PAUL SCHERRER INSTITUT

PSI

Wir schaffen Wissen – heute für morgen

Paul Scherrer Institut

Derek Feichtinger

HPC-CH forum on Scheduler and Queuing Systems – May 18th 2011

Conference Information

You can download the conference presentations by going to

http://indico.psi.ch/

Choose “Conferences Link” and go to the entry for today's date, May 18th, and the HPC-CH event.

Choose “Timetable” on the left menu, and then “detailed view” on the top for seeing all contributions and links to the slides.

PSI HPC/HTC resources

Local Resources

- Merlin cluster (104 cores, 3-4 years old).
  - Currently in planning phase for new Merlin cluster (order of 400 cores)
- LCG Tier-3 for the LHC/CMS experiment (160 cores)
- Some smaller clusters dedicated to research groups

Remote Resource

- PSI did a buy-in to a small part of the CSCS Cray XT5 “Rosa” system. This is the principal resource for our “no-nonsense” HPC power users.

Answers to Michele's Questions

What are the usage policies and how do you translate them in allocation algorithms? What about over allocation, over use, under use, discounts etc.?

For the smaller PSI clusters we have until now fairly trivial queuing policies based on per user fair shares. This will change with the intended consolidation of a number of smaller clusters

Answers to Michele's Questions

What scheduler and queuing systems are you using? Why did you choose this particular system?

We use SGE on all local clusters. Before, PSI had used LSF but went away from it due to high license costs (and we only needed fairly trivial scheduling).

SGE is feature rich, used to be well documented, and used at many sites. Also among grid sites there is an increasing number of sites wanting to use it

The big centers KIT and IN2P3 made both an evaluation of several batch systems where SGE emerged as best suited (also very performant)

Selecting a new batch system at CC-IN2P3, B. Chambon, HEPIX 2011
Grid Engine setup at CC-IN2P3

Answers to Michele's Questions

What is most important to you: price, performance, reliability, support?

Reliability

Answers to Michele's Questions

What are the costs of scheduler and queuing systems ?
Are open source products really cheaper than proprietary? What are you experiences?

Total cost of ownership is difficult to estimate.
Configuring batch systems requires significant manpower effort. Configurations evolve together with the systems.

Open source is preferred based on possibility to extend functionality or integrate with local infrastructure.
Also, protects from effects of company mergers...

Answers to Michele's Questions

What could be possible metrics to measure the performance of queuing systems?

- Jobs that can be scheduled per second
- Total queue-size system is still able to handle
  - Is significant slowdown with growing queue observed (e.g. Maui scheduler polling PBS every few seconds and parsing/scheduling whole job queue)?

Answers to Michele's Questions

What are your experiences in the integration of heterogenous systems?

At PSI we have until now worked within smaller rather homogeneous clusters – so, not much experience.

But investigating now, since new consolidated cluster will require more complex policies

Answers to Michele's Questions

How much should be hidden and shown to the end user? Is it better to have a single generic queue or to have specialized queues?

If one can implement an intelligent enough policy, naturally the preference would be one generic queue (e.g. like PBS routing queues in standalone PBS).

Answers to Michele's Questions

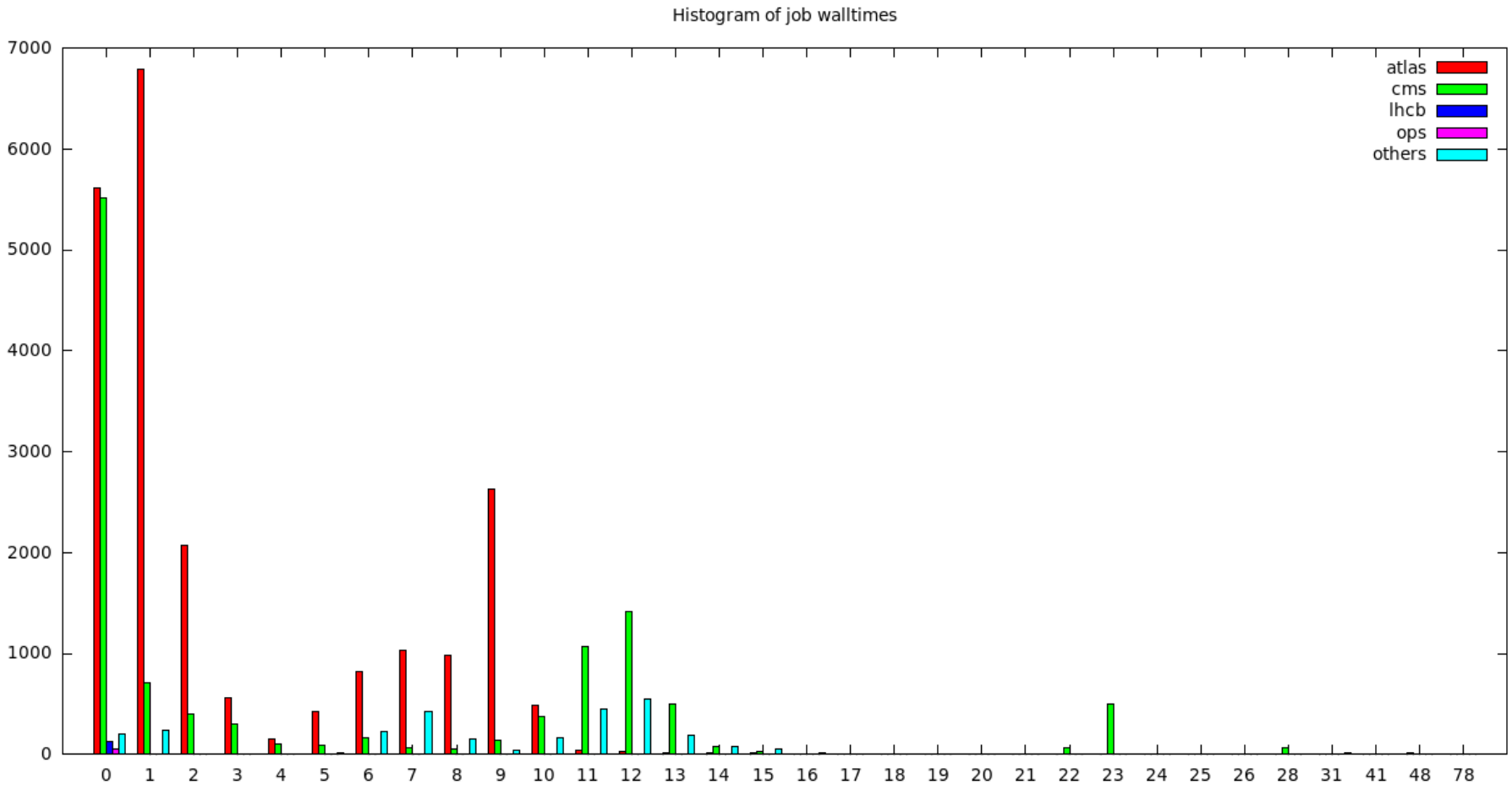
How do you interpret usage accounting and priority settings?

I guess the question is how accounting is done and how the adherence of the real system to the stated policy is checked for:

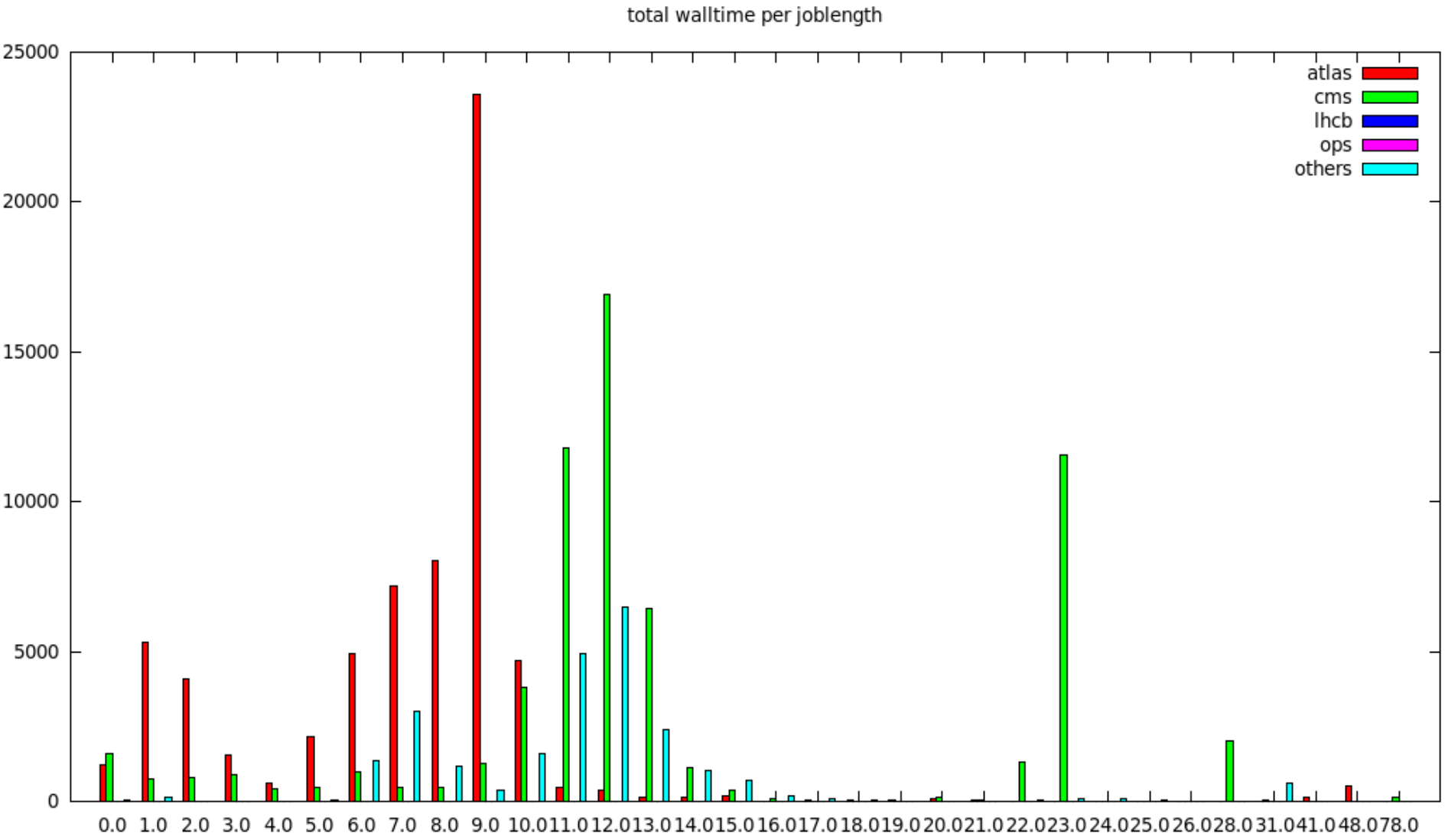
The batch system should offer detailed logs and reporting tools (q.v. Fabio Martinelli's talk on SGE ARCO)

- Useful feature: Maui (probably also Moab) offered possibility to feed job traces (job signatures) into the scheduler for simulating scheduling.

Batch system reporting tools

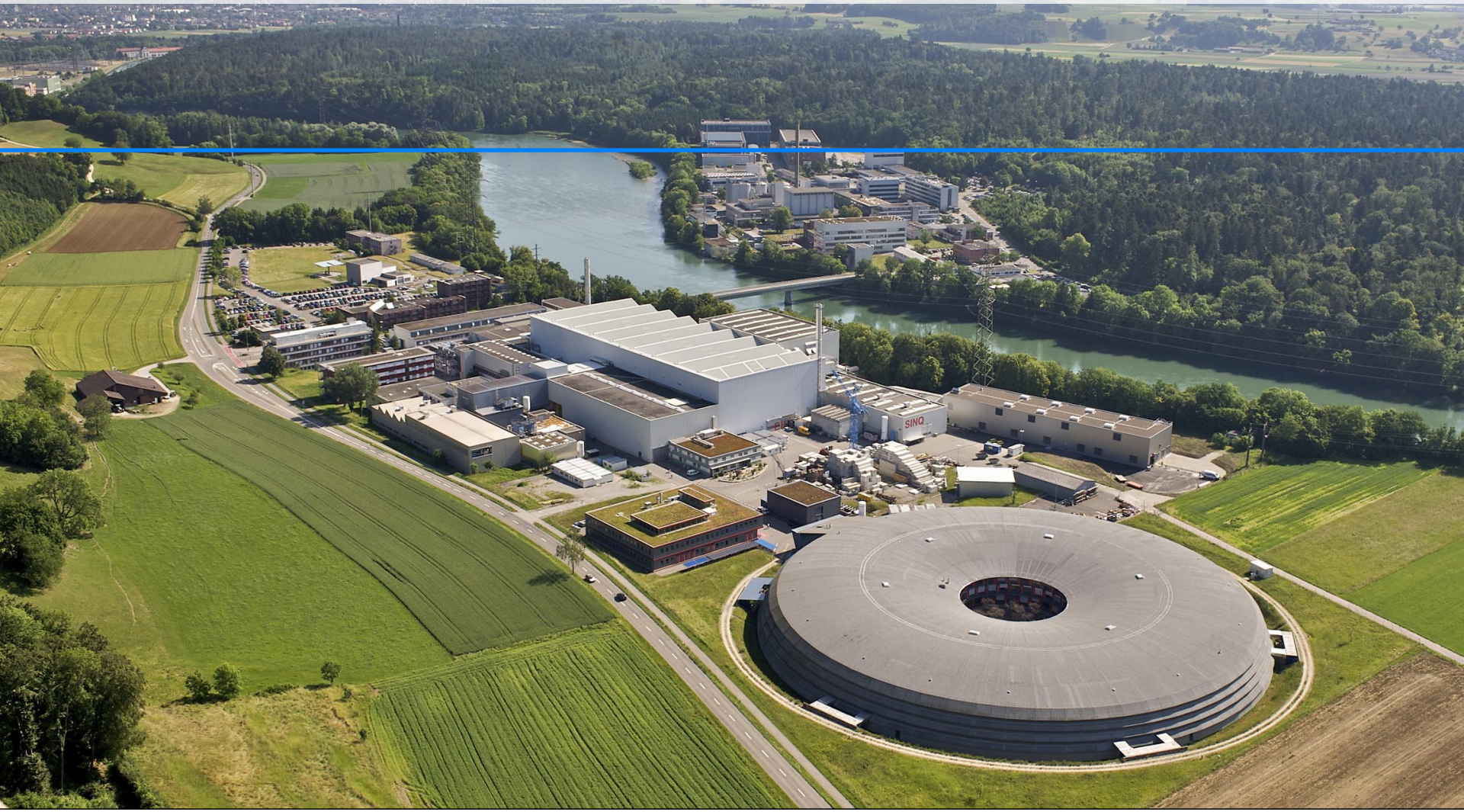


Batch system reporting tools



Thank you for your attention!

QUESTIONS?



Thanks to

The whole AIT team of PSI