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

2015/10/16

Upgrade to UGE 8.2: Positive effects at CC-IN2P3

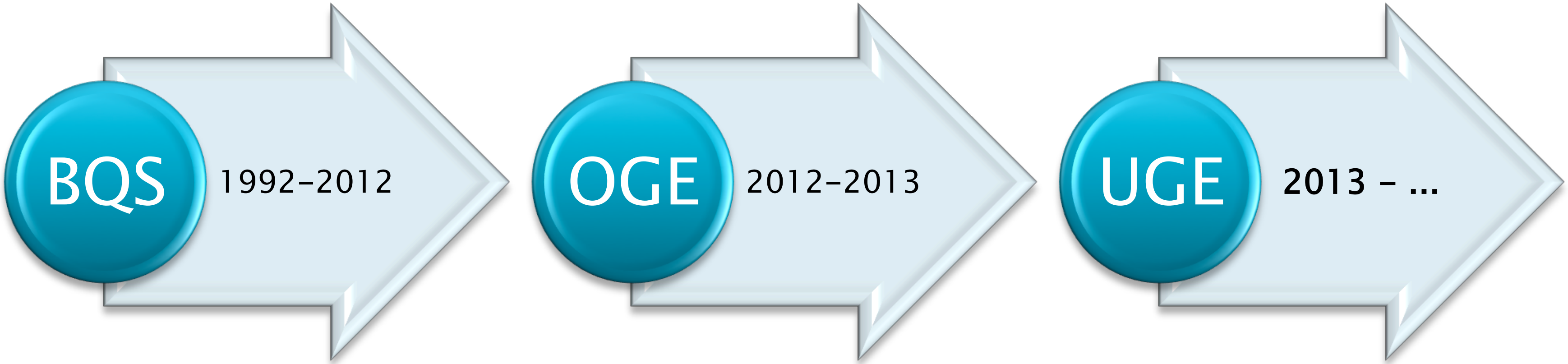
Vanessa HAMAR

On behalf of CC batch team

Overview

- History
- Configuration
- UGE Version 8.2.1 - Improvements
- Future Plans
- Conclusions

History


BQS 1992-2012

OGE 2012-2013

UGE 2013 – ...

HEP World faced new requirements

- multicores
- interactive
- increase of the needs
- Virtual machines

Support

- Oracle support was not satisfactory

Now, is it time to evaluate new options? Stay with UGE?

Our team

1 FTE

2 System Administrators

4 Batch Administrators

1 User Support

BATCH TEAM

Configuration - Nodes

Common

- Operating System = Scientific Linux 6
- UGE Version = 8.2.1

Servers : Master & Shadow

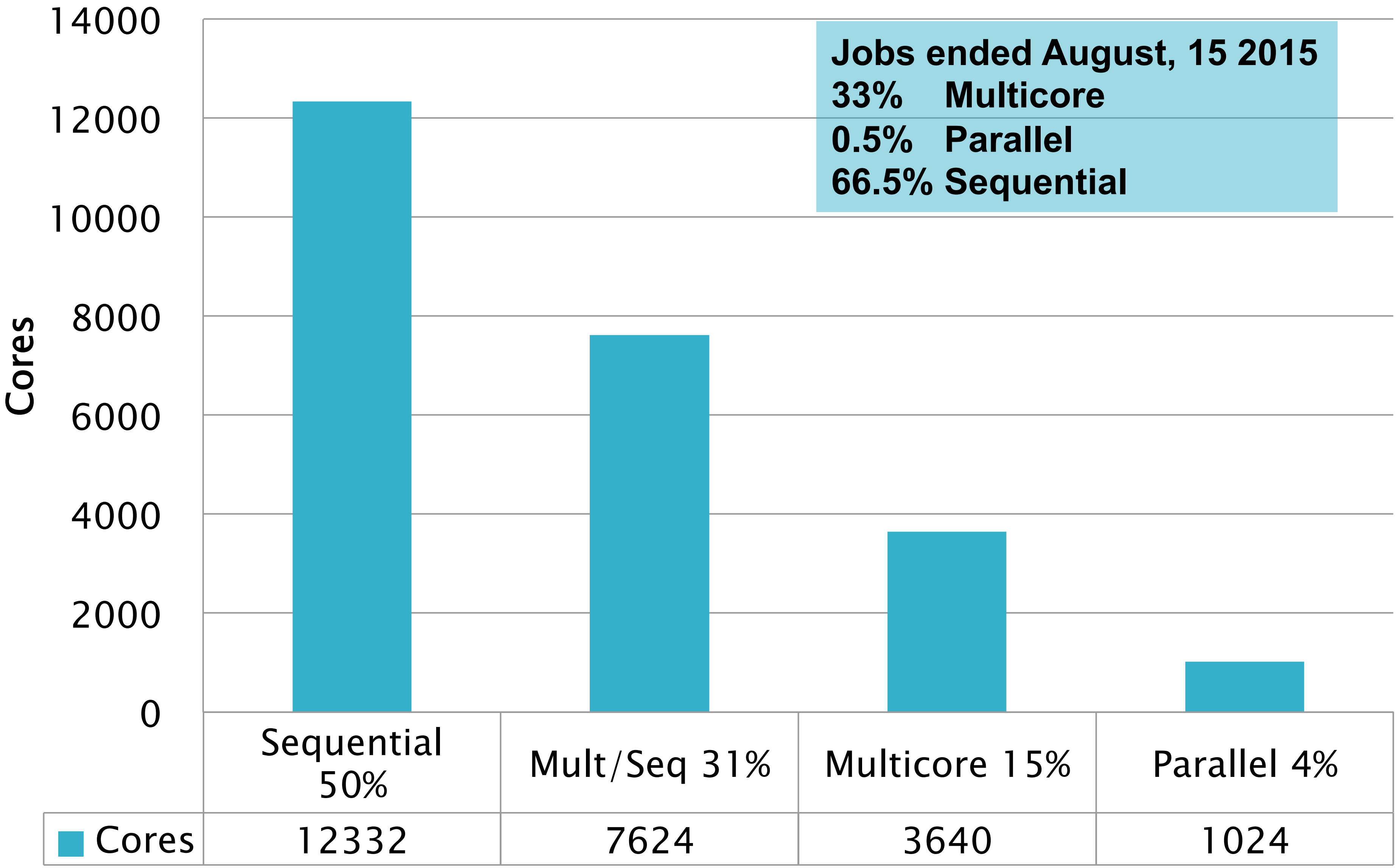
- PostgreSQL: Spooling, ARCo
- Qmaster process: automatic restart
- Externals tools for:
  - Messages:
    - Daily rotation files on master
  - Accounting:
    - Current file only last 7 days
    - One file per month.

Configuration - Nodes

Execution hosts / Worker nodes

- Common directories shared using AFS File system
- Trace files are maintained, but files older than 5 days are deleted (keep_active = true)
- Local development
  - AFS token renewal
  - GPFS access control to allow or deny job submission according complex specification
    - Kernel module used by automounter
  - Prolog / Epilog scripts
    - XFS quota used to manage local disk spaces
    - Copy job outputs to user’s HOME

One instance for all our needs



Jobs ended August, 15 2015
33% Multicore
0.5% Parallel
66.5% Sequential

	Sequential 50%	Mult/Seq 31%	Multicore 15%	Parallel 4%
▇ Cores	12332	7624	3640	1024

Configuration - Functionalities

- Resources are distributed according commitments with experiments in a fair way (Share tree policy)
- Resource regulations: as execution hosts, storage elements and databases (Complex, Resource Quota Sets (RQS))
- Administrators can adjust jobs priorities according to particular needs from users (Priority and override tickets)

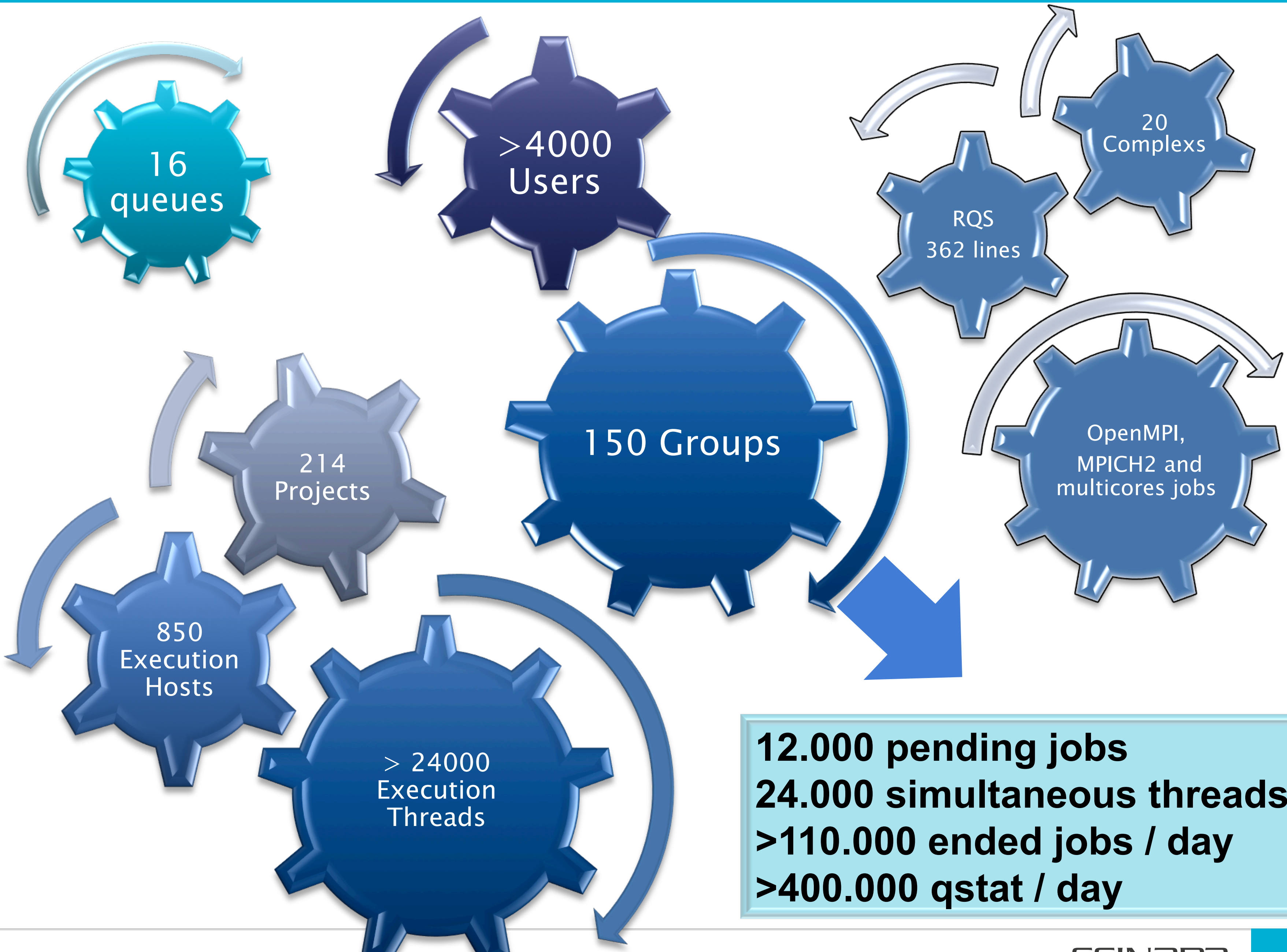
Configuration - Functionalities

- Execution hosts classification depending of “Load Formula” (Load sensors: disk space and memory usage)

- Scheduler limitations in order to avoid master blockage:
  - SCHEDULER_TIMEOUT = 300 seconds
  - MAX_SCHEDULING_TIME = 100 seconds (before 200 seconds)
  - MAX_DISPATCHED_JOBS = 100 seconds (before 200 seconds)

- User jobs submission are validated only in interactive machines to force core binding. (Job Submission Verifier (JSV))

Some numbers



12.000 pending jobs
24.000 simultaneous threads
>110.000 ended jobs / day
>400.000 qstat / day

UGE Version 8.2.1 - Improvements

- **Architecture**
  - Decoupling read-write and read-only threads improved time required for:
    - job submission
    - scheduling performance
    - job dispatching
    - overall cluster throughput

Up to 64 reader-threads

  - **In our case:**
    - More memory was added to our servers
    - 2 read threads
  - **Positive effects:**
    - Server stability improved, better response times

Commands performance = 5x faster

UGE Version 8.2.1 - Improvements

- **Job Accounting**
  - Job timestamps are recorded in milliseconds
  - Job deletions logs enhanced
  - Used qsub commands are logged
  - Supports full 32bit job ID numbers with a configurable rollover
    - **Before 9 999 999**
    - **Now 9 999 999 999**
  - **Positive effects:**
    - Job account more precise
    - Better traceability of deleted jobs
    - Time increased between rollovers, user job priority not affected

UGE Version 8.2.1 - Improvements

Request limits

- Requests that are sent by command line clients might get rejected when a limit is exceeded.

- **In our case:**
  - Qsub max 200 by second
  - Qstat max 200 by second
  - Qdel max 30 by second

- **Positive effect:**
  - System better performance when limits are applied for each command
  - Load system charge reduced

UGE Version 8.2.1 - Improvements

- **Job Resource Control**
  - Users can now specify dynamically runtime limits for jobs

  - **In our case**:
    - Is available

- **Cluster Diagnosis**
  - Annotations for queue state changes can be logged
  - Details about event clients have been added

  - **Positive effects**:
    - Logs about queue state changes
    - Users and hosts that trigger certain commands can be identified

UGE Version 8.2.1 - Improvements

- **Short Jobs**
  - Better short jobs management
  - **In our case:**
    - Server is not sensible when bunch of short jobs are submitted to the cluster
  - **Positive effect:**
    - Avoid scheduler performance impact and/or degradation
    - Allows users to run big amounts of repetitive tasks

Working on

- **NVIDIA GPU integration**
  - Add resources to the configuration is very easy
    - Declare the resource as a complex
    - Load Sensor specific added
    - Currently doing tests

- **Testing the new version 8.3.1**
  - Use cgroups managed directly by the batch system
  - Docker / Containers
  - Manual preemption
  - Different Resource Requests for Master and Slave Tasks of Parallel Jobs
  - Lost job Detection

Future plans

- GPU machines in production (Dec 2015 – Feb 2016)
- Update to version 8.3.1 (Dec 2015 – Mars 2016)
- Evaluate new tools like:
  - Unicloud
  - Docker / Containers
- Set number of jobs by user

Conclusions

- Configuration of our cluster is complex
  - Local and grid users
  - Diversity of requirements
  - Diversity of hardware

- UGE is a product in constant evolution
  - Changes are easy to apply
  - Our requirements are included in the product road map
  - UNIVA is dedicating a developer to our requests
  - Users support is quick and precise

Questions

- Thank you !