.metacentrum.cz`

Parallel job scheduling problem



Parallel job scheduling problem



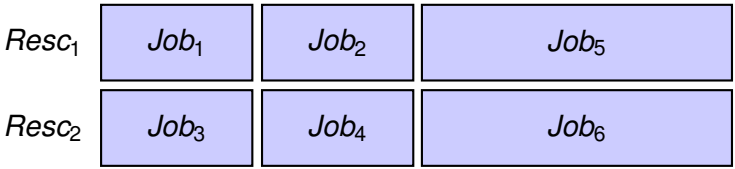
Problem specifics in GRID context

- multi-dimensional
  - each job can request a large set of resources
  - CPU, Memory, GPU, Scratch, Licenses,...
- on-line
  - jobs are not known until they arrive into the system
  - at any time any amount of jobs from any user can arrive
- non-clairvoyant
  - we only have upper bounds for job run times
  - jobs appearing as 24 hour long can easily end in 10 minutes

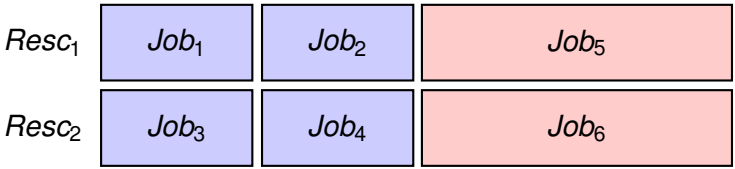
Wait-time and Slowdown

- most commonly used "quality" indicators
- both suffer from similar flaws
  - large sets of jobs from single user
  - different representations of equivalent requests

Example of bad evaluation



Example of bad evaluation



Fairness

- many different representations
- commonly used approach is fairshare
  - priority balancing algorithm
  - using the system decreases priority

$$\forall i; \lim_{time \to \infty} Usage(User_i) = TotalUsage \times DesignatedFraction(User_i)$$

Fundamentals

- quality based of user satisfaction
- modeling user expectations

The model

- each user given "bandwidth"
  - CPU 8 $\frac{core}{second}$
  - memory 16 $\frac{GB}{second}$
- for each job a deadline is calculated according to available bandwidth

Deadlines example

4 jobs

- 4 CPU cores
- 8 GB RAM
- 4 hour runtime

Deadlines:

- 4 hours for first and second job
- 8 hours for third and fourth job

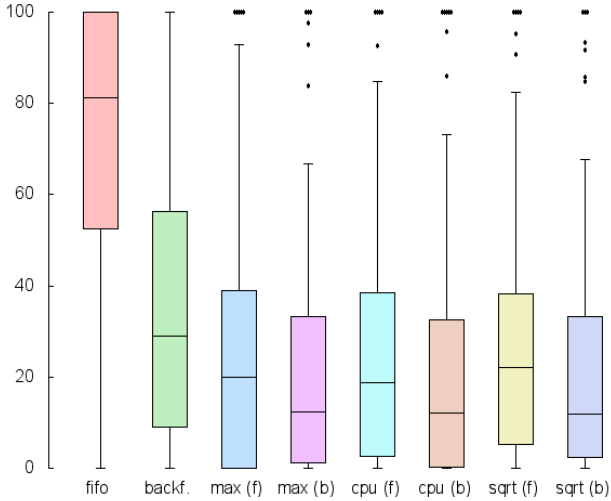
Alea - the Grid Simulation Environment

- created by RNDr. Dalibor Klusáček, Ph.D.
- uses real data sets from CERIT and MetaCentrum

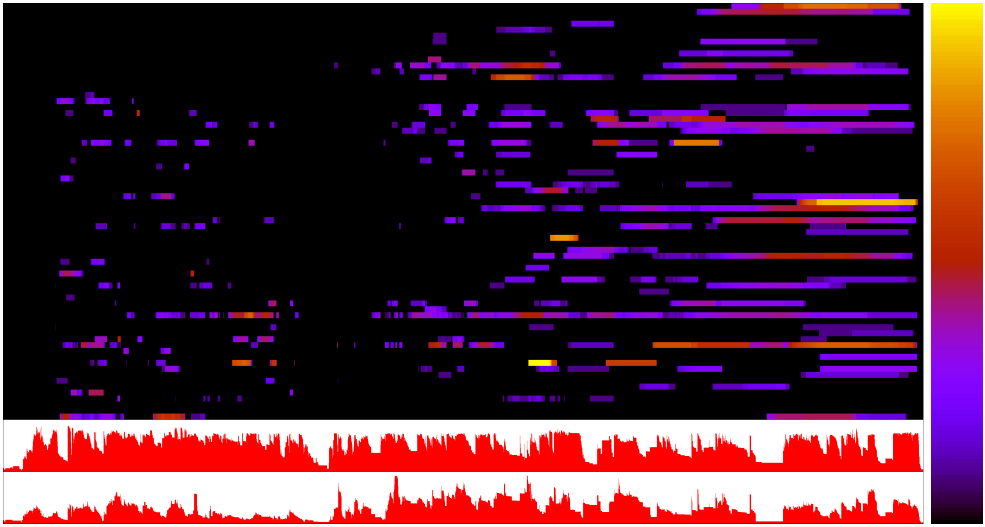
Scheduling algorithms

- trivial FIFO
- FIFO with backfilling
- combinations with fairshare variants

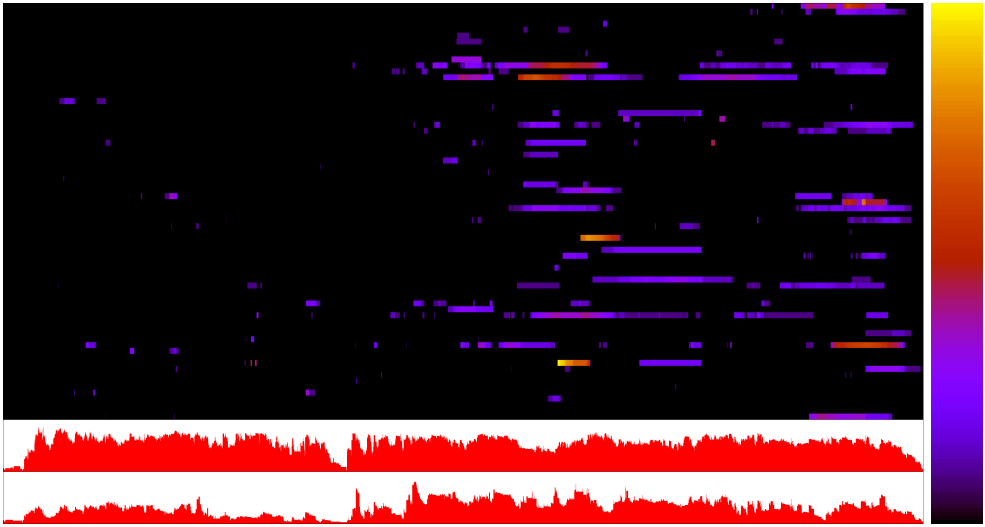
Overview graphs



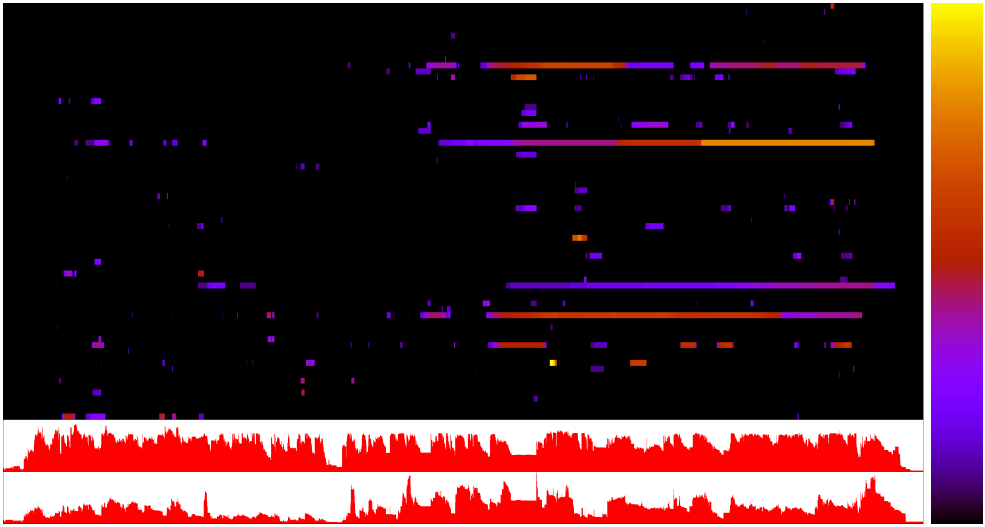
Trivial FIFO



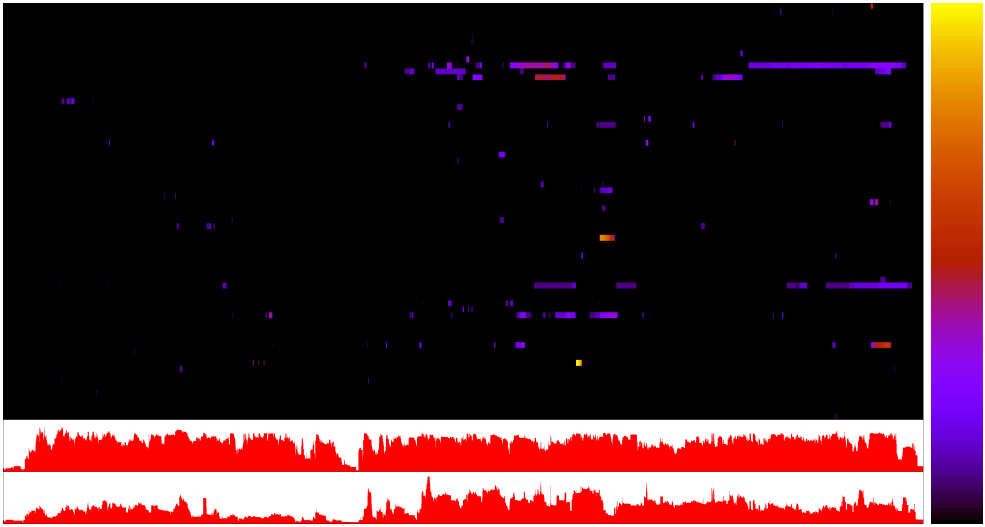
FIFO with backfilling



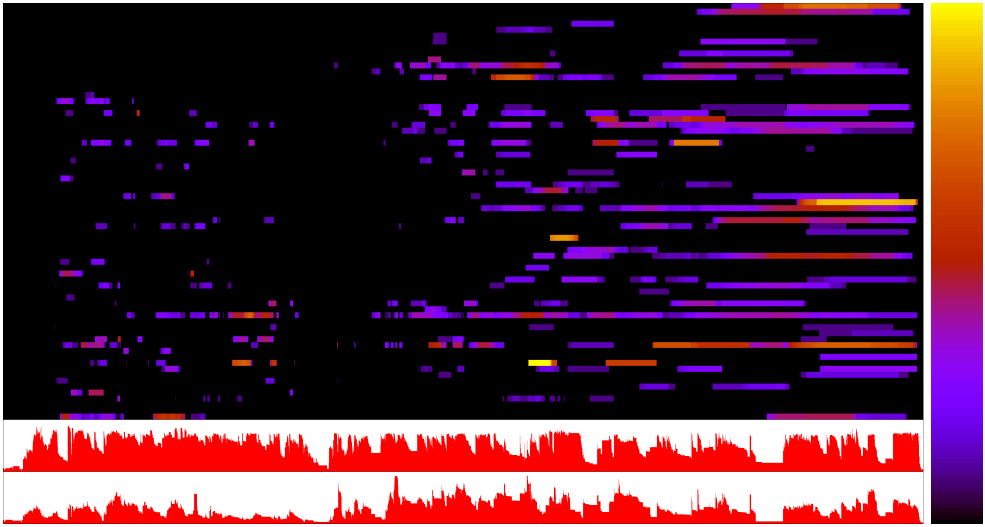
FIFO with fairshare



Fairshare and backfilling



Trivial FIFO



Future work

- users with different priorities

Future work

- users with different priorities
- users with different tolerance towards deadline violation

Future work

- users with different priorities
- users with different tolerance towards deadline violation
- users with special access to particular machines

Summary

- parallel job scheduling in grids

Summary

- parallel job scheduling in grids
- measuring quality of algorithms for job scheduling

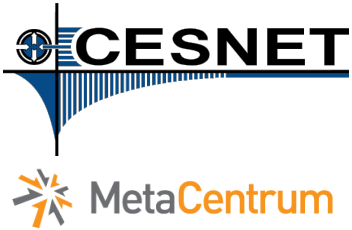
Summary

- parallel job scheduling in grids
- measuring quality of algorithms for job scheduling
- model for quantifying the quality schedules

Summary

- parallel job scheduling in grids
- measuring quality of algorithms for job scheduling
- model for quantifying the quality schedules
- examples of real measurements


FACULTAS ARTIS INFORMATICÆ
UNIVERSITAS MASARYKIANA


CESNET
MetaCentrum


sc
cerit
scientific cloud