PDSF User Meeting

- PDSF performance
- Announcements
- AOB

PDSF Shutdown in 2 months (soft date)

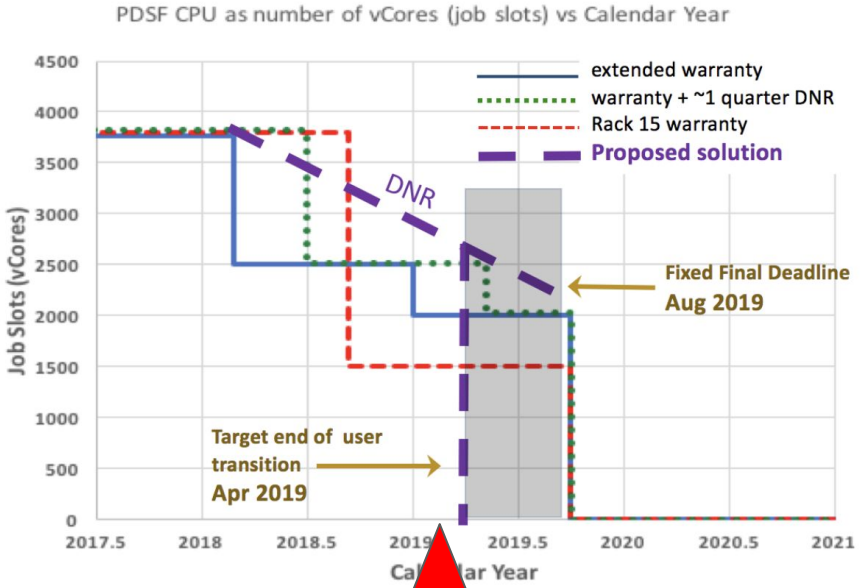
PDSF/Mendel retirement schedule

Instead of retiring components piecemeal, we propose a simpler model : keep all hardware DNR till agreed date, then shut down all at once.

While we considered scenarios to retire individual racks at different times based on support contracts, with service nodes distributed all over the racks move of all these nodes was going to cause NERSC staff and PDSF experts significant effort and PDSF users service interruptions.

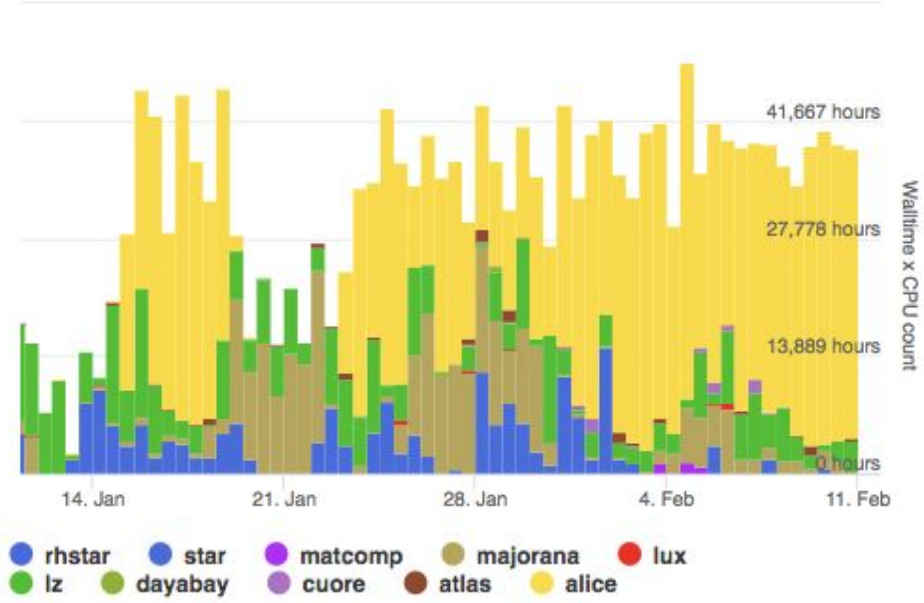
Assume services running on Mendel end April 2019.

Contingency period of 5 months to Aug. 2019



SLURM CPU*h aggregated over last month

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



Summary for **Jan 11, 2019 (14:00)** to **Feb 11, 2019 (1**

The average and standard deviation refer to the avera each series across the date-time range selected. The specified.

Series	Total	Total (%)
rhstar	185,088	9.3%
star	0	0.0%
matcomp	3,742	0.2%
majorana	285,643	14.4%
lux	2,548	0.1%
lz	310,353	15.7%
dayabay	2,262	0.1%
cuore	6,817	0.3%
atlas	12,097	0.6%
alice	1,172,361	59.2%
Total	1,980,913	100.0%

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

PDSF system changes

- pd2015 (pdsfdtn1) - shut down for good
- 8 nodes (x 16 CPU) converted to -p nucori
  - Open for end-user testing and validation
- 5 nodes are DNR

Existing Slurm Shifter queues

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	240	2 days			0	
short	shifter	288	5 hours			0	
nucori	shifter	128	2 days			0	Cori-like OS
realtime	shifter	128	4 hours	50		0	share common hardware
debug	shifter	128	30 min		2	10	

Slurm defaults

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: 2.5 GB (was 4 , changed on June 13)
- Used cpu*h half decay time: 4 days (was 14 days)
- Cap of 250 jobs/user is enforced, pseudo-user alicesgm is an exception
- --mem → cpus conversion is 2.5 GB for all partitions, except -p long, -p short use 5 GB

Load test for -p nucori (1)

Test 1 : RAM bound , sbatch --mem 4GB
Task: 1 cpu, 4.4GB virt RAM, 3.2 res RAM, IO: 1MB/min/task

11 concurrent tasks

```
  PID USER      PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND
37930 balewski  20   0 4416468 3.225g   2772 R 100.00 5.131   1:00.06 vet3.exe
38036 balewski  20   0 4416464 3.225g   2788 R 100.00 5.131   0:43.80 vet3.exe
37993 balewski  20   0 4416468 3.225g   2680 R 94.118 5.131   0:57.35 vet3.exe
37999 balewski  20   0 4416468 3.225g   2752 R 94.118 5.131   0:56.98 vet3.exe
38005 balewski  20   0 4416468 3.226g   2800 R 94.118 5.131   0:50.38 vet3.exe
38011 balewski  20   0 4416468 3.225g   2796 R 94.118 5.131   0:48.72 vet3.exe
38017 balewski  20   0 4416468 3.225g   2792 R 94.118 5.131   0:47.57 vet3.exe
38023 balewski  20   0 4416464 3.225g   2780 R 94.118 5.131   0:46.76 vet3.exe
38030 balewski  20   0 4416468 3.225g   2780 R 94.118 5.131   0:45.85 vet3.exe
38042 balewski  20   0 4416468 3.225g   2772 R 94.118 5.131   0:42.72 vet3.exe
38049 balewski  20   0 4416468 3.226g   2808 R 94.118 5.131   0:41.60 vet3.exe
38057 balewski  20   0   34144   3320   2568 R  5.882 0.005   0:00.02 top
balewski@mc0122:~/prj/tmpPdsf> free -g
             total       used       free     shared    buffers     cached
Mem:            62         49         13          0          0          8
-/+ buffers/cache:         40         21
Swap:            0          0          0
```

Loaded node occupancy

$ scontrol show node mc0122
CfgTRES=cpu=32,mem=61000M
AllocTRES=cpu=28,mem=56G

Performance:

- 300 tasks launched
- Occupancy : 11 tasks/node
- All tasks completed w/o crash
- Execution time:
  - 240 tasks 25-26 min
  - 60 task [30-45] min

Load test for -p nucori (2)

Test 2 : CPU bound , sbatch --mem 2GB
Task: 1 cpu, 2.3GB virt RAM, 1.2 res RAM, IO: 20MB/min/task

16 tasks

```
  PID USER      PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND
97587 balewski  20   0 2368468 1.272g   2792 R 100.00 2.024   7:32.56 vet3.exe
97613 balewski  20   0 2368468 1.272g   2796 R 100.00 2.024   7:25.44 vet3.exe
96938 balewski  20   0 2368468 1.272g   2800 R 100.00 2.024   9:19.66 vet3.exe
97278 balewski  20   0 2368468 1.272g   2772 R 100.00 2.024   8:32.03 vet3.exe
97286 balewski  20   0 2368468 1.272g   2800 R 100.00 2.024   8:35.30 vet3.exe
97294 balewski  20   0 2368468 1.272g   2680 R 100.00 2.024   8:28.26 vet3.exe
97302 balewski  20   0 2368468 1.272g   2752 R 100.00 2.024   8:22.16 vet3.exe
97596 balewski  20   0 2368468 1.272g   2760 R 100.00 2.024   7:38.26 vet3.exe
97604 balewski  20   0 2368468 1.272g   2868 R 100.00 2.024   7:27.88 vet3.exe
97621 balewski  20   0 2368468 1.272g   2796 R 100.00 2.024   7:25.34 vet3.exe
97691 balewski  20   0 2368468 1.272g   2760 R 100.00 2.024   6:51.59 vet3.exe
97764 balewski  20   0 2368468 1.272g   2800 R 100.00 2.024   6:44.74 vet3.exe
96947 balewski  20   0 2368468 1.272g   2792 R 94.118 2.024   9:08.54 vet3.exe
97197 balewski  20   0 2368468 1.272g   2752 R 94.118 2.024   8:39.98 vet3.exe
97834 balewski  20   0 2368468 1.272g   2796 R 94.118 2.024   6:07.14 vet3.exe
97981 balewski  20   0 2368468 1.272g   2752 R 94.118 2.024   4:05.84 vet3.exe
98091 balewski  20   0   34144   3336   2572 R 17.647 0.005   0:00.04 top
RAM check
             total       used       free     shared    buffers     cached
Mem:            62         46         15          0          0         20
-/+ buffers/cache:         25         37
Swap:            0          0          0
```

Loaded node occupancy
$ **scontrol** show node mc0122

CfgTRES=cpu=32,mem=61000M
AllocTRES=cpu=32,mem=32G

Performance:
- 1000 tasks launched
- Occupancy : 16 tasks/node
- 11 tasks crash at start
- 16 task crashed mid run
- 973 tasks completed , all with execution time 25-26 min

Proposed PDSF change

Convert 50 % of PDSF to -p nucori

We would convert all non-Haswell nodes

Conversion date : Wednesday Feb 20 - it will take 2-3 working days (Pending OK from Craig)

Since the conversion process is manual (costly) there is no coming back

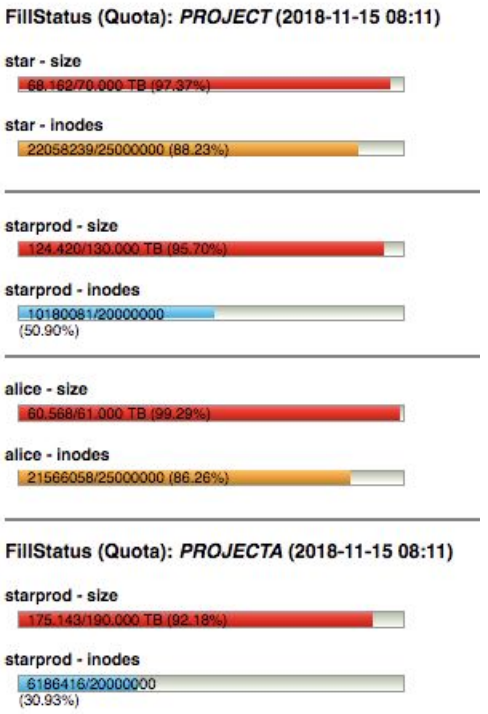
/project(a) utilization - snapshot

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

./quota_pdsf.py

```
2019-02-11_15.21 project(a) usage for PDSF users, ver 1.1
```

		space usage	space quota	space percent	inode usage	inode quota	inode percent
alice	project	62438	63488	98	21693443	25000000	86
alice	projecta	-	-	-	-	-	-
star	project	69670	71680	97	22189038	25000000	88
star	projecta	-	-	-	-	-	-
dayabay	project	850207	870400	97	126759502	150000000	84
dayabay	projecta	977618	1126400	86	6581268	10000000	65
majorana	project	61119	81920	74	4342734	6000000	72
majorana	projecta	52709	61440	85	8635759	15000000	57
atlas	project	90580	103424	87	29834606	40000000	74
atlas	projecta	192293	256000	75	28147352	40000000	70
lz	project	36406	51200	71	3385938	40000000	8
lz	projecta	238218	256000	93	4793924	40000000	12
lux	project	35905	102400	35	6868515	8000000	85
lux	projecta	217246	245760	88	62248403	70000000	88
cuore	project	29575	30720	96	2295620	1000000	229
cuore	projecta	36925	51200	72	268035	6000000	4

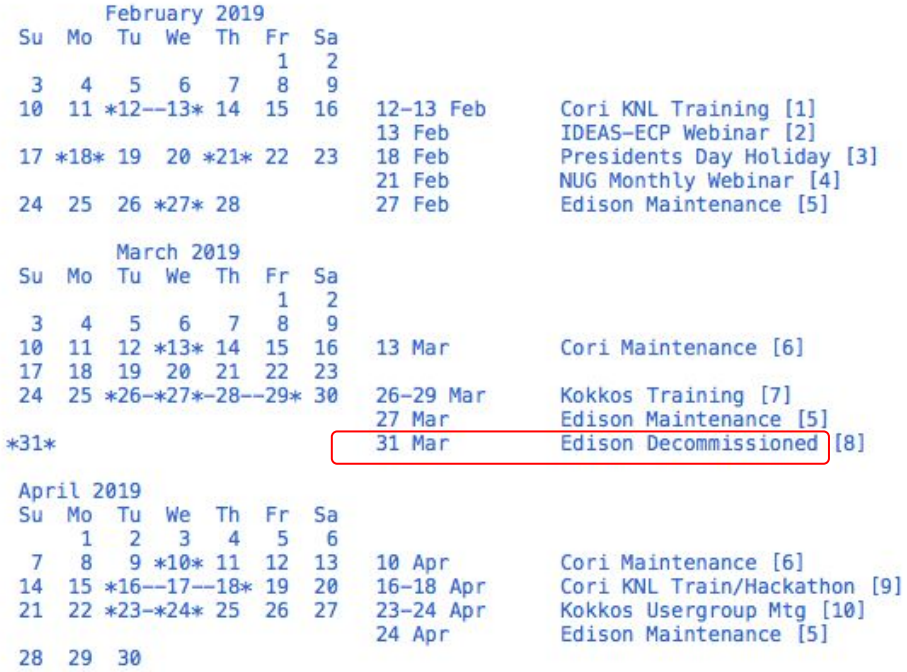


Announcements

Bi-weekly office hours Feb 18, Mar 4,59-4016A

PDSF user meeting: Tuesday, March 12

New KNL ‘low’ queue with 50% discount



February 2019

Su	Mo	Tu	We	Th	Fr	Sa		
					1	2		
3	4	5	6	7	8	9		
10	11	*12--	--13*	14	15	16	12-13 Feb	Cori KNL Training [1]
							13 Feb	IDEAS-ECP Webinar [2]
17	*18*	19	20	*21*	22	23	18 Feb	Presidents Day Holiday [3]
							21 Feb	NUG Monthly Webinar [4]
24	25	26	*27*	28			27 Feb	Edison Maintenance [5]

March 2019

Su	Mo	Tu	We	Th	Fr	Sa		
					1	2		
3	4	5	6	7	8	9		
10	11	12	*13*	14	15	16	13 Mar	Cori Maintenance [6]
17	18	19	20	21	22	23		
24	25	*26-	-*27*	-28-	-29*	30	26-29 Mar	Kokkos Training [7]
							27 Mar	Edison Maintenance [5]
31							31 Mar	Edison Decommissioned [8]

April 2019

Su	Mo	Tu	We	Th	Fr	Sa		
	1	2	3	4	5	6		
7	8	9	*10*	11	12	13	10 Apr	Cori Maintenance [6]
14	15	*16-	-17-	-18*	19	20	16-18 Apr	Cori KNL Train/Hackathon [9]
21	22	*23-	-*24*	25	26	27	23-24 Apr	Kokkos Usergroup Mtg [10]
							24 Apr	Edison Maintenance [5]
28	29	30						