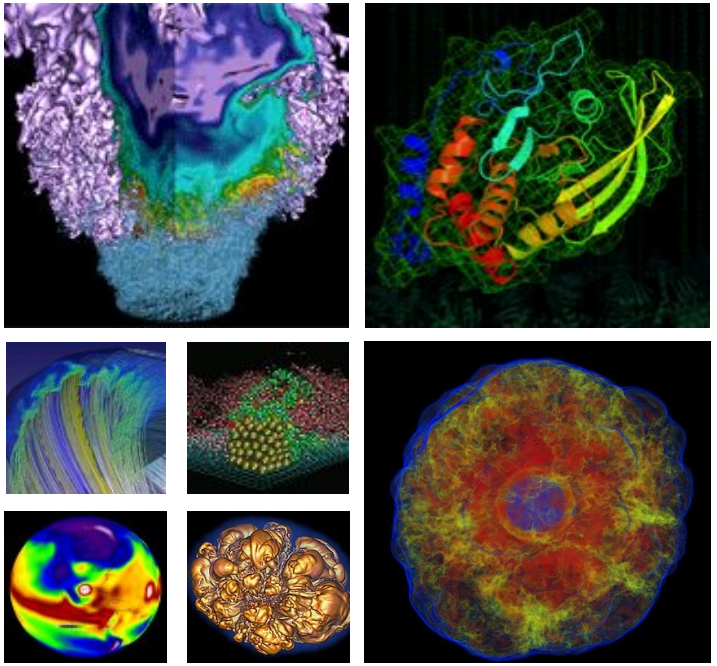
Impact of Large Jobs and Reservations on Cray Systems



Yun (Helen) He, Emily Zhang, Woo-Sun Yang, NERSC

Cray User Group Meeting, May 9, 2019

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

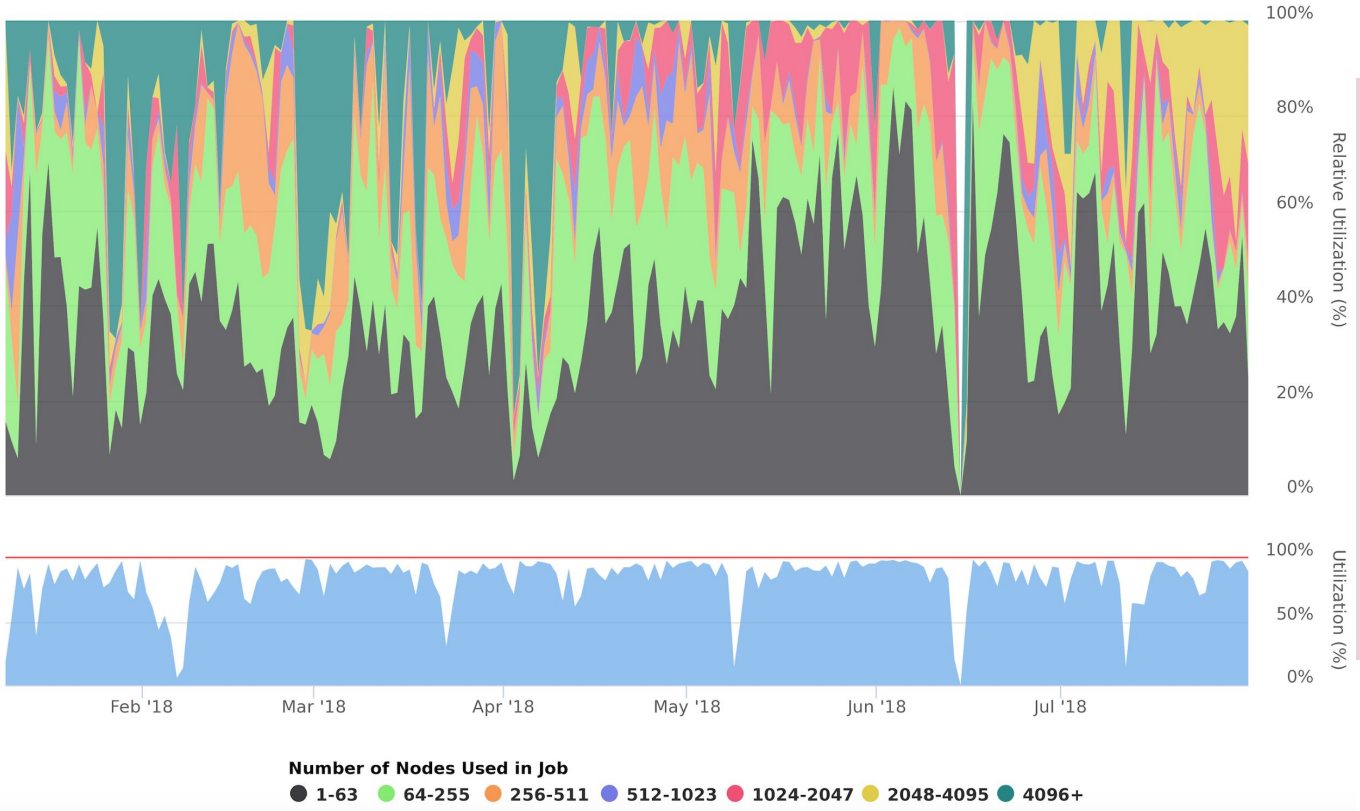
Outline

- Introduction and goal for drain analysis
- Drain analysis definitions
- Impact from large jobs
- Impact from large reservations
- Hand-holding users running large jobs
- Impact from system monitoring benchmark runs
- New “flex” QOS to mitigate the impact

NERSC production Cray systems

- Cori (Cray-XC40)
  - 9,688 KNL nodes
  - 2,388 Haswell nodes
- Edison (Cray-XC30)
  - 5,586 Ivybridge nodes
- Slurm batch scheduler used on both systems
  - QOS: regular, debug, premium, interactive, shared, realtime, etc.
  - Long queue backlogs

Cori KNL job size distribution, 1/10-7/31, 2018



~30% on KNL are jobs use >1,024 nodes

Utilization impacted by system maintenances and scheduling holes from large jobs and large reservations, even with queue backlogs.

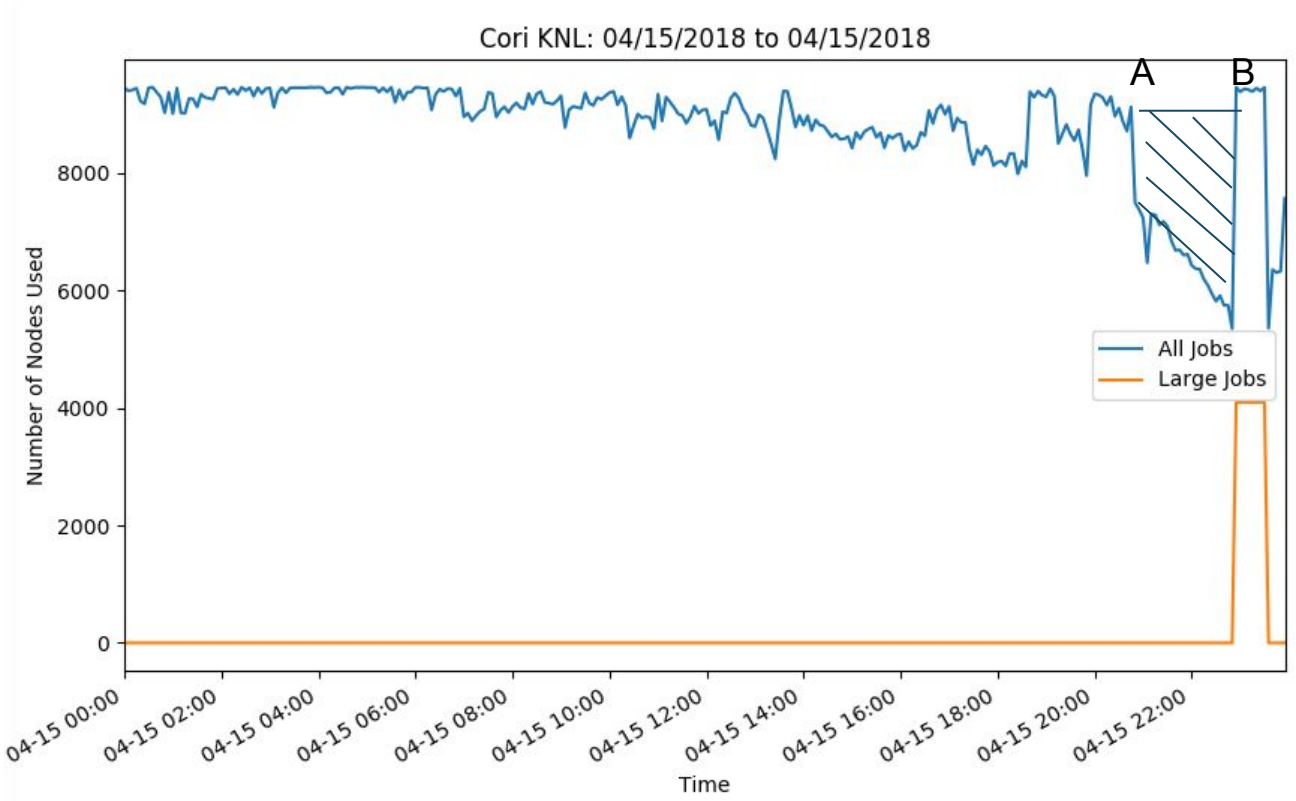
Introduction and Goal

- When the batch scheduler gathers nodes for a large job or reservation, the system “drains” and affects utilization:
  - Fewer and fewer nodes can be used for other jobs except small and short “backfill” jobs that won’t affect the start time of the large job.
- **The goal of this study is understand the impact from large jobs and system reservations on Cori and Edison to system utilizations.**
- We also investigated the impact of routinely run large SSP benchmarks on Cori, especially on Haswell.
- **How could these results guide us to improve scheduling policies and help users?**
- We used Slurm jobs database and python script for analysis.

Definitions

- **Large Job or Large Reservation:** a “regular” QOS job that uses or a reservations that requests:
  - > 2,048 Cori KNL nodes
  - or > 512 Cori Haswell nodes
  - or > 1,024 Edison nodes
  - Back-to-back reservations are counted as 1 due to no additional drain is needed.
- **95% Threshold**: a predefined node count for perceived “normal” usage on each architecture for the “regular” QOS jobs.
  - 9,000 for Cori KNL (total 9,482 nodes)
  - 1,680 for Cori Haswell (total 1,772 nodes)
  - 5,150 for Edison (total 5,421 nodes)

Example drain analysis of a large Job



The shaded area is the **Drain Node Hours**. This is the integrated area of "time duration" times "difference of total nodes used from threshold". (5 min interval)

Time duration from A to B is the **Drain Time**.

Drop-off Loss is the percentage of the above Drain Time over Total Available node hours during the time period for this plot.

Drop-off Ratio is Node hours used by this job / Drain Time. The larger this ratio is, the worthier the drain is.

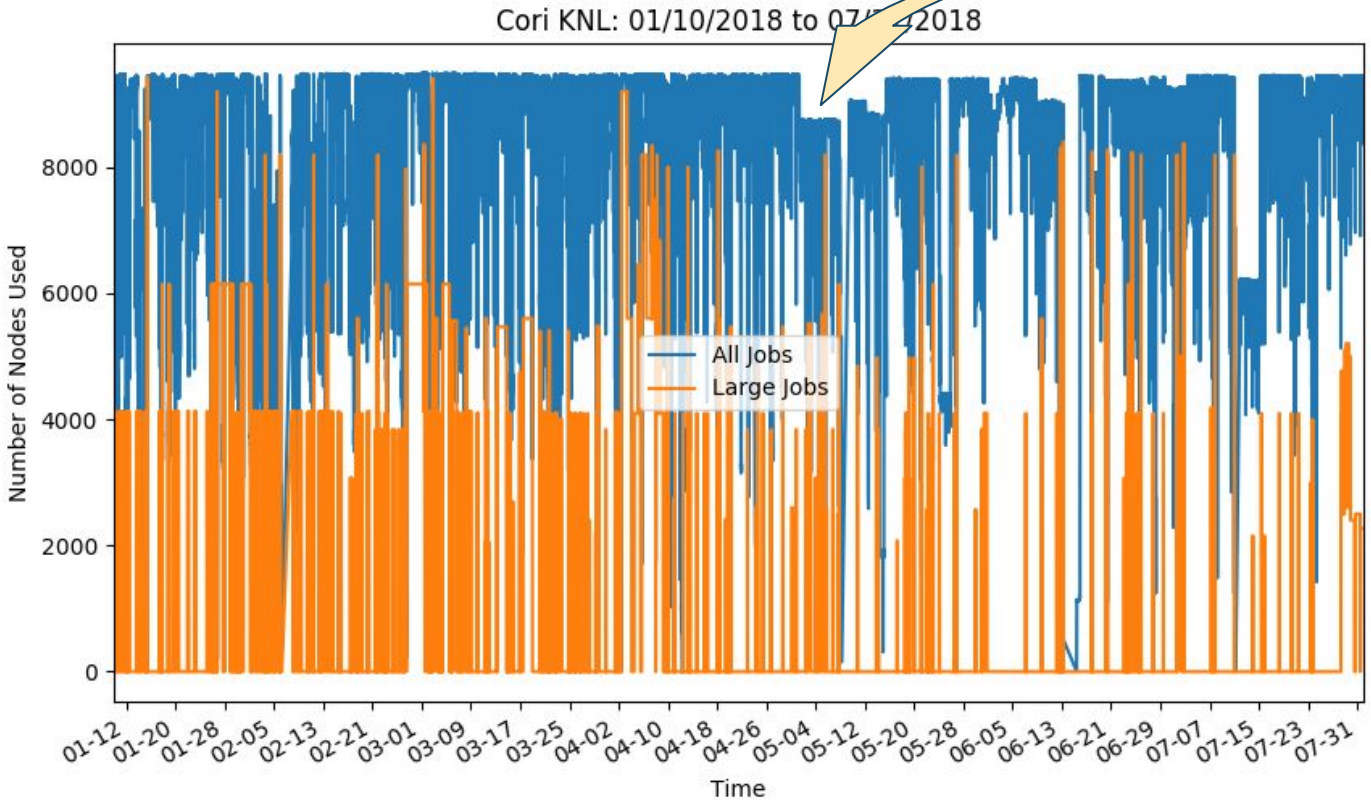
For all large jobs during a time period

$$\textbf{Drop-off Ratio} = \frac{\Sigma\,(\text{Large Jobs' Node Hours})}{\Sigma\,(\text{Total Drain Node Hours})}$$

$$\textbf{Drop-off Loss} = \frac{\Sigma\,(\text{Total Drain Node Hours})}{[\,\Sigma\,(\text{Total Used Node Hours}) + \Sigma\,(\text{Todal Drain Node Hours})\,]} \times 100\%$$

Adapted from Mark Thornburg, 2014

Drain analysis of large Cori KNL jobs



Not counted in overall Drop-off calculation for large jobs run during this period of system degradation since total nodes used always < Threshold

Drain analysis of large Cori jobs

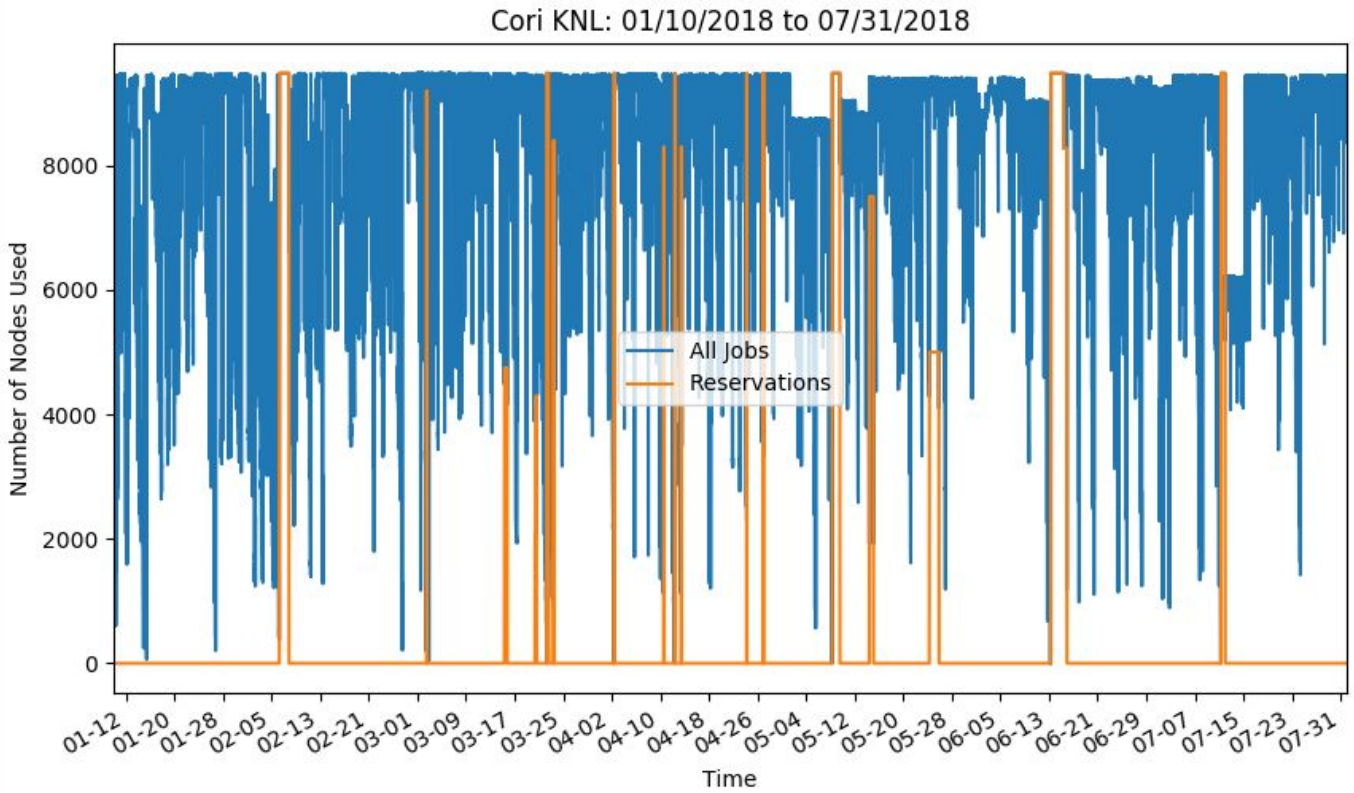
Cori Haswell, 1/10-7/31/2018

Total Drain Time	170,684 Node Hours
Drop-off Ratio	1.48
Drop-off Loss	2.1% of Machine
Average Drop-off Time	**1.12 Hours**
Average Job Length	0.82 Hours
Average Job Size	1,005.07 Nodes
Total Number of Large Jobs	361 Jobs
Total Number of Jobs	430,877 Jobs

Cori KNL, 1/10-7/31/2018

Total Drain Time	2,684,876 Node Hours
Drop-off Ratio	1.70
Drop-off Loss	**6.5% of Machine**
Average Drop-off Time	**1.62 Hours**
Average Job Length	**1.67 Hours**
Average Job Size	**4,263.34 Nodes**
Total Number of Large Jobs	591 Jobs
Total Number of Jobs	968,930 Jobs

Drain analysis of large Cori KNL reservations



Full system reservations used for scheduled maintenances:
- Feb 6-7
- Mar 22
- May 8-9
- Jun 13-15
- Jul 11

And other large user reservations

Drain analysis of large Cori reservations

Cori Haswell, 1/10-7/31/2018

Total Drain Time	28,234 Node Hours
Drop-off Loss	0.3% of Machine
Average Drop-off Time	**3.47 Hours**
Average Reservation Length	22.40 Hours
Average Reservation Size	2,164.5 Nodes
Total Number of Large Reservations	8 Reservations
Total Number of Jobs	430,882 Jobs

Cori KNL, 1/10-7/31/2018

Total Drain Time	526,150 Node Hours
Drop-off Loss	**1.3% of Machine**
Average Drop-off Time	**6.12 Hours**
Average Reservation Length	**10.32 Hours**
Average Reservation Size	**8,005.56 Nodes**
Total Number of Large Reservations	18 Reservations
Total Number of Jobs	968,935 Jobs

Drop-off time and drop-off loss

	Large Jobs			Large Reservations		
	Drop-off Time	Drop-off Loss (Per Job Loss)	#jobs / Average Large Job Size	Drop-off Time	Drop-off Loss (Per resv Loss)	#reservations / Average Resv Size
Cori KNL	1.67 hr	**6.5%** (0.0110%)	591 / 4,263 nodes	**6.12 hr**	1.3% (0.0722%)	18 / 8,006 nodes
Cori Haswell	1.12 hr	2.1% (0.0058%)	361 / 1,005 nodes	3.47 hr	0.3% (0.0375%)	8 / 2,164 nodes
Edison	2.01 hr	2.3% (0.0106%)	216 / 2,051 nodes	8.65 hr	0.6% (0.1%)	6 / 5,603 nodes

Observations: large jobs and reservations

- **Cori KNL**:
  - average large job size is 4,263 nodes, average drop-off time is **1.67 hr**
  - average large reservation is 8,006 nodes, average drop-off time is **6.12 hr**
- **Cori Haswell**:
  - average large job size is 1,005 nodes, average drop-off time is **1.12 hr**
  - average large reservation is 2,164 nodes, average drop-off time is **3.47 hr**

- The average **drop-off time and drop-off loss** per reservation is much larger than those per large job, due to average reservation sizes being much larger.
  - **Reservations are expensive**

Jobs of highest & lowest drain-worthiness

User	#jobs	Average job	Drop-off Ratio	Total Node Hours	Total Drain Time
pxxxx	9.00	2472.00	107.31	268493.00	2502.00
hxxxx	4.00	4096.00	46.68	6162.00	132.00
jxxxx	5.00	2151.00	32.10	17657.00	550.00
nxxxx	2.00	5000.00	30.47	80783.00	2651.00
hxxxx	21.00	6144.00	18.67	1267376.00	67898.00
mxxxx	5.00	3400.00	8.59	37715.00	4393.00
uxxxx	28.00	6631.00	5.41	1729949.00	319754.00
lxxxx	15.00	5768.00	0.17	11989.00	70101.00
fxxxx	73.00	3840.00	0.16	32178.00	196464.00
lxxxx	3.00	6144.00	0.15	4438.00	29545.00
axxxx	24.00	4096.00	0.10	10316.00	103259.00
axxxx	17.00	5024.00	0.09	6086.00	67020.00
lxxxx	7.00	4827.00	0.08	5993.00	77886.00
mxxxx	9.00	4186.00	0.05	1264.00	26142.00

We analyzed all user large jobs and obtained individual users drop-off ratios.

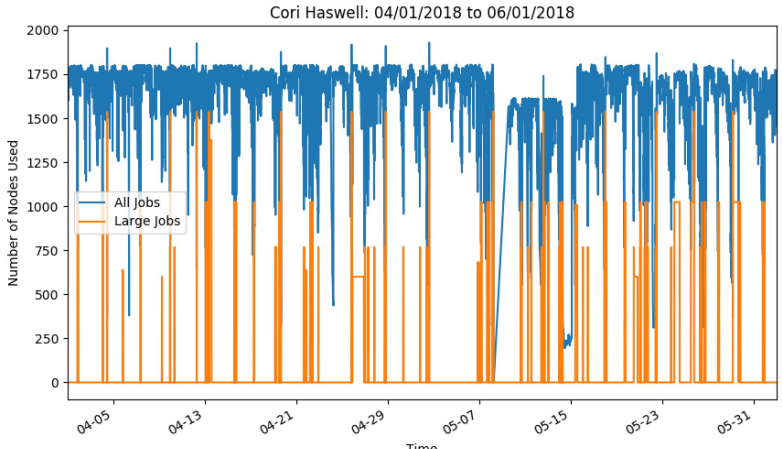
We contacted each user at the bottom of the table individually

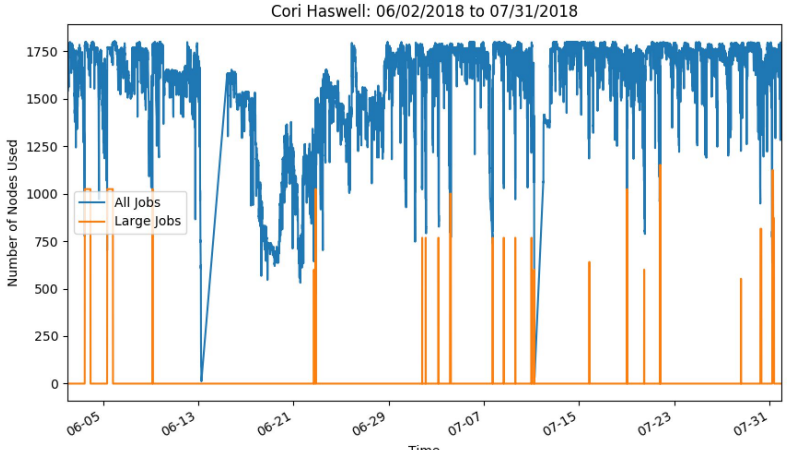
-- helped debug job failures
-- suggested bundle jobs and order jobs from large to small

to reduce drain cost and increase drain worthiness.

“fbench” jobs analysis

- Staff user “fbench” runs regularly Cori SSP benchmarks. Does it have large impact on draining compared to other large jobs, especially on Haswell? Some benchmarks use close to >90% available regular QOS Haswell nodes.
- We changed from weekly runs in production to monthly runs dedicated under reservation from 6/1/2018.
- We performed two-month large jobs analysis before and after this change.





fbench jobs *vs.* non-fbench jobs on Haswell

4/1 - 6/1 fbench only

Total Drain Time	27,856 Node Hours
Drop-off Loss	**1.1% of Machine**
Average Drop-off Time	**0.75 Hours**
Average Job Length	0.07 Hours
Average Job Size	1226.67 Nodes
Total Number of Large Jobs	**67 Jobs**

4/1 - 6/1 non fbench

Total Drain Time	57,773 Node Hours
Drop-off Loss	**2.4% of Machine**
Average Drop-off Time	**1.66 Hours**
Average Job Length	1.80 Hours
Average Job Size	907.98 Nodes
Total Number of Large Jobs	**103 Jobs**

Observations: “fbench” SSP jobs

- No strong evidence that “fbench” large SSP jobs affected system draining significantly.
  - Average fbench job size is larger than that of non-fbench jobs.
  - Average Drop-off Time and per job Drop-off Loss are smaller.
  - The above is due to the **optimal ordering of fbench SSP jobs** from biggest to smallest at submission which achieved minimal draining needed.
  - After June 1, the average Drop-off Time (no fbench jobs) is larger than before.
- **It is management decision** whether we run SSP Haswell benchmarks regularly (~weekly) or only during maintenances (~monthly)
  - 1.1% Drop-off Loss from 67 jobs ran 4/1-6/1.
  - **Do we want to waste 1.1% to get weekly data or not?**
    - **No. So we are running monthly dedicated during system maintenances only since then.**

What type of jobs are good for backfill

- The average drop-off time is related to the job size and wall time limit we need as backfill jobs.
- Variable-time jobs and low charge factors for small short jobs may help with getting more backfill to reduce the drain impact.
  - Variable-time jobs are jobs that request a time-min request and a regular wall time limit. The scheduler is free to allocate time duration between these two limits.
  - Attractive since it reduces wait time in the queue
  - Even more attractive if combines with queue discount

New “flex” QOS to mitigate drain impact

- Encourage users to submit more small short jobs to be eligible as “backfill” jobs to run during system drains
- We added a new “flex” qos on KNL on April 22, 2019
  - For user jobs that can produce useful work with a relatively short amount of run time before terminating, such as jobs capable of checkpointing and restarting where left off
  - Helps to improve throughput by submitting jobs that can fit into cracks in Slurm job scheduling
  - Required to use “--time-min” of <= 2hrs, max “--time” is 12 hrs
  - Free of charge during the first month
  - 4700 jobs ran in 2 weeks as of May 8. We will examine the effect of “flex” later.

Summary

- System drains caused by large jobs or large reservations are very costly to system utilizations.
  - The larger the size of a job or reservation is, the larger the drop-off time usually is, which could be up to 6 hrs on KNL.
  - Due to large number large KNL jobs ran, the overall Drop-off Loss adds up to 6.5%.
- The drain analysis helped us to mitigate these impacts by:
  - Encouraging users to analyze and optimize their behaviors in running large jobs.
  - Verifying staff benchmarking did not cause significant drain (1.1%).
  - Adding new “flex” qos to help getting more small and short jobs to fill the drain gaps.

Acknowledgement

- The drain analysis algorithm is inspired by and adapted from the algorithm developed by the Cray intern Mark Thornburg in 2014.

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Thank You