Centre de Calcul de l’Institut National de Physique Nucléaire et de Physique des Particules

(Short) status report on the multicore jobs @CCIN2P3

S.Poulat, S. Gadrat

September, 9, 2014

Outline

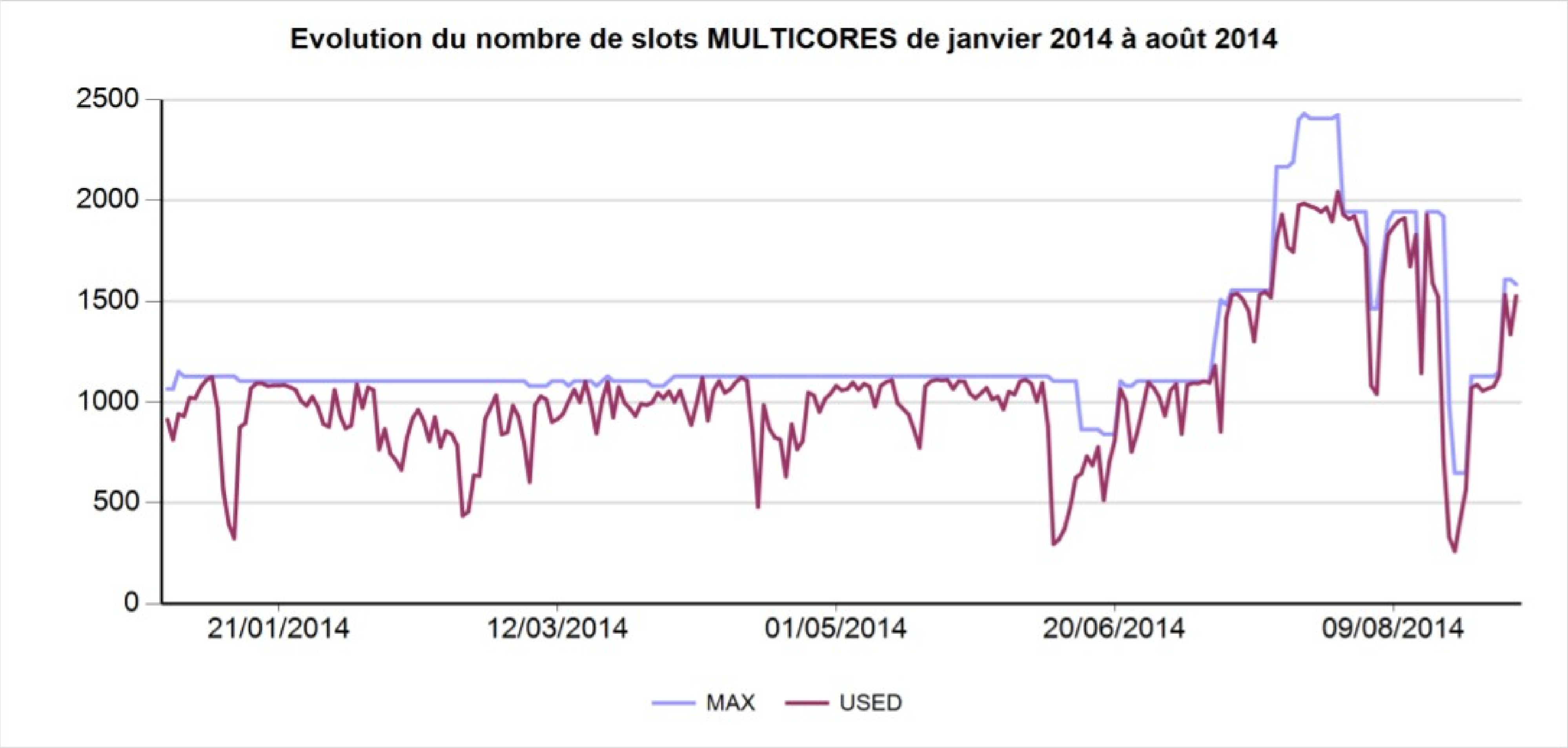
- Setup
- Issues
- Perspectives

A detailed status report has been made at the last HEPiX Spring 2014@Annecy (given by Suzanne Poulat) :
https://indico.cern.ch/event/274555/session/14/contribution/16/material/slides/0.pdf

Current setup

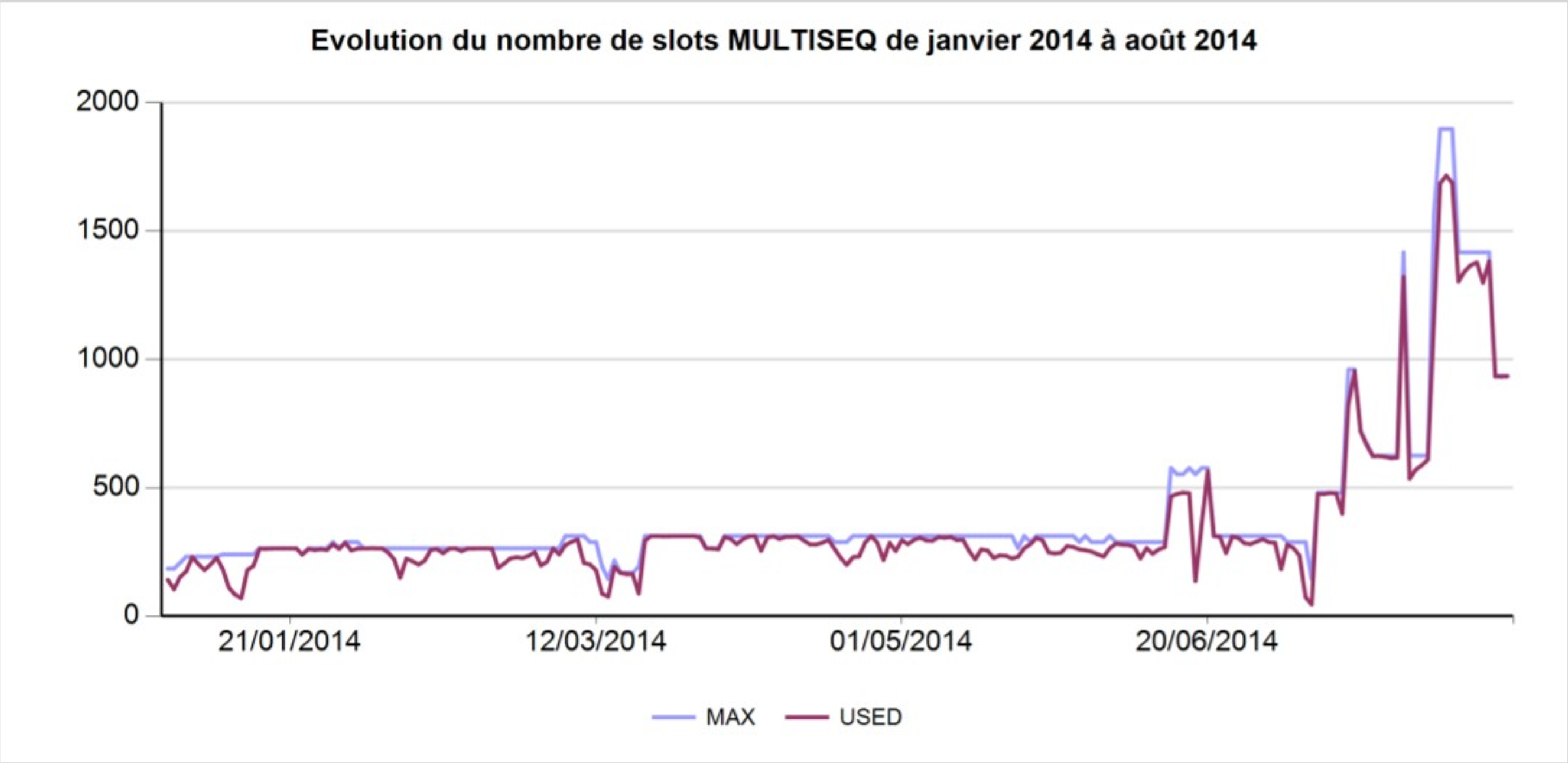
- No resource reservation: use of dedicated WNs
- Between 500 and 2500 cores used for multicore jobs
- 14 groups doing mc: 2 through the grid (ATLAS and CMS)
- Nodes are split into 3 groups :
  - multicore: used for multicore jobs (single jobs not allowed)
  - multiseq: used for both multicore and single core jobs
  - sequential: used for single core jobs only (15000 slots)
- WNs in the pure multicore farm depend on requests (required manual intervention)
- Close collaboration with the users is required in order to plan the resources a bit in advance

Occupancy of the pure multicore farm



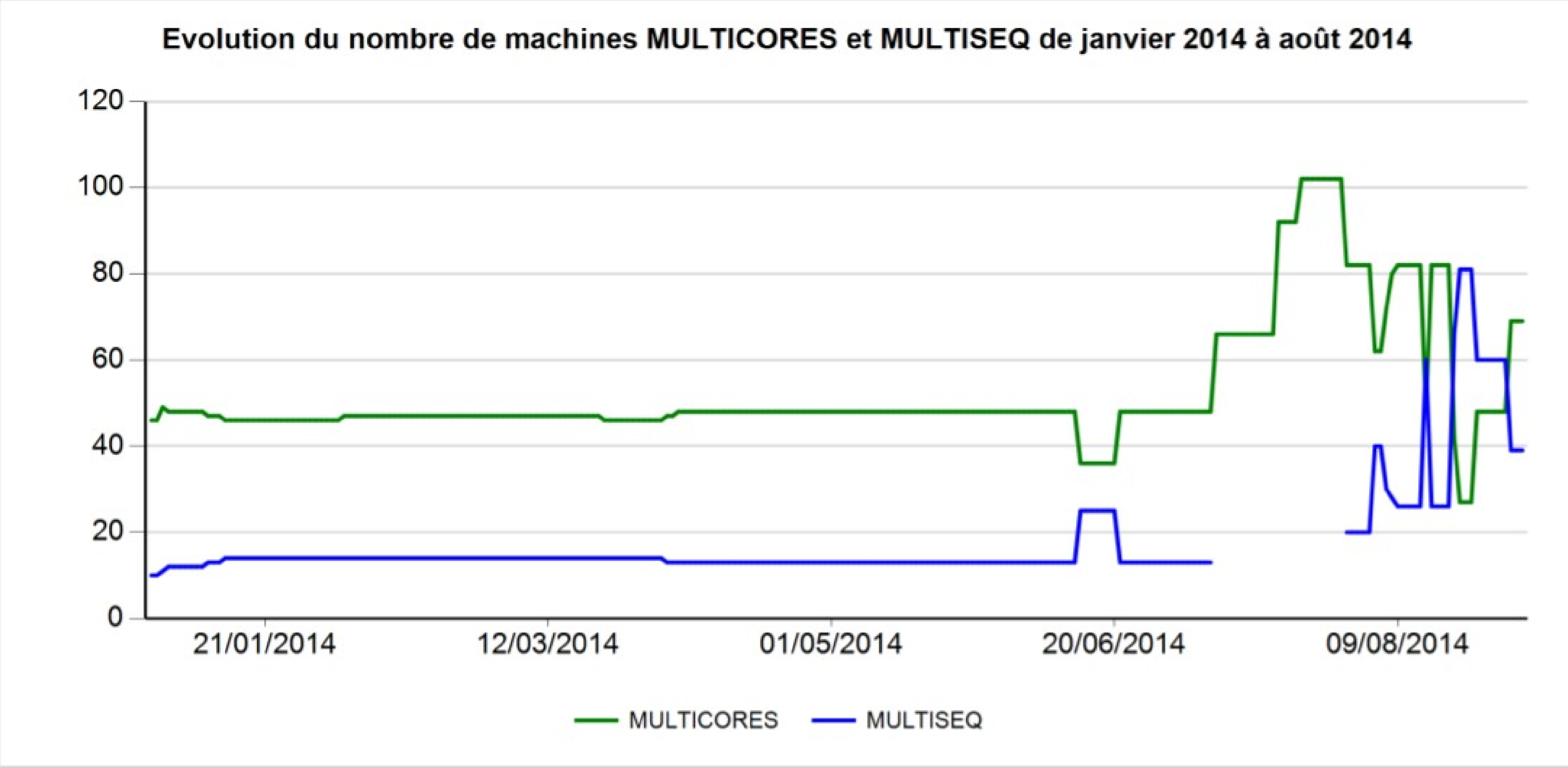
- only multicore jobs allowed, occupancy strongly depends on the users needs.
- Waste of CPU when moving WN from the multiseq farm to the multicore one, as single core jobs are not allowed anymore (drain of the single core jobs to allow multicore).

Occupancy of the multiseq farm



- With single core jobs allowed, the occupancy is ok.
- When moving WN from the multicore farm to the multiseq one, single core jobs allowed to keep a good occupancy.

Distribution of the multicore ans multiseq WNs



- Each WNs has 24 cores
- Repartition between the 2 farms depends of the users needs

Issues which drove the current setup

- Complex configuration of GridEngine (due to the different shared services like storage, and more than 100 groups with specific needs) :
  - 20 differents queues
  - > 300 RQS (limitations required to preserve the back-end services)

- Usual scheduling pass is about 20 to 30 seconds
- Test to mix multicore and single core jobs in the same farm, which required resource reservation :
  - This can increase up to x10 the sched. pass (due to resource reservation)
  - Which leads to system unstabilities and CPU waste

- That's why we are running for now with dedicated WNs

Perspectives

- We are working on the simplification of our GridEngine configuration
- In order to mix all jobs, switching on reservation for jobs could be done using a script (or CRON) just like what is done @KIT