



Nikhef Multicore Experience

Jeff Templon

Multicore TF 2014.04.01

J. Templon
Nikhef
Amsterdam
*Physics Data Processing
Group*

disclaimer

- It may be possible to do this with native torque and maui features
 - Standing reservations
 - Partitions
- Couldn't figure it out
 - Docs very poor (cases too simple)
 - Nobody answers questions
- Hence wrote something; tried to keep it ASAP (as small/simple as possible)

Summary

- With a bit of scriptology, working system
- Performance is adequate for us
- Entropy is important here too : 32 vs 8 cores per node

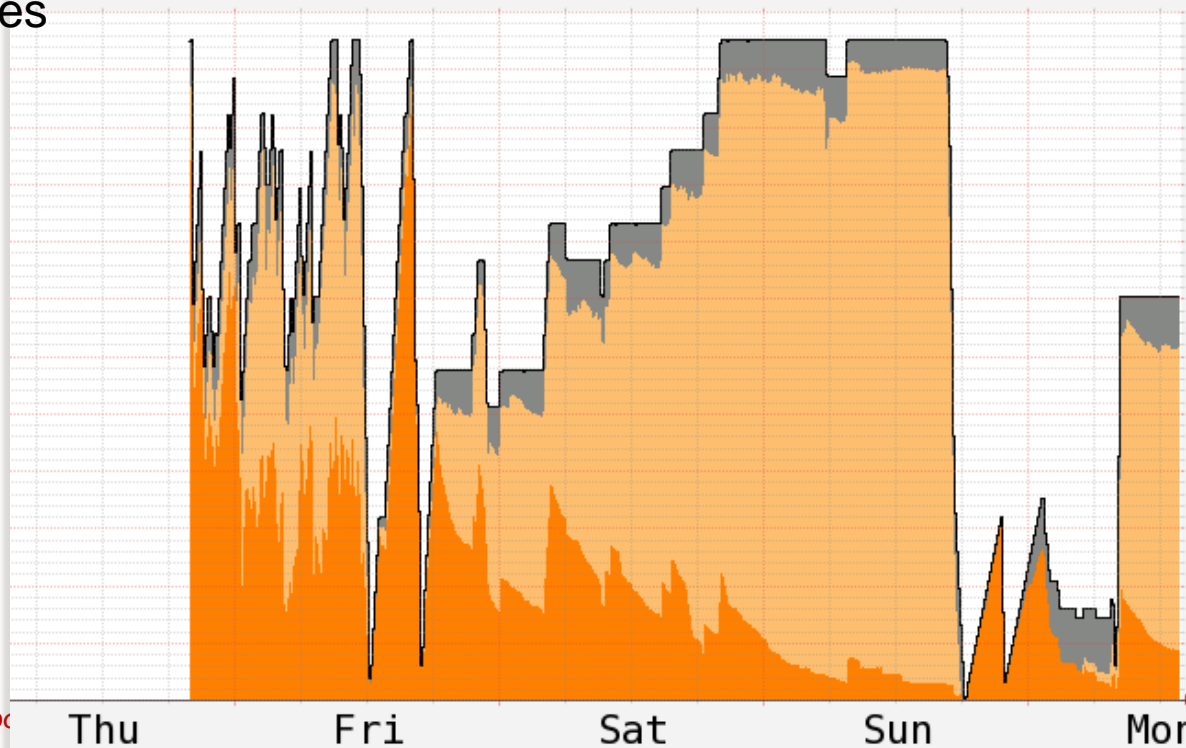
Contents

- Results
- How we did it
- Why we did it like we did
- Random observations and musings

Mc pool performance

Weekend of 29 march 2014

600 cores



Legend:
■ unused (9.7)
■ atlmc (62.0)
■ nonmc (28.3)

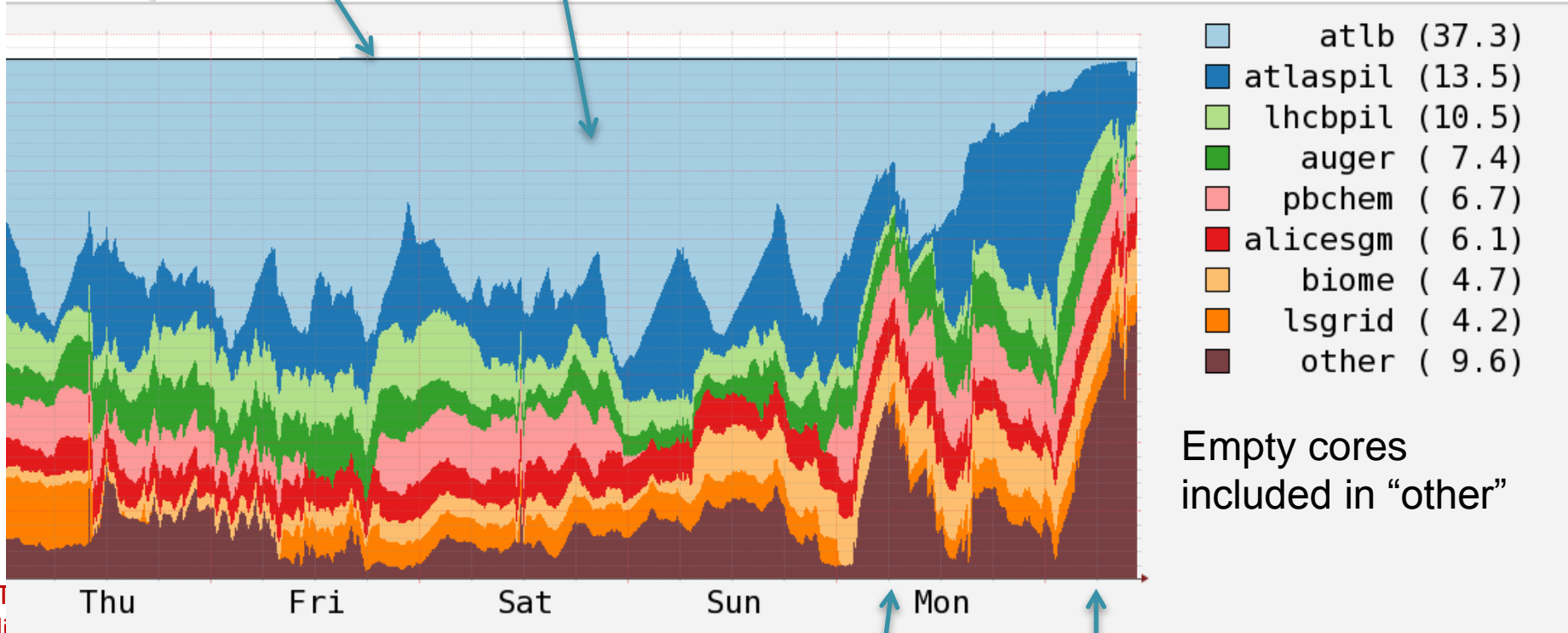
J. Templon
Nikhef
Amsterdam
Physics Data Proc
Group

Multicore jobs are 'atlb'; lots of single core atlb jobs too.

3800 cores

The whole farm

Weekend of 29 march 2014



Empty cores included in "other"

Farm > 98% full with two exceptions

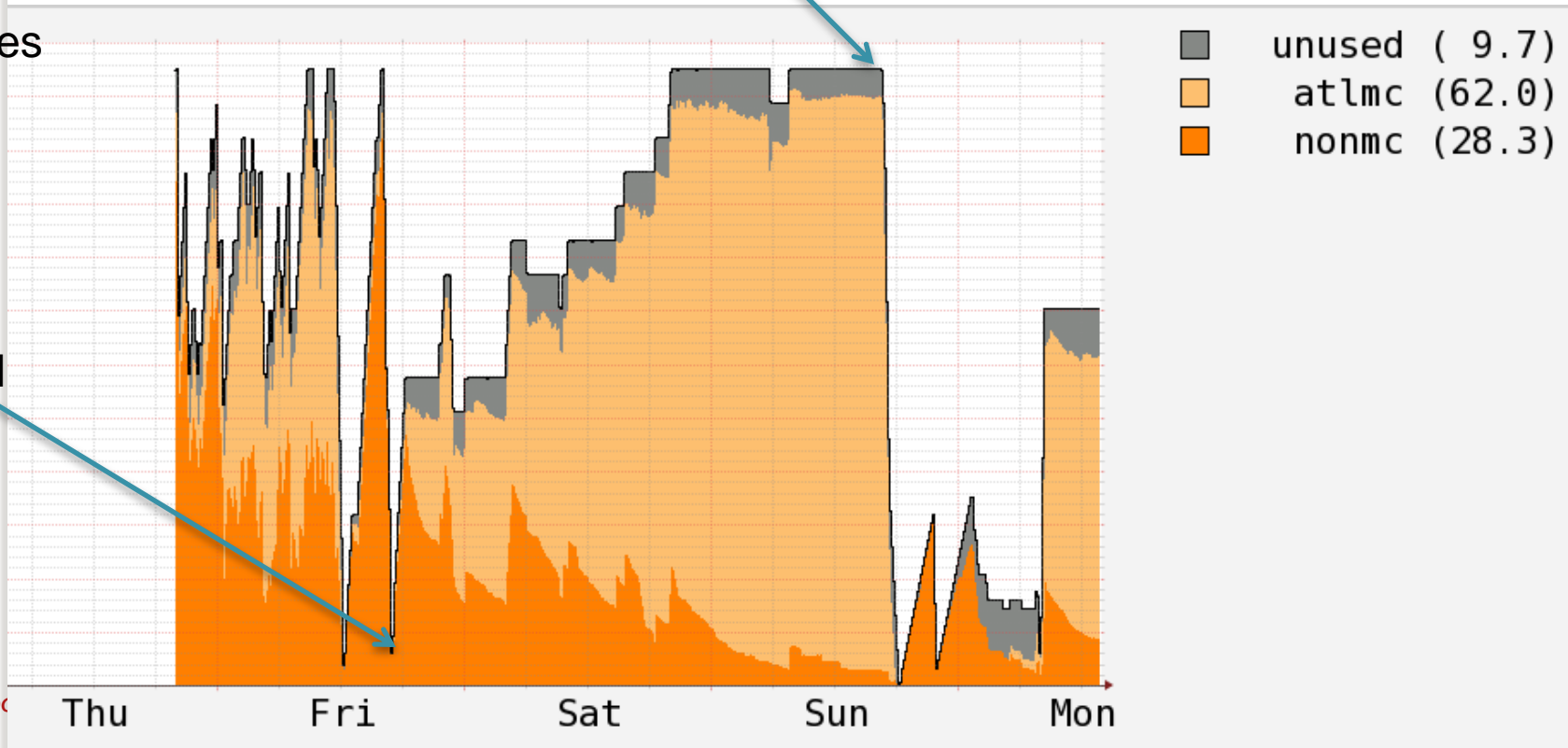
J.T
Ni
Amsterdam
Physics Data Processing
Group

48 h : grabbed all 18 nodes
Kept total farm at 98.5% occupancy
At 48 h point : 528 atlas-mc cores
21 non-mc cores, 27 unused cores

Pool max : 18 32-core boxes
576 cores total

600 cores

Started here

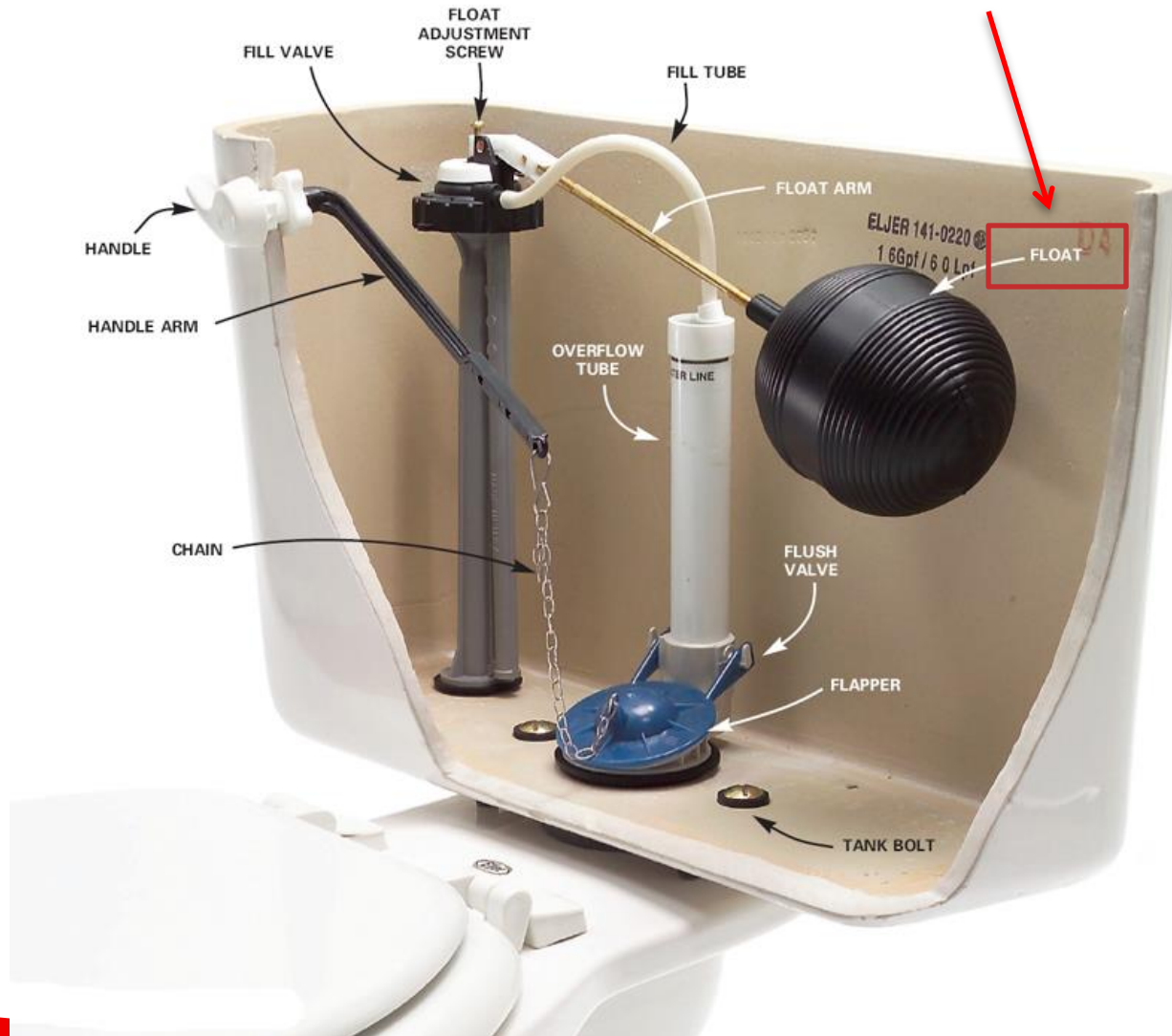


J. Templon
Nikhef
Amsterdam
Physics Data Proc
Group

How we did it

- Torque node properties:
 - Normal jobs want “el6” property
 - Make ATLAS mc jobs look for ‘mc’ property
 - Separation done by diff. queue for mc vs rest
- Small cron job that observes situation and adjusts node properties as needed

Cron job is a 'float' for the mc pool level in the 'tank' of our farm



J. Templon
Nikhef
Amsterdam
Physics Data Processing
Group

For 32-core nodes

- Soft limit of 7 nodes “draining”
- Hard limit of 49 unused slots
- If below both limits : put another node in pool
- Soft exceeded, hard not : do nothing
- Hard limit exceeded: put node(s) back in generic pool
- Every ten minutes

Node recovery

- In case no more atlasmc jobs:
 - 8(+) slots free on node: check on next run
 - 10 min later still 8+ free: add to list marked for returning to generic pool
 - Put half of marked nodes back in pool, check the other half again on next run
- Conservative : protect pool against temporary glitches.
- Worst case : 60 minutes until all slots back in generic pool

32-core case

Slow ramp as machines become 100% mc

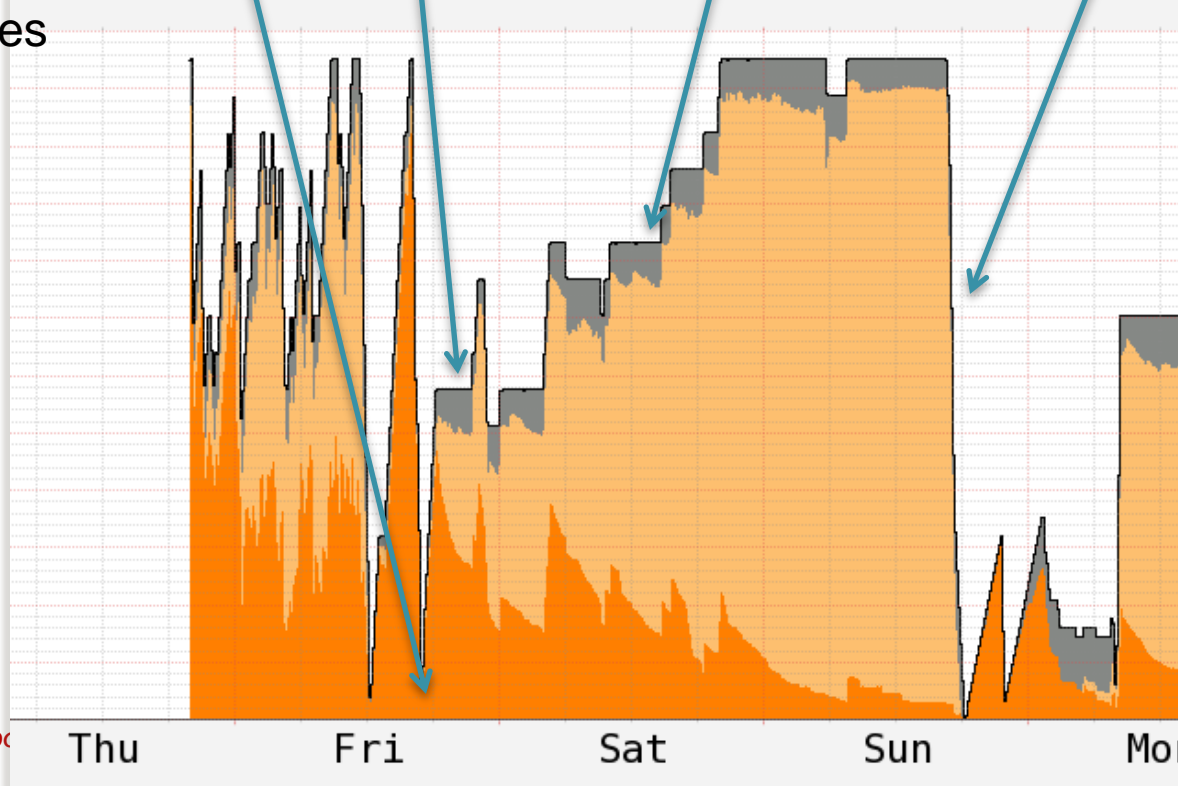
Mc pool performance

Reached 7 draining nodes

Disabled job starts : pool "flushed"

Turned on float

600 cores



Legend:
grey: unused (9.7)
light orange: atlmc (62.0)
orange: nonmc (28.3)

J. Templon
Nikhef
Amsterdam
Physics Data Proc
Group

For 8-core nodes

- Soft limit of 16 nodes “draining”
- Hard limit of 49 unused slots
- If below both limits : put another node in pool
- Soft exceeded, hard not : do nothing
- Hard limit exceeded: put node(s) back in generic pool
- Every ten minutes

8-core case

Give back nodes due to max unused Mc pool performance

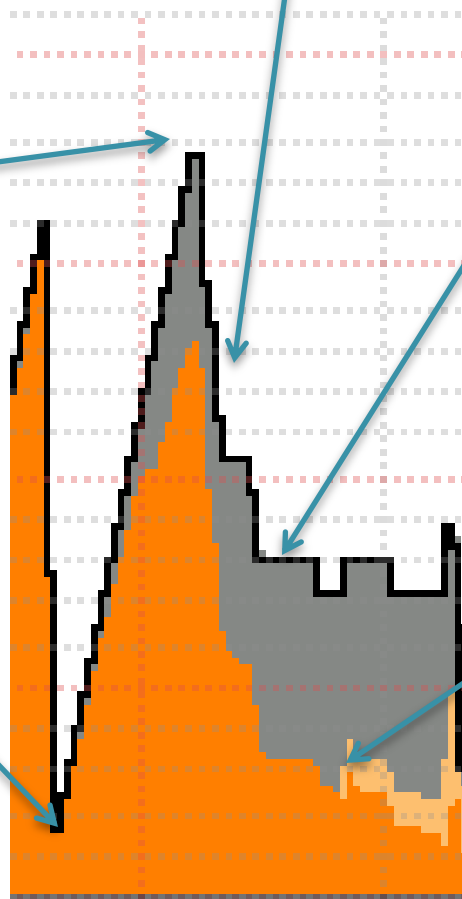
Stabilize at 10 pool nodes

Peak at 22 nodes / 176 cores

1st mc job : 7 hrs after start

Pool : 72 8-core boxes
576 cores total

Turned on float

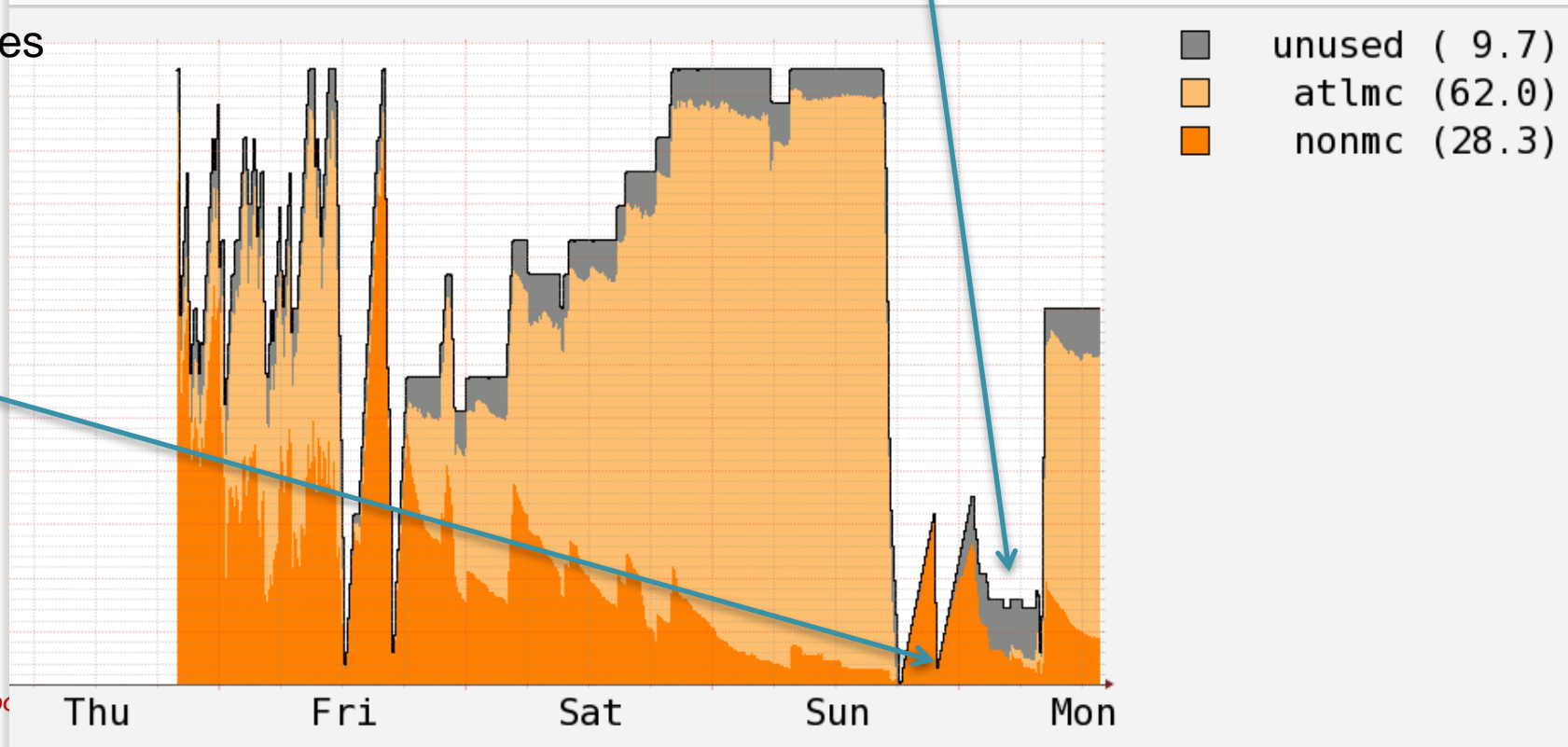


J. Templon
Nikhef
Amsterdam
Physics Data Processing
Group

10 h : grabbed 10 nodes
Kept same farm inoccupancy (49 core lim)
At 10 h point : 80-core pool subset
24 atlas-mc cores
18 non-mc cores, 38 unused cores

600 cores

Started here



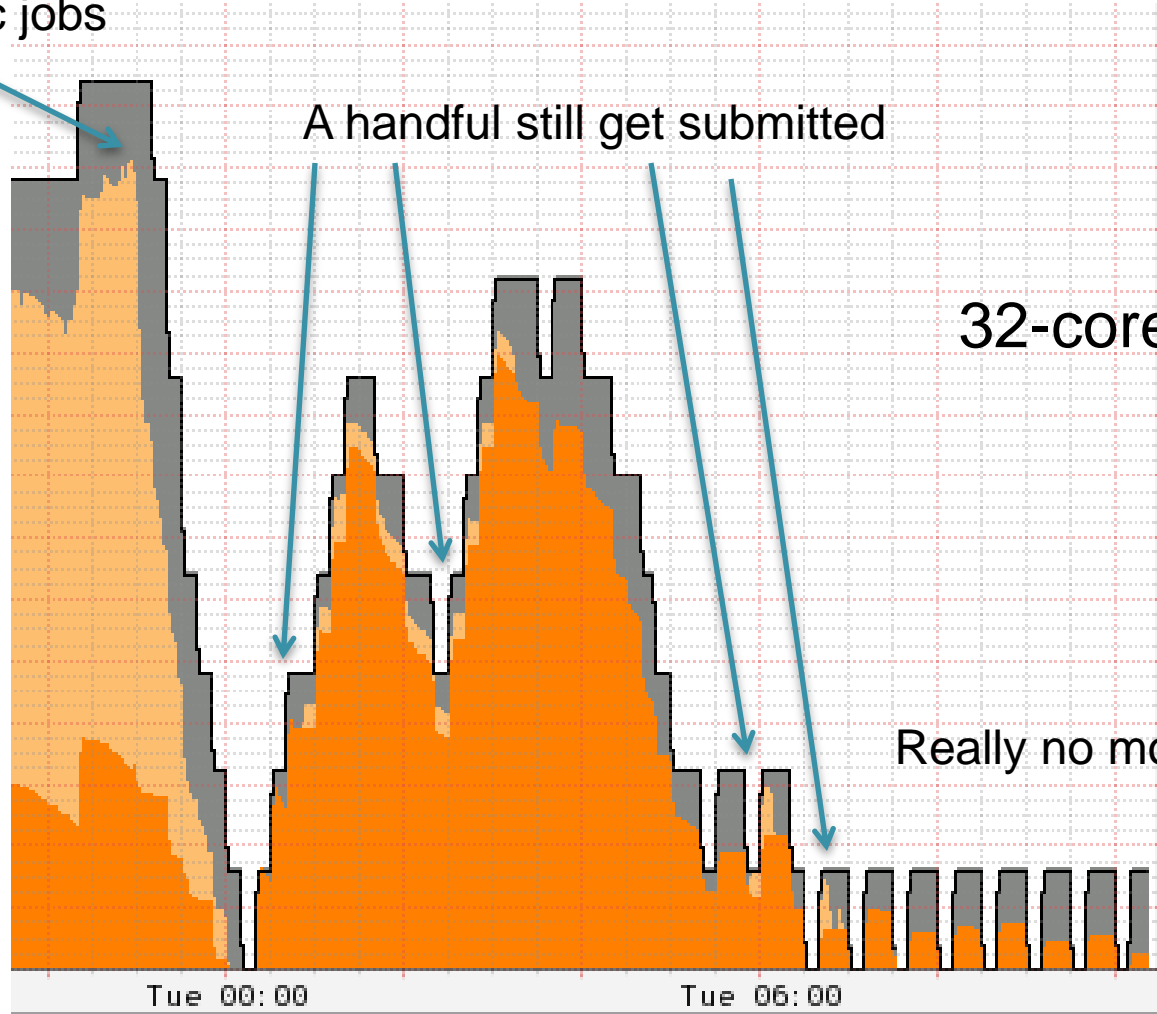
Pool : 72 8-core boxes = 576 cores total

Why we did it like this

- Separate pool : avoid the ‘ops job’ (or other higher prio job) takes 1 of my 8 slots and destroys ‘mc slot’
- Floating pool boundary w/ policies for filling and draining the tank:
 - Avoid too many empty slots during filling
 - Avoid empty slots if supply of mc jobs consistently (10+ minutes) dries up
 - Protect against short stops (eg maui server restart!)

Holding the handle down

No more waiting mc jobs



32-core case

Really no more waiting

J. Templon
Nikhef
Amsterdam
Physics Data Processing
Group

32 vs 8 core (I)

- 8 core case : 1st mc job after 7 hrs
- 32 core case : 1st mc job after 16 minutes
- ***Not just 4 times faster***
- Entropic effect
 - 8 cores: all cores of box must free up
 - 32 cores : 8 of 32 cores must free up
 - Time to free up entire box ~ same 8 or 32
 - Much shorter time to free up 1/4 of box

32 vs 8 core (II)

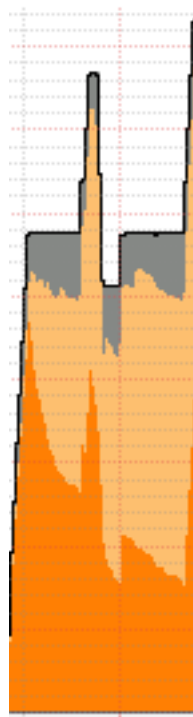
- 8 core case : in 7 hr, 1 mc job (8 cores)
- 32 core case : in 7 hr, 18 mc jobs (144 cores)

4 times container size
18 times the result

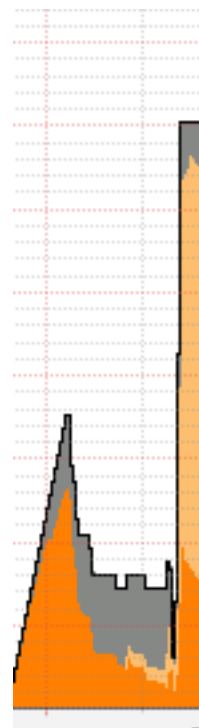
32 vs 8 core

- 32 core case clearly easier due to entropy
- Sites where all machines 8 or fewer cores?
 - Review cycle time ... much more conservative on giving back a drained node
 - Probably more aggressive node acquisition
 - Semi-dedicated manual system might be better





VS



Random musings

- System is not at all smart
 - By design ... being smart is really hard, esp with bad information
 - Often a simple (passively stable) system has vastly better “poor behavior” and only slightly worse “optimal behavior”
- Still some simple tweaks possible
 - E.g. give back most recent grabbed node instead of random one

Check your own farm

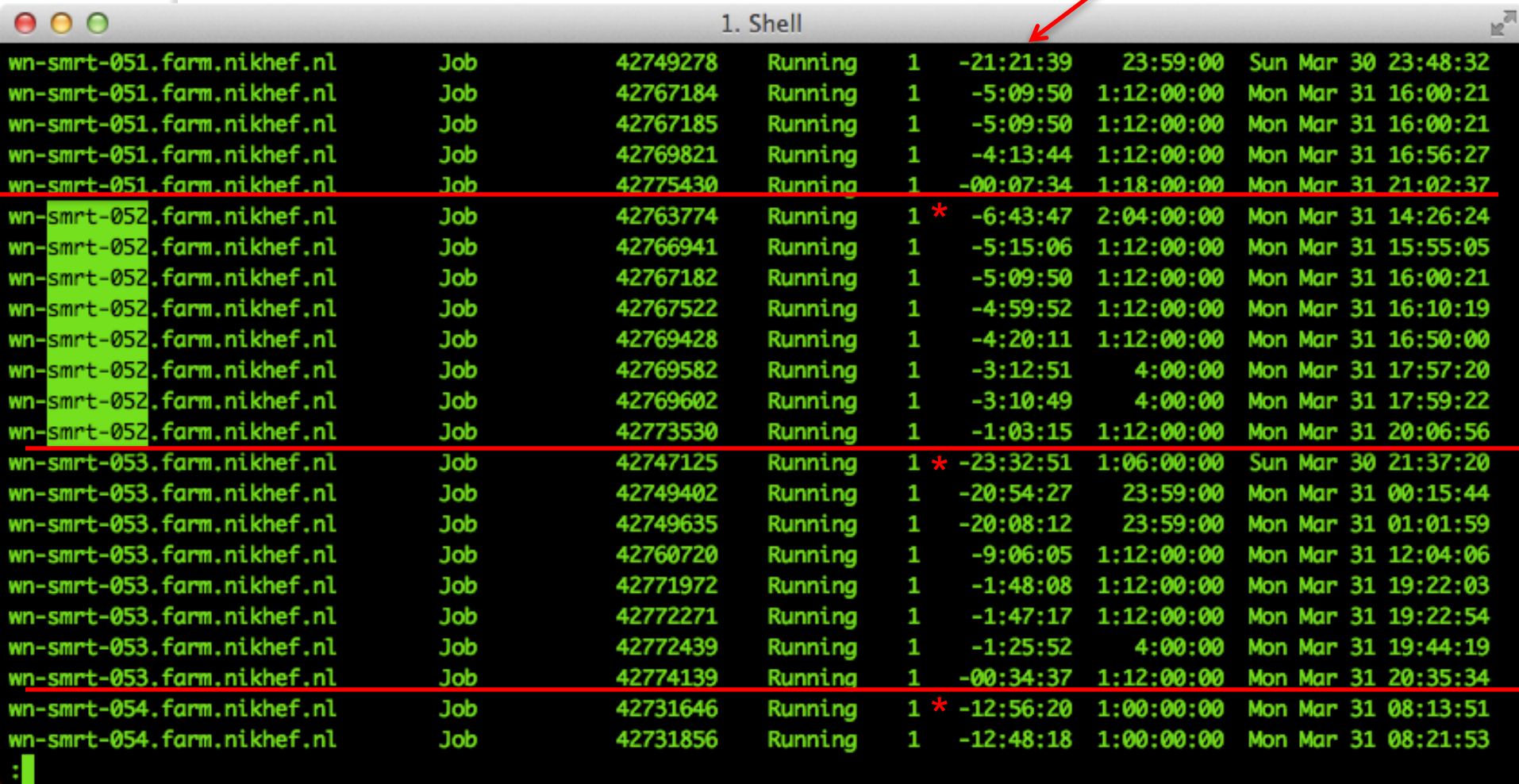
- **The** MC Question : if I don't start any new jobs, which current job will end last?
- Statistically, same as question : which current job started first? *Fully utilized node looks same both directions in time*
- Can answer 2nd question with maui
 - `showres -n | grep Running | sort -k1 -k6nr`
- Careful with sort order (not always ok)

Start to drain all 147 now :
1st start in about 7 hrs

8 core nodes

How long ago job started

• Of 147 'smrt' class 8-core nodes, the best:



1. Shell

| | | | | | | | |
|----------------------------|-----|----------|---------|-----|-----------|------------|---------------------|
| wn-smrt-051.farm.nikhef.nl | Job | 42749278 | Running | 1 | -21:21:39 | 23:59:00 | Sun Mar 30 23:48:32 |
| wn-smrt-051.farm.nikhef.nl | Job | 42767184 | Running | 1 | -5:09:50 | 1:12:00:00 | Mon Mar 31 16:00:21 |
| wn-smrt-051.farm.nikhef.nl | Job | 42767185 | Running | 1 | -5:09:50 | 1:12:00:00 | Mon Mar 31 16:00:21 |
| wn-smrt-051.farm.nikhef.nl | Job | 42769821 | Running | 1 | -4:13:44 | 1:12:00:00 | Mon Mar 31 16:56:27 |
| wn-smrt-051.farm.nikhef.nl | Job | 42775430 | Running | 1 | -00:07:34 | 1:18:00:00 | Mon Mar 31 21:02:37 |
| wn-smrt-052.farm.nikhef.nl | Job | 42763774 | Running | 1 * | -6:43:47 | 2:04:00:00 | Mon Mar 31 14:26:24 |
| wn-smrt-052.farm.nikhef.nl | Job | 42766941 | Running | 1 | -5:15:06 | 1:12:00:00 | Mon Mar 31 15:55:05 |
| wn-smrt-052.farm.nikhef.nl | Job | 42767182 | Running | 1 | -5:09:50 | 1:12:00:00 | Mon Mar 31 16:00:21 |
| wn-smrt-052.farm.nikhef.nl | Job | 42767522 | Running | 1 | -4:59:52 | 1:12:00:00 | Mon Mar 31 16:10:19 |
| wn-smrt-052.farm.nikhef.nl | Job | 42769428 | Running | 1 | -4:20:11 | 1:12:00:00 | Mon Mar 31 16:50:00 |
| wn-smrt-052.farm.nikhef.nl | Job | 42769582 | Running | 1 | -3:12:51 | 4:00:00 | Mon Mar 31 17:57:20 |
| wn-smrt-052.farm.nikhef.nl | Job | 42769602 | Running | 1 | -3:10:49 | 4:00:00 | Mon Mar 31 17:59:22 |
| wn-smrt-052.farm.nikhef.nl | Job | 42773530 | Running | 1 | -1:03:15 | 1:12:00:00 | Mon Mar 31 20:06:56 |
| wn-smrt-053.farm.nikhef.nl | Job | 42747125 | Running | 1 * | -23:32:51 | 1:06:00:00 | Sun Mar 30 21:37:20 |
| wn-smrt-053.farm.nikhef.nl | Job | 42749402 | Running | 1 | -20:54:27 | 23:59:00 | Mon Mar 31 00:15:44 |
| wn-smrt-053.farm.nikhef.nl | Job | 42749635 | Running | 1 | -20:08:12 | 23:59:00 | Mon Mar 31 01:01:59 |
| wn-smrt-053.farm.nikhef.nl | Job | 42760720 | Running | 1 | -9:06:05 | 1:12:00:00 | Mon Mar 31 12:04:06 |
| wn-smrt-053.farm.nikhef.nl | Job | 42771972 | Running | 1 | -1:48:08 | 1:12:00:00 | Mon Mar 31 19:22:03 |
| wn-smrt-053.farm.nikhef.nl | Job | 42772271 | Running | 1 | -1:47:17 | 1:12:00:00 | Mon Mar 31 19:22:54 |
| wn-smrt-053.farm.nikhef.nl | Job | 42772439 | Running | 1 | -1:25:52 | 4:00:00 | Mon Mar 31 19:44:19 |
| wn-smrt-053.farm.nikhef.nl | Job | 42774139 | Running | 1 | -00:34:37 | 1:12:00:00 | Mon Mar 31 20:35:34 |
| wn-smrt-054.farm.nikhef.nl | Job | 42731646 | Running | 1 * | -12:56:20 | 1:00:00:00 | Mon Mar 31 08:13:51 |
| wn-smrt-054.farm.nikhef.nl | Job | 42731856 | Running | 1 | -12:48:18 | 1:00:00:00 | Mon Mar 31 08:21:53 |

32 core nodes

Whole node takes a day : only need 8 cores to get 1st job

| Job ID | Job Name | Job ID | Status | Cores | Start Time | End Time | Start Date | End Date | Start Time | End Time |
|----------------------------|----------|----------|---------|-------|------------|------------|------------|----------|------------|----------|
| wn-knal-007.farm.nikhef.nl | Job | 42769764 | Running | 1 | -4:55:57 | 1:12:00:00 | Mon Mar 31 | 16:36:33 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42769785 | Running | 1 | -4:37:46 | 1:12:00:00 | Mon Mar 31 | 16:54:44 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42769798 | Running | 1 | -4:37:17 | 1:12:00:00 | Mon Mar 31 | 16:55:13 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42769837 | Running | 1 | -4:35:15 | 1:12:00:00 | Mon Mar 31 | 16:57:15 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42769920 | Running | 1 | -4:32:06 | 1:12:00:00 | Mon Mar 31 | 17:00:24 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42769943 | Running | 1 | -4:46:12 | 1:18:00:00 | Mon Mar 31 | 16:46:18 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42771558 | Running | 1 * | -3:42:05 | 1:12:00:00 | Mon Mar 31 | 17:50:25 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42772782 | Running | 1 | -2:37:20 | 2:04:00:00 | Mon Mar 31 | 18:55:10 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42774142 | Running | 1 | -1:39:04 | 2:04:00:00 | Mon Mar 31 | 19:53:26 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42774158 | Running | 1 | -1:36:04 | 1:18:00:00 | Mon Mar 31 | 19:56:26 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42773517 | Running | 1 | -1:25:08 | 1:12:00:00 | Mon Mar 31 | 20:07:22 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42773598 | Running | 1 | -1:24:17 | 1:12:00:00 | Mon Mar 31 | 20:08:13 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42774103 | Running | 1 | -1:00:03 | 1:12:00:00 | Mon Mar 31 | 20:32:27 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42774135 | Running | 1 | -00:57:47 | 1:12:00:00 | Mon Mar 31 | 20:34:43 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42774173 | Running | 1 * | -00:52:17 | 1:12:00:00 | Mon Mar 31 | 20:40:13 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42774680 | Running | 1 | -00:48:19 | 1:12:00:00 | Mon Mar 31 | 20:44:11 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42774706 | Running | 1 | -00:46:48 | 1:12:00:00 | Mon Mar 31 | 20:45:42 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42774718 | Running | 1 | -00:41:54 | 1:12:00:00 | Mon Mar 31 | 20:50:36 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42775361 | Running | 1 | -00:38:58 | 1:18:00:00 | Mon Mar 31 | 20:53:32 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42775044 | Running | 1 | -00:17:54 | 1:12:00:00 | Mon Mar 31 | 21:14:36 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42775642 | Running | 1 | -00:04:57 | 1:18:00:00 | Mon Mar 31 | 21:27:33 | | |
| wn-knal-007.farm.nikhef.nl | Job | 42775643 | Running | 1 | -00:04:57 | 1:18:00:00 | Mon Mar 31 | 21:27:33 | | |
| wn-knal-009.farm.nikhef.nl | Job | 42749341 | Running | 1 * | -21:29:39 | 23:59:00 | Mon Mar 31 | 00:02:51 | | |
| wn-knal-009.farm.nikhef.nl | Job | 42749410 | Running | 1 | -21:15:33 | 23:59:00 | Mon Mar 31 | 00:16:57 | | |
| wn-knal-009.farm.nikhef.nl | Job | 42749458 | Running | 1 | -21:00:28 | 23:59:00 | Mon Mar 31 | 00:32:02 | | |
| wn-knal-009.farm.nikhef.nl | Job | 42749461 | Running | 1 | -20:59:58 | 23:59:00 | Mon Mar 31 | 00:32:32 | | |