PDSF Users Meeting

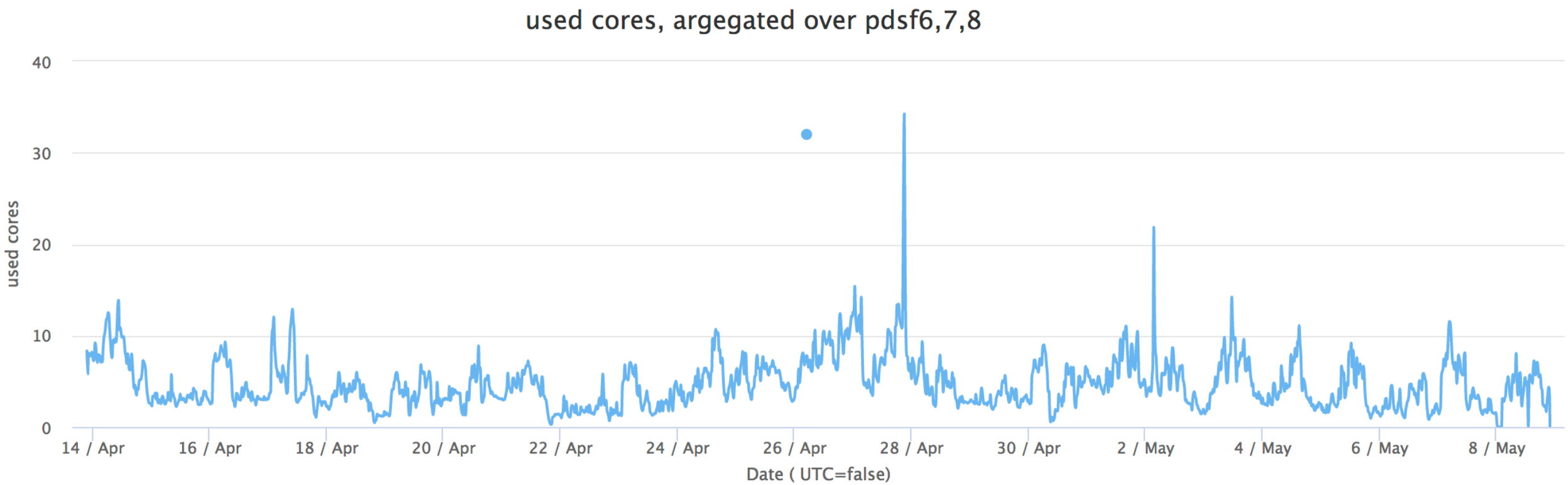
* PDSF-2 performance (UGE+CHOS)
* announcements
* AOB
* PDSF-3 evaluation (SLURM+Shifter)
* user production on Cori

May 9, 2017

Jan Balewski

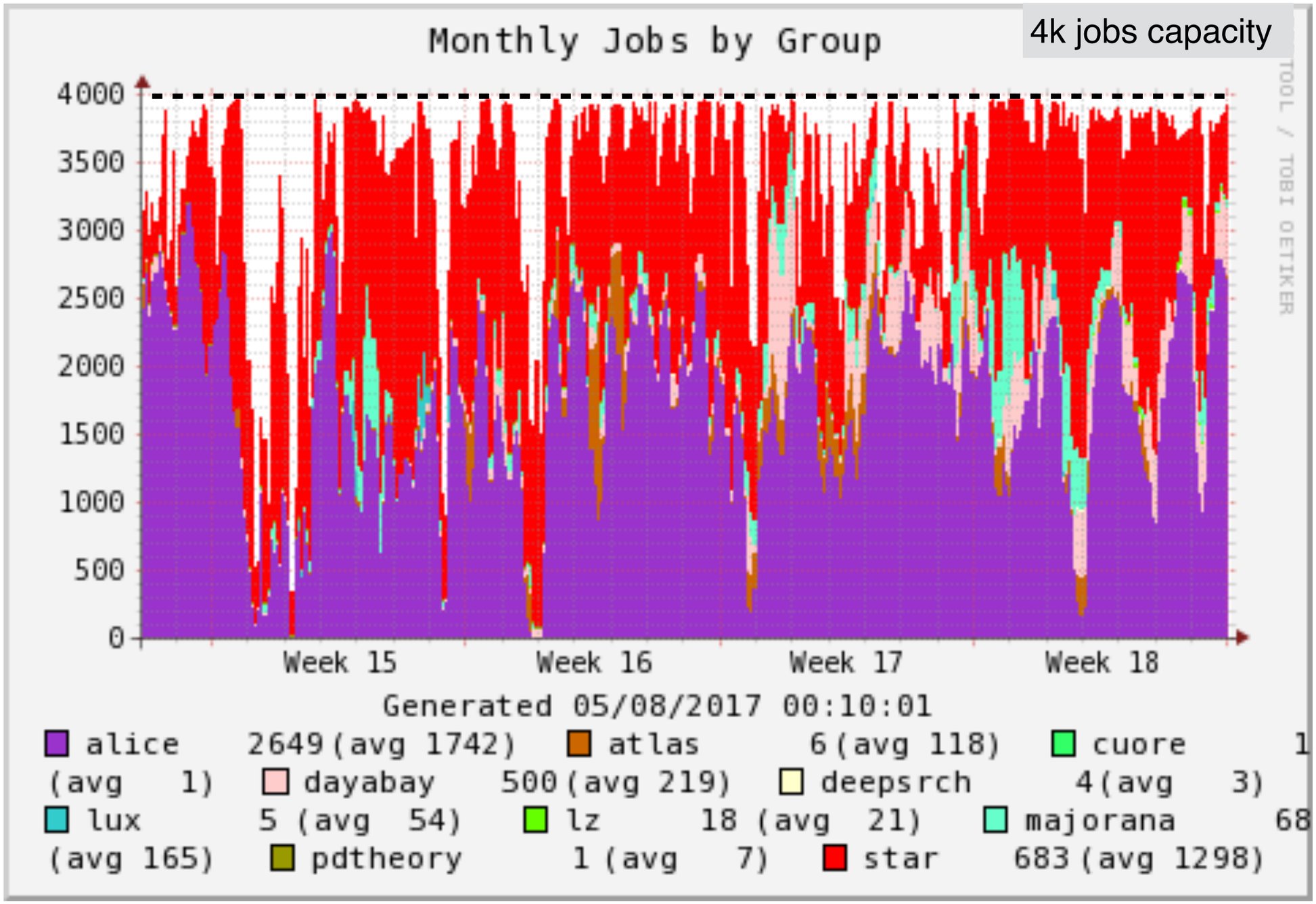
aggregated load on PDSF interactive nodes

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



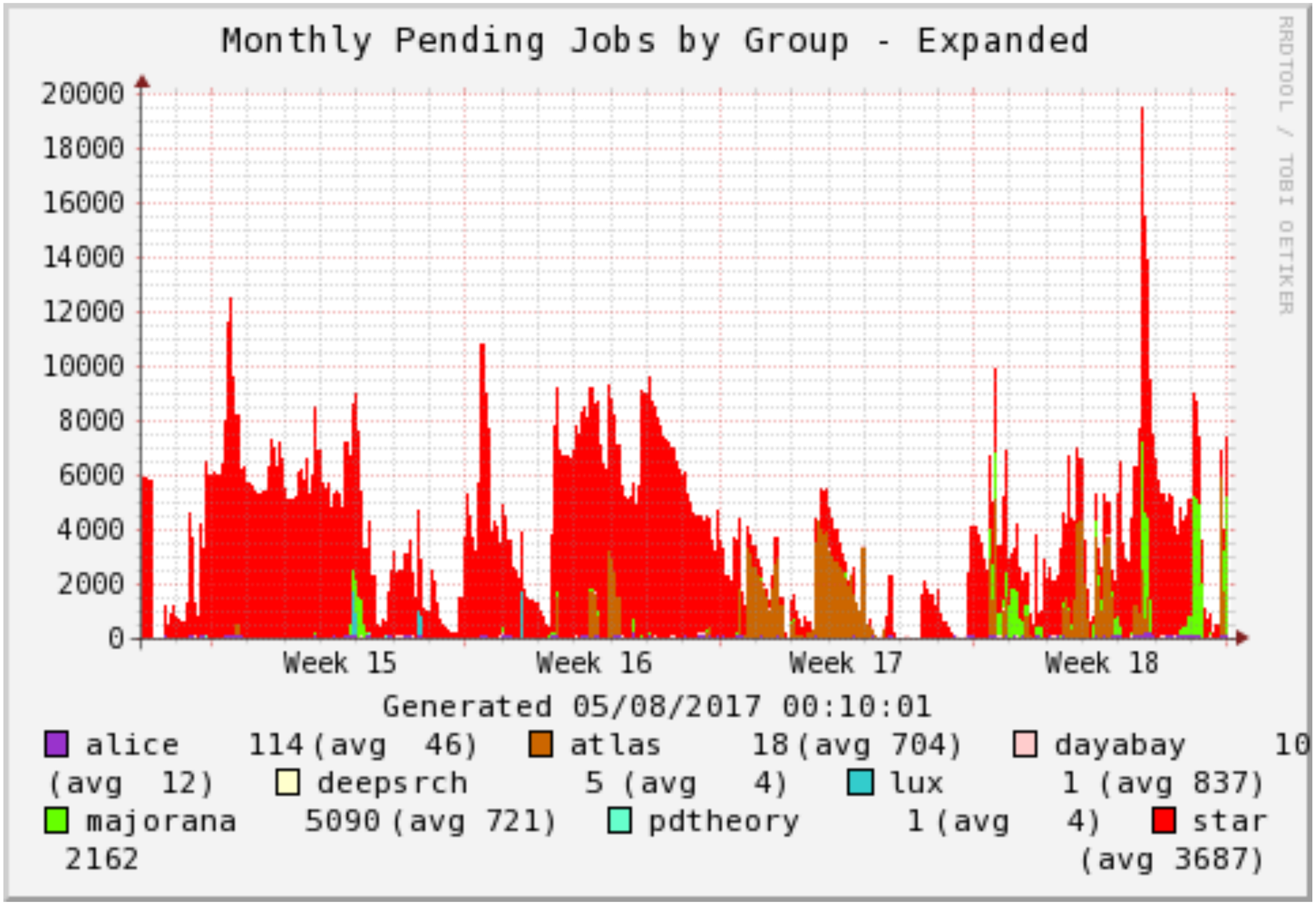
CPU Utilization is not great

- not enough jobs submitted and/or
- few STAR user submit enough jobs but they are throttled because xrootd can't keep up w/ job demand



UGE queue load

~ enough submitted jobs, most of the time



Utilization of PDSF compute nodes

May average
Served 3345 vCores
Capacity 4000 vCores
CPU utilization 0.84

April average
Served 3364 vCores
Capacity 4000 vCores
CPU utilization 0.76

UGE usage report cron-generated: 2017-05-08_00.41, tc
end-time was set to 2017-05-08 00:13, DB collName=u

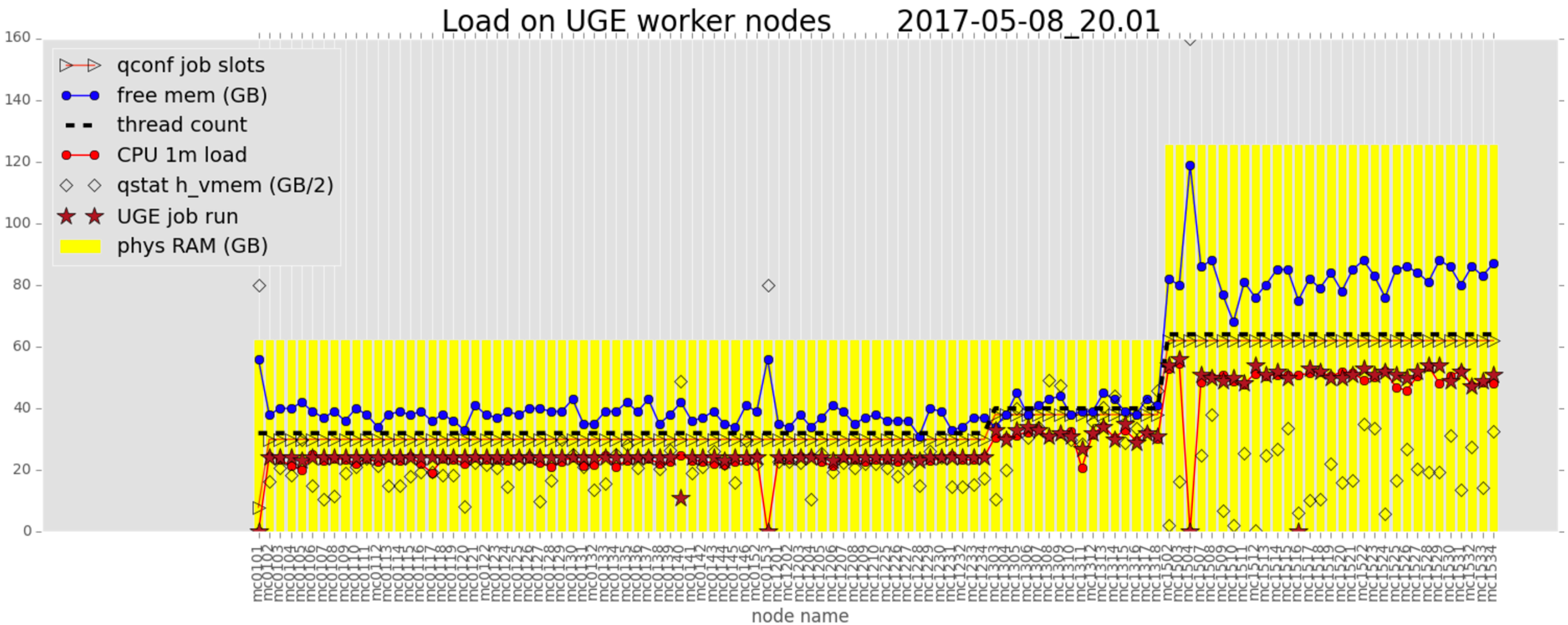
user \ timeRange	sum last 7_days	sum last 30_days
0-all-users : owner	64.5 (cpu*year) , 335768 jobs wallT fract=1.000, CPU eff=0.79	275.0 (cpu*year) , 1071967 jobs wallT fract=1.000, CPU eff=0.84
alice : project	28.1 (cpu*year) , 24839 jobs wallT fract=0.436, CPU eff=0.69	137.0 (cpu*year) , 186237 jobs wallT fract=0.498, CPU eff=0.73
atlas : project	1.0 (cpu*year) , 152149 jobs wallT fract=0.015, CPU eff=0.68	6.1 (cpu*year) , 257047 jobs wallT fract=0.022, CPU eff=0.71
dayabay : project	5.5 (cpu*year) , 3727 jobs wallT fract=0.086, CPU eff=0.99	14.6 (cpu*year) , 35134 jobs wallT fract=0.053, CPU eff=0.98
dybprod : owner	0.0 (cpu*year) , 0 jobs wallT fract=0.000, CPU eff=0.00	0.4 (cpu*year) , 1826 jobs wallT fract=0.001, CPU eff=0.95
dybspade : owner	0.0 (cpu*year) , 0 jobs wallT fract=0.000, CPU eff=0.00	0.0 (cpu*year) , 0 jobs wallT fract=0.000, CPU eff=0.00
hack : owner	5.5 (cpu*year) , 3727 jobs wallT fract=0.086, CPU eff=0.99	14.1 (cpu*year) , 33216 jobs wallT fract=0.051, CPU eff=0.98
lux : project	0.1 (cpu*year) , 339 jobs wallT fract=0.001, CPU eff=0.90	0.7 (cpu*year) , 17943 jobs wallT fract=0.002, CPU eff=0.97
lz : project	0.2 (cpu*year) , 570 jobs wallT fract=0.002, CPU eff=1.00	0.2 (cpu*year) , 4237 jobs wallT fract=0.001, CPU eff=0.94
majorana : project	4.4 (cpu*year) , 44578 jobs wallT fract=0.069, CPU eff=0.89	10.5 (cpu*year) , 116657 jobs wallT fract=0.038, CPU eff=0.90
star : project	25.2 (cpu*year) , 109143 jobs wallT fract=0.391, CPU eff=0.86	105.2 (cpu*year) , 450113 jobs wallT fract=0.383, CPU eff=0.92
staremb : owner	0.0 (cpu*year) , 0 jobs wallT fract=0.000, CPU eff=0.00	11.9 (cpu*year) , 10877 jobs wallT fract=0.043, CPU eff=1.00
zmarshal : owner	0.0 (cpu*year) , 6 jobs wallT fract=0.000, CPU eff=0.16	0.0 (cpu*year) , 29 jobs wallT fract=0.000, CPU eff=0.13

end-time was set to 2

user \ timeRange	sum last 30_days
0-all-users : owner	276.5 (cpu*year) , 1140416 jobs wallT fract=1.000, CPU eff=0.76
alice : project	129.2 (cpu*year) , 167883 jobs wallT fract=0.467, CPU eff=0.68
atlas : project	7.8 (cpu*year) , 109082 jobs wallT fract=0.028, CPU eff=0.76
dayabay : project	18.9 (cpu*year) , 34902 jobs wallT fract=0.068, CPU eff=0.35
dybprod : owner	0.0 (cpu*year) , 0 jobs wallT fract=0.000, CPU eff=0.00
hack : owner	18.7 (cpu*year) , 34377 jobs wallT fract=0.068, CPU eff=0.34
lux : project	1.5 (cpu*year) , 18492 jobs wallT fract=0.005, CPU eff=0.62
lz : project	0.1 (cpu*year) , 5581 jobs wallT fract=0.000, CPU eff=0.82
majorana : project	12.0 (cpu*year) , 238318 jobs wallT fract=0.043, CPU eff=0.74
star : project	106.3 (cpu*year) , 562115 jobs wallT fract=0.385, CPU eff=0.92
staremb : owner	14.5 (cpu*year) , 8624 jobs wallT fract=0.052, CPU eff=1.00
zmarshal : owner	0.0 (cpu*year) , 31 jobs wallT fract=0.000, CPU eff=0.19

UGE CPU utilization

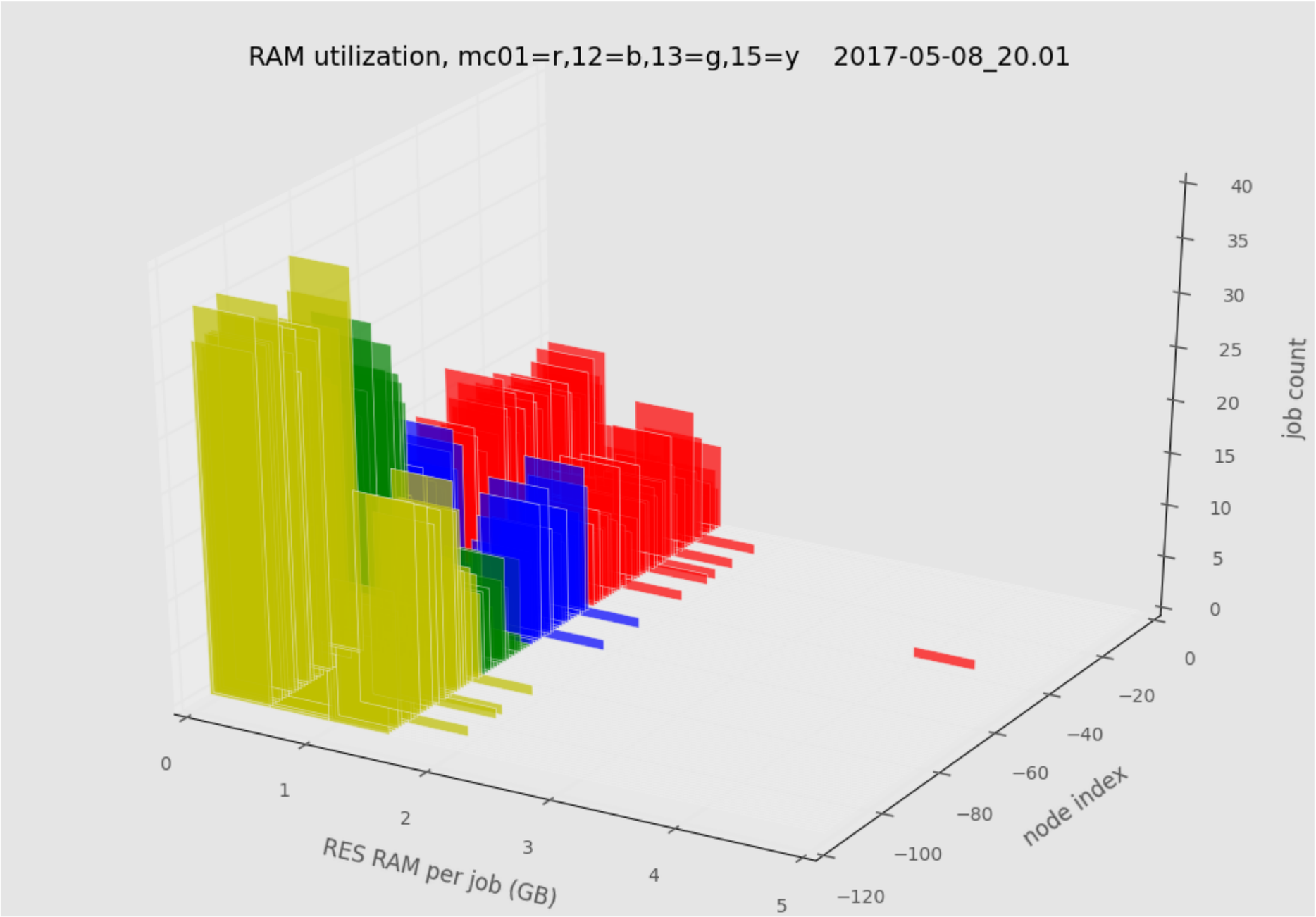
115 UGE nodes, 2300 phys cores ,UGE+CHOS (PDSF-2)
3 nodes ,64 phys cores, SLURM+Shifter (PDSF-3)



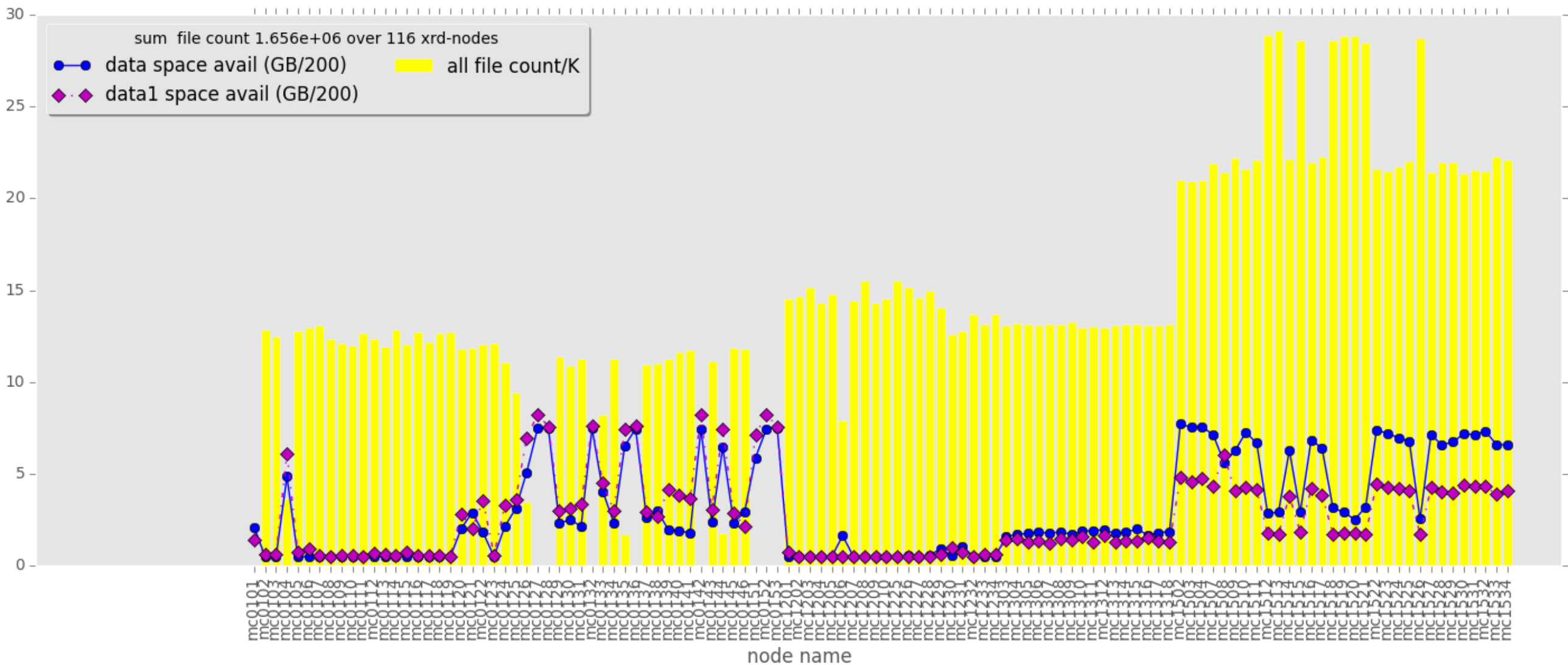
More details: http://portal.nersc.gov/project/mpccc/balewski/ugeLoad/latest/

RAM/job distribution

RAM/job <1 GB



xrootd files re-distribution



/project utilization

http://portal.nersc.gov/project/star/jthaeder/diskUsage/overview/indexExt.html

STAR is out of comfort zone

FillStatus (Quota): *PROJECT* (2017-05-08 20:05)

star - size

55.214/70.000 TB (78.87%)

star - inodes

19600324/20000000 (98.00%)

---

starprod - size

118.558/130.000 TB (91.19%)

starprod - inodes

8925968/20000000
(44.62%)

---

alice - size

42.231/60.976 TB (69.25%)

alice - inodes

16575071/25000000 (66.30%)

FillStatus (Quota): *PROJECTA* (2017-05-08 20:05)

starprod - size

182.502/190.000 TB (96.05%)

starprod - inodes

6309124/20000000
(31.54%)

Announcements

Bi-weekly office hours 12:30 -2:30pm
Thursday, May 11, 25, 59-4016-CR

PDSF Steering committee : May 11, 4-5pm
PDSF user meeting

- Tuesday, June 13, 11am - 12pm, 59-3034-CR

Outages

—Outage for Quarterly Maintenance: May 16–18
—Memorial Day Holiday (No Consulting or Account Support): May 29

Files on Edison and Cori scratch are subject to the systems' purge policies, in which files are automatically deleted:
- On Edison, files not accessed in 8 weeks are eligible to be purged.
- On Cori, files not accessed in 12 weeks are eligible to be purged.

PDSF-3 evaluation

3 nodes ,64 phys cores, SLURM+Shifter+CVMFS (setup by James+Bhupender+Doug ~10 days ago)

Tested workflows run by Jan, confirmed output by experts

LZ w/ CVMFS : interactive + batch, code from Alden --> DEMO
lzap scripts/validations/PulseFinderValidation.py

Majorana , no private code in the image

```
mjd_pulse_basis_generator -d P42538A -r 16836 16840
```

STAR : interactive event processing

root4star -b -q bfc.C'(1,5,"DbV20160418 P2014a pxlHit istHit btof mtd mtdCalib BEmcChkStat -evout CorrX OSpaceZ2 OGridLeak3D -hitfilt", "/global/projecta/projectdirs/starprod/daq/2015/pau200_bht1_adc/st_physics_adc_16150045_raw_5500055.daq")'

PDSF-3 LZ w/ CVMFS batch - DEMO

```
ssh -A -Y afan@pdsf

module use /chos/common/nsg/slurm/modulefiles/rhel6
module load slurm

# check your SLURM account
sacctmgr show assoc where user=$USER

- - - -   Hellow-world SLURM job
cd forJan/PhotonDetection/
batch hello2.slr
>>>Submitted batch job 2109

sqs - shows status of the job

# check the output:
cat slurm-2109.out
done-jan0-lz


 - - - - - - LZ_reco job  run the LZ reconstruction package. It uses cvmfs.- - - - - -
batch lzRecoOne.slr
>>> Submitted batch job 2111


tail -f slurm-2111.out
```

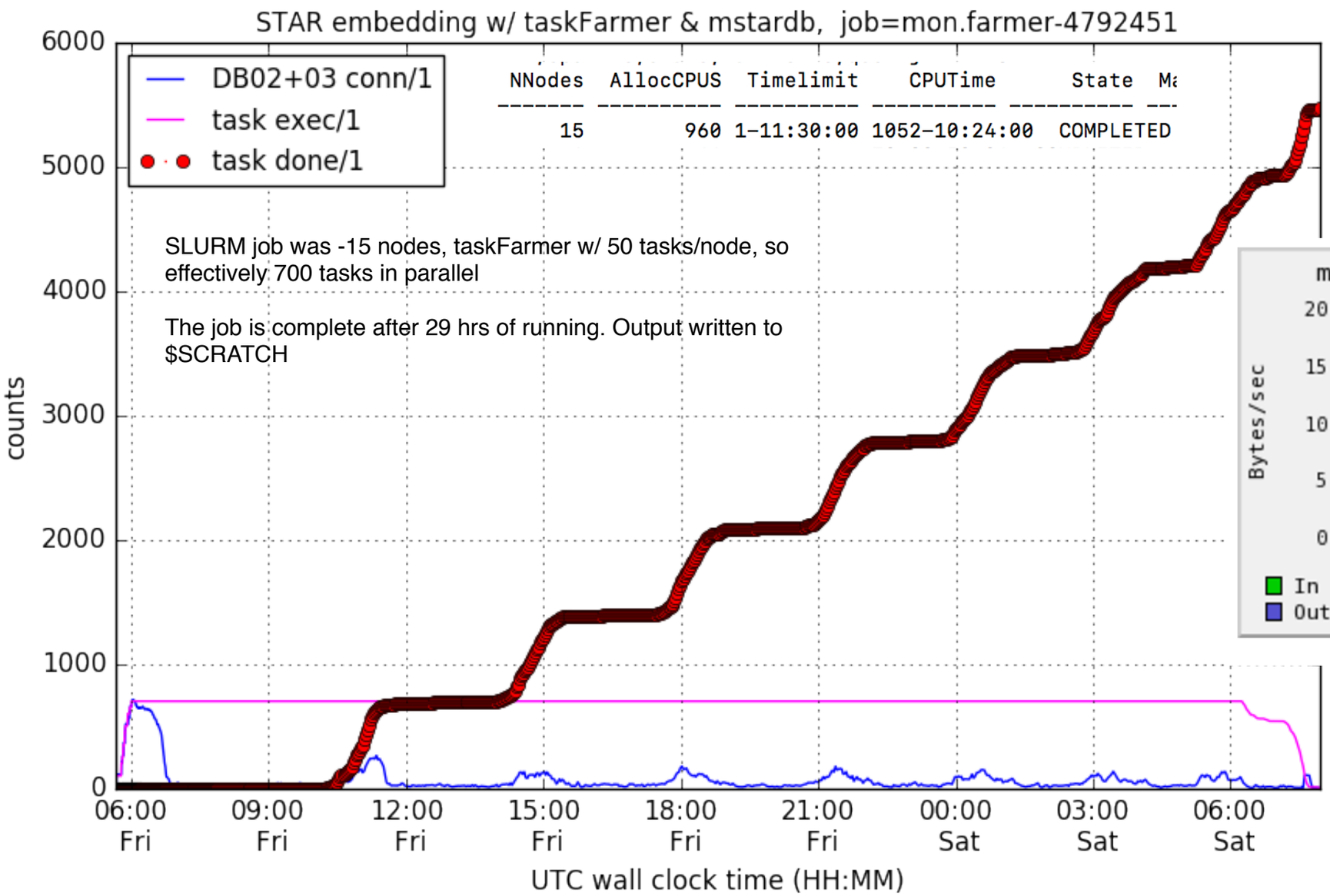
STAR embedding on Cori in Shifter/SL64

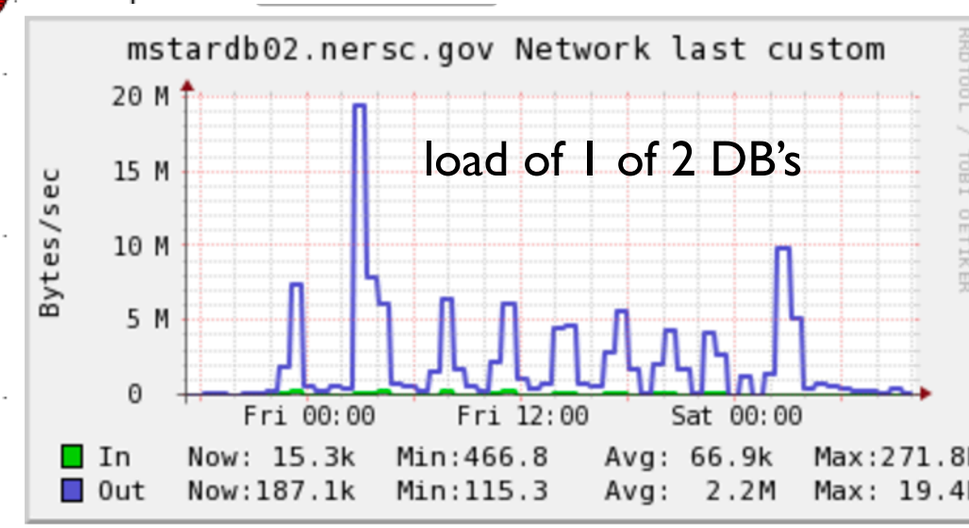
60 r4s task per node

```
free -g
             total       used       free     shared    buffers     cached
Mem:           126         73         52          0          0         10
-/+ buffers/cache:         62         63
Swap:            0          0          0
top - 23:20:07 up 3 days,  9:24,  0 users,  load average: 59.52, 59.48, 54.48
Tasks: 677 total,  62 running, 615 sleeping,   0 stopped,   0 zombie
%Cpu(s): 41.1 us,  0.2 sy,  0.0 ni, 58.6 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem:  13216380+total, 76871760 used, 55292052 free,   100120 buffers
KiB Swap:        0 total,        0 used,        0 free. 10958920 cached Mem

  PID USER      PR  NI    VIRT    RES    SHR S   %CPU  %MEM     TIME+ COMMAND
    1  16995 balewski  20   0 1484276 1.074g  93276 R 100.00 0.852  39:53.75 root4star
    2  17777 balewski  20   0 1495544 1.095g  97052 R 100.00 0.869  38:44.61 root4star
    3  20928 balewski  20   0 1458492 1.051g  95424 R 100.00 0.834  39:06.72 root4star
    4  24046 balewski  20   0 1477820 1.070g  96360 R 100.00 0.849  39:15.17 root4star
    5  26774 balewski  20   0 1544608 1.142g  93240 R 100.00 0.906  38:21.16 root4star
    6   9801 balewski  20   0 1477340 1.070g  95764 R 100.00 0.849  40:12.45 root4star
    7   9802 balewski  20   0 1513796 1.109g  98040 R 100.00 0.880  39:59.50 root4star
    8  10072 balewski  20   0 1514864 1.107g  98828 R 100.00 0.878  40:15.63 root4star
    9  10353 balewski  20   0 1500460 1.093g  97844 R 100.00 0.867  40:33.31 root4star
   10  10732 balewski  20   0 1474472 1.069g  96984 R 100.00 0.848  40:12.61 root4star
   11  11126 balewski  20   0 1464792 1.055g  94736 R 100.00 0.837  34:58.26 root4star
   12  11516 balewski  20   0 1467348 1.055g  93572 R 100.00 0.837  32:38.22 root4star
   13  11907 balewski  20   0 1489552 1.082g  98036 R 100.00 0.859  39:49.53 root4star
   14  12297 balewski  20   0 1463256 1.056g  96176 R 100.00 0.838  37:36.07 root4star
   15  12688 balewski  20   0 1500488 1.098g  97752 R 100.00 0.871  40:12.60 root4star


   51  28717 balewski  20   0 1475208 1.066g  95408 R 100.00 0.846  37:34.33 root4star
   52  29106 balewski  20   0 1529152 1.122g  94064 R 100.00 0.890  38:09.74 root4star
   53  29494 balewski  20   0 1490808 1.086g  96168 R 100.00 0.861  36:11.60 root4star
   54  29889 balewski  20   0 1500088 1.097g  97308 R 100.00 0.870  37:53.78 root4star
   55  30337 balewski  20   0 1504052 1.095g  95376 R 100.00 0.869  37:53.20 root4star
   56  30669 balewski  20   0 1480480 1.075g  97076 R 100.00 0.853  38:39.54 root4star
   57  31058 balewski  20   0 1449216 1.038g  93300 R 100.00 0.824  37:20.52 root4star
   58  31446 balewski  20   0 1484268 1.073g  92828 R 100.00 0.852  38:02.71 root4star
   59  31926 balewski  20   0 1461488 1.054g  96168 R 100.00 0.836  37:08.48 root4star
   60  32224 balewski  20   0 1463356 1.056g  96260 R 100.00 0.838  37:36.12 root4star
```

STAR embedding on Cori by STAR user (Xianglei)

Long SLURM job (15 nodes, 30hrs, premium queue) for a full-scale embedding production of two FSETs at Cori





embedding workflow on UGE - on PDSF-2

procure

```
pdsf6 $ cat preparexml.sh
#!/bin/sh

perl StRoot/macros/embedding/get_embedding_xml.pl \
 -trg production_pp200_2015 -production P16id -lib SL16d \
 -mixer StRoot/macros/embedding/bfcMixer_Tpx.C -prodname P16idpp200 \
 -r 20163401 \
 -geantid 169 -particle Psi2SMuMu -mode FlatPt \
 -trigger 470602 -trigger 480602 -trigger 490602 \
 -z 200.0 -vrcut 100.0 \
 -ymin -0.8 -ymax 0.8 -ptmin 0 -ptmax 12.0 \
 -mult 8 \
 -local \
 -daq /global/projecta/projectdirs/starprod/daq/2015/pp200_mtd \
 -tag /global/projecta/projectdirs/starprod/tags/2015/pp200_mtd
```

submit: few * 600 jobs

```
pdsf6 $ cat subalot.sh
#!/bin/sh

fset="200"
nset="1"

eset=$(($fset+$nset))
for (( i=$fset ; $i - $eset ; i++ ))
do
echo "./submit.csh $i"
./submit.csh $i
done
```

```
pdsf6 $ cat embed_template_P16id.xml
<!-- Generated by StRoot/macros/embedding/get_embedding_xml.pl on Thu Dec  1 14:35:45 PST 2016 -->
<?xml version="1.0" encoding="utf-8"?>
<job maxFilesPerProcess="1" fileListSyntax="paths" simulateSubmission="false">

<command>
<!-- Set Log file -->
setenv LOGFILE /global/projecta/projectdirs/starprod/log/P16id/LOG/$JOBID.log
touch ${LOGFILE}

<!-- Setup permissions -->
umask 2 &gt;&gt;&amp; ${LOGFILE}
chgrp starprod . &gt;&gt;&amp; ${LOGFILE}
chmod g+s . &gt;&gt;&amp; ${LOGFILE}

<!-- Load library -->
starver SL16d &gt;&gt;&amp; ${LOGFILE}

<!-- Set tags file directory -->
setenv EMBEDTAGDIR /global/projecta/projectdirs/starprod/tags/2015/pp200_mtd &gt;&gt;&amp; ${LOGFILE}
```

embedding workflow on Cori

procure

```
pdsf6 $ cat  prepEmbedTaskList.sh
#!/bin/bash

python ./prepEmbedTaskList.py \
 -trg production_pp200_2015 \
 -production P16id -lib SL16d \
 -mixer StRoot/macros/embedding/bfcMixer_Tpx.C -prodname P16idpp200 \
 -r 20163401 \
 -geantid 169 -particle Psi2SMuMu -mode FlatPt \
 -trigger 470602 -trigger 480602 -trigger 490602 \
 -z 200.0 -vrcut 100.0 \
 -ymin -0.8 -ymax 0.8 -ptmin 0 -ptmax 12.0 \
 -mult 8 \
 -local \
 -daq /global/projecta/projectdirs/starprod/daq/2015/pp200_mtd \
 -tag /global/projecta/projectdirs/starprod/tags/2015/pp200_mtd \
 -fSetRange 901-902 \
 -nEve 10 \
# -outPath /global/project/projectdirs/mpccc/balewski/out1 \
# -endGameOpt
```

submit: few * 600 jobs

```
$head starFarmer.slr
#!/bin/bash
#SBATCH -c64 -C haswell # 123 GB RAM
...
myTaskList=starTask_20163401_901_902.list
myTaskCSH=r4sTask_bfc.csh
...
# prepare sandbox
wrkDir=$SCRATCH/starFarm-$SLURM_JOBID
...
./farmerQAutil.sh  $myTaskList
```

```
 $cat  r4sTask_embed.csh
#!/bin/tcsh
# this script produced from template on: May 09 2017 09:19AM
...
  set STAR_VER = SL16d
    set PATH_DAQ = /global/projecta/projectdirs/starprod/daq/2015/pp200_mtd
    set PATH_TAG = /global/projecta/projectdirs/starprod/tags/2015/pp200_mtd

   env | grep  SHIFTER_RUNTIME

    setenv GROUP_DIR /common/star/star${SCHOS}/group/
    source $GROUP_DIR/star_cshrc.csh
    starver $STAR_VER

    echo 'my new STAR ver='$STAR'  test root4star '
    root4star -b -q
...
```

```
$cat starTask_20163401_901_902.list
balewski@cori08:~/tutorNersc/2017-04-cori-starExp> head starTask_20163401_901_9
shifter  /bin/tcsh  ${WRK_DIR}/r4sTask_embed.csh st_mtd_adc_16115053_raw_100000
st_mtd_adc_16115053_raw_1000007_fset901.taskLog
shifter  /bin/tcsh  ${WRK_DIR}/r4sTask_embed.csh st_mtd_adc_16098010_raw_250001
st_mtd_adc_16098010_raw_2500011_fset901.taskLog
shifter  /bin/tcsh  ${WRK_DIR}/r4sTask_embed.csh st_mtd_adc_16102046_raw_100001
st_mtd_adc_16102046_raw_1000010_fset901.taskLog
shifter  /bin/tcsh  ${WRK_DIR}/r4sTask_embed.csh st_mtd_adc_16083058_raw_250000
st_mtd_adc_16083058_raw_2500009_fset901.taskLog
shifter  /bin/tcsh  ${WRK_DIR}/r4sTask_embed.csh st_mtd_adc_16099042_raw_550003
st_mtd_adc_16099042_raw_5500037_fset901.taskLog
shifter  /bin/tcsh  ${WRK_DIR}/r4sTask_embed.csh st_mtd_adc_16069028_raw_200000
st_mtd_adc_16069028_raw_2000009_fset901.taskLog
shifter  /bin/tcsh  ${WRK_DIR}/r4sTask_embed.csh st_mtd_adc_16085057_raw_450000
st_mtd_adc_16085057_raw_4500009_fset901.taskLog
```