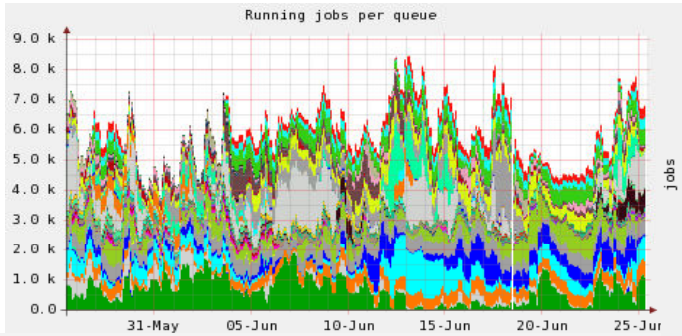
Practical aspects of multi-core job submission at CERN

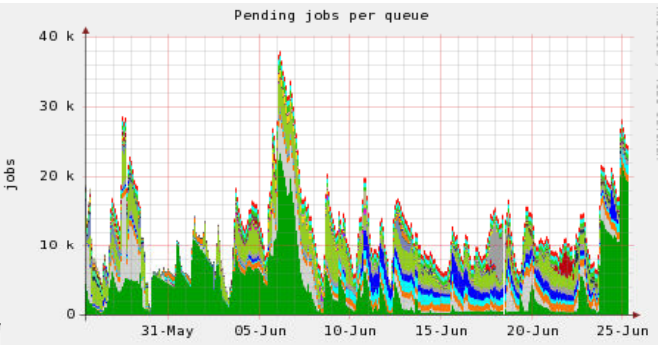
Ricardo Silva
CERN
IT-FIO-FS

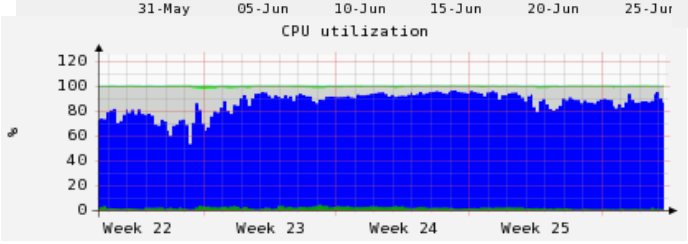
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

The batch system at CERN

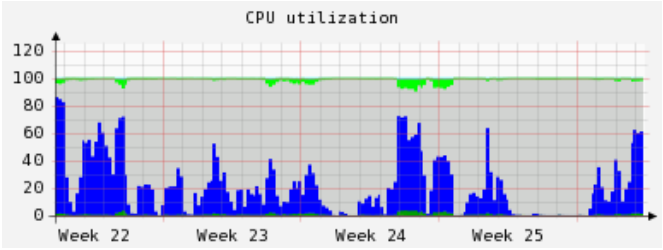
- Platform LSF
- In general no free capacity in the public SLC4 resources (> 10k pending jobs)







- But there are significant unused dedicated resources

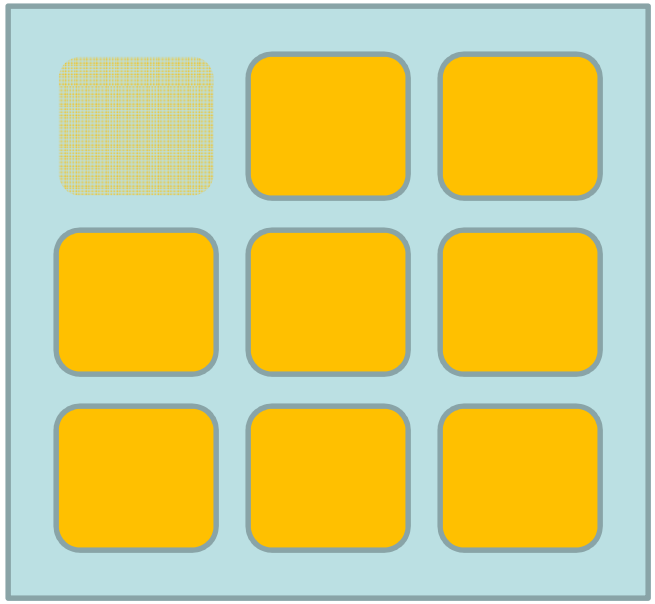


CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Job slot model

- As a general rule we have one job slot per core + 1: in a 8 core machine 9 job slots (2 CPUs, 4 cores per CPU)

- Resource allocations (i.e. memory) are based on a job slot model:
  2 GB RAM / core



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Submitting multi-core jobs (1)

- Without telling the system about it:

```
bsub my_8thread_job.sh
```

  - Only 1 job slot is allocated to the job
    - 1 core / 2GB of RAM
  - There are other 7 jobs running on the same 8 cores, the machine will be quickly overloaded
  - LSF will stop scheduling new jobs to any machine if the utilization (load) is above a certain threshold
    - No new jobs will be started on the machine until the utilization is below the threshold again, but these job slots will be reported as “free” (implications for WLCG)
  - The jobs will be killed based on the resource limits defined for 1 job slot (at CERN this happens at 4GB at the moment)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Submitting multi-core jobs (2)

- And telling the system about it:

```
bsub -n 8 -span[hosts=1] my_8thr_job.sh
```

  - 8 job slots are allocated to the job
    - 8 cores / 16 GB of RAM
  - The **`-span`** option ensures the 8 job slots/CPUs are in the same node)
  - LSF knows the job wants to use the 8 cores and will start reserving cores on the same node before starting the job.
    - The job slots will be marked as being used

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Conclusions

- The system is already capable of correctly handling multi-core jobs, if it knows how many cores the job needs
- Our experience supporting these jobs is very limited
  - We need to understand the effects of large numbers of multi-core jobs on the scheduling (there is the possibility of wasting resources). We will need to tune the scheduling parameters.
  - For effective backfilling we need **accurate runtime estimation** (users should use **`-We`** option)
- LSF is in principle also capable of supporting MPI applications but the system is not optimized (ex: infrastructure, GB ethernet)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Questions?

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it