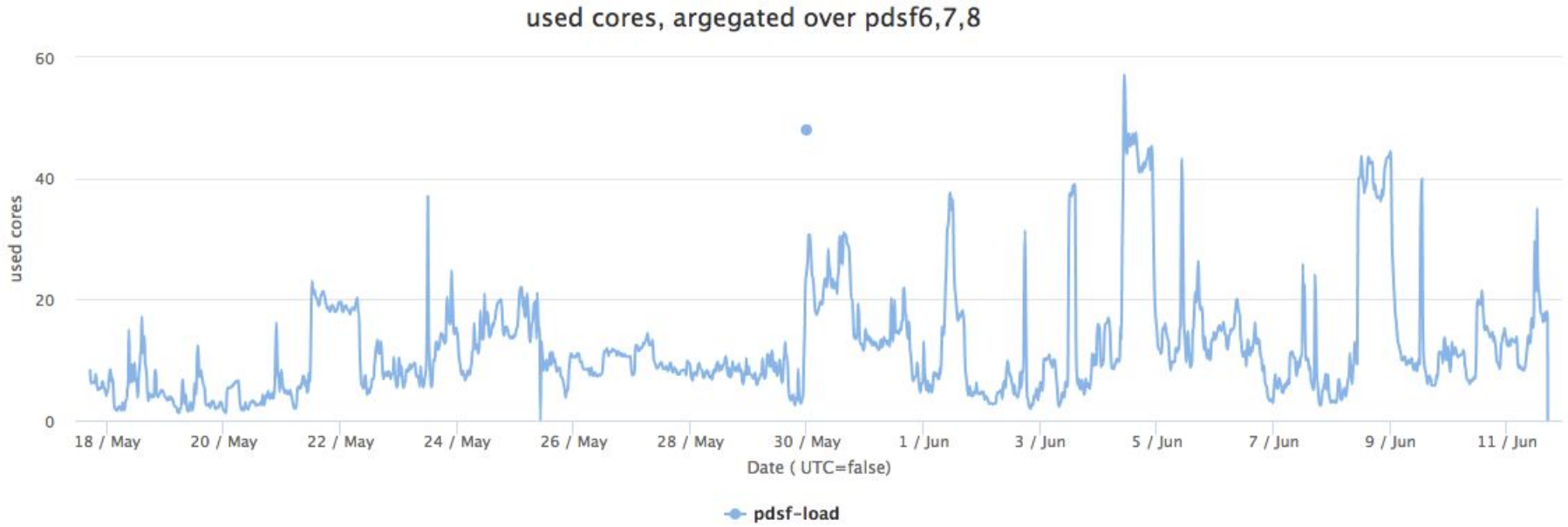
PDSF User Meeting

- PDSF performance
- Announcements
- AOB
- Action items

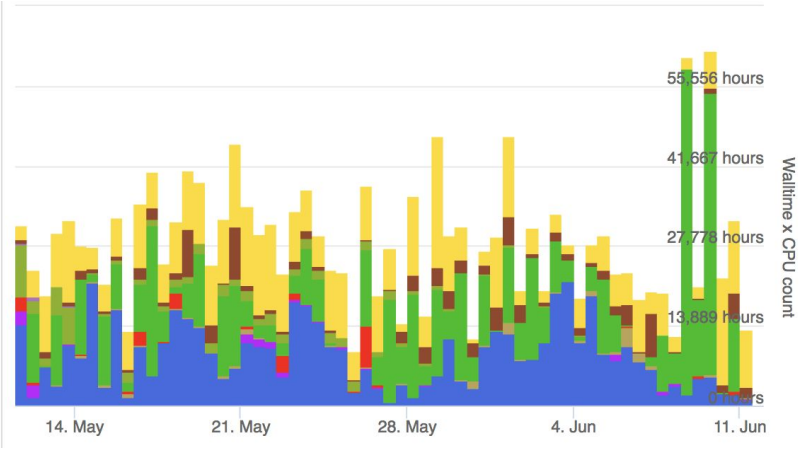
aggregated load on PDSF interactive nodes

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



SLURM CPU*h aggregated over last month

SLURM : **completed** jobs in last month
http://portal.nersc.gov/project/mpccc/ebasheer/jobbygroup.php



Summary for **May 11, 2018 (18:00)** to **Jun 11, 2018 (21:00)**

The average and standard deviation refer to the average and standard dev
series across the date-time range selected. The values are all in units of th

Series	Total	Total (%)	Average
rhstar	537,796	30.4%	2,159.8
star	0	0.0%	0.0
matcomp	13,481	0.8%	54.1
majorana	19,088	1.1%	76.7
lux	22,714	1.3%	91.6
lz	525,663	29.7%	2,111.1
dayabay	66,597	3.8%	268.5
cuore	1,246	0.1%	5.0
atlas	100,745	5.7%	404.6
alice	481,487	27.2%	1,933.7
Total	1,768,817	100.0%	7,103.7

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

Existing Slurm Shifter queues (unchanged)

partition	OS provider	TotalCPUs	Time limit	MaxJobPA (per account)	MaxJobPU (per user)	Relative priority	remarks
shared-chos	chos	3352	2 days		250	0	Share 94% of hardware
alice	chos	3224	2 days		1500	0	
realtime-chos	chos	128	4 hours	50		0	
debug-chos	chos	128	30 min		2	10	
long	shifter	384	2 days			0	
short	shifter	288	5 hours			0	
realtime	shifter	128	4 hours	50		0	share common hardware
debug	shifter	128	30 min		2	10	

The capacity of queues is as shown below, note **totalCPus** denotes max number of jobs when 2 tasks can run on a single physical core. This is disabled by default and user must specify: #SBATCH --mem1900M --oversubscribe
to actually run that way.

List QOS limits :pdsf6 $ **sacctmgr show qos format=Name,Priority,GrpTRES,MaxTRESPU,MaxTRESPA%100,MaxJobsPA,MaxJobsPU**

List timeouts: pdsf6 $ **sinfo -a ; scontrol show partition -a short**

Slurm defaults (unchanged)

The following changes were enforced on Slurm all queues :

- Use true RAM per node
- compute high mem use tax : nCPU= mem/4 GB
- Num job slots per node: 2x phys cores, but …
- Each job locks 1 physical core by default but is available by using --oversubscribe
- Default mem per task: 4 GB (no changes)
- Used cpu*h half decay time: 4 days (was 14 days)
- Cap of 250 jobs/user is enforced, pseudo-user alicesgm is an exception

New working point set on April 10, 2018

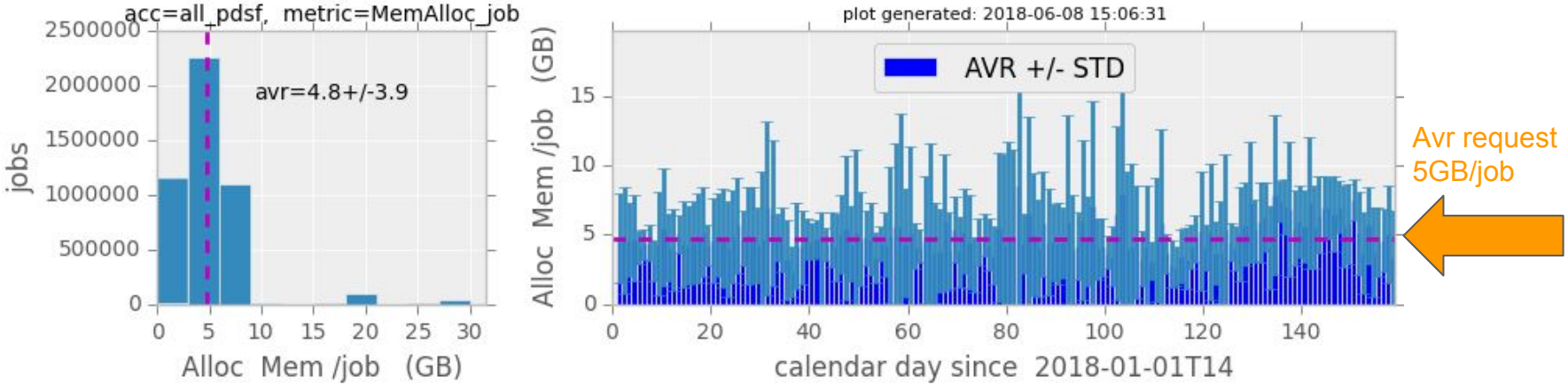
TIP: If you are certain your jobs need less than 1.9 GB of RAM, you can add this line to your slurm job description:
#SBATCH --mem1900M --oversubscribe
and most likely it will double the number of running jobs(if free slots are available), but they will run 50% slower
- your yield/wall hour can increase by 20-30%.

Use of --mem >4.0 GB locks CPU cores

FYI, PDSF consists of nodes w/ **4 GB RAM/CPU**:

- 16 CPUs & 64 GB RAM
- 32 CPUs & 128 GB RAM (Haswell)

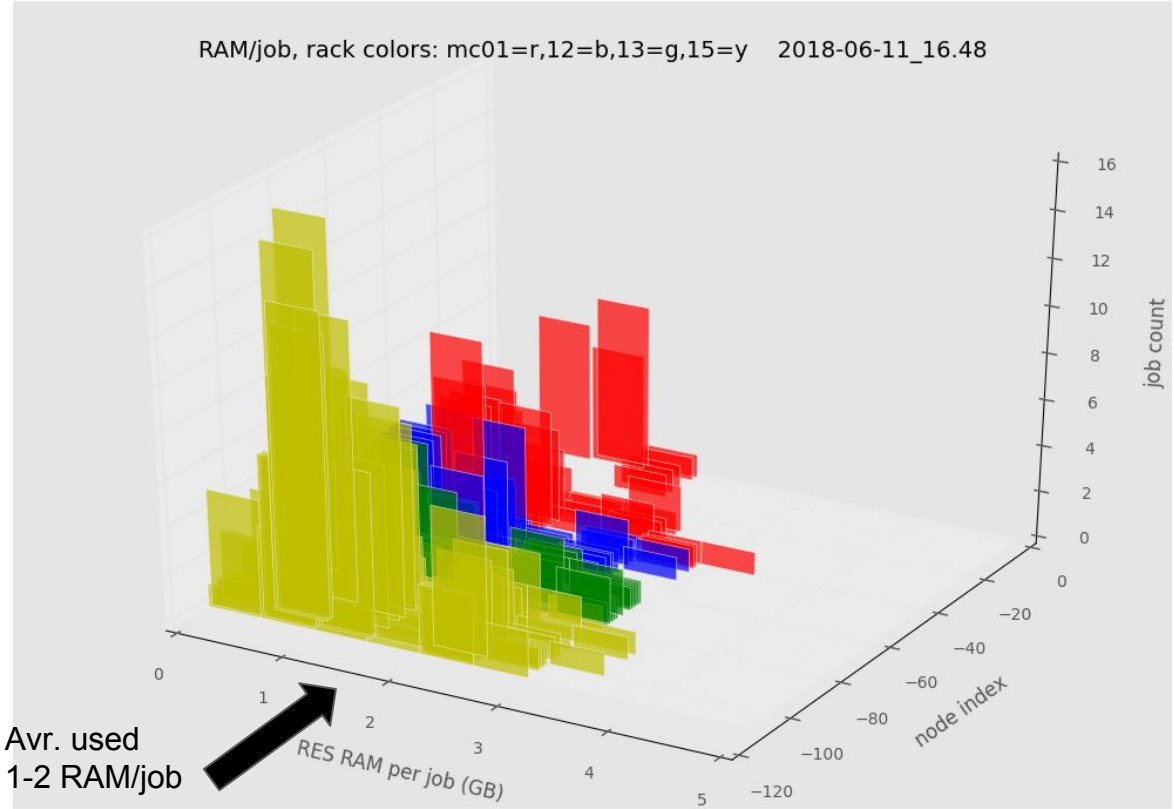
A job : --mem 8GB locks 2 CPUs, 16 such jobs locks full Haswell
If your task needs 1.3 GB of RAM - we could run 40-50 of your tasks on Haswell !



How much RAM per job is actually needed?

Users ask for 3-5 times more RAM they need.

→ PDSF utilization plunges



PDSF jobs profile in 2018 by project

Used Max RAM

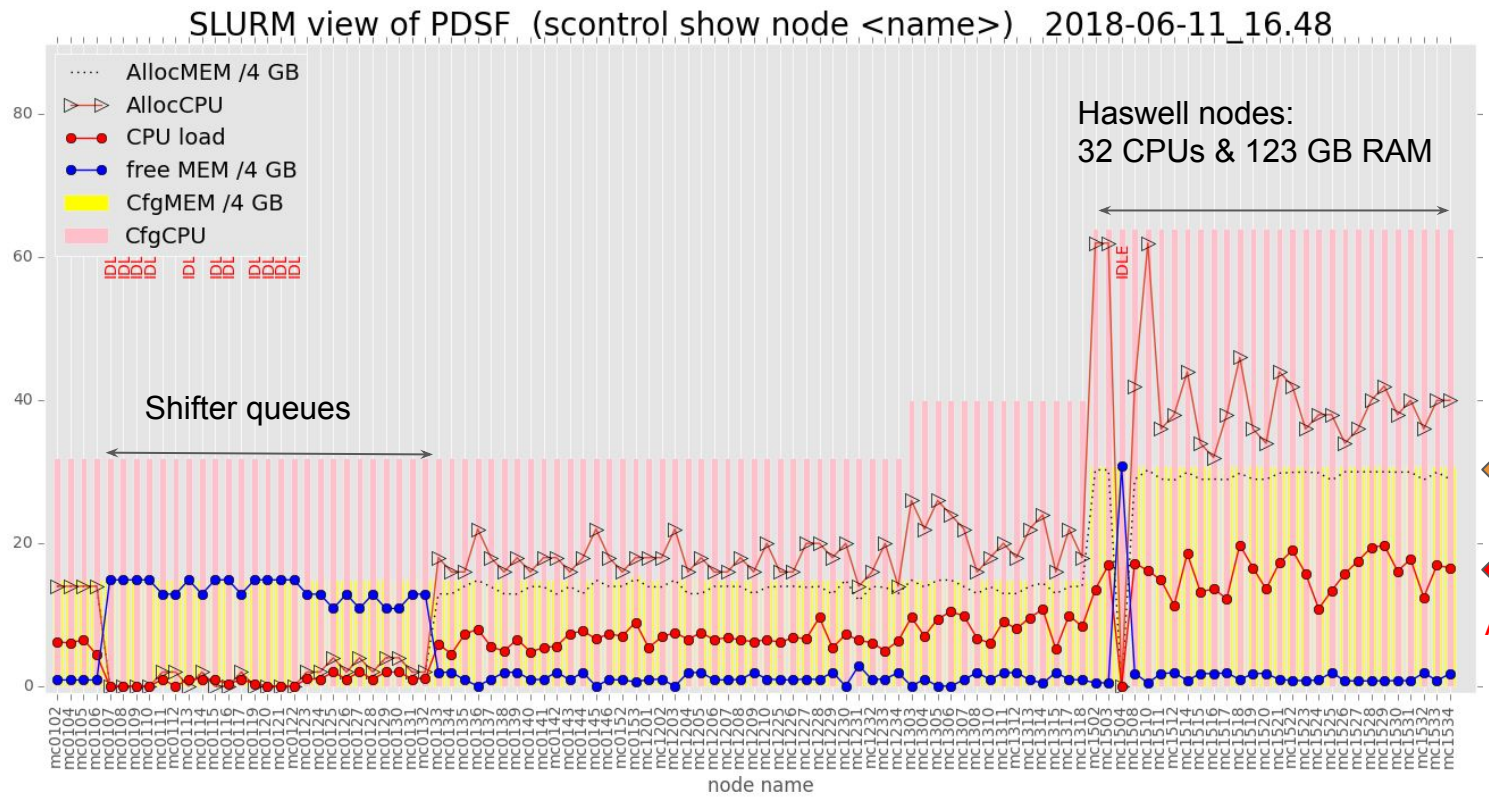
Asked RAM is 3-5 larger

Recent PLOT , produced 2018-06-08_15.15

account + period (plots)	CPUs_job avr +/- std (vCores)	MaxRSS_job avr +/- std (GB)	MaxRSS_wallT avr +/- std (GB)	MemAlloc_job avr +/- std (GB)	cpu2wall_job avr +/- std (eff)	cpu2wall_wallT avr +/- std (eff)	nodeFail_job avr +/- std	wallH_job avr +/- std (hours)
alice 158 days	1.5 +/- 0.5	0.7 +/- 1.1	2.1 +/- 1.1	5.3 +/- 1.9	0.5 +/- 0.4	0.8 +/- 0.3	7.8 +/- 3.3	2.7 +/- 4.9
rhstar 158 days	1.4 +/- 1.0	0.4 +/- 0.6	0.7 +/- 0.6	4.8 +/- 4.1	0.7 +/- 0.4	0.8 +/- 0.3	2.1 +/- 2.1	1.4 +/- 2.6
dayabay 158 days	1.2 +/- 0.5	0.4 +/- 0.9	0.5 +/- 1.0	3.7 +/- 1.8	0.5 +/- 0.4	0.4 +/- 0.4	2.7 +/- 2.6	3.0 +/- 5.0
majorana 158 days	1.4 +/- 0.6	0.2 +/- 0.4	0.7 +/- 0.8	3.7 +/- 2.6	0.4 +/- 0.4	0.7 +/- 0.4	6.2 +/- 6.5	0.3 +/- 1.1
atlas 158 days	1.6 +/- 1.1	0.6 +/- 0.9	0.9 +/- 0.9	5.2 +/- 5.8	0.6 +/- 0.3	0.7 +/- 0.3	1.7 +/- 0.0	0.3 +/- 0.8
lz 158 days	1.4 +/- 1.2	0.9 +/- 1.3	1.1 +/- 1.4	4.7 +/- 2.2	0.8 +/- 0.4	1.0 +/- 0.1	12.3 +/- 11.3	2.9 +/- 4.6
lux 158 days	2.3 +/- 5.1	0.4 +/- 4.0	0.8 +/- 3.9	4.9 +/- 3.7	0.6 +/- 0.4	0.9 +/- 0.3	15.3 +/- 0.1	1.1 +/- 3.1
cuore 158 days	1.0 +/- 0.2	0.4 +/- 0.6	0.6 +/- 0.7	3.3 +/- 1.4	0.7 +/- 0.4	0.9 +/- 0.3	0.0 +/- -7.0	1.4 +/- 2.6
all_pdsf 158 days	1.4 +/- 1.1	0.5 +/- 1.0	1.2 +/- 1.2	4.8 +/- 3.9	0.6 +/- 0.4	0.8 +/- 0.3	3.9 +/- 5.7	1.4 +/- 3.2
account + period (plots)	CPUs_job avr +/- std (vCores)	MaxRSS_job avr +/- std (GB)	MaxRSS_wallT avr +/- std (GB)	MemAlloc_job avr +/- std (GB)	cpu2wall_job avr +/- std (eff)	cpu2wall_wallT avr +/- std (eff)	nodeFail_job avr +/- std (h)	wallH_job avr +/- std (hours)

- Detailed plots: http://portal.nersc.gov/project/mpccc/balewski/tryAny/tmp-jeff3/

PDSF load - SLURM snapshot

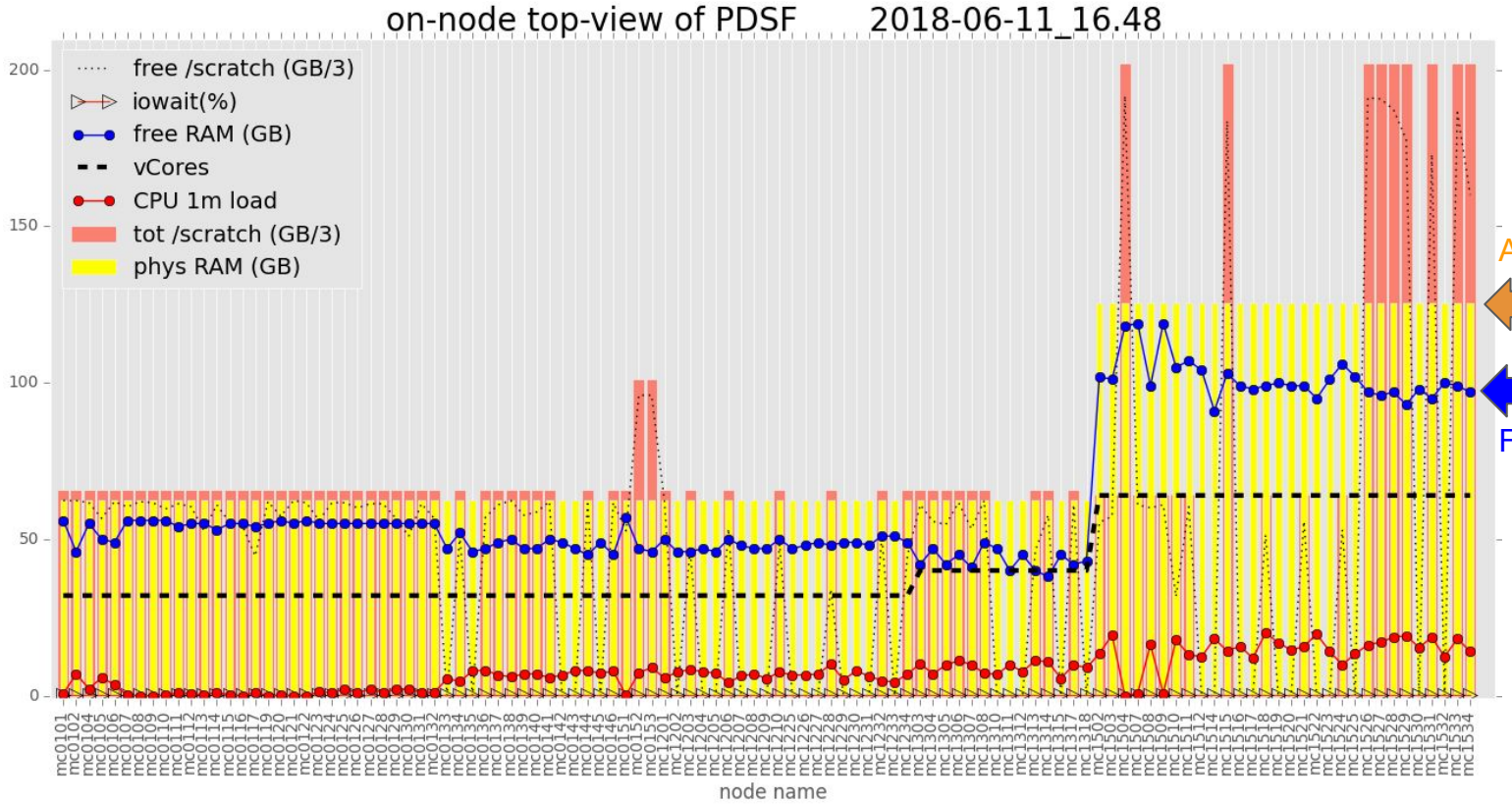


Possible load: 32

Achieved load: 16

System is underutilized

PDSF load - node snapshot



Only ¼ of available RAM is actually used

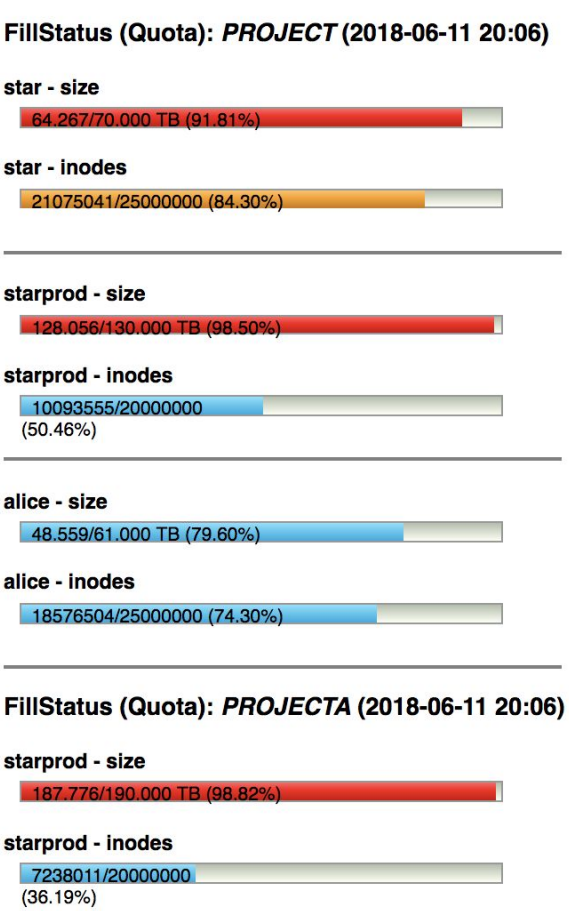
/project(a) utilization - snapshot

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           844341    870400       97      125840581  150000000        83
balewski@cori06:~>  prjaquota dayabay
             ---------- Space (GB) ---------     ------------- Inode --------------
Project          Usage     Quota    Percent        Usage      Quota     Percent
---------------  --------- --------- ---------    ---------- ---------- ----------
dayabay           886813   1126400       78       6369886   10000000        63

balewski@cori06:~>  prjquota majorana
             ---------- Space (GB) ---------     ------------- Inode --------------
Project          Usage     Quota    Percent        Usage      Quota     Percent
---------------  --------- --------- ---------    ---------- ---------- ----------
majorana          38491     40960       94       3611135    4000000        90
balewski@cori06:~>  prjaquota majorana
             ---------- Space (GB) ---------     ------------- Inode --------------
Project          Usage     Quota    Percent        Usage      Quota     Percent
---------------  --------- --------- ---------    ---------- ---------- ----------
majorana          57834     61440       94       6465561   10000000        64
```

Add Lz next time.



Past (May) Action Items

- Implement rolling upgrade of CVMFS to 2.5.0 - done
- Keep Snapshot of Job Stats twice a month , ~done, last 6 months is posted here: http://portal.nersc.gov/project/mpccc/balewski/tryAny/tmp-jeff3/
- Bump of Prio for the Alice users temporary - done, undone
- Turn off MaxMemPerCPU off for ALICE partition - done
- Bump alicesgm Priority using static QOS permanent -done

Action Items (June)

- PDSF shares will be adjusted to 2018 values, need final shares from Jeff
- Default RAM per job will be reduced from 4 GB to 2.5 GB/job next Tuesday
- Jan will send reminder to users to evaluate & reduce needed RAM
- The capacity & use of Shifter partitions was discussed, large PDSF shareholders do not need it, no decision, to be continued at the SC meeting on June 22

How much RAM my job used ?

You can verify how much RAM you completed Slurm job needed by executing one this commands:

a) for one job (e.g. 153 MB for jobId=21115):

```
pdsf7 $ sacct --format=job,start,maxrs -j 21115
       JobID         MaxRSS
12394_15.ba+ 2017-12-04T23:47:27    152,584K
```

b) for all jobs run by user=alicesgm since midnight today

```
pdsf7 $ sacct --format=job,start,maxrs -u alicesgm | grep K
5237842.bat+ 2018-06-12T05:54:28   3987488K
5238047.bat+ 2018-06-12T05:55:28   2530456K
5238207.bat+ 2018-06-12T06:15:31   4336212K
```

c) for all jobs run by all STAR user since June 7

```
pdsf7 $ sacct --starttime 2018-06-07  --format=job,start,account,maxrs -a -A rhstar | grep K
5030343.bat+ 2018-06-06T23:10:34     rhstar    381080K
5030371.bat+ 2018-06-06T23:01:57     rhstar    381984K
5030384.bat+ 2018-06-06T23:10:36     rhstar    384376K
5030385.bat+ 2018-06-06T23:01:57     rhstar    385280K
```

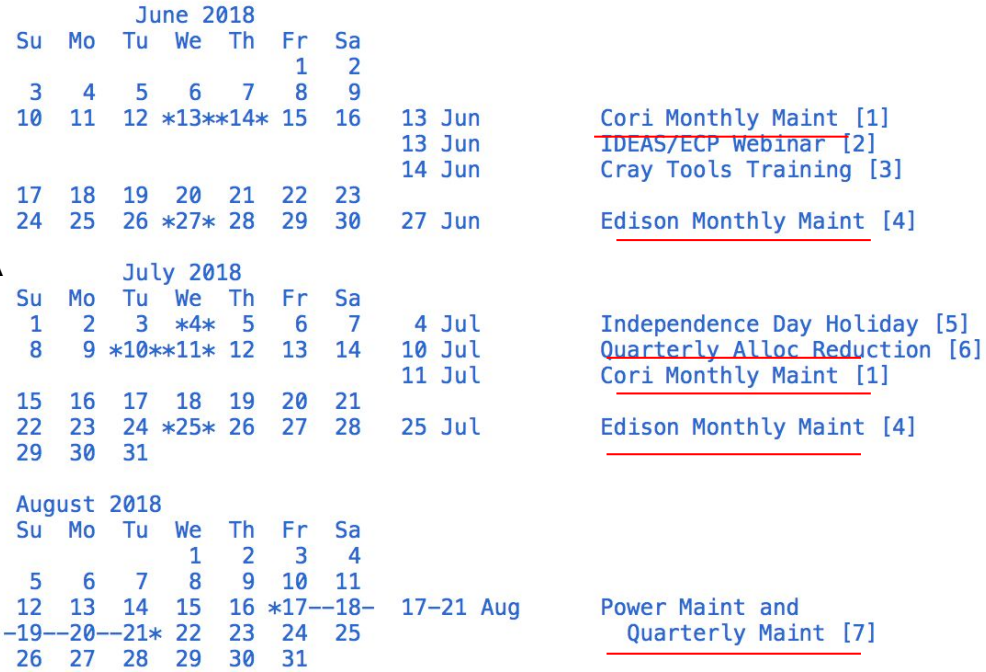
Announcements

Slurm security patch applied 5/30/2018

PDSF Steering Committee meeting: June 22

Bi-weekly office hours June 21, July 5, 59-4016A

PDSF user meeting: Tuesday, July 10



June 2018

Su	Mo	Tu	We	Th	Fr	Sa
					1	2
3	4	5	6	7	8	9
10	11	12	*13*	*14*	15	16
17	18	19	20	21	22	23
24	25	26	*27*	28	29	30

13 Jun	Cori Monthly Maint [1]
13 Jun	IDEAS/ECP Webinar [2]
14 Jun	Cray Tools Training [3]
27 Jun	Edison Monthly Maint [4]

July 2018

Su	Mo	Tu	We	Th	Fr	Sa
1	2	3	*4*	5	6	7
8	9	*10*	*11*	12	13	14
15	16	17	18	19	20	21
22	23	24	*25*	26	27	28
29	30	31				

4 Jul	Independence Day Holiday [5]
10 Jul	Quarterly Alloc Reduction [6]
11 Jul	Cori Monthly Maint [1]
25 Jul	Edison Monthly Maint [4]

August 2018

Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	*17*	*18*
19	*20*	*21*	22	23	24	25
26	27	28	29	30	31	

17-21 Aug	Power Maint and Quarterly Maint [7]