

Workload management

FJPPL Computing Workshop

Nadia LAJILI, Suzanne POULAT

& CCIN2P3 Batch Team

- History
- Workload management
- Pros and cons of GE
- Incident management
- Batch and cloud

History



- BQS : Home made batch system
- OGE : 2011-2013
- Univa GE : since 2013
- SL6 – UGE 8.1.6
- ~23000 virtual cores : one instance for all our needs

**ORACLE
support
was not
satisfactory**

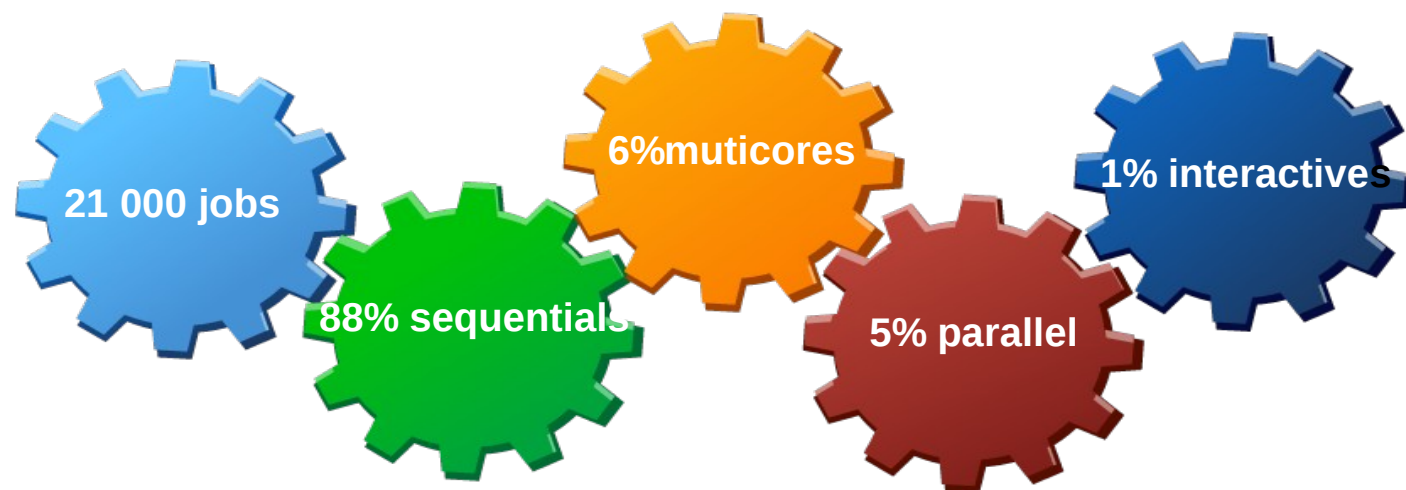
**“HEP world” faced
new requirements :**

- VM,
- multicores,
- interactive
- increase of the needs

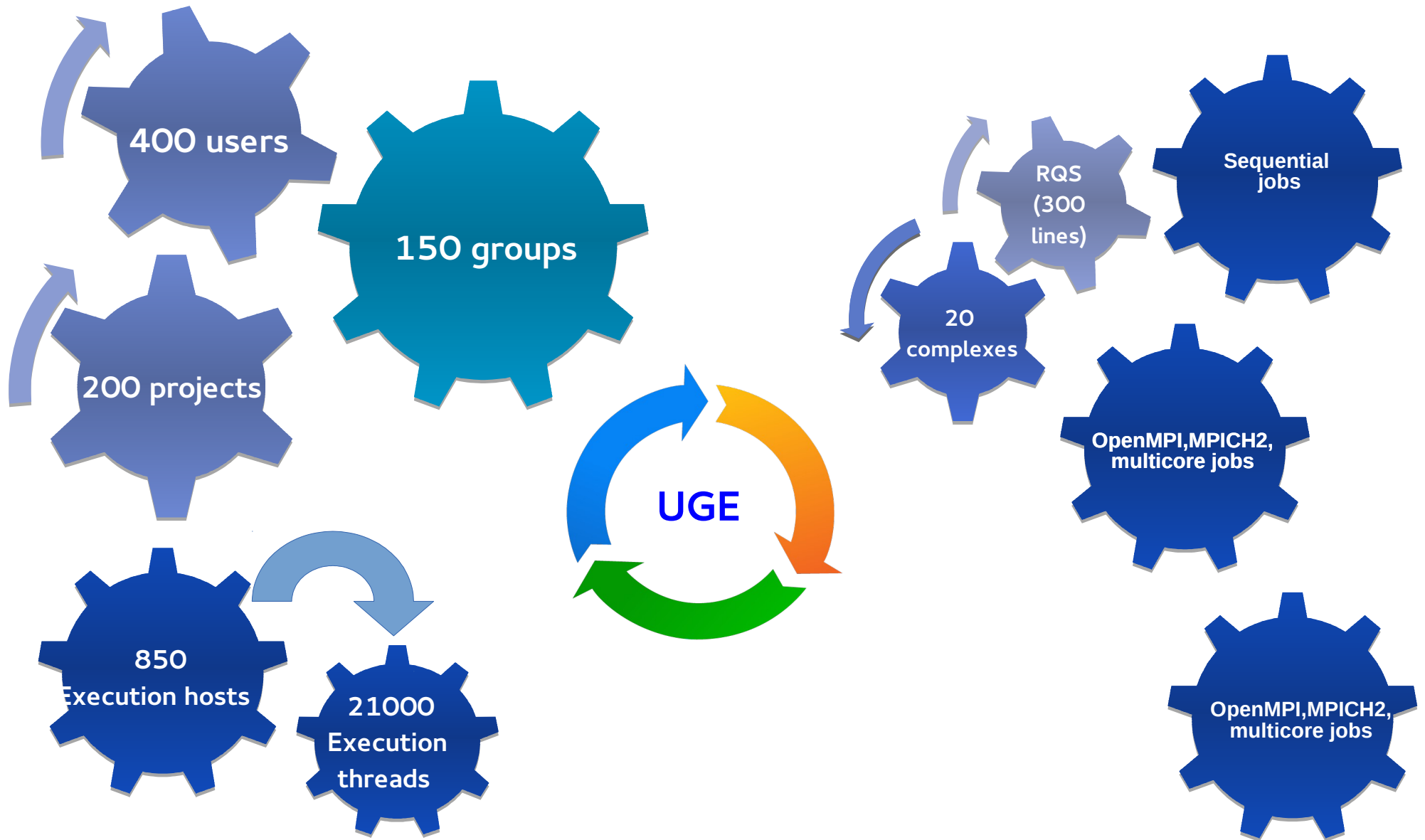
**1 FTE for
Administra
tion &
Operation**

- **Jobs**

- ~12 000 pending jobs, some are array jobs : 40 000 pending tasks
- ~21 000 running jobs, some are parallel : 22 000 used slots
- > 110 000 ended jobs / day
- > 600 000 qstat / day



Overview



- To regulate the load on the system :

Global configuration

- Fair share (two levels) on group & on projects (200) using <share tree policy>
- Job flow regulation : via complexes (20) and intensive usage of Resource Quota Sets (340 lines)
- Scheduler limitations : SCHEDULER_TIMEOUT, MAX_SCHEDULING_TIME, MAX_DISPATCHED_JOBS
- Manual adjustment for cluster CPU optimization using <override policy> to increase the priority of pending jobs

Batch resources configuration

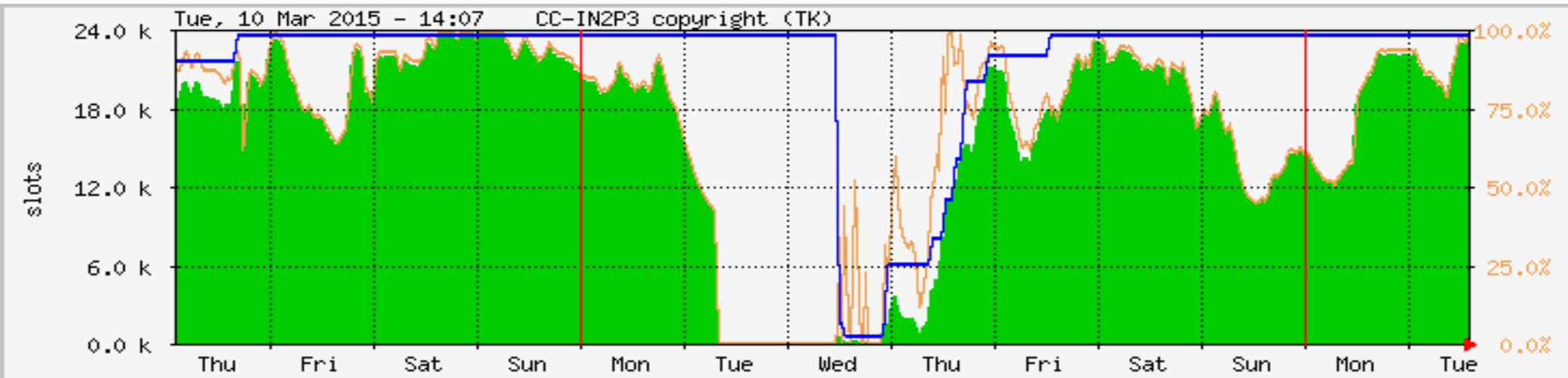
- Load sensors for disk space and memory usage, integration in “load formula”

The way to check if the worker is able to run jobs

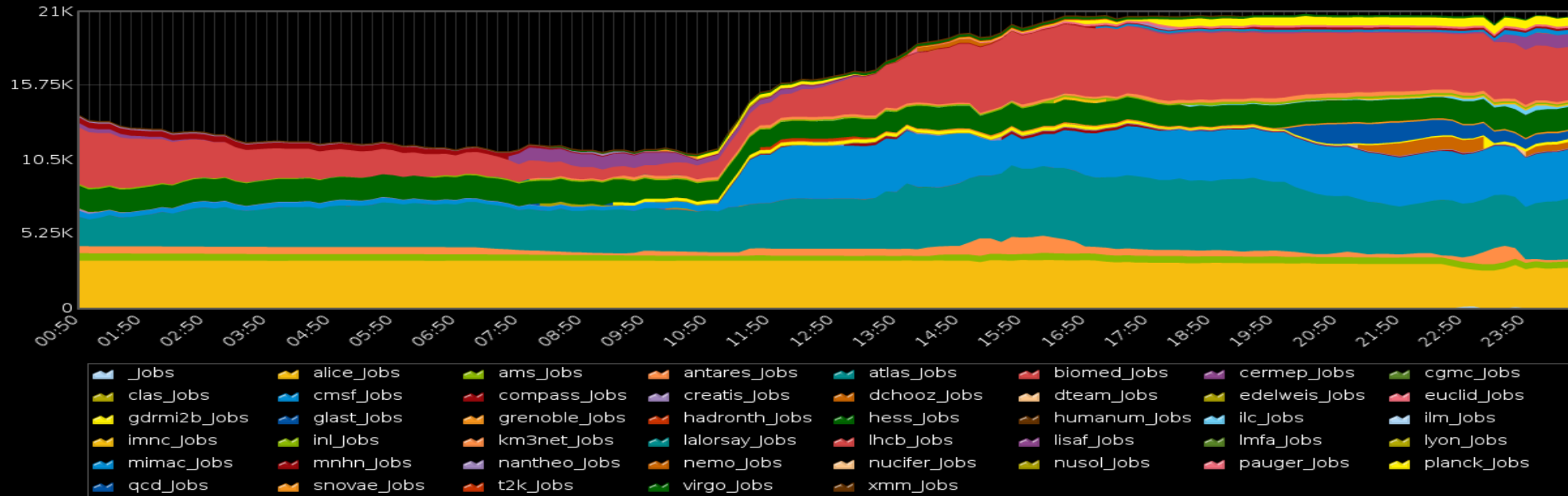
- Fix the global number of slots and the number of slots by queue according to hardware limit of the worker node

- To prevent the system from heavy load :
 - **Monitoring platform based on NAGIOS**
 - Nagios probes for master monitoring (load)
 - Nagios probes checking job efficiency
 - **Home made monitoring tools**
 - Dedicated process parsing GE master error emails, block and notify users having problematic jobs
 - Cron jobs producing hourly statistics on jobs efficiencies
 - **Web portals**
 - Main parameters of the system are available (Kibana)
 - Many web pages on cluster usage per group (MRTG)
 - Real-Time metrics on Running/Ended jobs, workers efficiency (SYMODO)

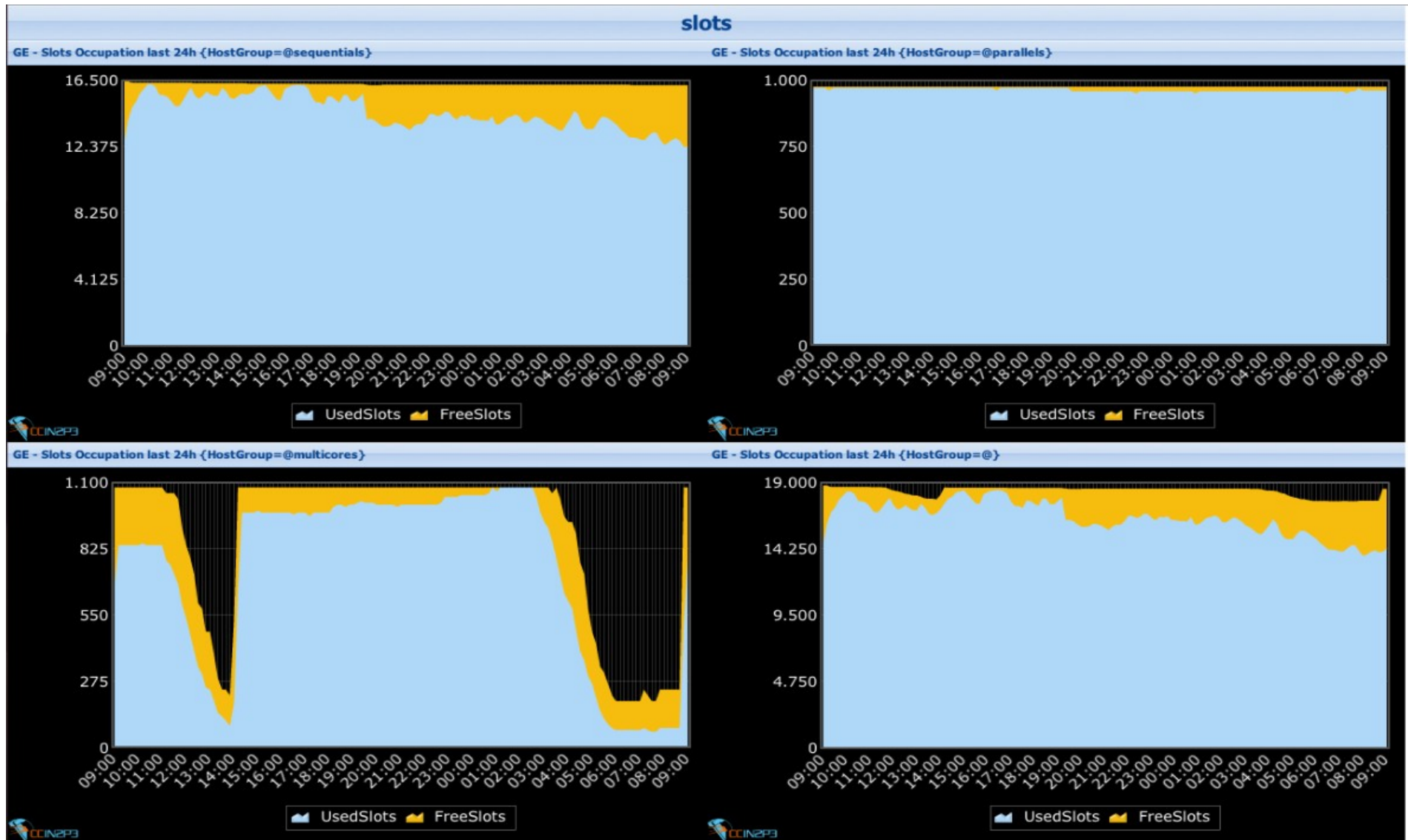
- Current production by MRTG/SYMOD



GE - Slots by group on last 24 hours
Running jobs



- Slots occupation by type (SYMMOD)



- BATCH Monitoring by NAGIOS

The screenshot displays the Nagios Core web interface in a Mozilla Firefox browser. The page title is "Nagios Core" and the URL is "https://ccnagios.in2p3.fr/nagios/". The interface includes a navigation sidebar on the left with sections for General, Current Status, Reports, and System. The main content area shows the "Service Status Details For Service Group 'ge-servicegroup'".

Current Network Status
Last Updated: Tue Mar 10 14:08:43 CET 2015
Updated every 90 seconds
Nagios® Core™ 3.4.1 - www.nagios.org
Logged in as Nadia Lajili

Host Status Totals

Up	Down	Unreachable	Pending
3	0	0	0

Service Status Totals

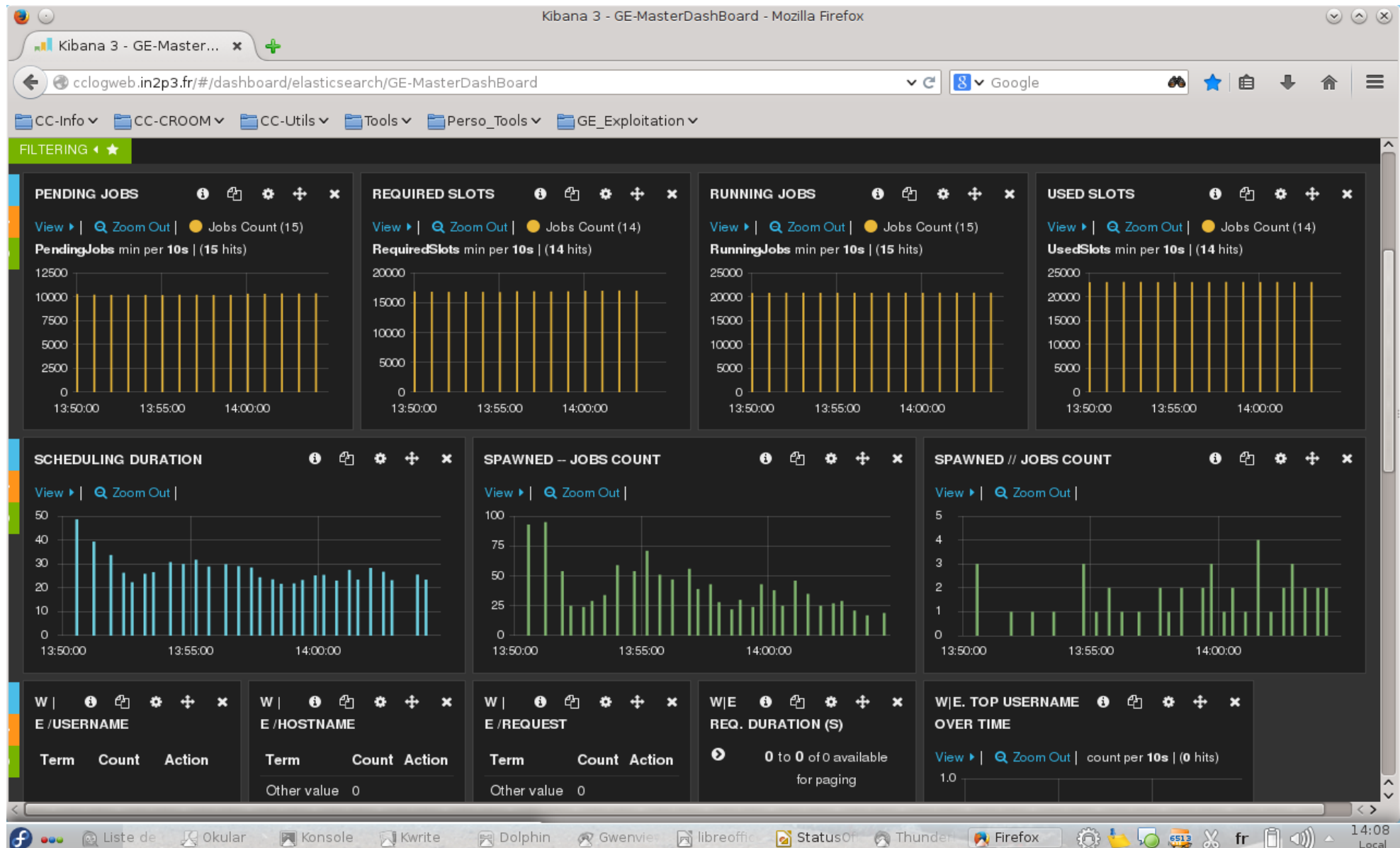
Ok	Warning	Unknown	Critical	Pending
28	0	0	3	0

Service Status Details For Service Group 'ge-servicegroup'

Limit Results: 100

Host	Service	Status	Last Check	Duration	Attempt	Status Information
VIRTUAL	GE Disabled queue	OK	2015-03-10 14:01:00	4d 21h 15m 51s	1/2	Only 26 DISABLED queue(s) for GE
	GE Error queue	OK	2015-03-10 14:06:25	4d 21h 15m 34s	1/2	OK : NO queue in ERROR state for GE
	GE Job Spawn	OK	2015-03-10 14:07:37	4d 11h 1m 6s	1/2	OK : Job execution (the spawn) is opened
	GE Job Submission by qsub	OK	2015-03-10 14:06:04	0d 11h 2m 39s	1/2	OK : Job Submission by qsub is opened
	GE Master status	OK	2015-03-10 14:05:28	4d 21h 23m 15s	1/2	Master GE [ccmgeli01.in2p3.fr] is OK
	GE Max Job Id before Rollover	CRITICAL	2015-03-10 13:55:39	19d 3h 18m 55s	2/2	(Return code of 127 is out of bounds - plugin may be missing)
ccmgeli01	check AFS volume COMMON	OK	2015-03-10 13:42:23	91d 2h 25m 58s	1/2	VOLUME OK - afs volume common.uge.prod.common mounted at /afs/in2p3.fr/common/uge/prod/common is 50% full
	check GE master load	OK	2015-03-10 14:07:48	91d 2h 25m 20s	1/2	OK - load average: 2.12, 2.14, 2.09
	check GE scheduling time	OK	2015-03-10 14:06:19	5d 20h 59m 24s	1/2	Current pass duration : 4s
	check GE services	OK	2015-03-10 14:07:25	4d 5h 21m 18s	1/2	DAEMONS OK - All grid engine daemons are running
	check VarRun partition	OK	2015-03-10 13:47:50	91d 3h 30m 25s	1/2	DISK OK - free space: /var 2495 MB (66% inode=87%):
	check backup partition	OK	2015-03-10 13:53:15	91d 3h 25m 0s	1/2	DISK OK - free space: /backup 16237 MB (79% inode=99%):
	check failover procedure	OK	2015-03-10 14:02:27	91d 2h 25m 57s	1/2	MASTER OK - Current grid engine qmaster is the expected one
	check file hierarchy	CRITICAL	2015-03-10 14:05:43	6d 5h 10m 51s	2/2	file /opt/sgel/ccin2p3/common/accounting is no more a link to /afs/in2p3.fr/common/uge/prod/common/accounting
	check local SPOOL partition	OK	2015-03-10 13:53:17	91d 3h 24m 55s	1/2	DISK OK - free space: /var/spool/sgel 18623 MB (97% inode=99%):
	check network traffic	OK	2015-03-10 13:57:28	91d 2h 28m 12s	1/2	NETWORK USAGE OK - (172 rKB/s, 687 wKB/s)

- Main system parameters and activity by Kibana



- **Pros of UGE**

- **Many functionalities :**
 - To distribute fairly resources
 - To regulate load on storage services (RQS, complexes)
 - To optimize cluster CPU usage
- Patch produced for our environment
- **Failover procedures** (using a shadow master or a cluster of shadows)
- **Stable service** (spooling in Postgres DB)
- **Accounting** (Ended jobs) in Postgres DB (ARCO)

- **Pros of UGE**

- **Good support** : reactivity and user needs
- **Visibility of the roadmap**, scalable software
- Regular bug corrections, new releases, improvements (thread RO)
- Active community (user forums, webinars)

- **Cons of UGE**

- A few incidents related to overload and bugs
- Can be disturbed by a large amount of short jobs **(better in 8.2.1?)**
- Very few tools to debug when the system gets in trouble
- Not possible to mix sequential and muticore jobs in the same @hostgroup
- In some cases (known bug) : lost of running jobs
- **RFE not yet implemented :**
 - Merge of qacct, qstat commands
 - Reject job submission when encountering impossible resource requirement specification
 - A way to grant a minimum of running jobs per user
 - Change task priority of an array job

- **Detection**

- **Nagios probe monitoring**

- GE main components (qmaster, shadow, scheduler...)
- GE global configuration
- Queue status (disabled, errors queue)
- The production (jobs short, failing ..)

- **Monitoring of the Master node (SMURF)**

- CPU, memory usage
- Disk I/O, filesystem, network, processes, load

- **Kibana portal**

- Checking GE main parameters (scheduling time, dispatched jobs...)
- GE global load activity (Jobs R&Q, qacct, qstat requests)

- **SYMODO**

- Produce information about pathological jobs

- **Solutions**

- Delete the failing jobs and lock the corresponding user account
- Remove problematic workers from production
- Restart completely the service
- Send request to support in case of misbehaviour

- **Cloud Integration in GE with UniCloud software**

- UniCloud tested in March 2014
- Possibilities tested :
 - Virtual machines as workers
 - Worker instantiation on the fly
- The integration to puppet is difficult (need a dedicated server)
- More suitable to deploy private CLOUD
- Poorly documented



- **Cloud integration in the batch with htcondor**

- Currently tested with ATLAS simulation jobs
 - It works : jobs are spawned on VM nodes by the scheduler
 - 20% overhead raised on the system

