Los Alamos
NATIONAL LABORATORY

EST. 1943

NNSA
National Nuclear Security Administration

Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA

LA-UR-19-29207

Monitoring Slurm with a Splunk App

LANL Workload Management Team

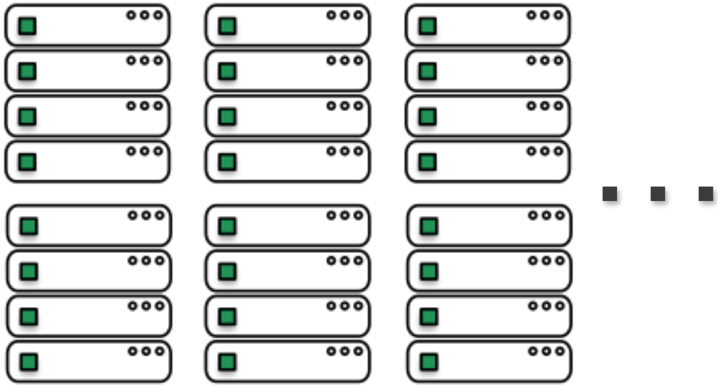
Nicole Dobson

18 Sept 2019

HIGH PERFORMANCE COMPUTING

The need for a better monitoring tool…

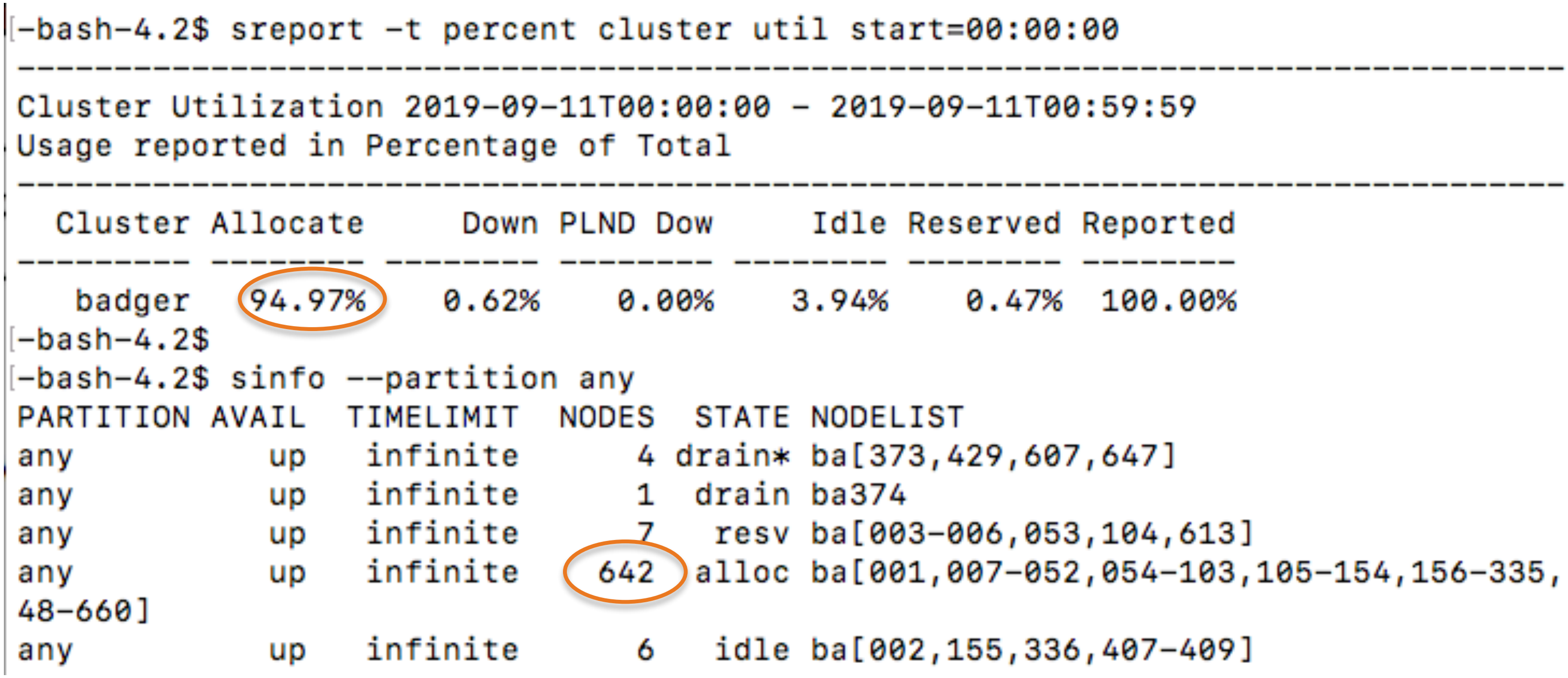
- Maintain functionality
- Multiple systems
- Response time



- Portable and easy for multiple systems
- Quick and easy detection
- Faster diagnosis

Using the command line

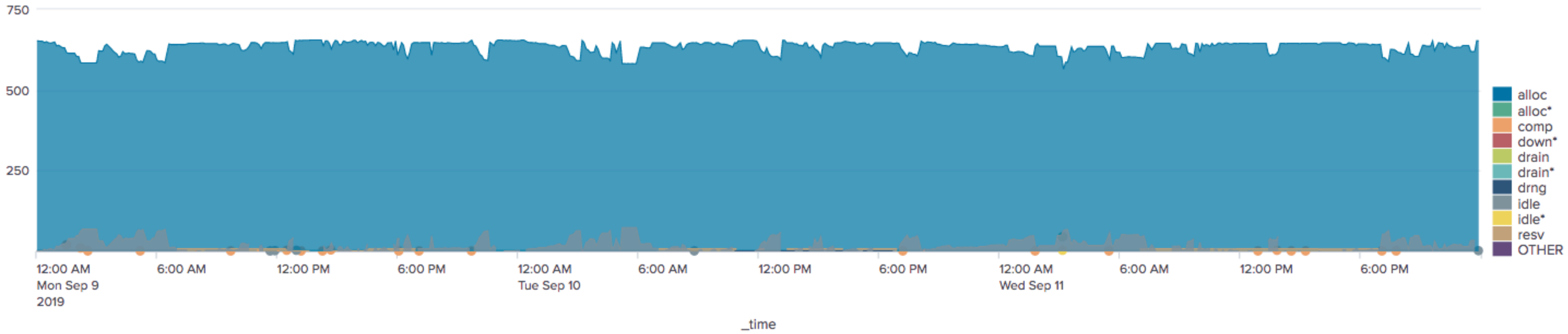
Say we just want to monitor utilization:



```
-bash-4.2$ sreport -t percent cluster util start=00:00:00
--------------------------------------------------------------------------------
Cluster Utilization 2019-09-11T00:00:00 - 2019-09-11T00:59:59
Usage reported in Percentage of Total
--------------------------------------------------------------------------------
  Cluster Allocate     Down PLND Dow     Idle Reserved Reported
--------- -------- -------- -------- -------- -------- --------
   badger   94.97%    0.62%    0.00%    3.94%    0.47%  100.00%
-bash-4.2$
-bash-4.2$ sinfo --partition any
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
any          up   infinite      4 drain* ba[373,429,607,647]
any          up   infinite      1  drain ba374
any          up   infinite      7   resv ba[003-006,053,104,613]
any          up   infinite    642  alloc ba[001,007-052,054-103,105-154,156-335,
48-660]
any          up   infinite      6   idle ba[002,155,336,407-409]
```

Using Splunk

We can expand on that single number to a visual representation that updates in real time



Now we can easily identify unexpected utilization metrics without having to constantly run a command and verify that the numbers it returns are within normal bounds

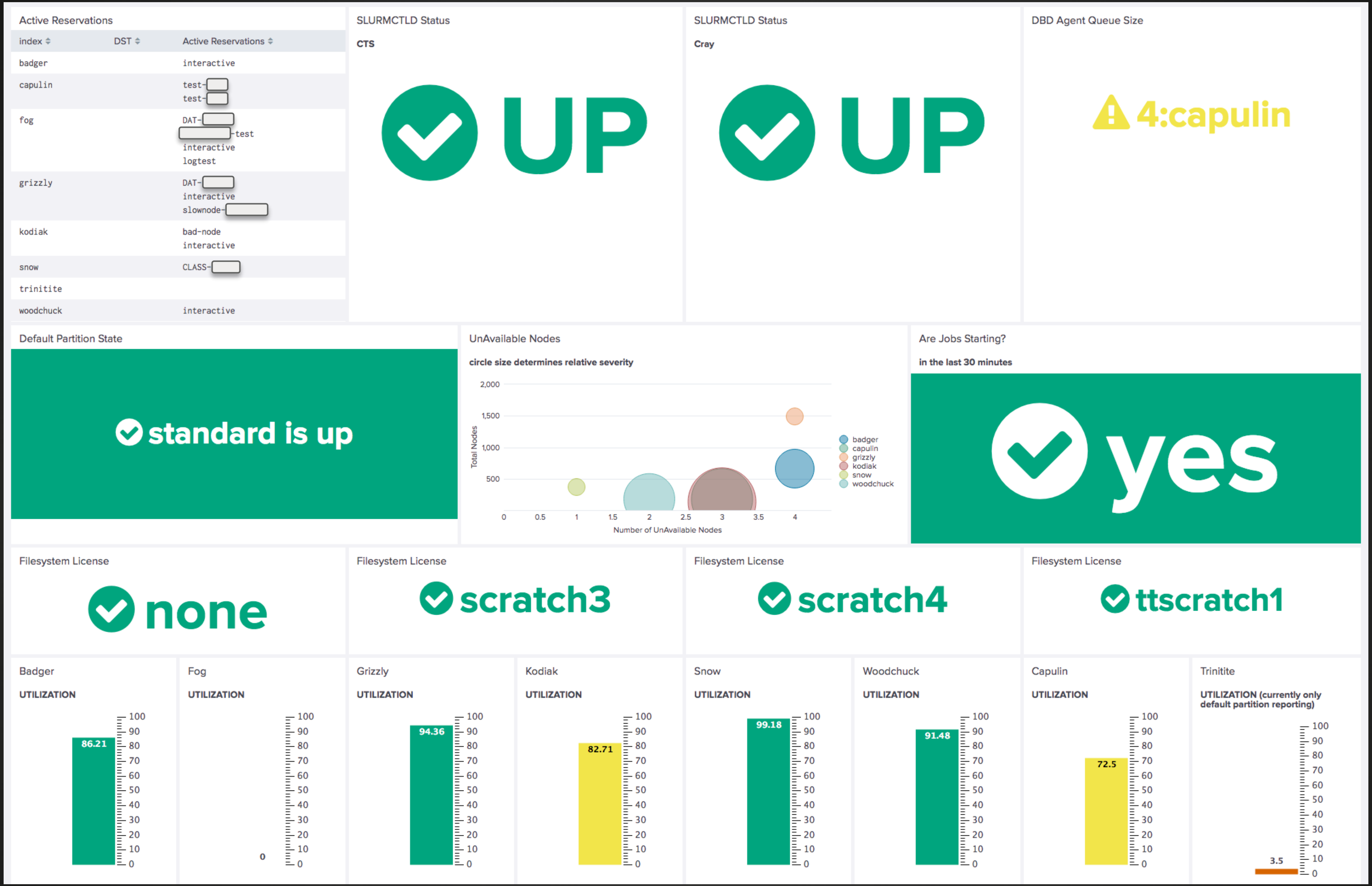
Why we use Splunk

What Splunk does:

- Ingests log messages and other log-like data
- Command box functions similarly to grep
- Allows visuals to be created and auto-updated all from that one command box
- Can create alerts on data and trends

Benefits:

- No need to scp data off of clusters to monitoring server in order to interpret it
- Splunk command box allows for grepping through logs same as command line
- One command to ingest data and create visual, not having to keep track of and maintain multiple scripts
- No need to set up cron job to look for and alert on conditions of interest, can alert on trends



Active Reservations

index	DST	Active Reservations
badger		interactive
capulin		test-[redacted] test-[redacted]
fog		DAT-[redacted] [redacted]-test interactive logtest
grizzly		DAT-[redacted] interactive slownode-[redacted]
kodiak		bad-node interactive
snow		CLASS-[redacted]
trinitite		
woodchuck		interactive

SLURMCTLD Status

CTS

✓ UP

SLURMCTLD Status

Cray

✓ UP

DBD Agent Queue Size

⚠ 4:capulin

Default Partition State

✓ standard is up

UnAvailable Nodes

circle size determines relative severity

Are Jobs Starting?

in the last 30 minutes

✓ yes

Filesystem License

✓ none

Filesystem License

✓ scratch3

Filesystem License

✓ scratch4

Filesystem License

✓ ttscratch1

Badger

UTILIZATION

86.21

Fog

UTILIZATION

0

Grizzly

UTILIZATION

94.36

Kodiak

UTILIZATION

82.71

Snow

UTILIZATION

99.18

Woodchuck

UTILIZATION

91.48

Capulin

UTILIZATION

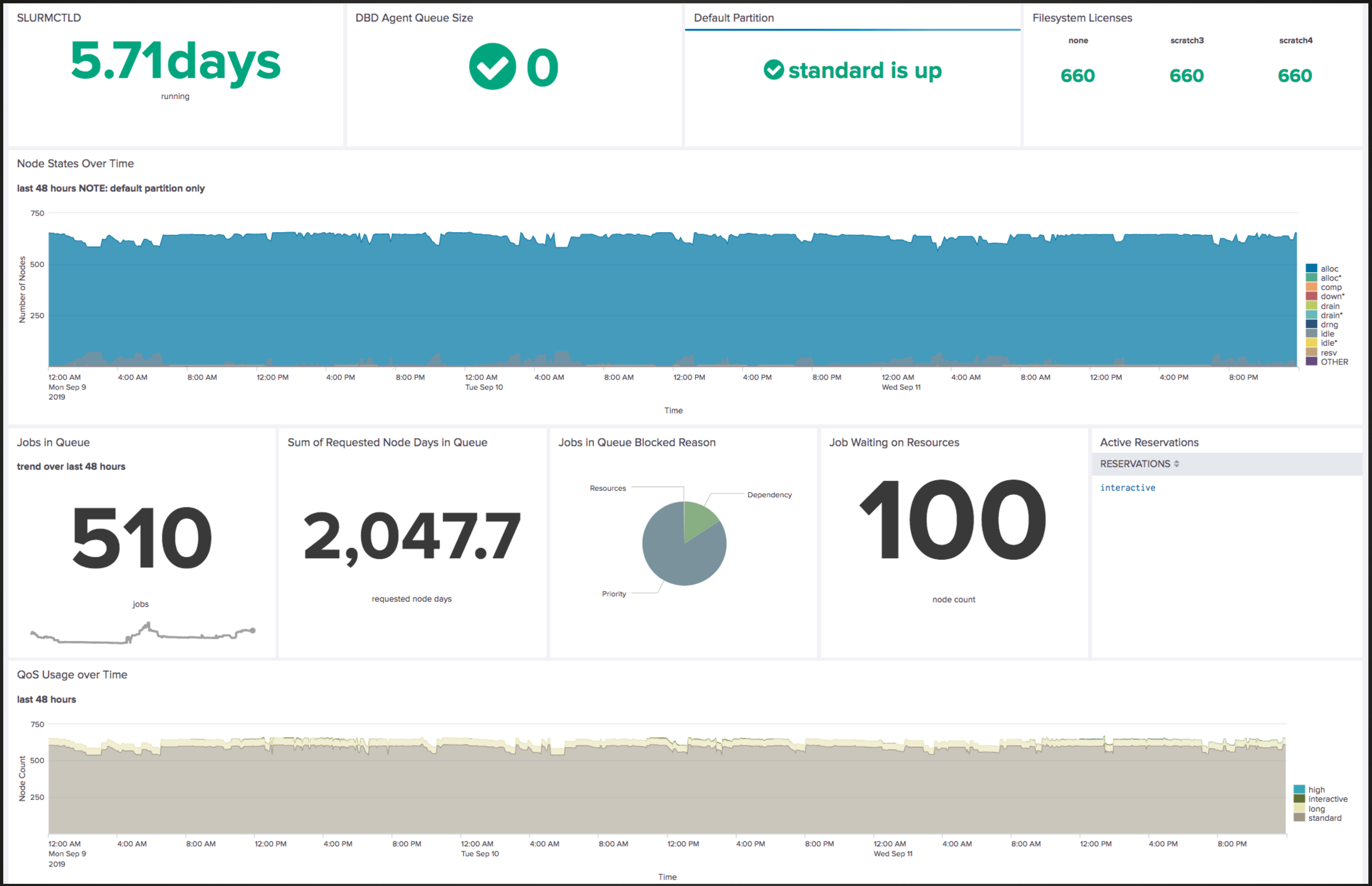
72.5

Trinitite

UTILIZATION (currently only default partition reporting)

3.5

All Clusters Monitoring Dashboard



SLURMCTLD

5.71days

running

DBD Agent Queue Size

✓ 0

Default Partition

✓ standard is up

Filesystem Licenses

none	scratch3	scratch4
660	660	660

Node States Over Time

last 48 hours NOTE: default partition only

Jobs in Queue

trend over last 48 hours

510

jobs

Sum of Requested Node Days in Queue

2,047.7

requested node days

Jobs in Queue Blocked Reason

Job Waiting on Resources

100

node count

Active Reservations

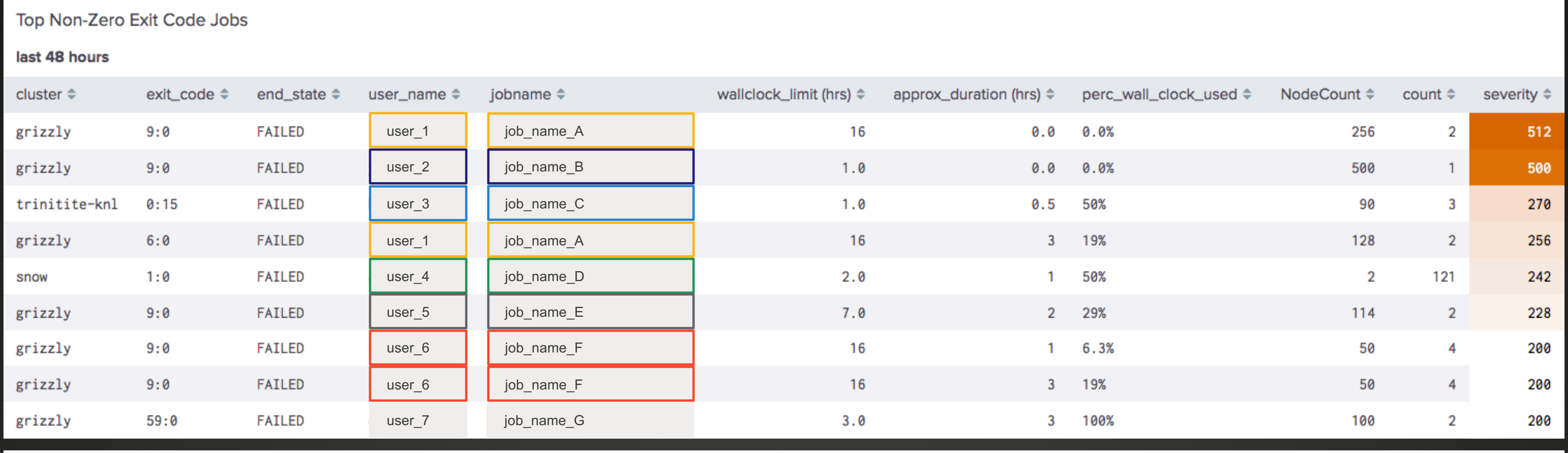
RESERVATIONS ⇕

interactive

QoS Usage over Time

last 48 hours

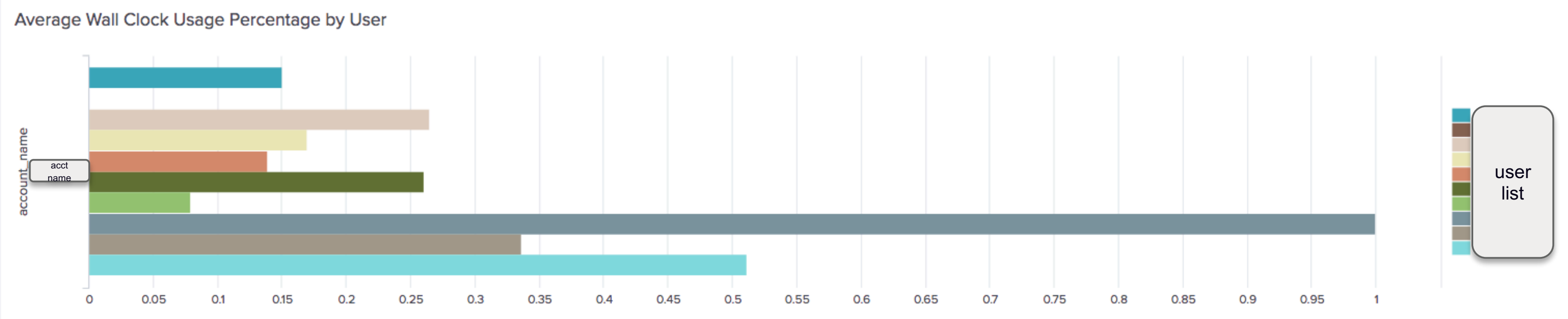
Single Cluster Dashboard



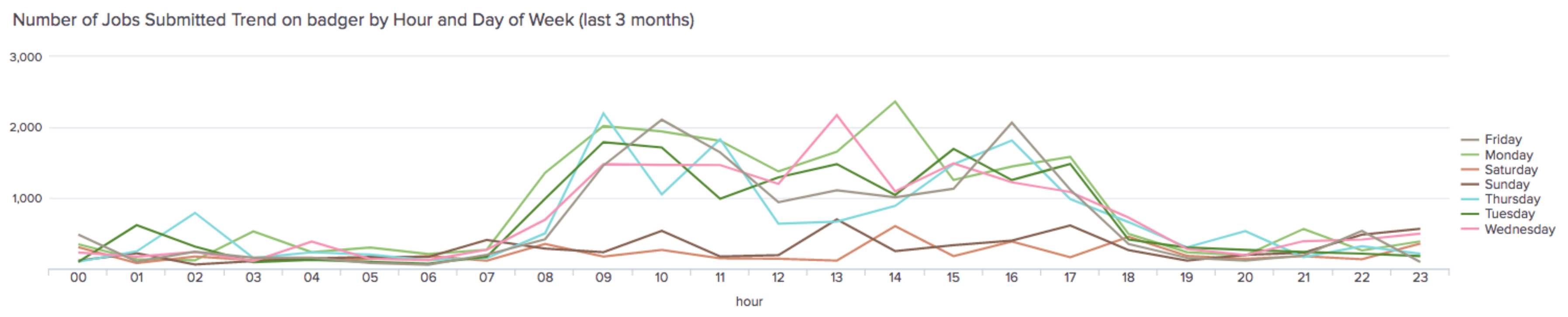
Top Non-Zero Exit Code Jobs

last 48 hours

cluster	exit_code	end_state	user_name	jobname	wallclock_limit (hrs)	approx_duration (hrs)	perc_wall_clock_used	NodeCount	count	severity
grizzly	9:0	FAILED	user_1	job_name_A	16	0.0	0.0%	256	2	512
grizzly	9:0	FAILED	user_2	job_name_B	1.0	0.0	0.0%	500	1	500
trinitite-knl	0:15	FAILED	user_3	job_name_C	1.0	0.5	50%	90	3	270
grizzly	6:0	FAILED	user_1	job_name_A	16	3	19%	128	2	256
snow	1:0	FAILED	user_4	job_name_D	2.0	1	50%	2	121	242
grizzly	9:0	FAILED	user_5	job_name_E	7.0	2	29%	114	2	228
grizzly	9:0	FAILED	user_6	job_name_F	16	1	6.3%	50	4	200
grizzly	9:0	FAILED	user_6	job_name_F	16	3	19%	50	4	200
grizzly	59:0	FAILED	user_7	job_name_G	3.0	3	100%	100	2	200



Average Wall Clock Usage Percentage by User



Number of Jobs Submitted Trend on badger by Hour and Day of Week (last 3 months)

Analysis Panels and Graphs

Log Messages and Data Sources

- slurmctld log messages
  - Reservation start and end
  - slurmctld running

- Custom made cron script
  - Slurm commands: sinfo, sdiag, squeue, scontrol …
  - Easy to maintain and add to, same across all clusters

- Job completion data
  - Epilog script ran at end of job reporting on data items

- Some supporting logs from other systems or software

Improved our maintenance procedures

Fine-tuned our policies

Quickly get a sense of health, normal pattern of use, and appropriate heartbeats

Questions?

HIGH
PERFORMANCE
COMPUTING

Over 70 years at the forefront of supercomputing

HPC

Over 70 years at the forefront of supercomputing