MySched

Andreas Gellrich

Scientific Computing Group Meeting
12 July 2012, DESY

HELMHOLTZ ASSOCIATION

DESY

DESY Grid Center: Grid + NAF

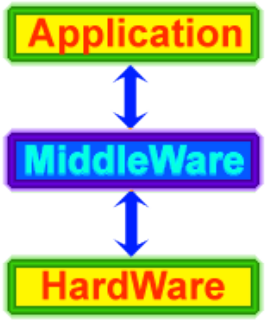
- Grid computing started at DESY in 2004:
  - DESY is the *home* of 10 VOs (site: DESY-HH)
  - (WLCG-)Tier-2 for ATLAS, CMS, and LHCb in Germany (Tier-1: GridKa)
  - HERMES / H1 / ZEUS; ILC / CALICE;
  - BELLE2; IceCube; BIOMED / W-ENMR

- One *complete generic* Grid infrastructure for *all* VOs
  - *Federated* resources w/ *opportunistic* usage ("*everybody profits*")
  - Flexible and scalable to support new VOs
  - Roughly 2/3 of the resources are currently used by the Tier-2 VOs

- Grid is *complemented* by the National Analysis Facility (NAF) [size: ~1 Tier-2]

DESY Grid Center: Resources at DESY-HH

- The Grid infrastructure is the largest Linux installation at DESY
- Grid services: (Core servers)
  - Servers: ~50
  - OS: SL 5.5/64-bit (x86_64)



- Computing Resources: (Computing Elements)(CE)
  - Compute nodes: 370 hosts, 808 procs, 3504 cores, 4784 job slots
    2GB RAM/slot, 15GB scratch space/slot
  - Processing power: ~38 kHEPSPEC
  - OS: SL 5.5/64-bit (x86_64)

- (Disk) Storage Resources: (Storage Elements)(SE)
  - Total: 4300 TB

Job classification

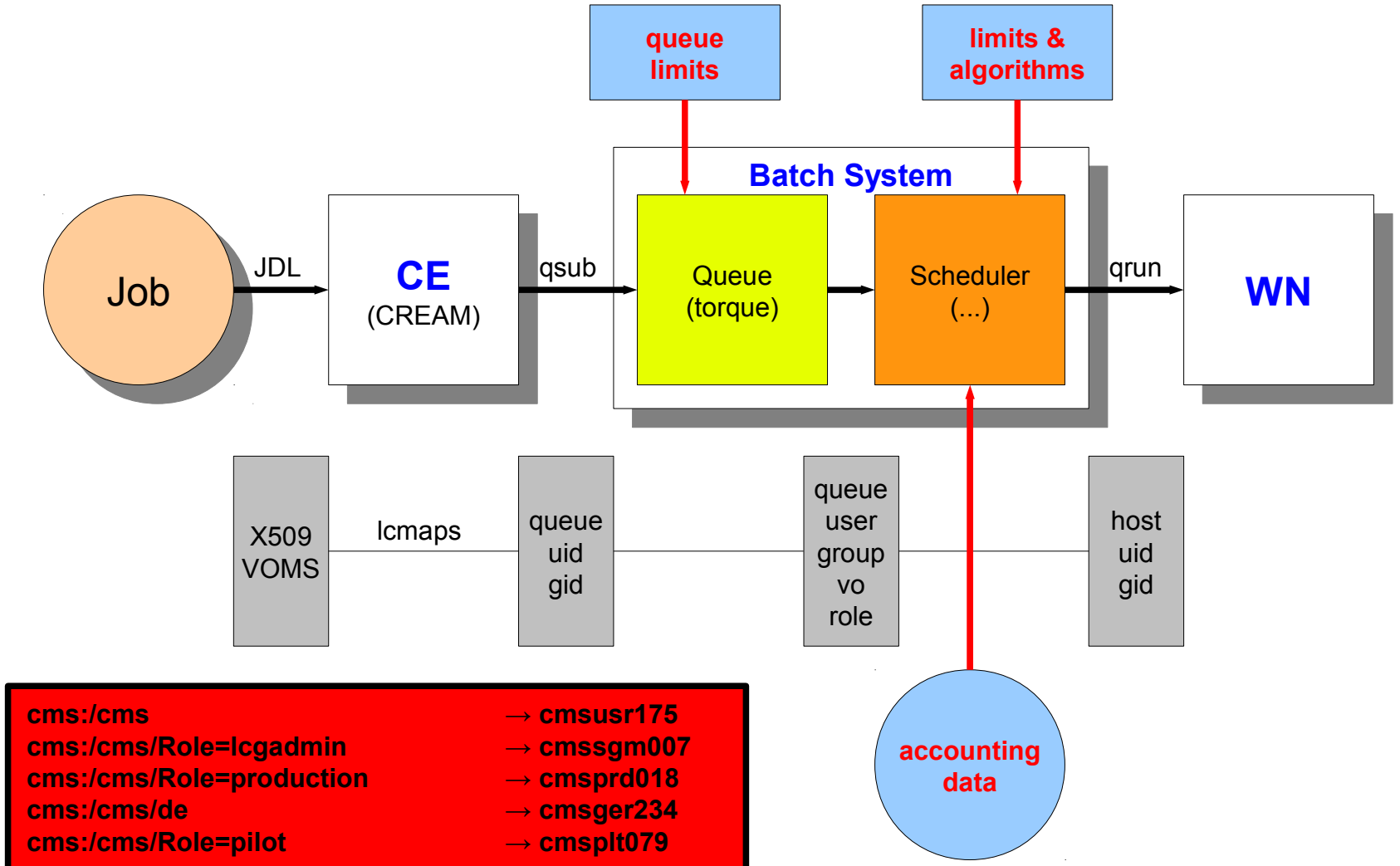
- Computing requests in HEP are parallelized on job-level - not in the application
- HEP jobs are independent and self-contained and can be treated individually

- wrsp. to resource requirements jobs can be classified:
  - Monte Carlo jobs are CPU-dominated
  - Analysis jobs are I/O-dominated and massively use the local disk

- Jobs contain the submitter's credentials as X509 VOMS-proxies
- VOMS-proxies are mapped to UID/GID on the Computing Elements (CE)

- The accounts are the key to distinguish and handle job classes

User credential mapping



Computing resources

- At Grid sites computing resources are deployed in *batch farms*
- The farm node are called *Worker Nodes* (WN)
- From the batch system point of view the Grid is a simple '*qsub*'

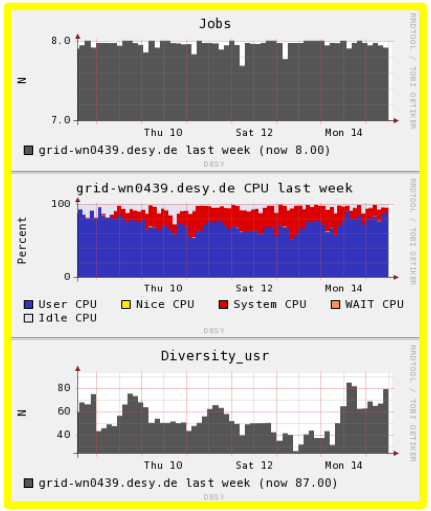
- WNs may be real or *virtual machines* (VM)
- WNs provide CPU-cores, memory, disk scratch space, a network link
- For the batch system each WN provides job slots; usually 1 slot per core

- DESY-HH WNs are heterogeneous reaching from 1 – 48 cores per WN with per slot >=2GB RAM, 20GB scratch space, and 1 GB per WN

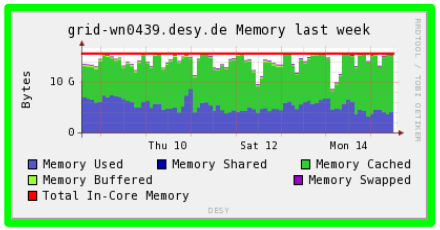
- single resources per WN must not be exhausted

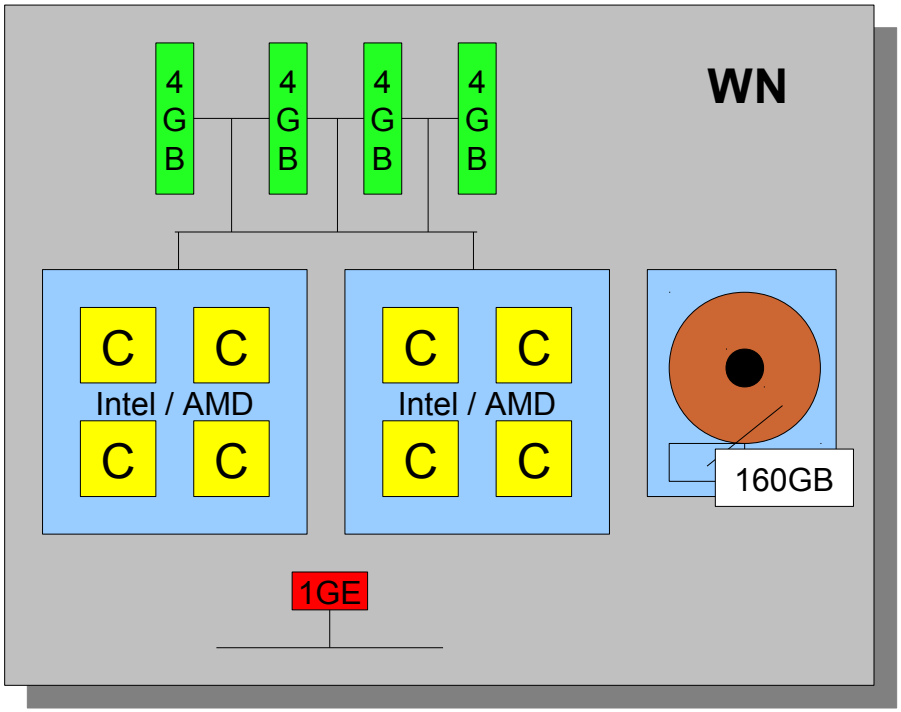
DESY

Computing resources cont'd



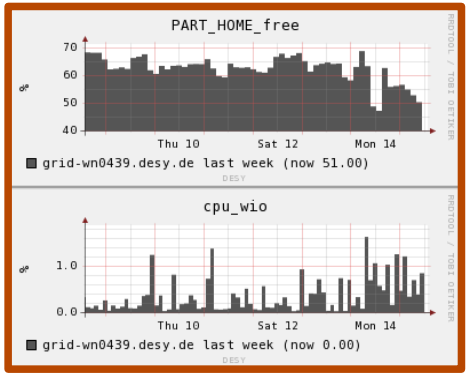
2 GB mem / slot

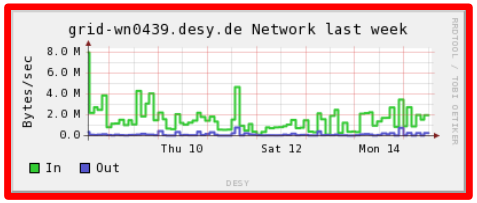




8 cores

20 GB scratch / job

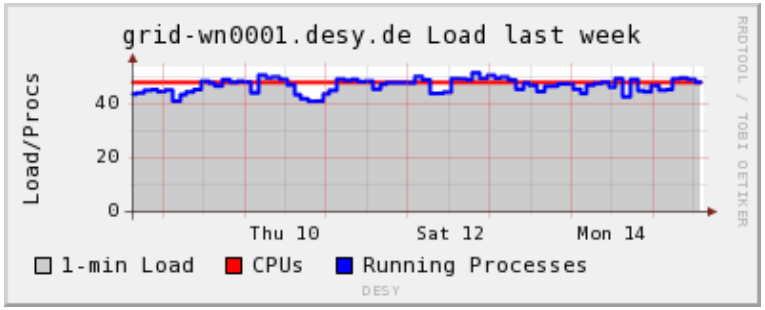


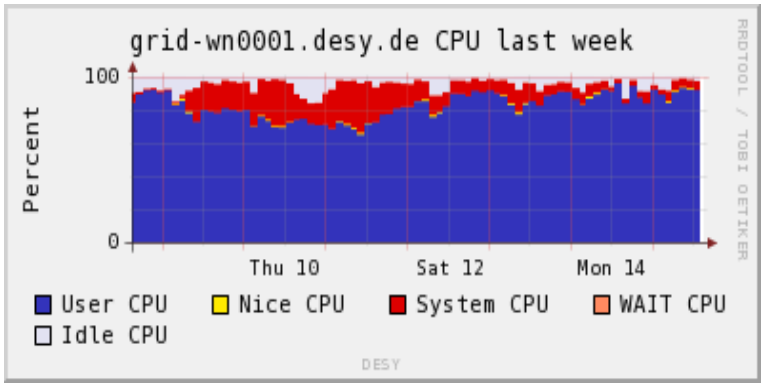


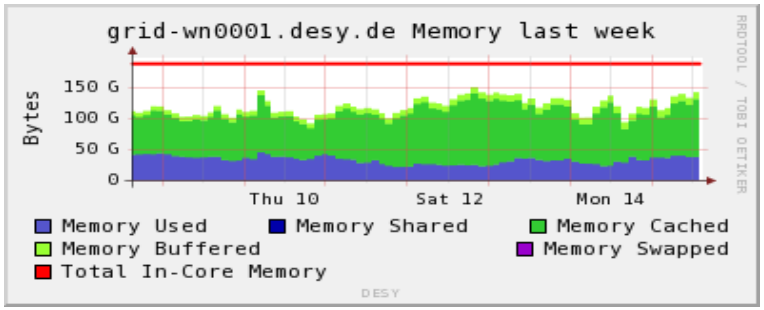
1 GE / 8 jobs

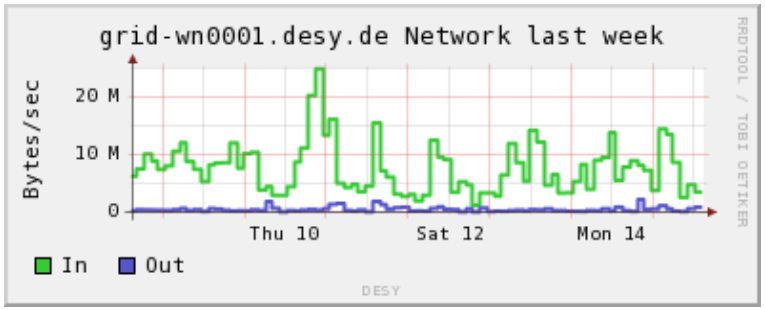
Computing resources cont'd

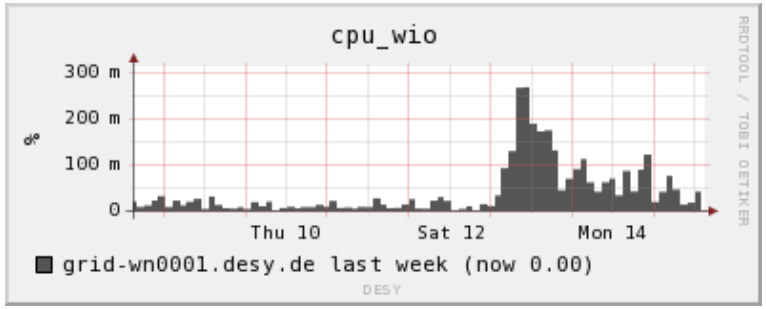
48-core WN











Queuing and scheduling

- Each batch system consists of a *queuing system* and a *scheduler*

- Jobs are enqueued by the CEs according to requirements (queue)
- Limits and rules may be individually applied to queues

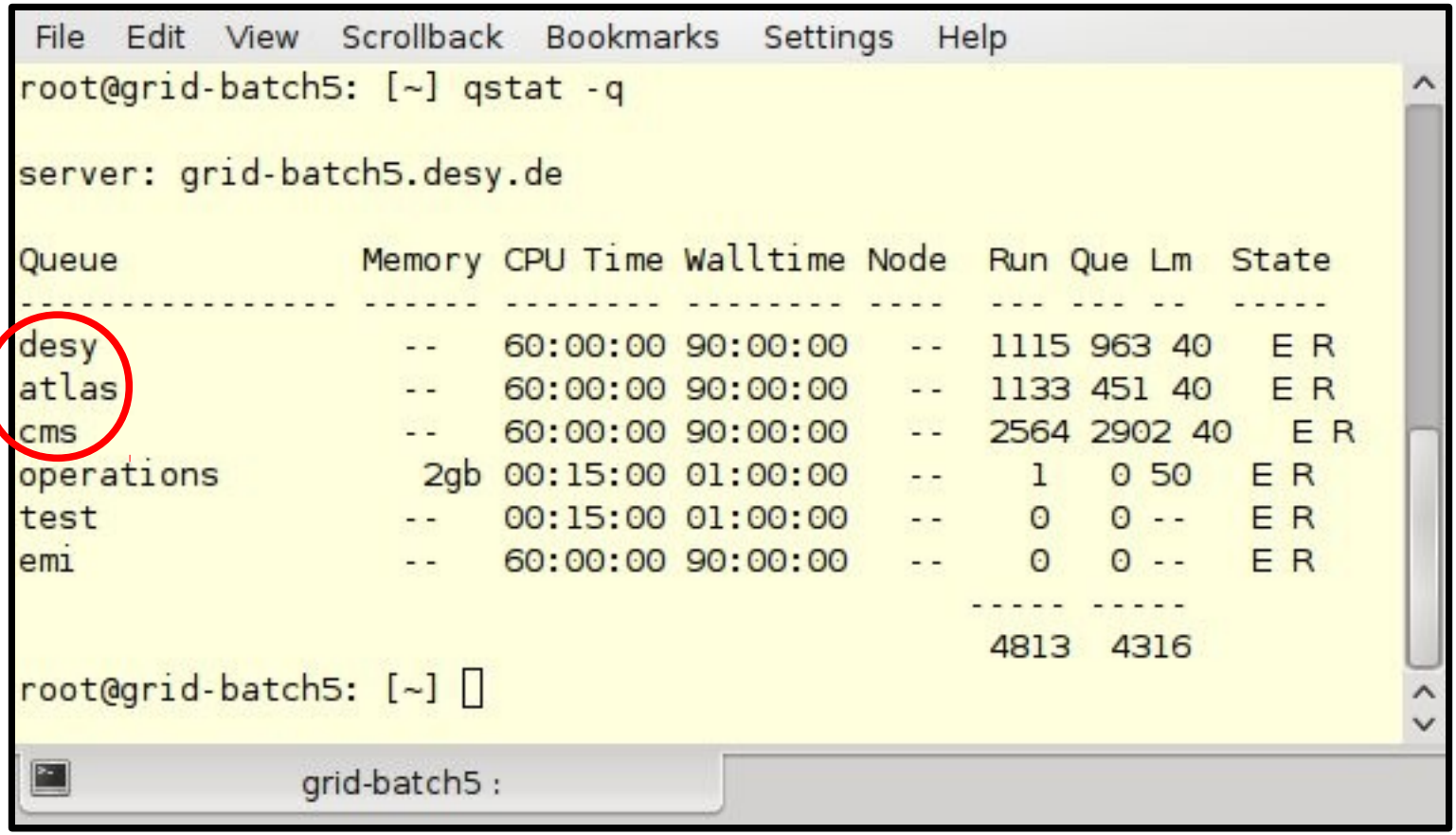
- The scheduler picks jobs of the queues and starts them on the WNs
- The scheduler bases the job distribution on configurable rules and limits
- The scheduler is the key to optimal utilization of computing resources

- gLite/EMI supports among others PBS/torque and maui
- DESY-HH: torque-2.5.7-7 / maui-3.2.6-p21

DESY

Queuing

- The DESY-HH batch system deploys a separate queue for each SE



Scheduling

- At DESY-HH as on many other sites maui was deployed as a scheduler

  "The Maui Scheduler is a policy engine which allows sites control over when, where, and how resources such as processors, memory, and disk are allocated to jobs. In addition to this control, it also provides mechanisms which help to intelligently optimize the use of these resources, monitor system performance, help diagnose problems, and generally manage the system."

- The scheduler maui has more features than currently needed in HEP
- Many sites have problem with the configuration …
- DESY-HH as well:
  - Instabilities
  - Blocking of submissions
  - Low occupancy
  - Configuration questions

A new scheduler (*MySched*)

- Occupancy with maui and afterwards:

Requirements to *MySched*

- Scalable (number of jobs; number of slots)
- Optimizing resource utilization (distribution of jobs to WNs)

- Light-weighted (CPU and memory usage)
- Configurable (config file)
- Tailored to HEP (job-parallel)

- Monitoring

- Limited development/maintenance time

Implementation

- To be based on the torque C-API (libtorque.so.2)
- Run in one process from a script as a cron
- Text logfile and monitoring info to a csv-file

- Base scheduling on a list of jobs and a list of nodes (WN)

- Configuration file with rules and limits

- Simple and obvious job distribution algorithm

- Monitoring via web

DESY

Limits and rules

- General:
  - jobMaxTotal (max number of jobs to be considered) [=30000]
  - jobMaxSubmit (max number of jobs to be submitted) [=1000]
  - maxNodeSub (max number of jobs to be submitted per node) [=2]
  - musecSleep (sleep after submission in musec) [=200000]
  - ndays (usage statistics of last n*24h) [=1]
  - maxDiversity (submit to node with max user diversity) [=on]

- By-pass ('*hot*'):
  - hotVo (*hot* VO) [="ops”]
  - hotGroup (*hot* group) [="desytst"]
  - hotRole (*hot* role) [="tst"]
  - hotUser (*hot* user) [="cmsusr165"]
  - hotQueue (*hot* queue) [="emi"]

Limits and rules cont'd

- VO/group
  - enable/disable (whole VO or group)
  - max (absolute number of jobs)
  - fraction (maximal fraction of number of jobs of online slots)
  - nodemax (maximal number of jobs per node=WN)
  - nodefrac (maximal fraction of number of jobs per node)
  - Type (meet type of node)
  - Share (relative usage with past time interval)

- Node
  - enable/disable (node)
  - type (node type)
  - queues (allow list of queues)

Algorithm

- Create node list
- Create job list
- Sort job list according to share (order jobs)
  - Treat job in list
  - Check each job for limits and rules (check limits)
  - Find appropriate WN (find Node)
  - Start job on WN (submit job)
  - Update node list

- Find node:
  - Online, not-busy, free slots, vo/group/user-on-node limits
  - Take node with least occupancy or max diversity

Algorith cont'd

File Edit View Scrollback Bookmarks Settings Help

```
gellrich@zitpcx5843: [~] cat x
Ordered by job share and usage of last 72h [20120711-103512]
-------------------------------------------------------------------------
 # Group        Q&W shar/% |      R    max diff/% | used/% diff/%
-------------------------------------------------------------------------
 1 bellesgm       6      2 |      0     97   -100 |      0   -100
 2 desysgm        1      1 |      0     29   -100 |      0   -100
 3 opsusr         1      0 |      0      2   -100 |      0    -79
 4 biomedusr     29      1 |     26     48    -46 |      1    -36
 5 atlasplt       5     10 |    347    504    -31 |      9    -16
 6 atlasger       3      0 |     18     13     38 |      0      4
 7 zeususr      673      2 |    400     96    317 |      2     14
 8 cmsprd       164     18 |   1306    850     54 |     23     28
 9 cmsusr      2812      4 |   1046    189    453 |      6     41
10 hermesusr    198      2 |    490     96    410 |      3     49
11 atlasprd     486     15 |    749    730      3 |     23     54
12 iceprd       172      1 |    127     39    226 |      1     54
-------------------------------------------------------------------------
TOTAL          4550     56 |   4799                     68
-------------------------------------------------------------------------
gellrich@zitpcx5843: [~]
```

gellrich : zsh

DESY

Demo

- `root@grid-batch5: [~] less mysched.conf`
- `root@grid-batch5: [~] less mysched.log`
- `http://grid-mon1.desy.de/mysched.html`

Outlook

- It's still a study not a product!

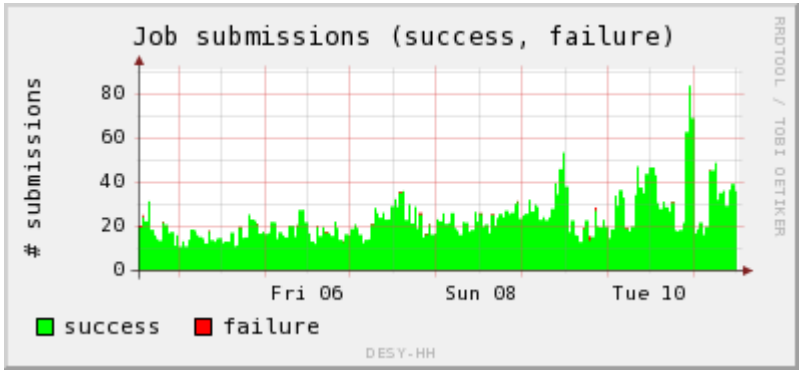
- It is hard to judge performance in details.
- Looking at inclusive data:
  - Occupancy plots
  - Timing
  - Resource utilization
- More analysis needed ...

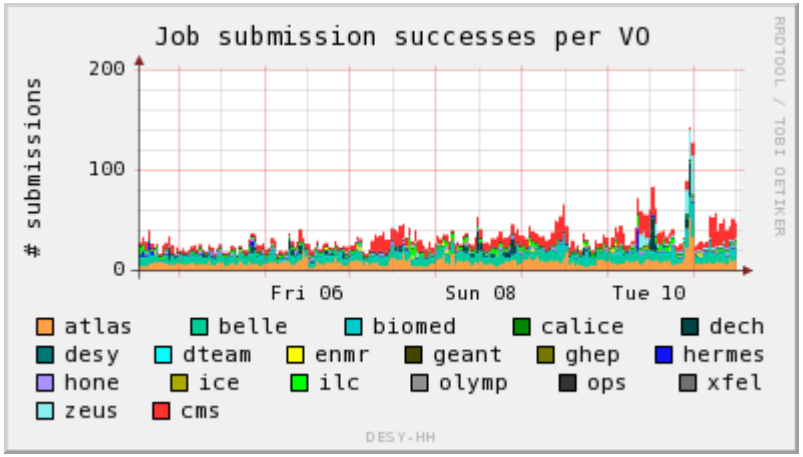
- Multi-core jobs are not treated ...

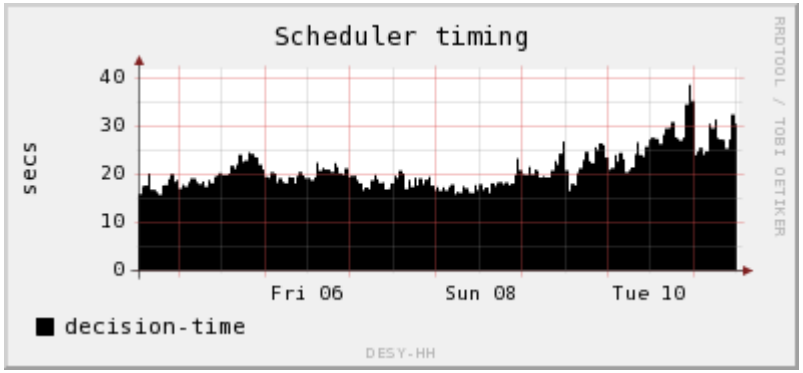
Appendix

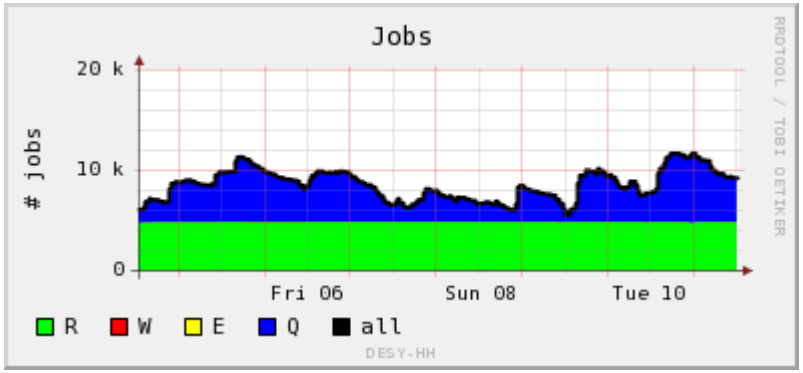
Monitoring: http://grid.desy.de/mysched.html

Statistics for 1 week:



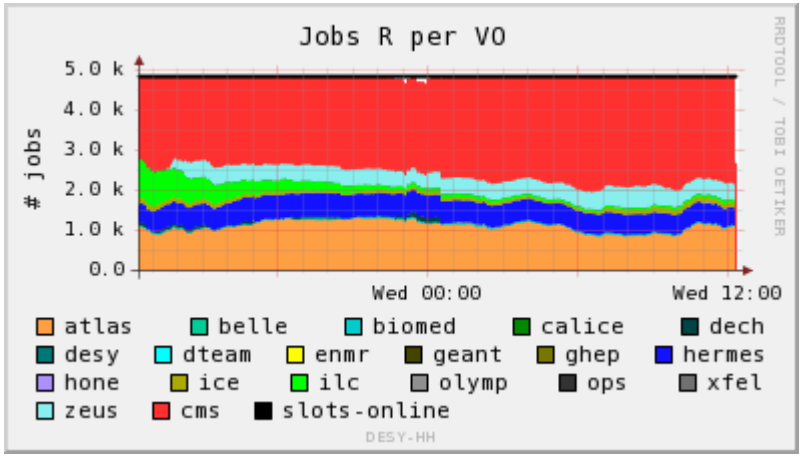


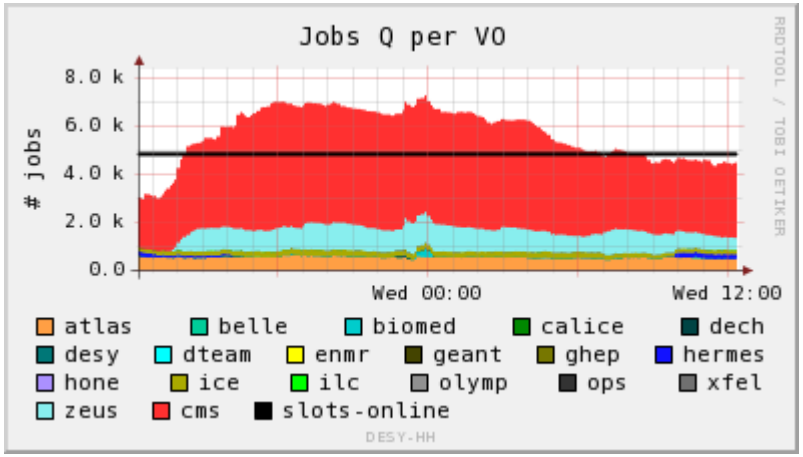




Monitoring: http://grid.desy.de/mysched.html

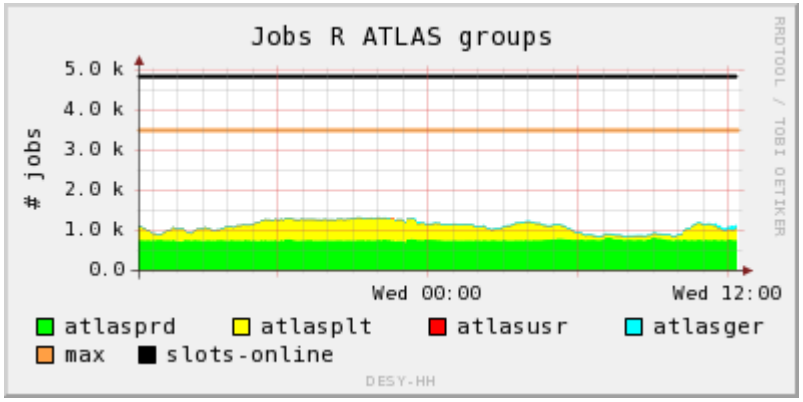
Jobs running/queuing in 24 hours:

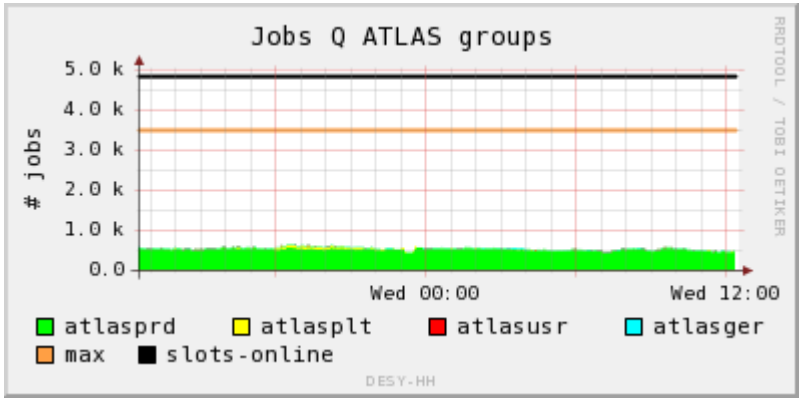


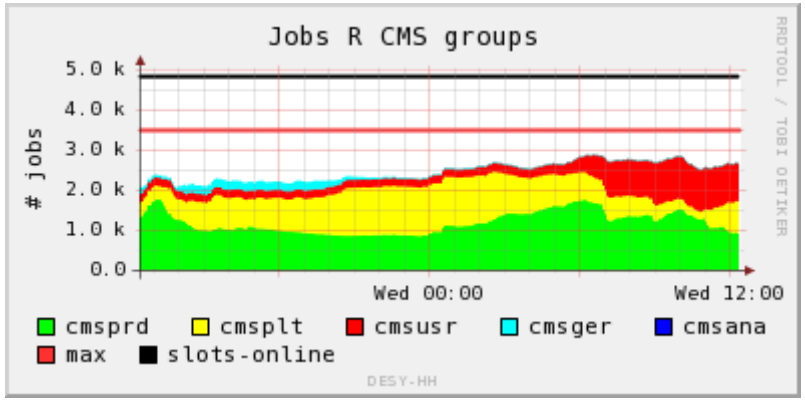


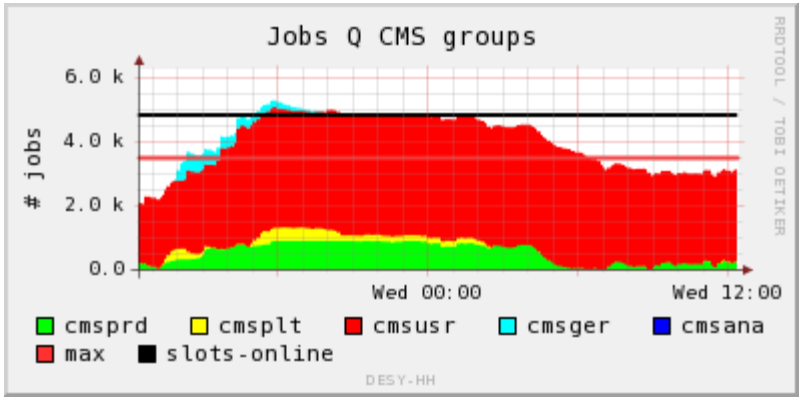
Monitoring: http://grid.desy.de/mysched.html

LHC jobs running/queuing per group in 24 hours:



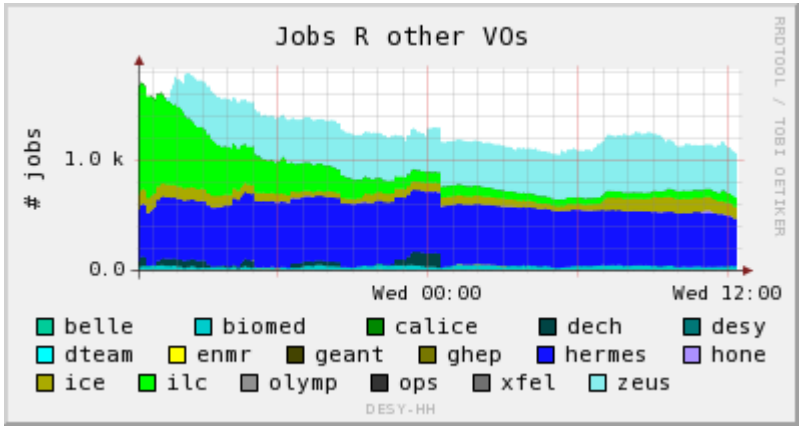


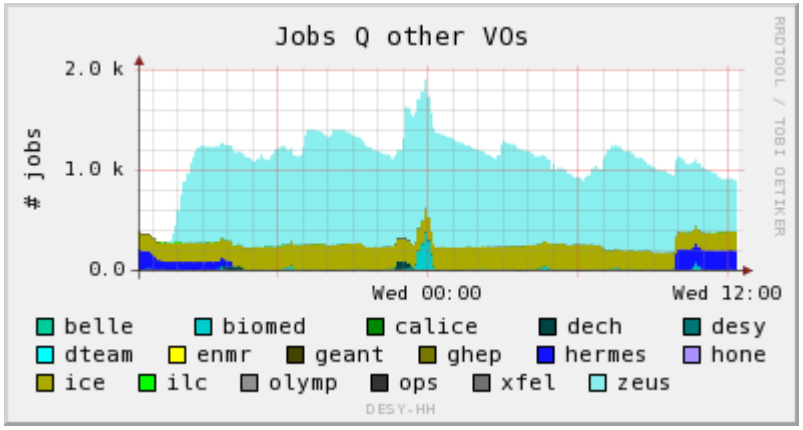




Monitoring: http://grid.desy.de/mysched.html

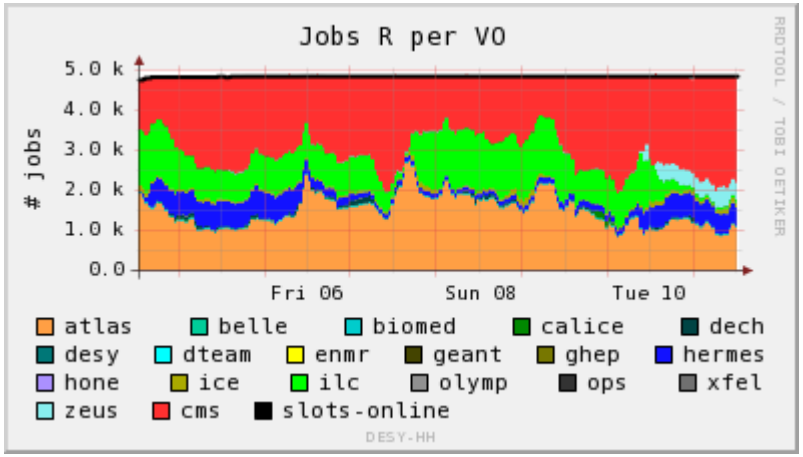
Non-LHC jobs running/queuing in 24 hours:

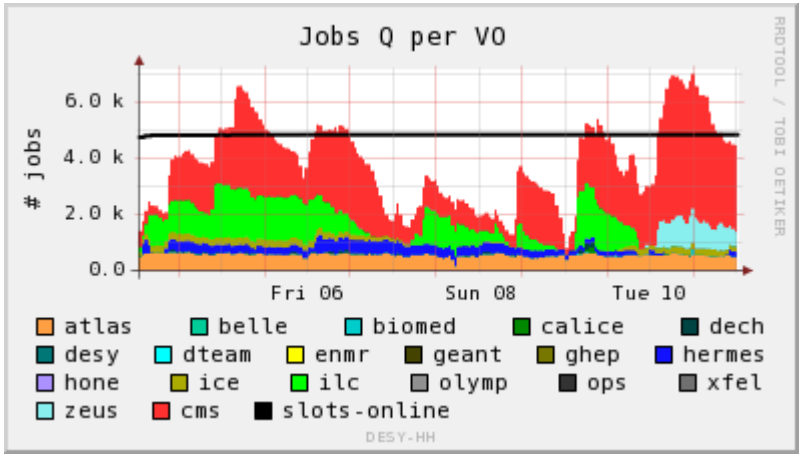




Monitoring: http://grid.desy.de/mysched.html

Jobs running/queuing in 1 week:





Monitoring: http://grid.desy.de/mysched.html

Jobs ended in 1 week: