PDSF Users Meeting

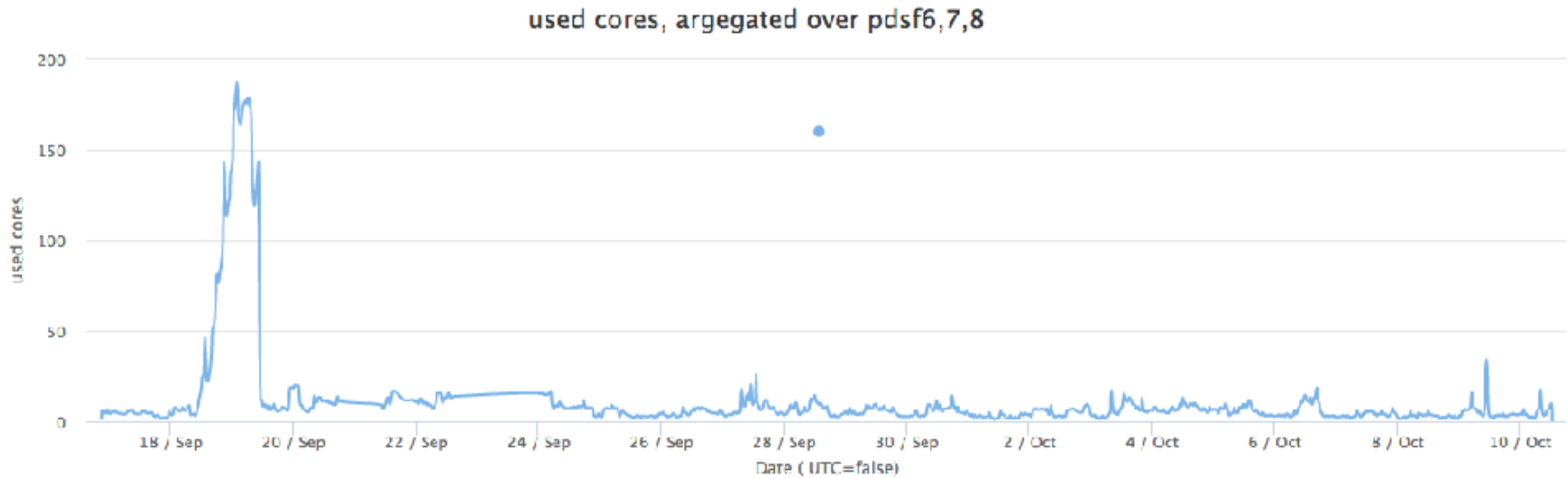
* PDSF performance
* announcements
* hardware utilization: PDSF vs. Cori
* improving PDSF performance
    - job slots vs. core count
    - benefit of hyper threading
        - fill up hyper-cores ‘as-is’ ?
* AOB

October 11, 2016

Jan Balewski

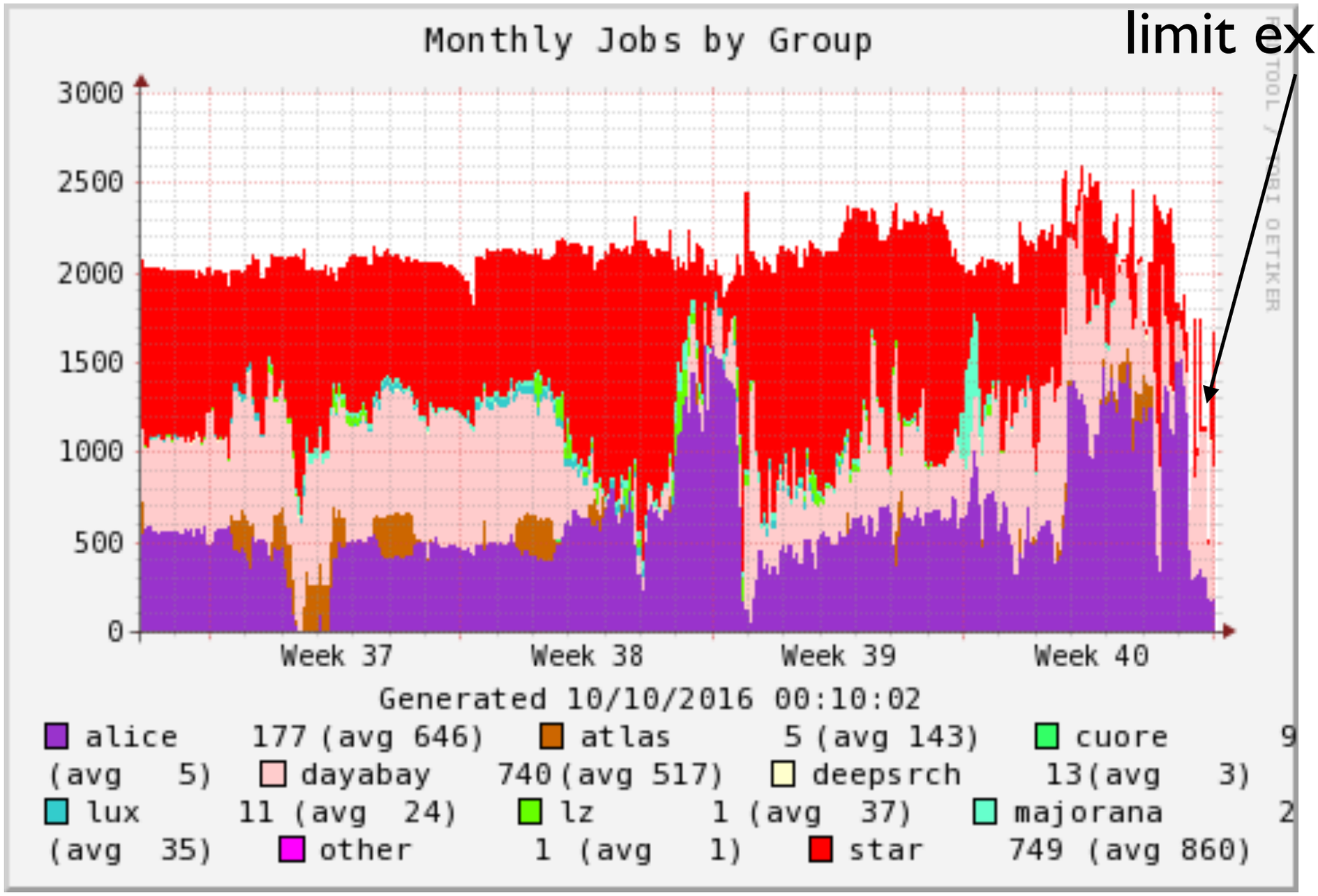
aggregated load on PDSF interactive nodes

https://portal-auth.nersc.gov/pdsf-mon/

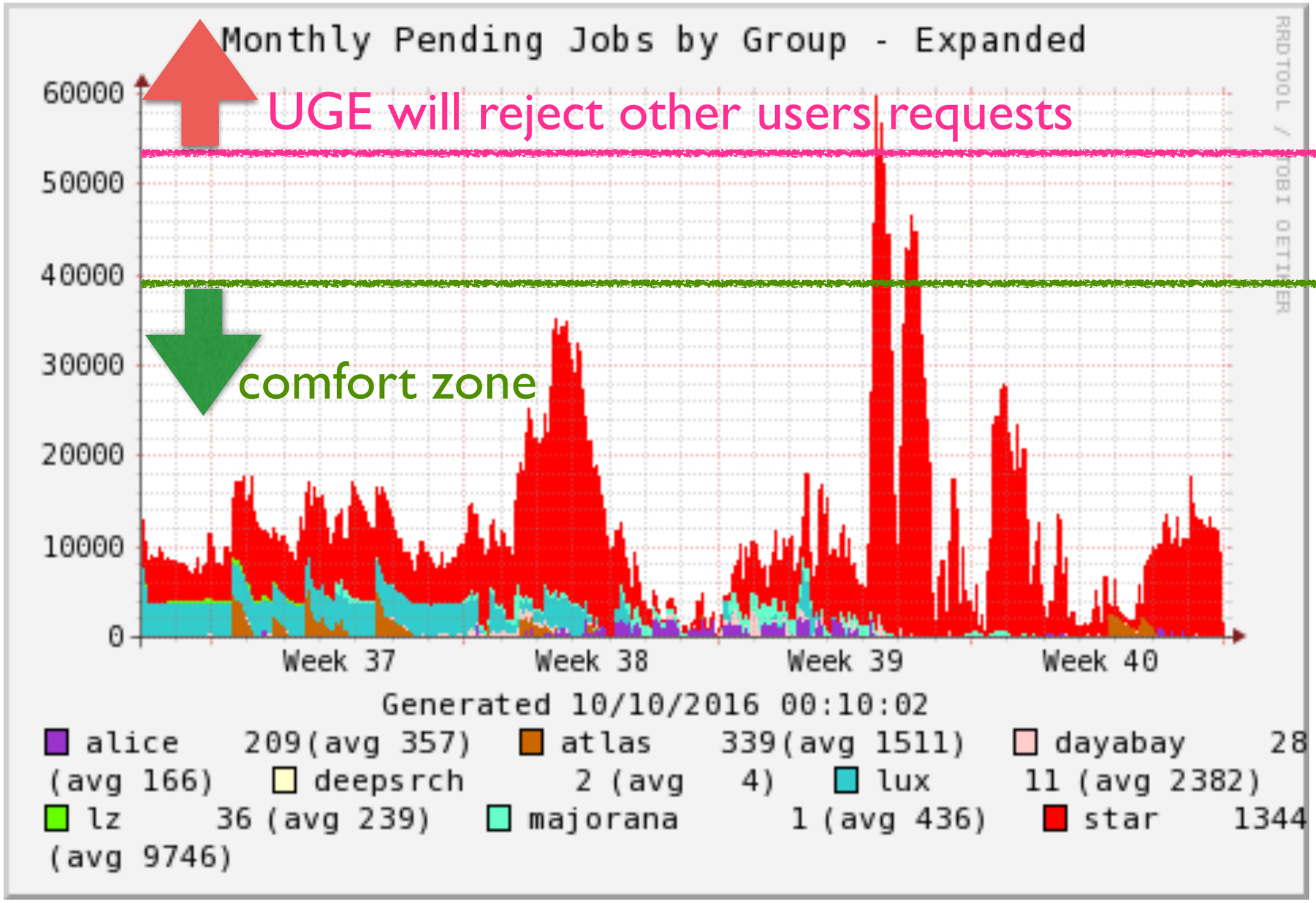


CPU Utilization

STAR inodes limit exhaust



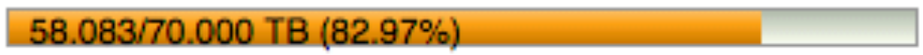
UGE queue load



Disc space utilization

FillStatus (Quota): *PROJECT* (2016-10-10 12:10)

star - size



star - inodes

---

starprod - size

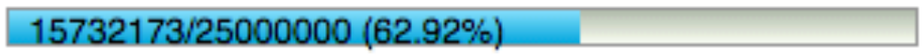
starprod - inodes

(49.98%)

---

alice - size

alice - inodes



FillStatus (Quota): *PROJECTA* (2016-10-10 12:10)

starprod - size

starprod - inodes

(31.37%)

STAR is too close to limits

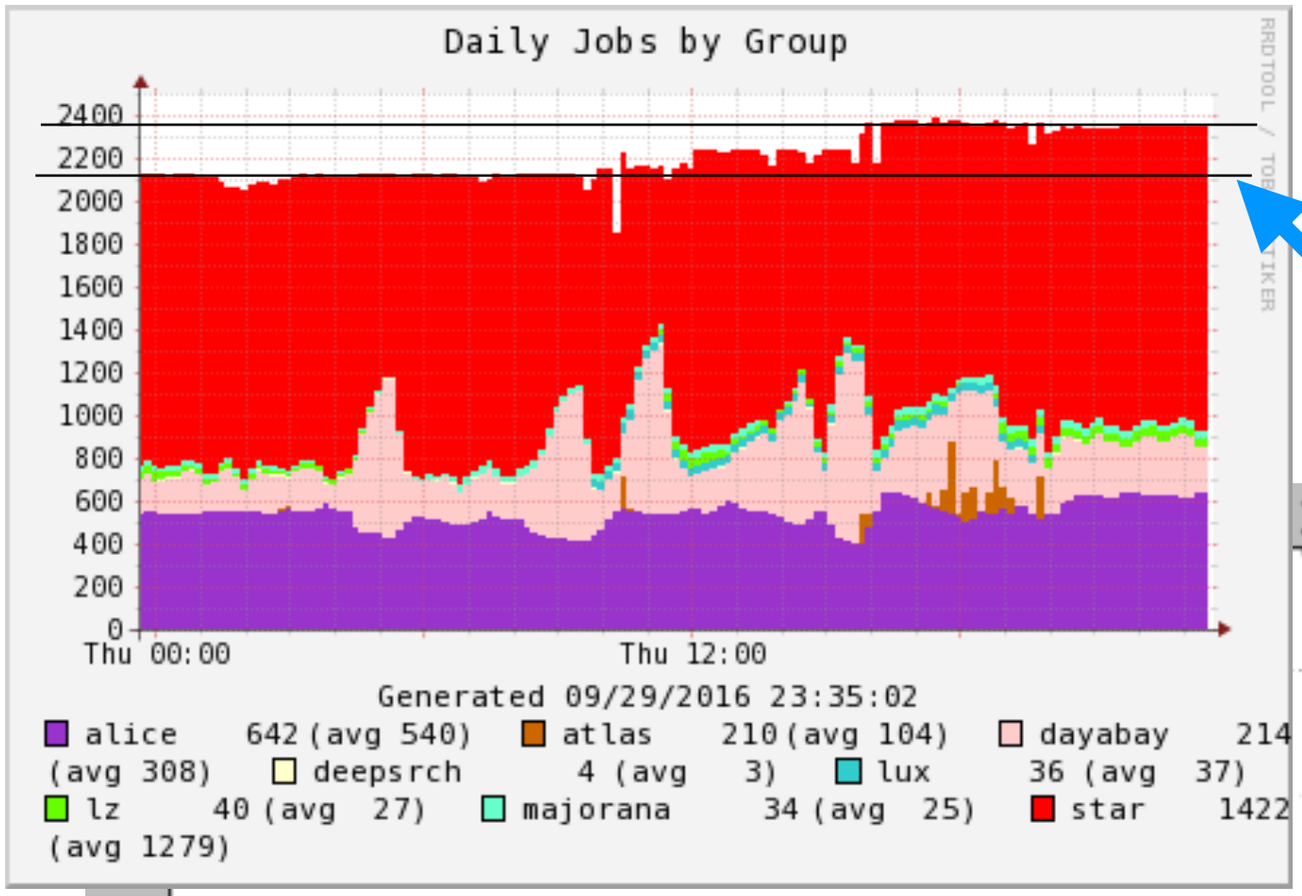
Announcements

Bi-weekly office hours 12:30 -2:30pm
Thursday, October 13 & 27, 59-4016-CR

PDSF users meeting

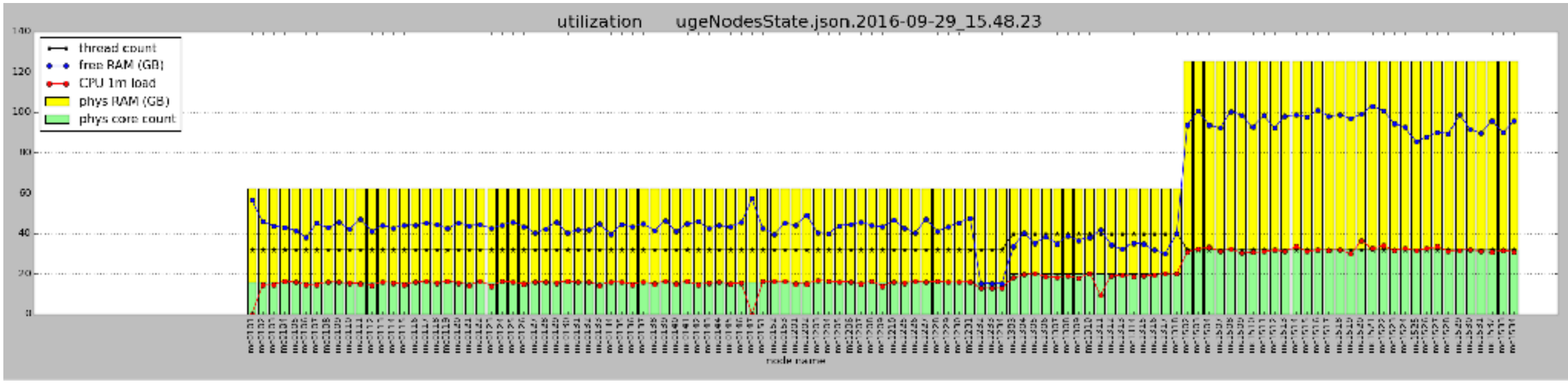
- Tuesday, November 8, 11:00 - 12:01pm 59-3034-CR

unused ~220 job slots, added



PDSF utilization by node

added ~220 jobs slots in crate15 on September 29



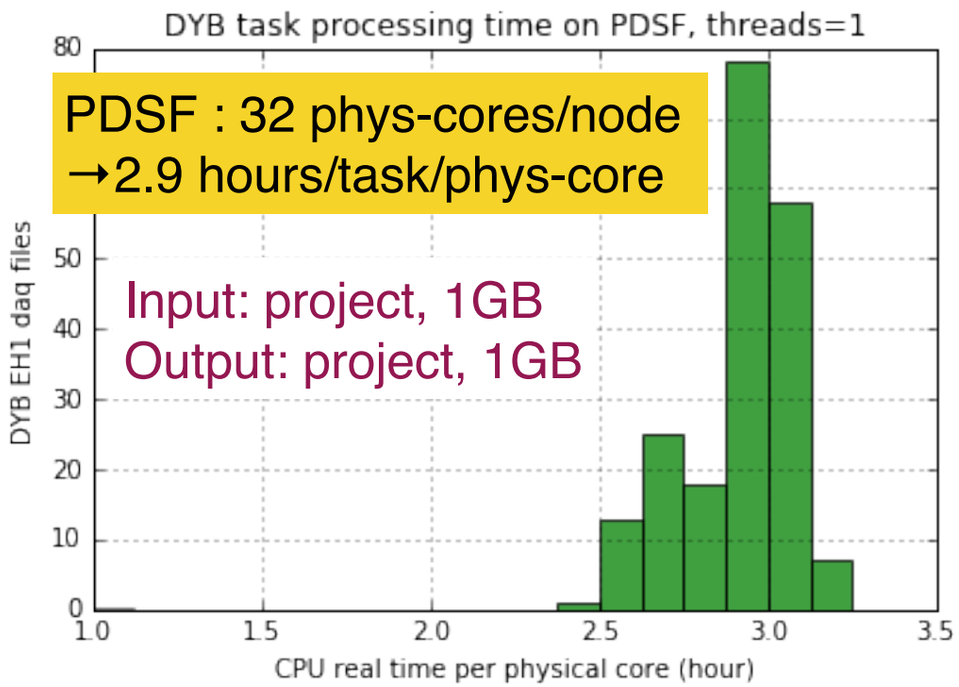
Advantage of hyper threading

200 similar DayaBay analysis tasks from detector “EH1”

PDSF worker node

```
mc1534 $ lscpu
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                32
On-line CPU(s) list:   0-31
Thread(s) per core:    1
Core(s) per socket:    16
Socket(s):             2
NUMA node(s):          2
Vendor ID:             GenuineIntel
CPU family:            6
Model:                 63
Stepping:              2
CPU MHz:               2301.000
BogoMIPS:              4589.04
Virtualization:        VT-x
L1d cache:             32K
L1i cache:             32K
L2 cache:              256K
L3 cache:              40960K
NUMA node0 CPU(s):     0-15
NUMA node1 CPU(s):     16-31
```

hyper threading not utilized



PDSF : 32 phys-cores/node
→2.9 hours/task/phys-core

Input: project, 1GB
Output: project, 1GB

grep for ‘real’ from time nywa.py bhla

Cori worker node

Each node has two sockets, each socket is populated with a 16-core **Intel "Haswell" processor at 2.3 GHz**

32 cores per node.

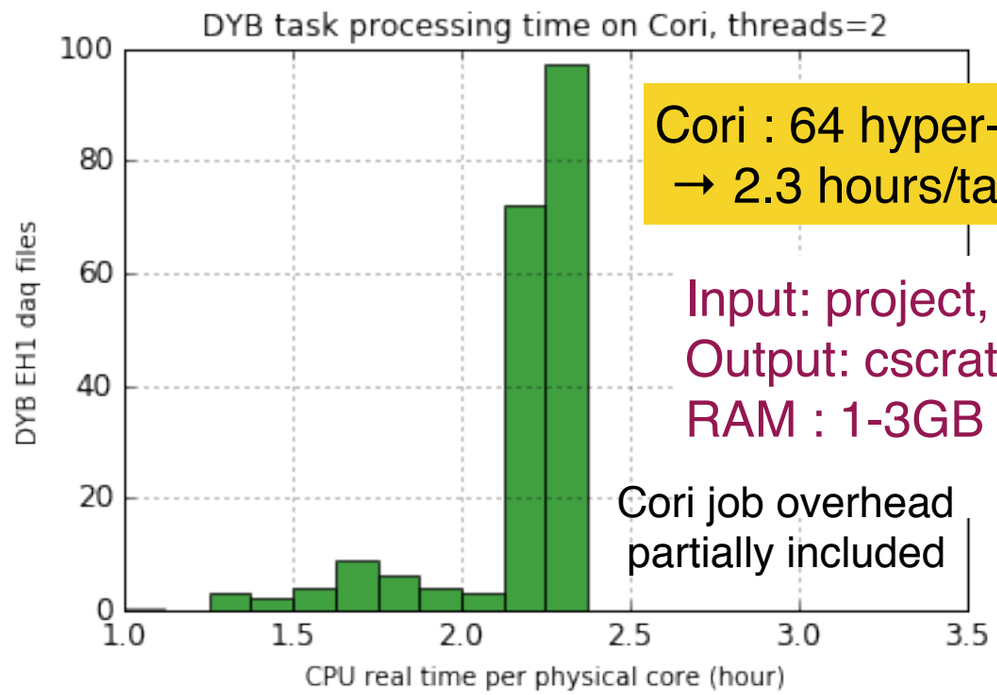
Each core has 1 or 2 user threads, and two 256 bits wide vector units

36.8 Gflops/core; 1.2 TFlops/node; 1.92 PFlops total (theoretical peak)

Each node has 128 GB DDR4 2133Mhz MHz memory (four 16 GB DIMMs per socket). 203 TB total aggregate memory.

Each core has its own L1 and L2 caches, with 64 KB (32 KB instruction cache, 32 KB data) and 256 KB, respectively; there is also a 40-MB shared L3 cache per socket

hyper threading fully utilized



Cori : 64 hyper-cores/node
→ 2.3 hours/task/phys-core

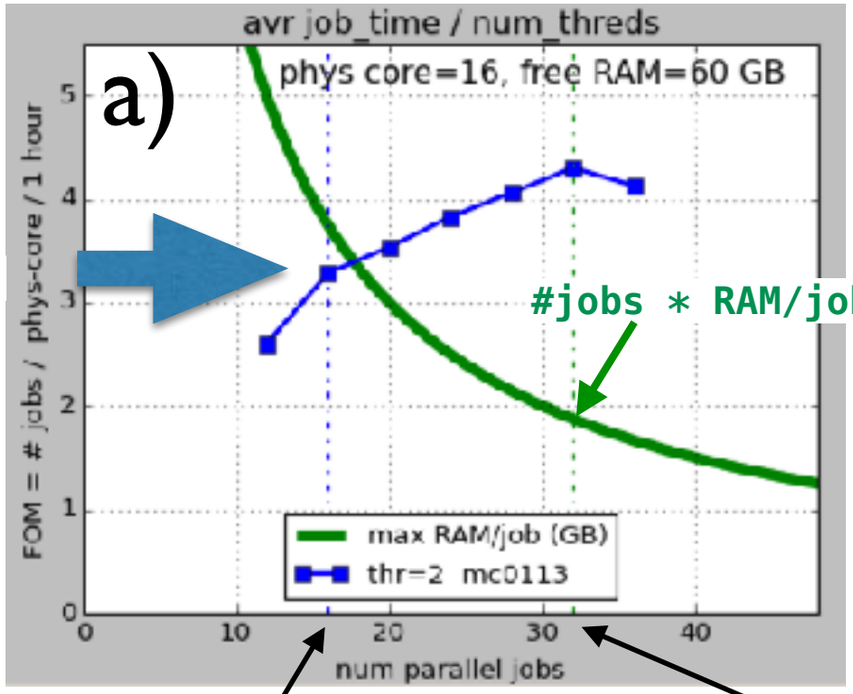
Input: project, 1GB
Output: cscratch, 1GB
RAM : 1-3GB

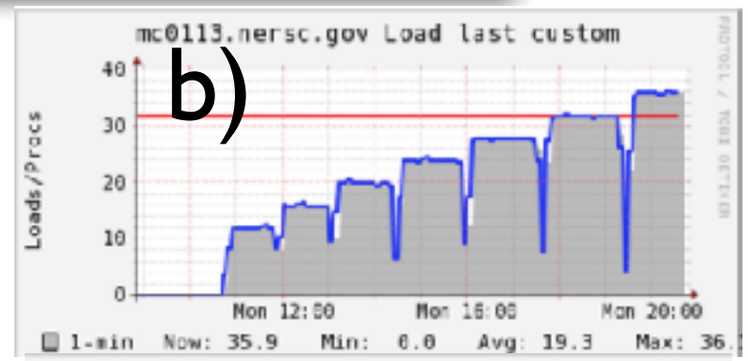
Cori job overhead partially included

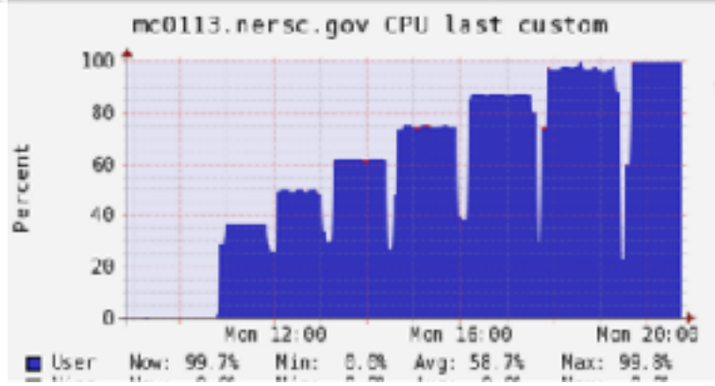
grep for `elapsed from` /usr/bin/time nuwa.py bhla

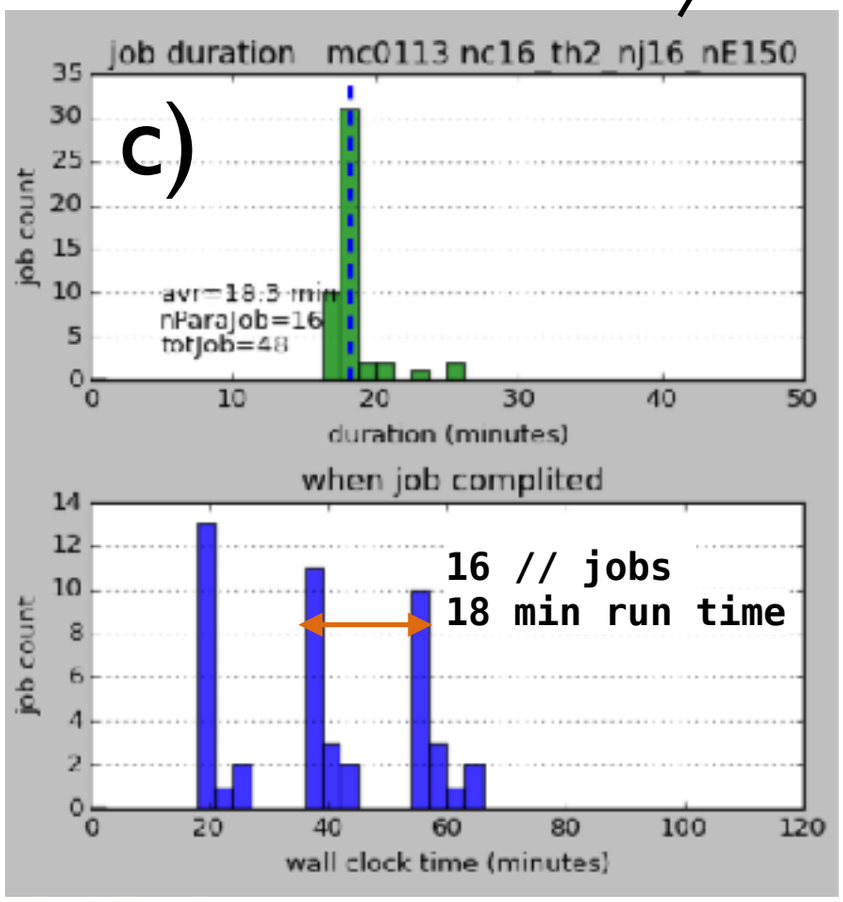
ramping up load on 16-core, 2-thread mode

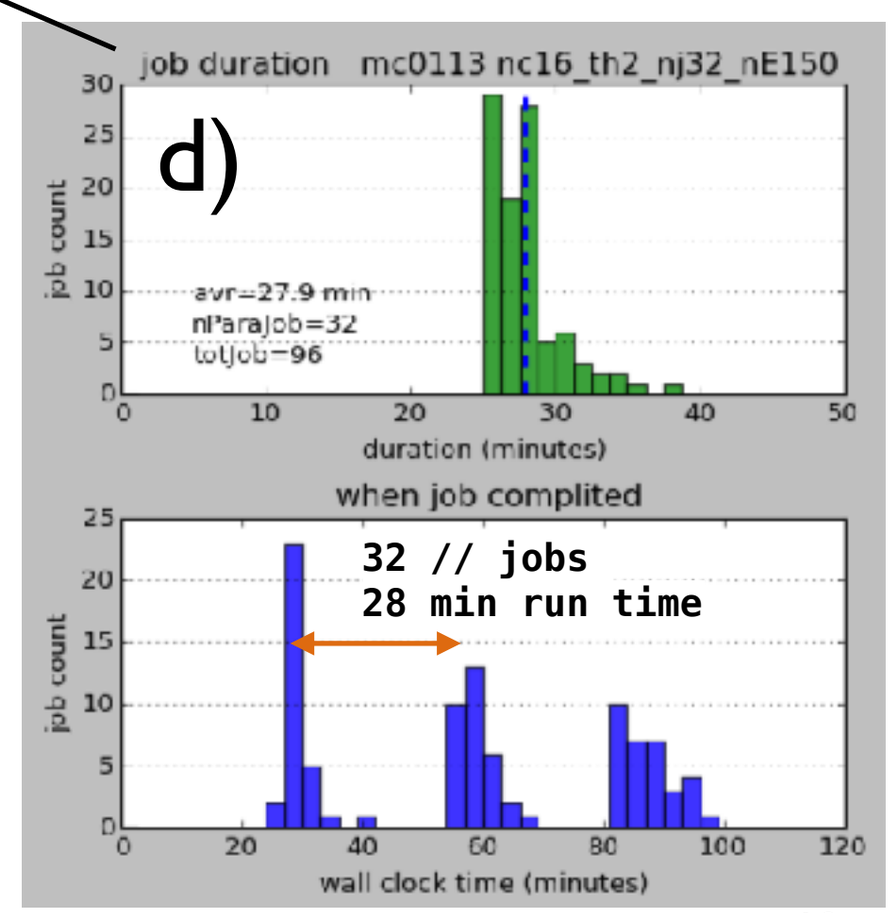
UGE working point











Efficiency of 3 types of PDSF nodes

32 jobs
1.9 GB RAM/job

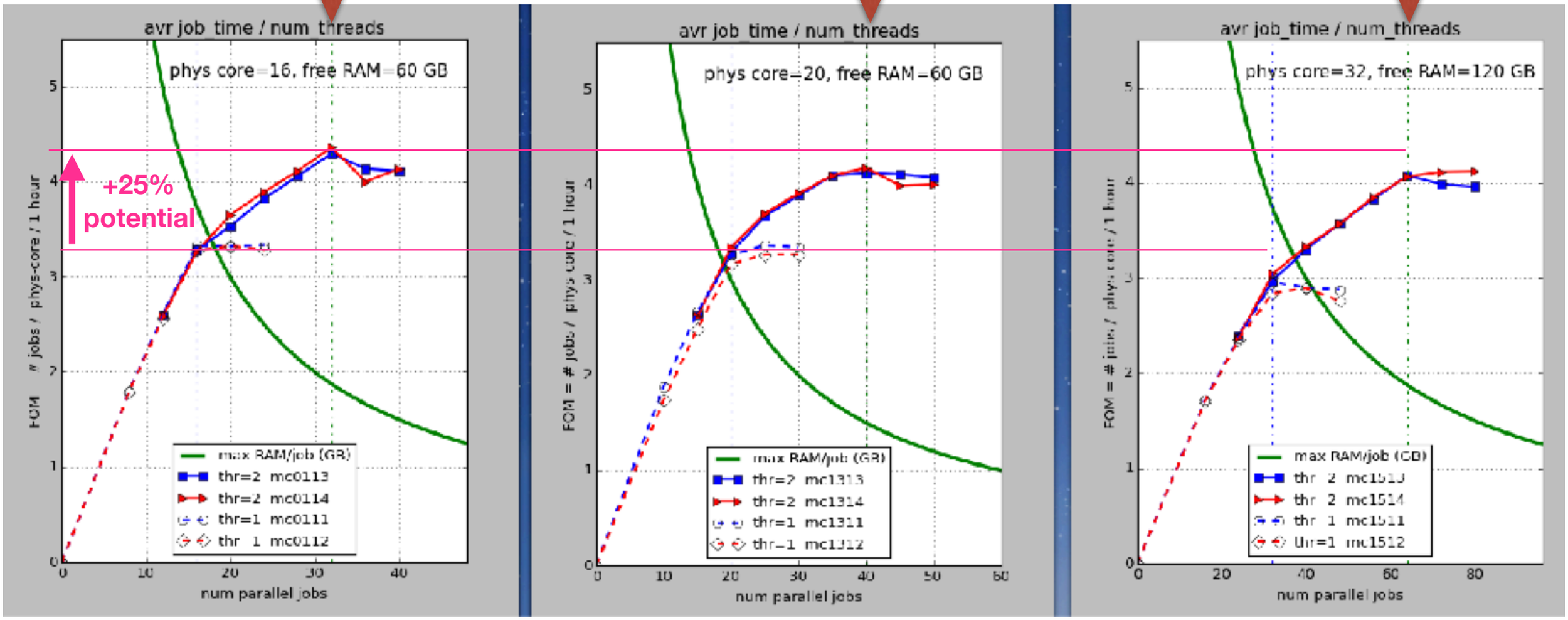
16-core blades
crates 01,11

40 jobs
1.4 GB RAM/job

20-core blades
crate 13

64 jobs
1.9 GB RAM/job

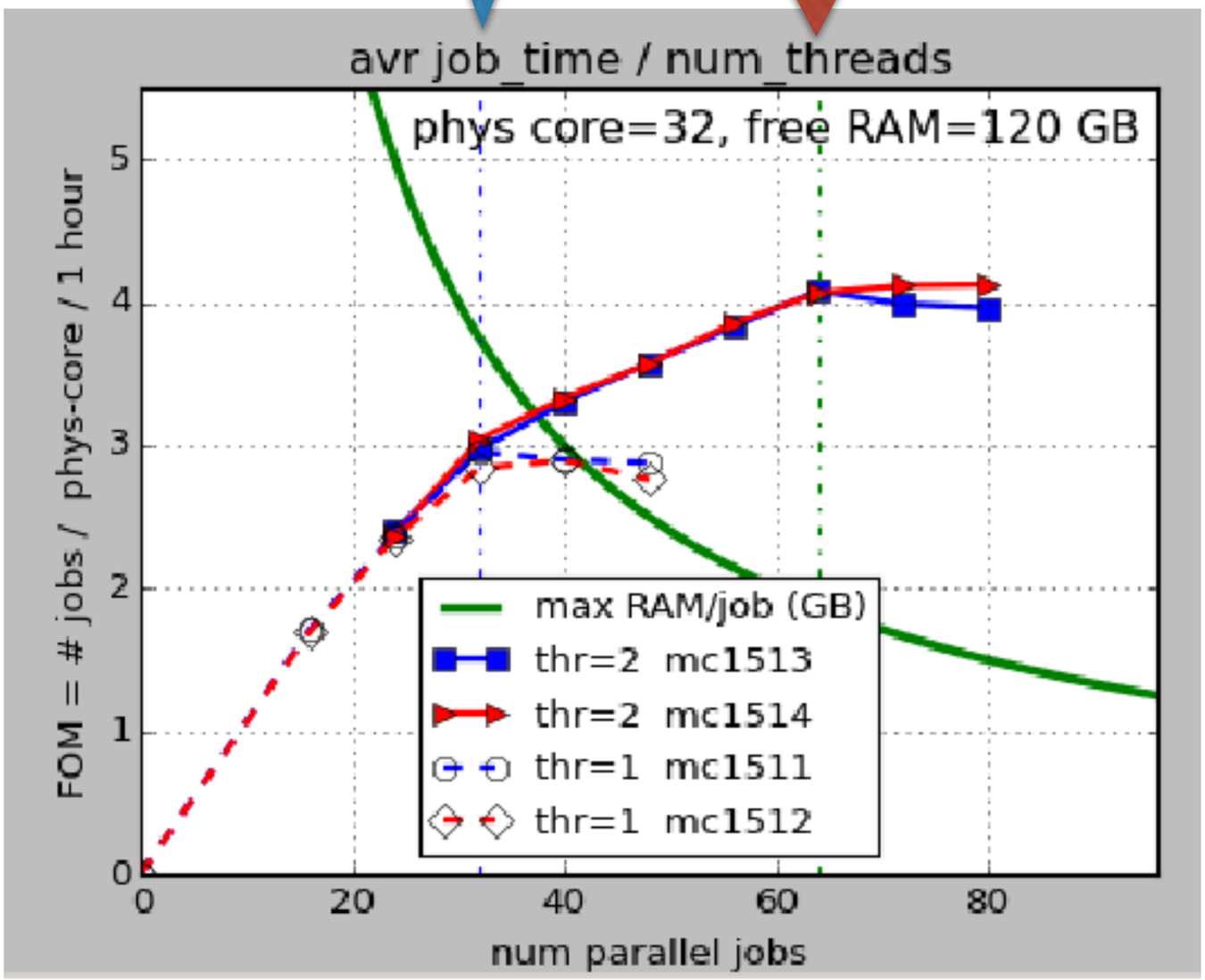
32-core blades
crate 15

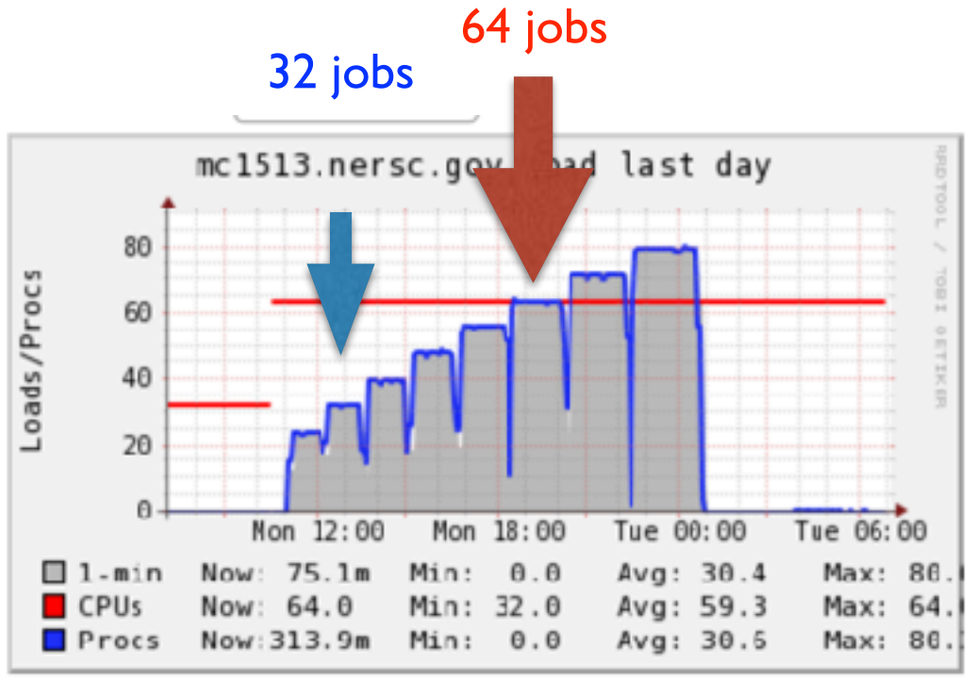


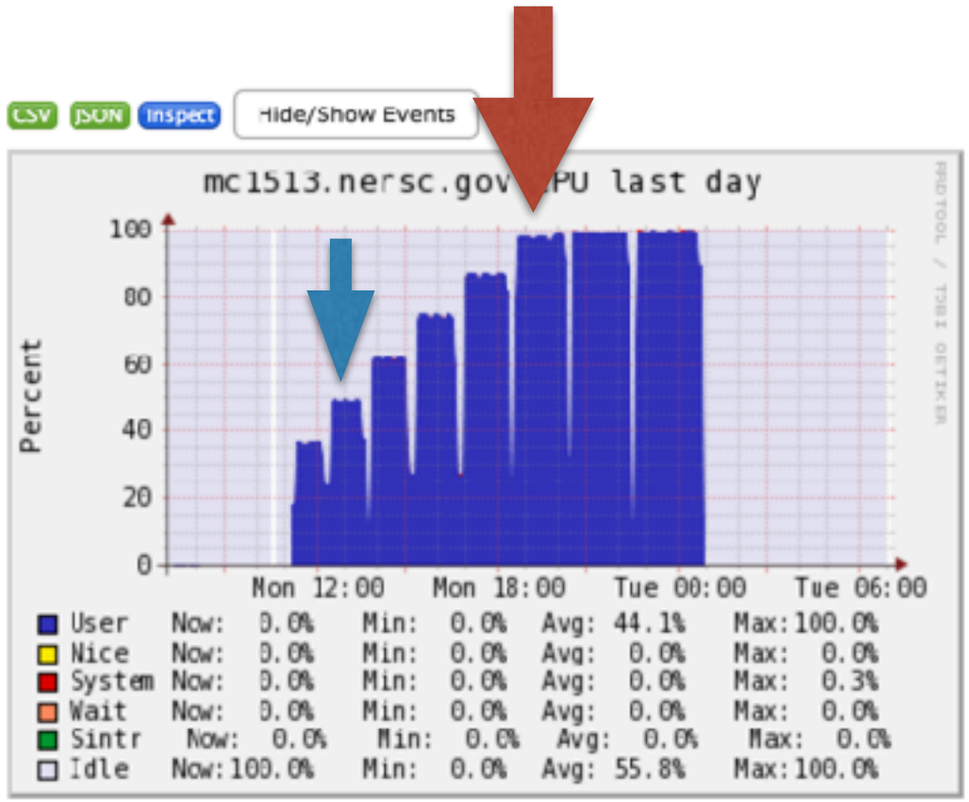
Jeff: HEPSPEC: 20 | 20 | 18

Yes, 32-core blades can handle 64 jobs.

single jobs uses 1.1GB RAM
executes fix amount of computation
runs for 20-30 minutes (depending on node congestion)
no read, no write





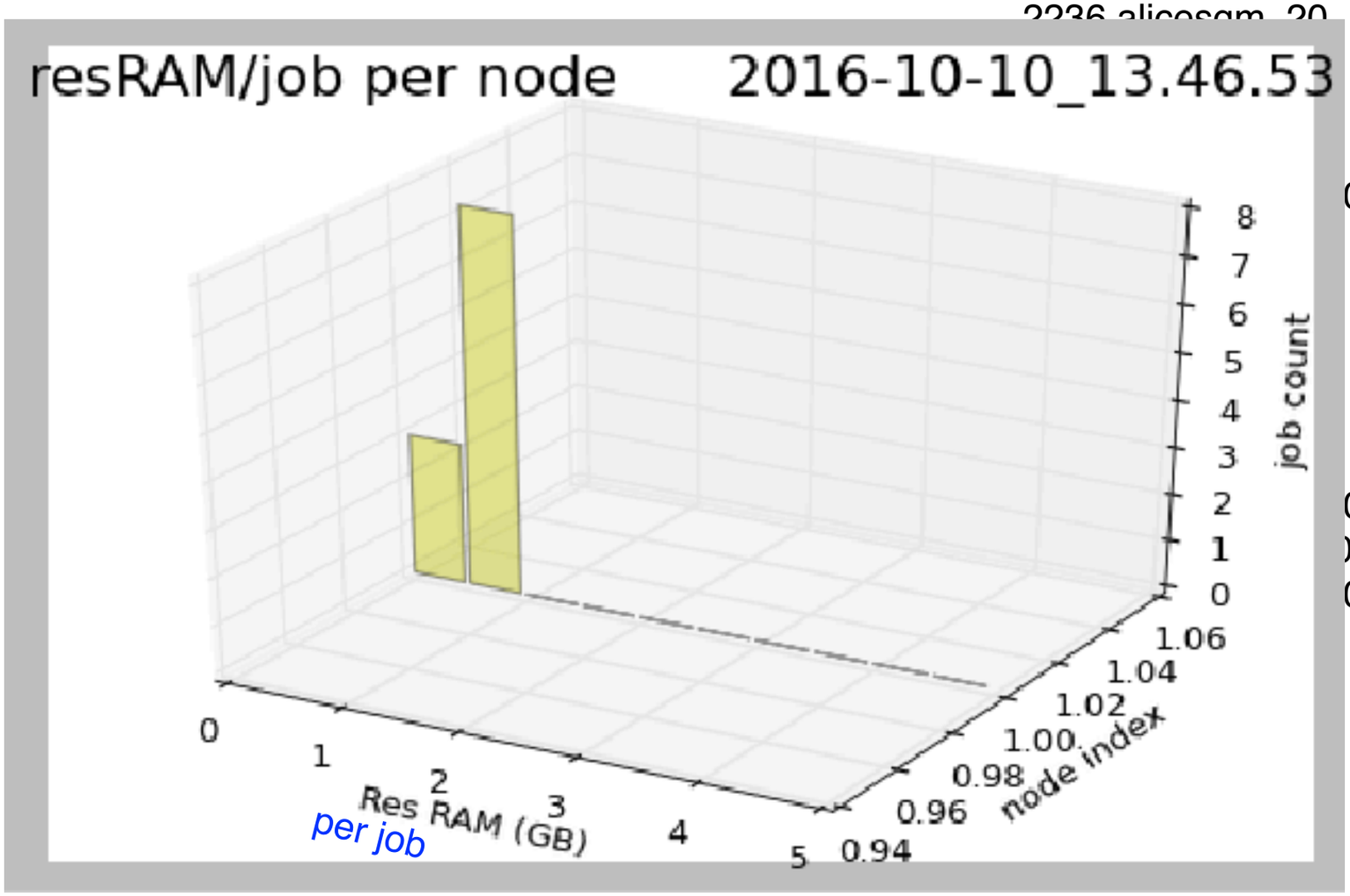


Typical 16-core UGE node analysis

node mc1229:
Physical: RAM=65GB, 16 cores
UGE max job slots: 16, RAM: 63 GB

Ground truth:
14 running jobs, 1min load 14.2
sum: res RAM=7.7 GB, virt RAM=19.5 GB
system free RAM=47.5 GB

UGE truth:
13 jobs running (mndl_prod:5, alice:8)
free RAM : 4.3 GB



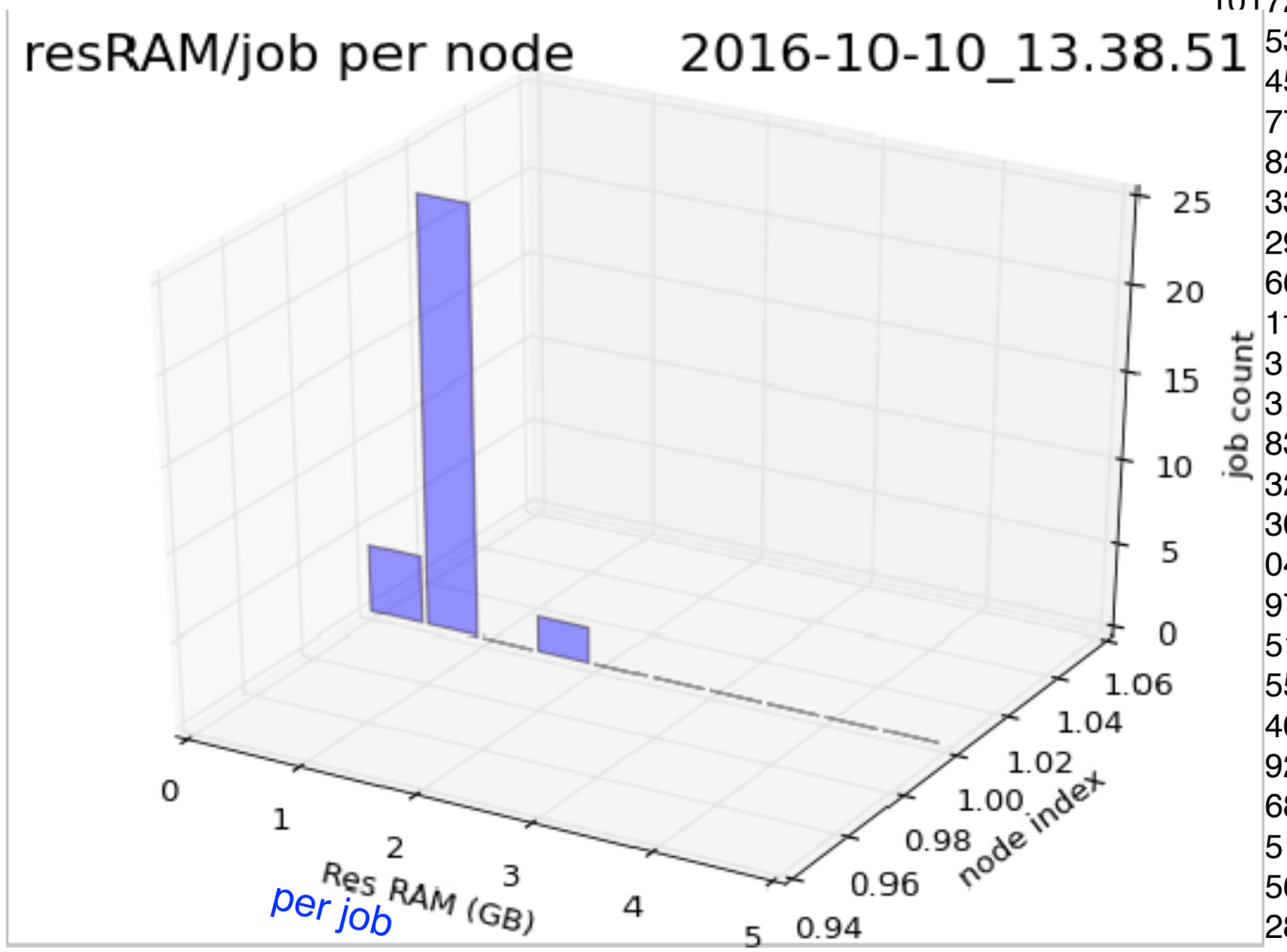

```
mc1229 $ top ibn1
top - 13:44:56 up 62 days,  4:08,  1 user,  load average: 14.17, 14.23, 13.9
Tasks: 872 total,  15 running, 857 sleeping,   0 stopped,   0 zombie
Cpu(s): 37.4%us,  0.7%sy,  0.0%ni, 61.6%id,  0.2%wa,  0.0%hi,  0.1%si,
Mem:  65529492k total, 60524944k used,  5004548k free,   598512k buffe
Swap:  2097148k total,   447428k used,  1649720k free, 43984176k cache

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COM
2236 alicesgm  20   0 2338m 821m  10m R 99.4  1.3 167:43.79 aliroot
    0 2350m 910m   9m R 99.4  1.4 185:42.13 aliroot
    0 2392m 910m   9m R 99.4  1.4 185:00.81 aliroot
    0 2350m 910m   9m R 99.4  1.4 184:49.57 aliroot
0  411m 266m  45m R 99.4  0.4  27:43.62 root4star
    0  244m 118m  42m R 99.4  0.2 352:48.71 root4star
    0 2370m 910m   9m R 99.4  1.4 182:54.40 aliroot
    0 2322m 910m   9m R 99.4  1.4 182:59.03 aliroot
    0 2366m 862m  10m R 99.4  1.3 179:10.66 aliroot
    0 2330m 874m  35m R 99.4  1.4 113:49.46 aliroot
    0  654m 379m  51m R 97.5  0.6 150:23.65 root4star
0  299m 149m  45m R 93.6  0.2   9:20.26 root4star
)  458m 307m  45m R 87.9  0.5  32:00.60 root4star
0  441m 291m  45m R 80.3  0.5  27:44.12 root4star
```

Oversubscribed 20-core UGE node analysis

node mc1308
Physical: RAM=65GB, 20 cores
UGE max job slots: 38, RAM: 180 GB

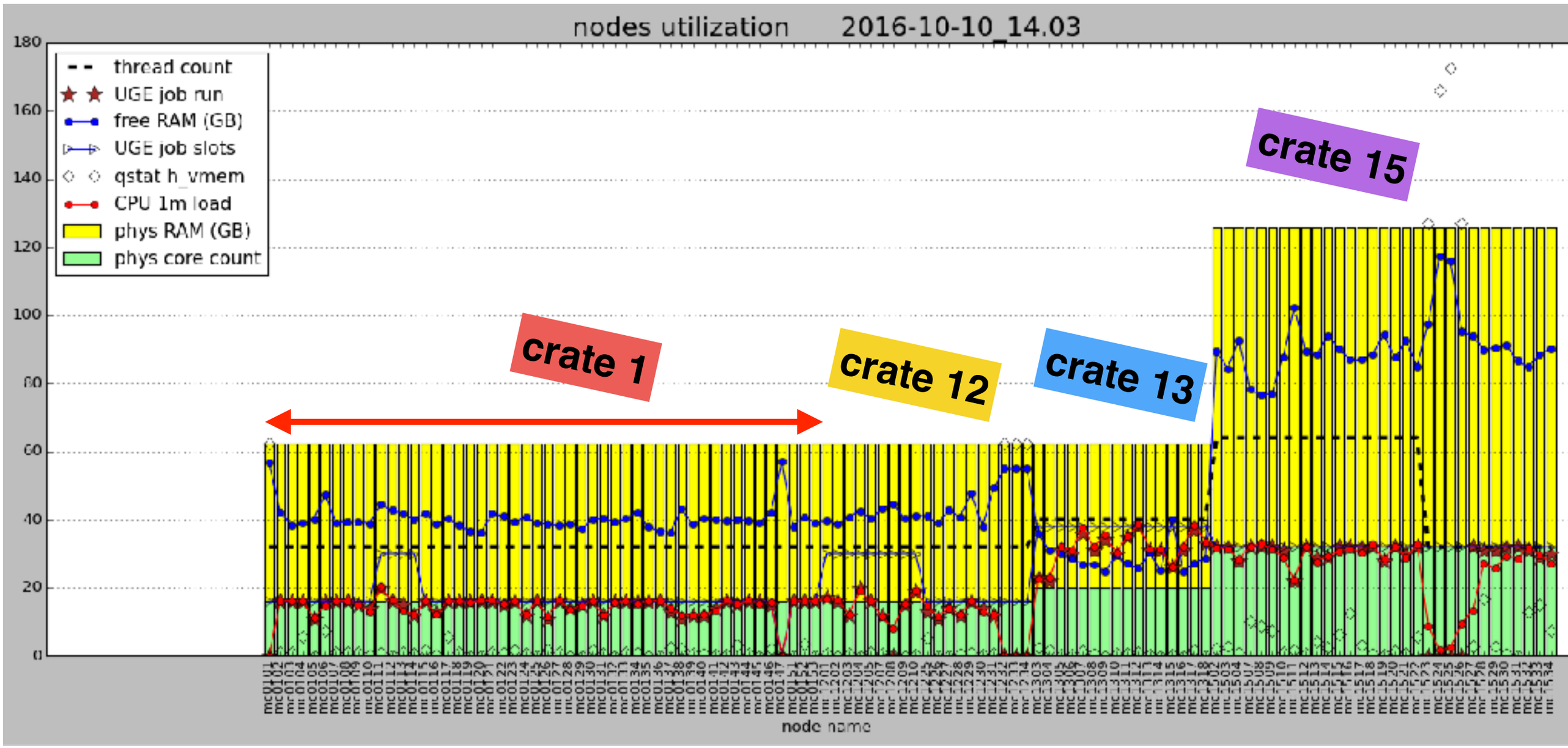
Ground truth:
31 running jobs, 1min load 31.7
sum: res RAM=25 GB, virt RAM=63 GB
system free RAM=47.5 GB

UGE truth:
13 jobs running (mndl_prod:11, alice:20)
free RAM : 0.9 GB




```
mc1308 $ top ibn1
top - 13:39:08 up 120 days, 20:15,  1 user,  load average: 31.73, 31.77, 31.76
Tasks: 1269 total,  25 running, 1244 sleeping,   0 stopped,   0 zombie
Cpu(s): 38.6%us,  0.9%sy,  0.0%ni, 60.3%id,  0.2%wa,  0.0%hi,  0.1%si,  0.0%st
Mem:  65477852k total, 65147236k used,   330616k free,   228028k buffers
Swap: 309297144k total,   798828k used, 308498316k free, 27515440k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
14437 hack      20   0 1946m 1.6g  34m R 103.3  2.5   8:45.83 ipython
 3804 alicesgm  20   0 2324m 859m 8428 R 101.4  1.3 247:40.55 aliroot
 5374 alicesgm  20   0 2380m 909m 8356 R 101.4  1.4 252:35.01 aliroot
 5818 alicesgm  20   0 2344m 794m  10m R 101.4  1.2 163:05.68 aliroot
 5930 alicesgm  20   0 2372m 781m  10m R 101.4  1.2 163:21.60 aliroot
 6349 alicesgm  20   0 2384m 860m 8444 R 101.4  1.3 246:31.37 aliroot
 6646 alicesgm  20   0 2344m 814m  22m R 101.4  1.3 144:16.30 aliroot
10172 hack      20   0 2009m 1.6g  34m R 101.4  2.5   9:44.65 ipython
   53 alicesgm  20   0 2342m 909m 8340 R 101.4  1.4 258:45.53 aliroot
   45 alicesgm  20   0 2312m 798m  29m R 101.4  1.2 143:03.99 aliroot
   77 alicesgm  20   0 2330m 824m  35m R 101.4  1.3 141:36.73 aliroot
   82 alicesgm  20   0 2354m 863m 8428 R 101.4  1.4 247:51.11 aliroot
   33 huangxj   20   0  638m 370m  62m R 100.0  0.6 142:28.56 root4star
   29 alicesgm  20   0 2362m 908m 8444 R 101.4  1.4 249:04.40 aliroot
   66 alicesgm  20   0 2326m 908m 8428 R 101.4  1.4 249:32.13 aliroot
   11 alicesgm  20   0 2316m 860m 8428 R 101.4  1.3 248:09.45 aliroot
    3 alicesgm  20   0 2324m 871m 8428 R 99.6  1.4 248:05.69 aliroot
    3 alicesgm  20   0 2310m 868m 8340 R 99.6  1.4 253:32.92 aliroot
   83 huangxj   20   0  591m 327m  62m R 99.6  0.5 143:37.84 root4star
   32 huangxj   20   0  628m 340m  62m R 99.6  0.5 142:37.93 root4star
   30 alicesgm  20   0 2346m 867m 8428 R 99.6  1.4 248:36.79 aliroot
   04 alicesgm  20   0 2320m 860m 8428 R 99.6  1.3 249:33.19 aliroot
   97 alicesgm  20   0 2334m 908m 8428 R 99.6  1.4 249:37.18 aliroot
   51 alicesgm  20   0 2434m 711m 8356 R 99.6  1.1 267:32.93 aliroot
   55 alicesgm  20   0 2358m 868m 8428 D 94.1  1.4 247:55.57 aliroot
   40 yiguo     20   0  250m 112m  44m D 92.2  0.2   3:31.21 root4star
   92 alicesgm  20   0 2306m 867m 8428 D 92.2  1.4 248:01.61 aliroot
   68 alicesgm  20   0 2412m 909m 8356 D 83.0  1.4 255:04.17 aliroot
    5 alicesgm  20   0 2348m 798m  10m D 81.1  1.2 164:31.14 aliroot
   56 alicesgm  20   0 2340m 821m  29m D 79.3  1.3 142:38.42 aliroot
   28 alicesgm  20   0 2322m 860m 8428 D 75.6  1.3 248:10.66 aliroot
```

All working UGE nodes : load



All working UGE nodes : res RAM