www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Slurm Site Report

Alejandro Lucero & Carles Fenoy

Barcelona, 9 October 2012

Introduction

- Barcelona SuperComputing Center (BSC) & National Supercomputing Center of Spain (RES)

- RES: Barcelona, Madrid, Valencia, Málaga, Santander, Zaragoza, La Palma, Las Palmas de Gran Canaria

RES & SLURM

- Moab license expiration → BSC as technical leader recommended to use Slurm as resource manager and scheduler
- This last year RES nodes have migrated from Moab/Slurm to Slurm systems
- Slurmdbd facilitates internal accounting and allows users to know how they are using resources

BSC & SLURM

- The Big one: MareNostrum
- SGI Altix 4700, SGI Altix UV-100
- Minotauro: 122 compute nodes (12 cores, 2 gpus)
  +2 computer nodes (8 cores, 4 gpus)
- CNAG: 100 compute nodes (8 cores)
- Montblanc Project: ARM cores

BSC & SLURM: MareNostrum

- Marenostrum2 disconnected this last September
- Marenostrum3 expected this Autumn
- Marenostrum2: Moab & Slurm
- MareNostrum3: ???

BSC & SLURM: Altix 4700

- Migrated from Moab/Slurm to Slurm
- Reservation of cores not supported by Slurm
- Fast & dirty patch supporting this feature …
- …though we followed another approach: virtual nodes with Slurm frontend configuration (limitations)
- This configuration could help for topology aware scheduling
- Working on affinity plugin being aware of virtual nodes (beta)

BSC & SLURM: Altix UV-100

- Installed this year and configured with Slurm
- No Slurm frontend so no reservations support (and no needed by now)
- Topology simpler than Altix 4700: scheduling doing well

BSC & SLURM: CNAG

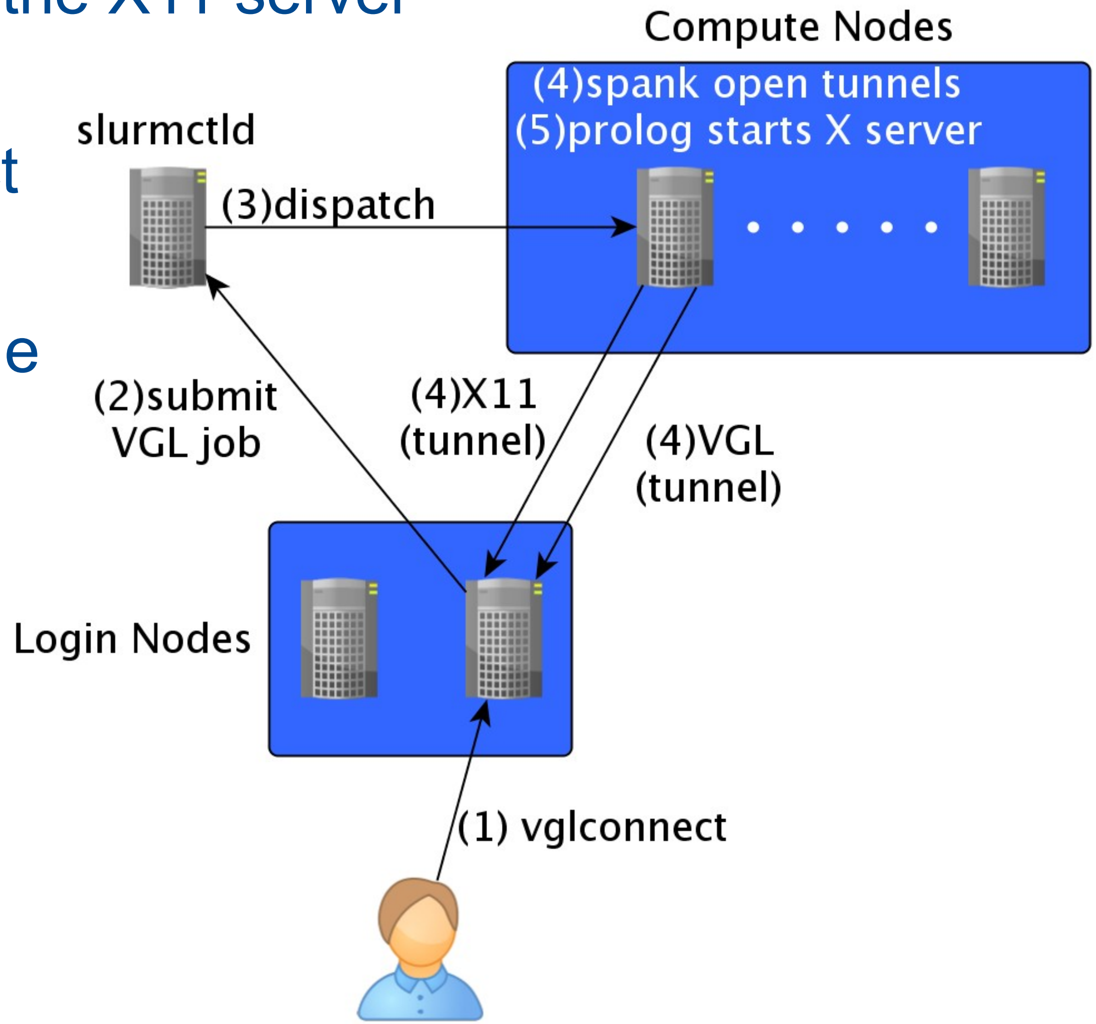
- Completly different usage than other BSC machines
- Goal is more HTC than HPC
- Thousands of jobs with dependencies: short jobs mostly sequential
- Scheduling is heavy
- New Slurm *sdiag* command implemented trying to bring to light how scheduling is doing
- Internal patches solving problems like old libc or “special” programs

BSC & SLURM: Minotauro

- GPU machine
- Slurm GRES patches
- Avoiding slurmctld crashes when GRES plugin misbehave
- Debugging by JOBID local patch
- Power management problems

BSC & SLURM: Minotauro Parallel Rendering

- VirtualGL used access the X11 server of the nodes
- Spank plugin to redirect X11 and virtualgl connection to login node



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC & SLURM: Id manager

Problem:

- Lots of user activations and deactivations
- Activations outside office hours not possible

Solution:

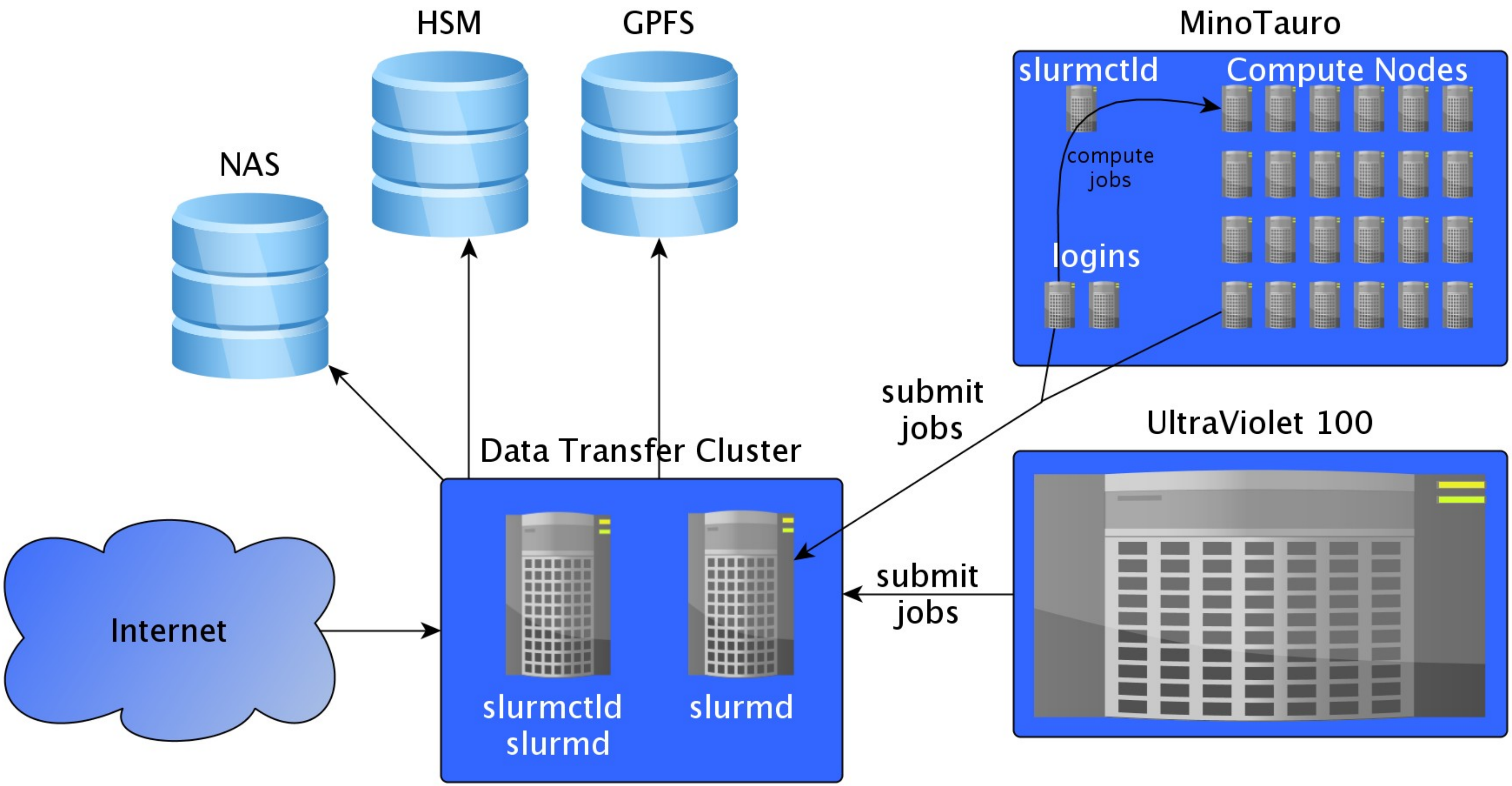
- Automatic system to add, modify or delete users from slurm in all clusters
- Diff current situation with support users database and applies updates.
- Avoids receiving lots of mails for user management

BSC & SLURM: data interface

- Not all filesystems are available on all clusters or nodes
- Copying data from one filesystem to another can take lot of time
- With the slurm copy system we avoid overloading some filesystems (tapes,...)
- Developed wrappers **dtcp**, **dttar**, **dtrsync** and **dtmv** to transparently interact with batch system

BSC & SLURM: data interface

- Cluster for transfering data between filesystems



BSC & SLURM: Future

- Power control / scheduling awareness
- Scalability
- Backfilling efficiency
- Updating a production system: critical patches control
- Network Aware Scheduling: Infiniband data

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you!

For further information please contact
alejandro.lucero@bsc.es
carles.fenoy@bsc.es