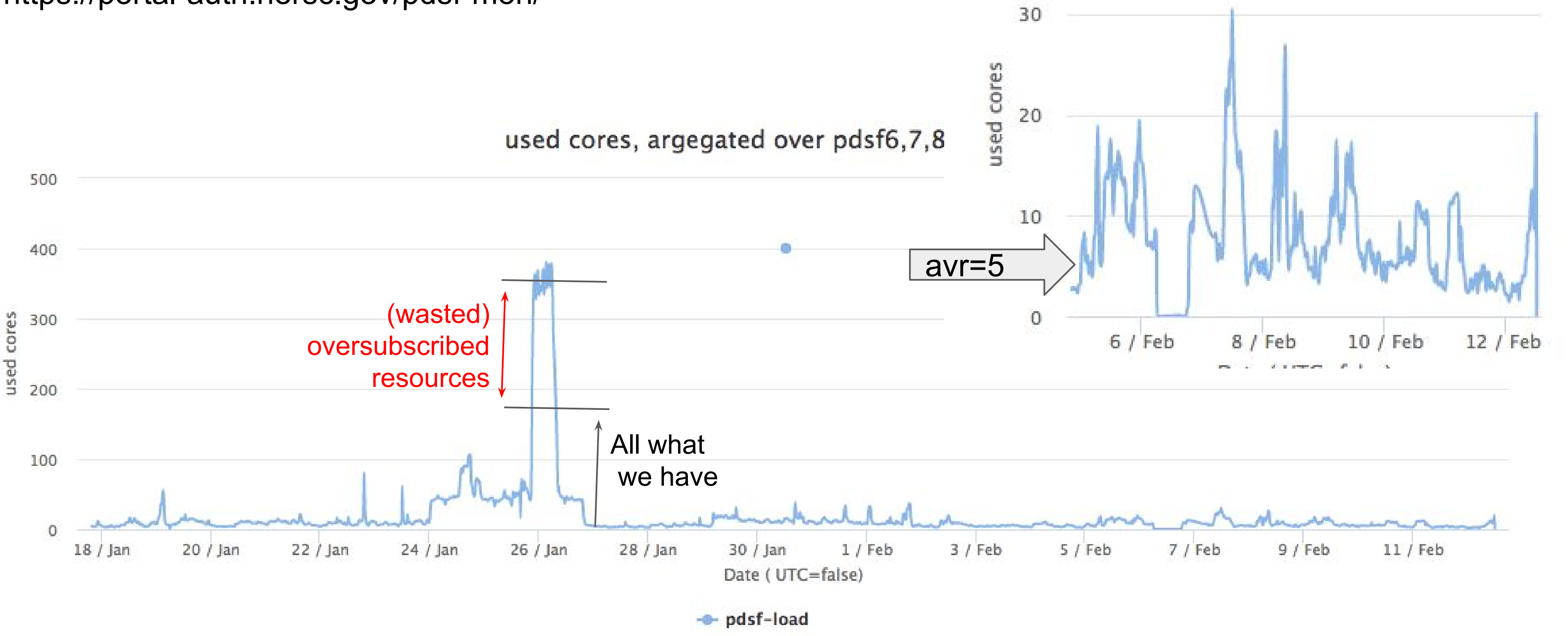
# PDF User Meeting

- PDSF performance
- Announcements
- New PDSF shares
- PDSF tips
- AOB: more Slurm queues at PDSF

# aggregated load on PDSF interactive nodes

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



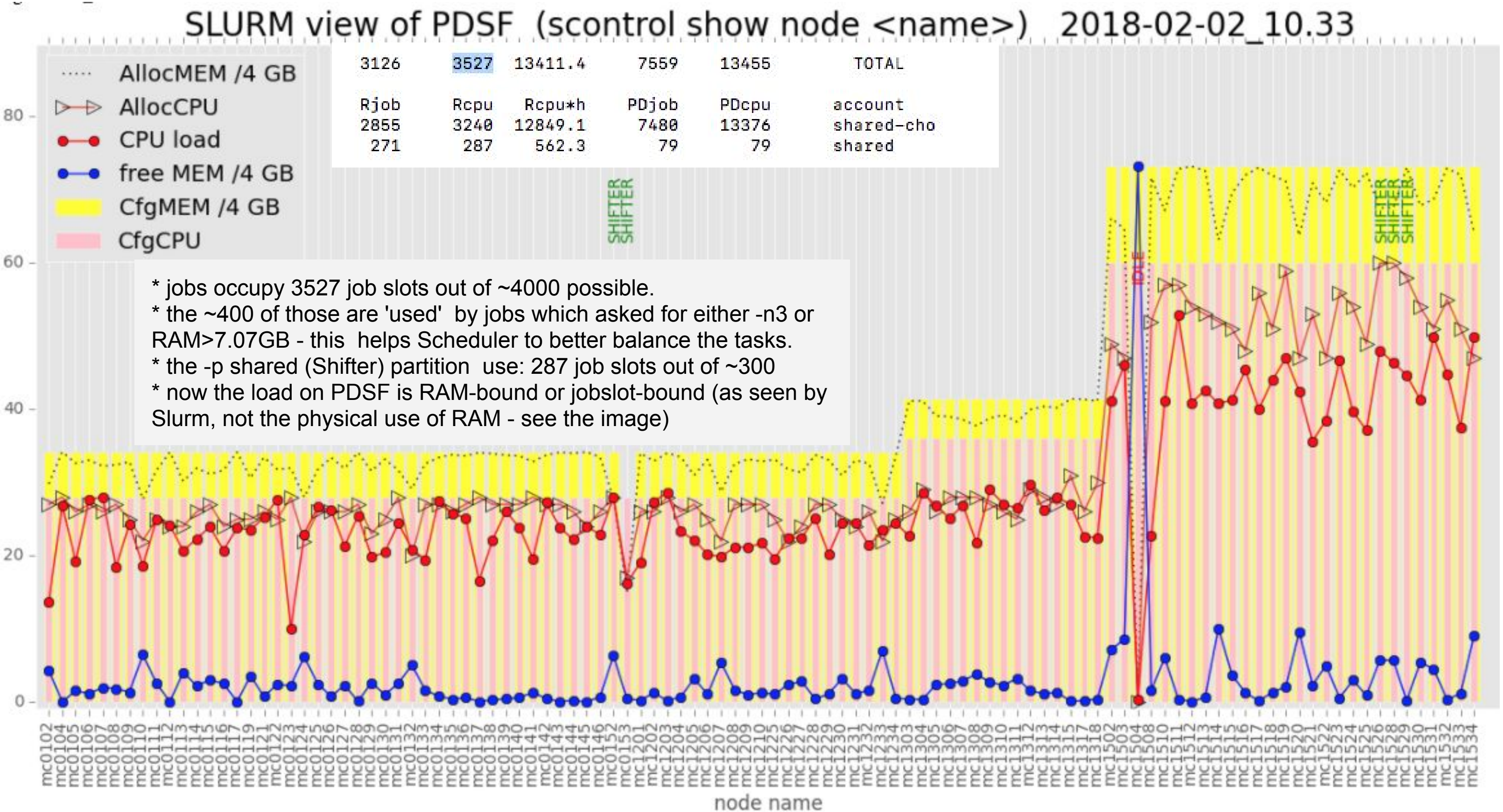
used cores, argegated over pdsf6,7,8

(wasted) oversubscribed resources

All what we have

avr=5

# PDSF load SLURM snapshot



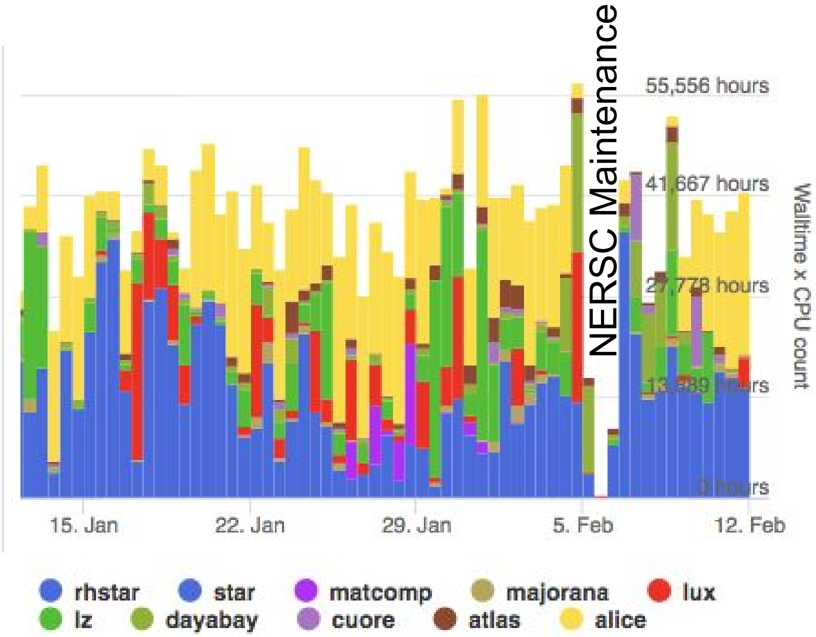
```
3126      3527  13411.4      7559     13455         TOTAL

Rjob      Rcpu    Rcpu*h     PDjob     PDcpu       account
2855      3240  12849.1      7480     13376       shared-cho
 271       287    562.3        79        79       shared
```

* jobs occupy 3527 job slots out of ~4000 possible.
* the ~400 of those are 'used' by jobs which asked for either -n3 or RAM>7.07GB - this helps Scheduler to better balance the tasks.
* the -p shared (Shifter) partition use: 287 job slots out of ~300
* now the load on PDSF is RAM-bound or jobslot-bound (as seen by Slurm, not the physical use of RAM - see the image)

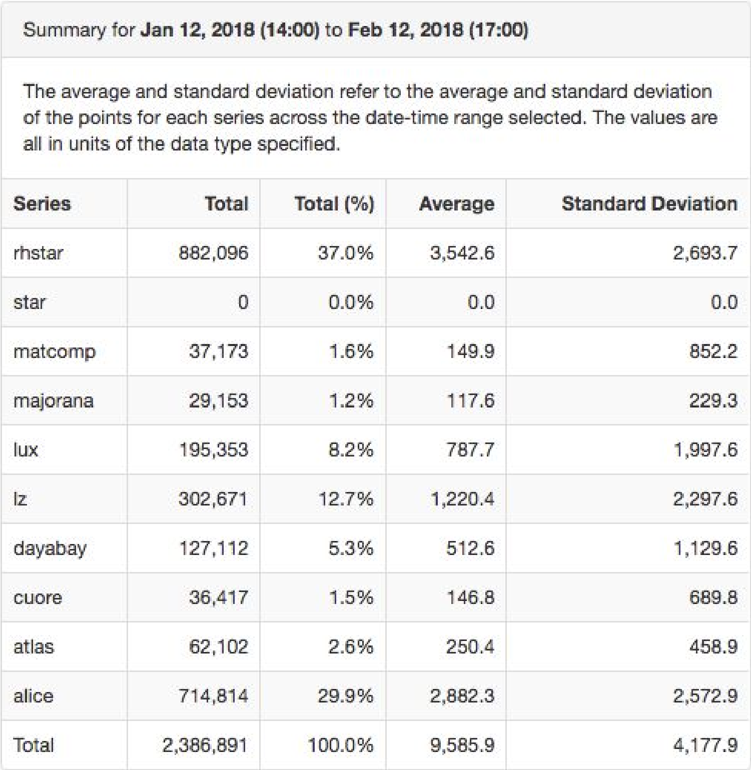
# SLURM CPU aggregated over last month

SLURM : **completed** jobs in last month

http://portal.nersc.gov/project/mpccc/ebasheer/jobbygroup.php



3800 jobs * 24 h = 91k cpu*h/day
→ 640k cpu*h /week
→ 2.7 M cpu*h per month



Summary for **Jan 12, 2018 (14:00)** to **Feb 12, 2018 (17:00)**

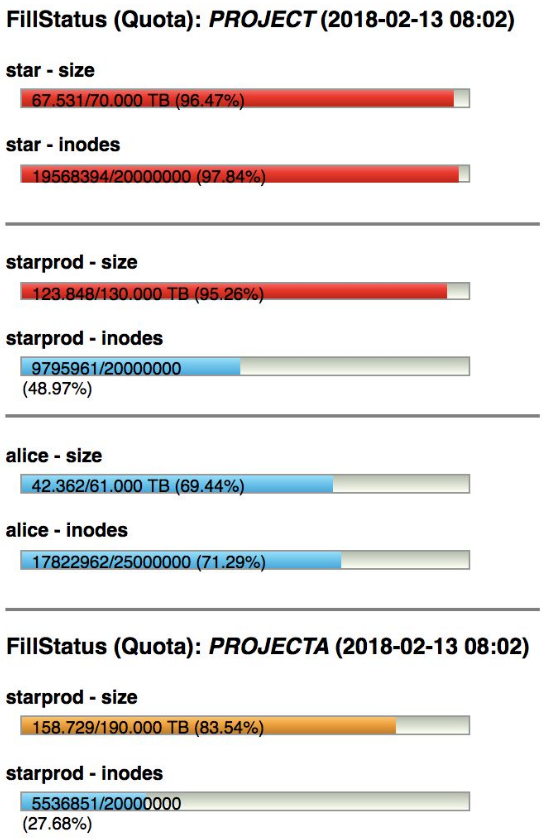
The average and standard deviation refer to the average and standard deviation of the points for each series across the date-time range selected. The values are all in units of the data type specified.

| Series | Total | Total (%) | Average | Standard Deviation |
|---|---:|---:|---:|---:|
| rhstar | 882,096 | 37.0% | 3,542.6 | 2,693.7 |
| star | 0 | 0.0% | 0.0 | 0.0 |
| matcomp | 37,173 | 1.6% | 149.9 | 852.2 |
| majorana | 29,153 | 1.2% | 117.6 | 229.3 |
| lux | 195,353 | 8.2% | 787.7 | 1,997.6 |
| lz | 302,671 | 12.7% | 1,220.4 | 2,297.6 |
| dayabay | 127,112 | 5.3% | 512.6 | 1,129.6 |
| cuore | 36,417 | 1.5% | 146.8 | 689.8 |
| atlas | 62,102 | 2.6% | 250.4 | 458.9 |
| alice | 714,814 | 29.9% | 2,882.3 | 2,572.9 |
| Total | 2,386,891 | 100.0% | 9,585.9 | 4,177.9 |

# /project(a)  utilization - snapshot

http://portal.nersc.gov/project/star/jthaeder/diskUsage/overview/indexExt.html
https://my.nersc.gov/data-mgt.php

```
cori12:~>  prjquota dayabay
           ---------- Space (GB) ---------    ------------- Inode --------------
Project        Usage      Quota    Percent       Usage       Quota       Percent
-------------  ---------  ---------  ---------   ----------  ----------  ----------
dayabay          848962     870400       97    101017027   150000000         67

balewski@cori10:~> prjaquota dayabay
           ---------- Space (GB) ---------    ------------- Inode --------------
Project        Usage      Quota    Percent       Usage       Quota       Percent
-------------  ---------  ---------  ---------   ----------  ----------  ----------
dayabay          836891     870400       96      6224434    10000000         62


balewski@cori10:~> prjquota majorana
           ---------- Space (GB) ---------    ------------- Inode --------------
Project        Usage      Quota    Percent       Usage       Quota       Percent
-------------  ---------  ---------  ---------   ----------  ----------  ----------
majorana          38824      40960       94      3279917     4000000         82

balewski@cori10:~> prjaquota majorana
           ---------- Space (GB) ---------    ------------- Inode --------------
Project        Usage      Quota    Percent       Usage       Quota       Percent
-------------  ---------  ---------  ---------   ----------  ----------  ----------
majorana          57875      61440       94      4456934    10000000         44
```



**FillStatus (Quota):** ***PROJECT*** **(2018-02-13 08:02)**

**star - size**

67.531/70.000 TB (96.47%)

**star - inodes**

19568394/20000000 (97.84%)

**starprod - size**

123.848/130.000 TB (95.26%)

**starprod - inodes**

9795961/20000000 (48.97%)

**alice - size**

42.362/61.000 TB (69.44%)

**alice - inodes**

17822962/25000000 (71.29%)

**FillStatus (Quota):** ***PROJECTA*** **(2018-02-13 08:02)**

**starprod - size**

158.729/190.000 TB (83.54%)

**starprod - inodes**

5536851/20000000 (27.68%)

# Announcements

PDF SC meet on Feb 13 to discuss timeline for Mendel shutdown

- we have 14 months of ‘business as usual’ before Mendel (compute nodes & services) powers off, Shifter+Cori is your best bet

Jan: do not wait till April 2019 to move to Cori.

Bi-weekly office hours Feb 14, March 1, 59-4016A

PDSF user meeting

- Tuesday, March 13

**PDSF shares in Slurm will change for 2018**

Outages :

14 Feb HPSS Scheduled Maintenance

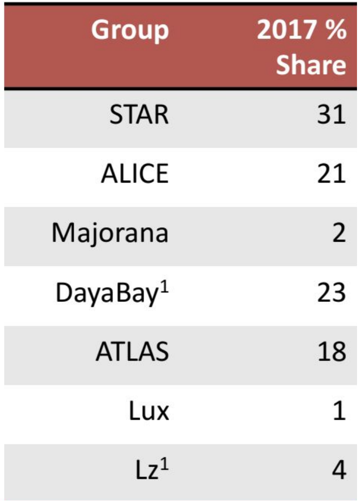
# PDSF shares

**PDSF shares are ‘redeemed’ as used CPU*hours with Slurm memory half-life of 2 weeks**

PDSF shares are NOT ‘guaranteed instant fraction of existing CPUs’ because we do not kill your running jobs if you happen to grab more tasks slots than your share.

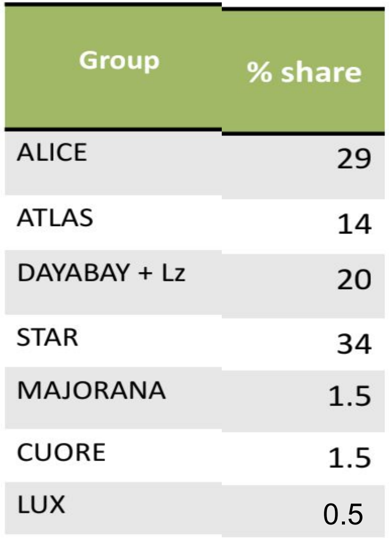
Instead, we count how much you have used and reduce priority **your experiment** afterwards. (accounting is per experiment, not per user)

Some users require more RAM or multiple CPUs per Slurm job - we account for this as well.
(see next slides)



| Group | 2017 % Share |
|---|---|
| STAR | 31 |
| ALICE | 21 |
| Majorana | 2 |
| DayaBay[1] | 23 |
| ATLAS | 18 |
| Lux | 1 |
| Lz[1] | 4 |

to be implemented soon for 2018



| Group | % share |
|---|---|
| ALICE | 29 |
| ATLAS | 14 |
| DAYABAY + Lz | 20 |
| STAR | 34 |
| MAJORANA | 1.5 |
| CUORE | 1.5 |
| LUX | 0.5 |

[1] Not shown: share was moved from DB to Lz during the year

# RAM usage and job ‘charge’ on PDSF Slurm

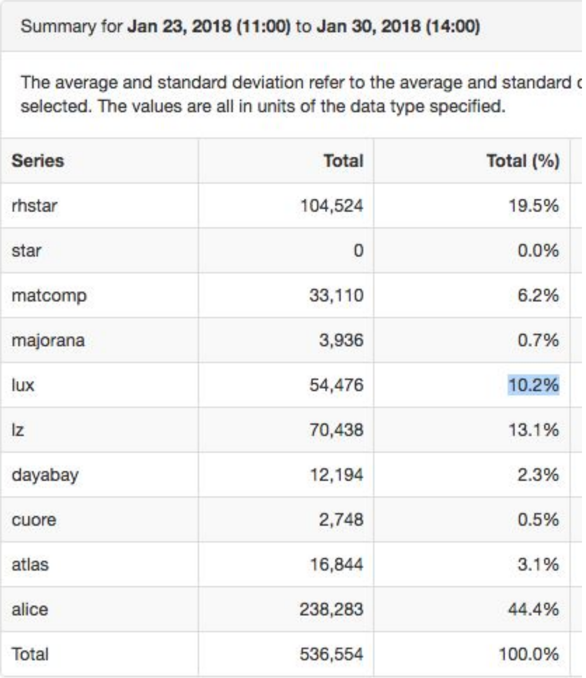
**2018 Guidelines**

1. default RAM per task is set at 4 GB - do not reduce it even if you know you need less
2. if your task exceeds the default RAM even for a fractions of a second
   Slurm may kill it , or swap will degrade performance (also for other tasks on the node)
3. if you need more RAM, than ask for as much as you need, up to 120 GB
4. if you request above 7.07 GB RAM than it will be automatically converted to
   a higher 'charge' : ‘n’ tasks =ceil(--mem RAM/ 7.07 GB).
5. you may request more vCores/task (e.g. #SLURM -n10) which will automatically change the default RAM to n*4 GB. You will be charged n x more for such job
6. You can request both: **#SLURM -n5 --ram 50 GB** will result with -n8 and 50 GB RAM limit
7. Tasks with larger ‘n’ are harder to schedule. ~Half of PDSF nodes support n <61, half n<29

**Georg : Slurm is running more stable since we added ‘7.07GB per task’ conversion.**

# Why my job is not running?

The share of LUX in PDSF is very low - LUX gets only 1%, this is average over weeks target use.
This page shows how much LUX have used (takes 30 sec to start)
http://portal.nersc.gov/project/mpccc/ebasheer/jobbygroup.php
Over last week LUX used 10% of PDSF:

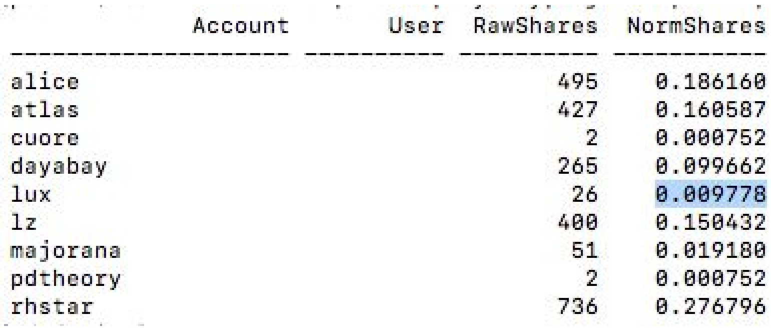


Summary for **Jan 23, 2018 (11:00)** to **Jan 30, 2018 (14:00)**

The average and standard deviation refer to the average and standard c
selected. The values are all in units of the data type specified.

| Series | Total | Total (%) |
|---|---|---|
| rhstar | 104,524 | 19.5% |
| star | 0 | 0.0% |
| matcomp | 33,110 | 6.2% |
| majorana | 3,936 | 0.7% |
| lux | 54,476 | 10.2% |
| lz | 70,438 | 13.1% |
| dayabay | 12,194 | 2.3% |
| cuore | 2,748 | 0.5% |
| atlas | 16,844 | 3.1% |
| alice | 238,283 | 44.4% |
| Total | 536,554 | 100.0% |

```
$pdsf06 sshare -A alice,rhstar,dayabay,majorana,atlas,lz,lux,cuore,pdtheory -l
```



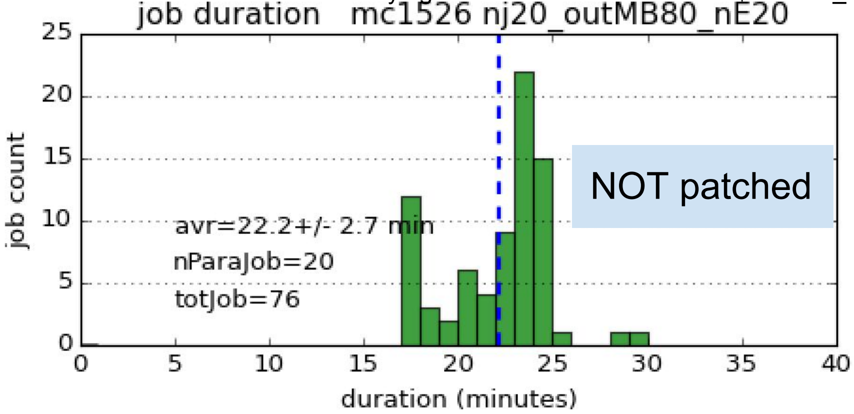
```
             Account       User  RawShares  NormShares
-------------------- ---------- ---------- -----------
alice                                  495    0.186160
atlas                                  427    0.160587
cuore                                    2    0.000752
dayabay                                265    0.099662
lux                                     26    0.009778
lz                                     400    0.150432
majorana                                51    0.019180
pdtheory                                 2    0.000752
rhstar                                 736    0.276796
```

**Disclaimer: LUX is used here as an example only.**

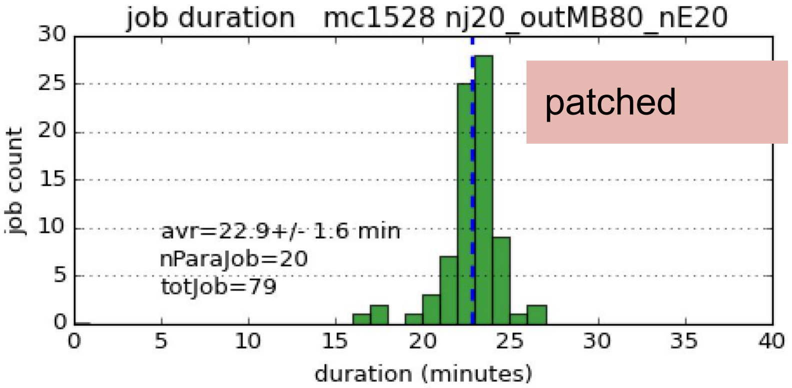
# Spectre patch : 80 MB/min/task performance

Tested with 1-core task doing CPU and writing 80 MB/min in bursts → ~5 GB/hour/task

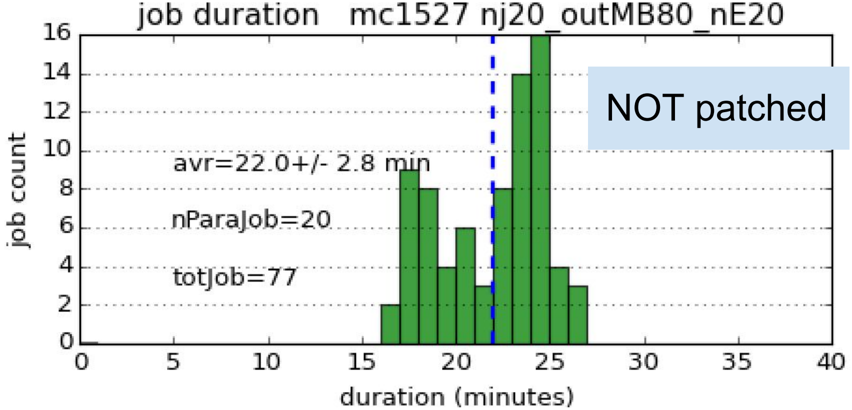
all 4 nodes were tested concurrently against local disc on the node: $SLURM_TMP



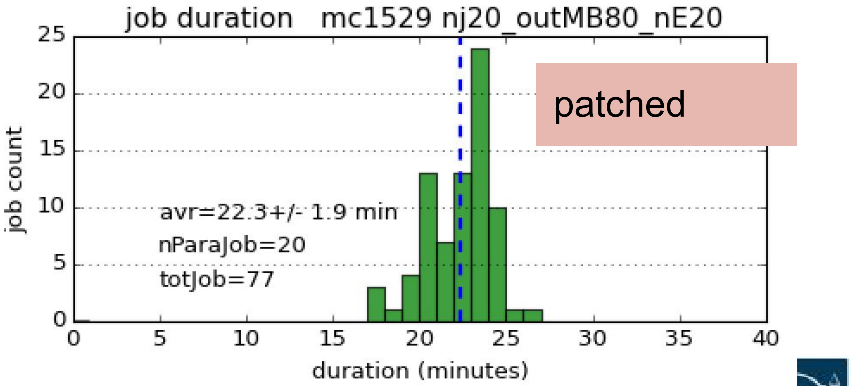
NOT patched



patched



NOT patched



patched

# STAR reconstruction on Cori  news article

http://www.nersc.gov/news-publications/nersc-news/science-news/2018/new-nersc-data-processing-framework-dramatically-cuts-reconstruction-time/

Collaborative work between NERSC & BNL

PHYSICS DATA PROCESSING AT NERSC DRAMATICALLY CUTS RECONSTRUCTION TIME

Demonstration Project Targets Nuclear Physics Data from STAR Experiment

# More Slurm queues

We have now: 2 queues: 48 h duration, priority ~PDSF shares, **3300 job-slots,**
The existing queues are in blue

Proposal: establish 5 queues: shares 'natural' in all queues, try it in ~2 weeks

1. **'-p shared-chos':** Big production, goal : latency O(1), throughput O(3), concurrency O(3), duration>5h
   Num slots=2550, maxTime=48h, only CHOS, (this exist, no changes except capacity)

2. **'-p shared-short'** User analysis, goal: latency O(2), throughput O(2), concurrency O(2), duration~[2-5]h
   Num slots=200, max 50 tasks per user, maxTime=5h, only Shifter

3. **'-p shared-long'** User analysis, goal: latency O(2), throughput O(2), concurrency O(2), duration 48h
   Num slots=300, maxTime=48h, only Shifter (this is current -p shared, only renamed)

4. **'-p realtime'** Real-time analysis, goal: latency O(3), throughput O(1), concurrency O(1), duration<4h
   Num slots=200, exclusive resource, max 50 tasks per group, maxTime=4h, only Shifter

5. **'-p debug'** goal: duration<30min, latency O(3), throughput O(1), concurrency O(1)
   Num slots=50, exclusive resource, max 2 tasks per user, maxTime=30m, only Shifter