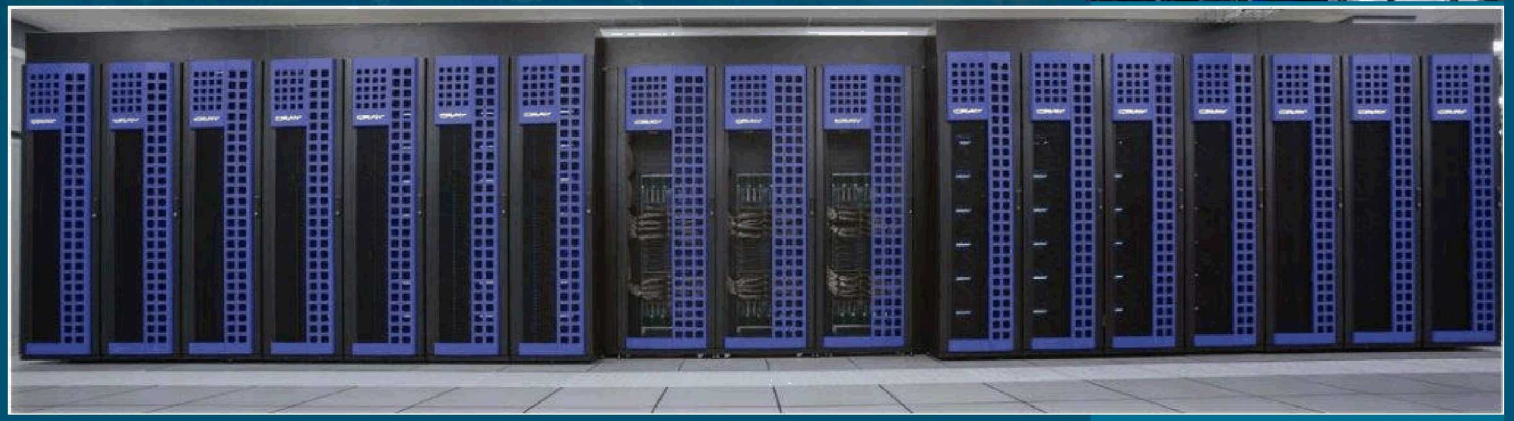
SAND2019-6760PE

OVIS Monitoring



PRESENTED BY

Edward Walsh

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

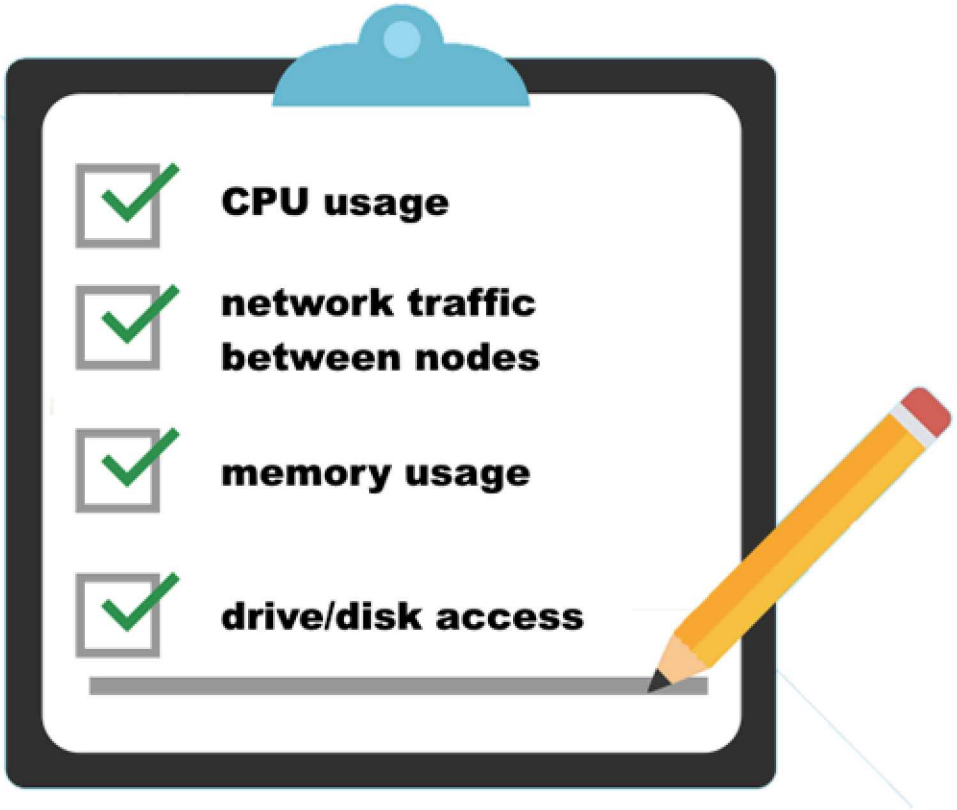
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

What does a supercomputer do?

Suppose we have an application that performs so many computations it would take days, weeks, or months to run on a desktop with just one CPU. A supercomputer can spread the work out over tens of thousands of CPUs. The CPUs run in parallel (at the same time).

https://pixabay.com/photos/children-win-success-video-game-593313/

What to monitor?



https://pixabay.com/en/icon-survey-check-form-mark-test-2967800/

Terminology

node: one motherboard that holds one or more CPUs, each CPU has one or more cores

job: a software script. Uses one or more nodes.

https://commons.wikimedia.org/w/index.php?title=Special:Search&limit=100&offset=240&profile=default&search=compute+node&searchToken=77i2v1yyltv7ut83lpt5r2t9v#/media/File:Tesla-NVIDIA_GPU_cluster_(3706444821).jpg

OVIS: real-time monitoring of clusters

13 year effort to develop light-weight distributed metric service

Acquire metrics from every node (CPU usage, network traffic density, # I/O calls, etc.).

Because nodes in a single cluster are identical, can use statistical analysis.

Characterize performance of the cluster. Can be used to identify anomalies that may turn into component failures.

Visualization makes it easy for a human to see anomalies.

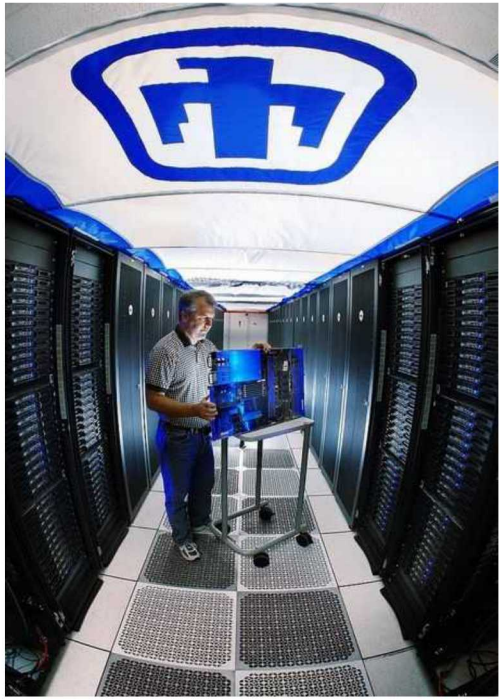
OVIS: elevator sales pitch

Developed for system admins, app developers, & system designers to gain real-time insight, or to view historical data, on resource utilization and bottlenecks(e.g., network bandwidth/hotspots, CPU utilization, Memory footprint/bandwidth). Awareness of how resources are being utilized, stressed or depleted due to the aggregate workload.

Examples:
Does job have a balanced load.
Is job stalled, in real time, waiting for a resource.

Low overhead on CPU & memory
to collect & store the data.

Is not a library that is linked into the job.
Doesn't perturb the job.



https://www.glassdoor.co.in/Photos/Sandia-National-Laboratories-Office-Photos-IMG3428368.htm

Winner of R&D 100 award (2015 year)

In 2015:
Other HPC performance monitoring systems collect 10-30 metrics every 5 to 10 minutes. OVIS can collect megabytes/minute of metric data continuously.

Is first HPC (High-Performance Computing) monitoring system that uses statistics to characterize behavior of nodes.

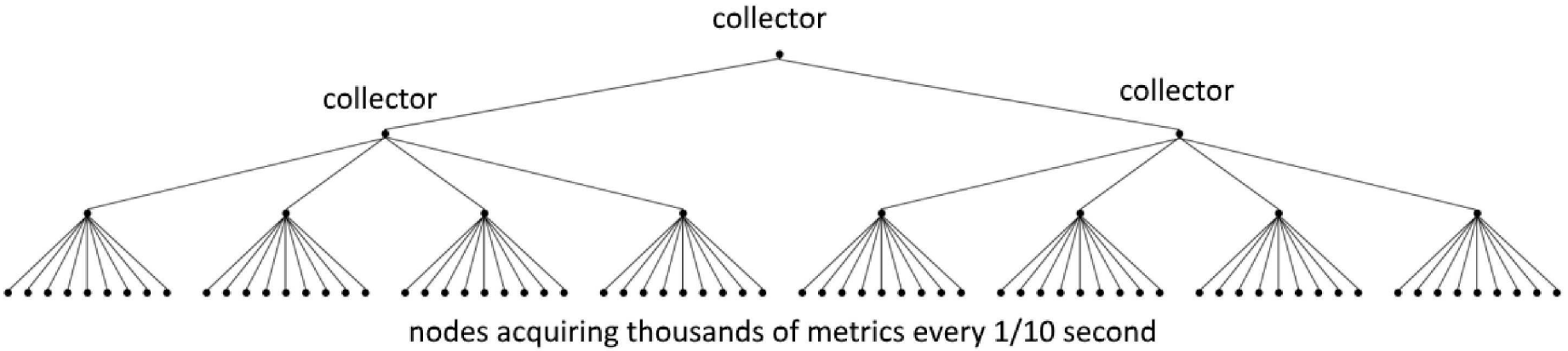
Before OVIS:
Monitoring system detected problems by comparing acquired data against threshold values. The system administrator is notified of a problem AFTER damage has occurred (e.g. after CPU temperature is too hot). OVIS can detect degradation of performance before damage occurs (e.g. CPU temperature is slightly warmer than neighboring CPUs).

OVIS: collecting the data

Data (memory usage, CPU load, network traffic, computed values, etc.) is acquired from nodes.

Data is pushed up to collectors.

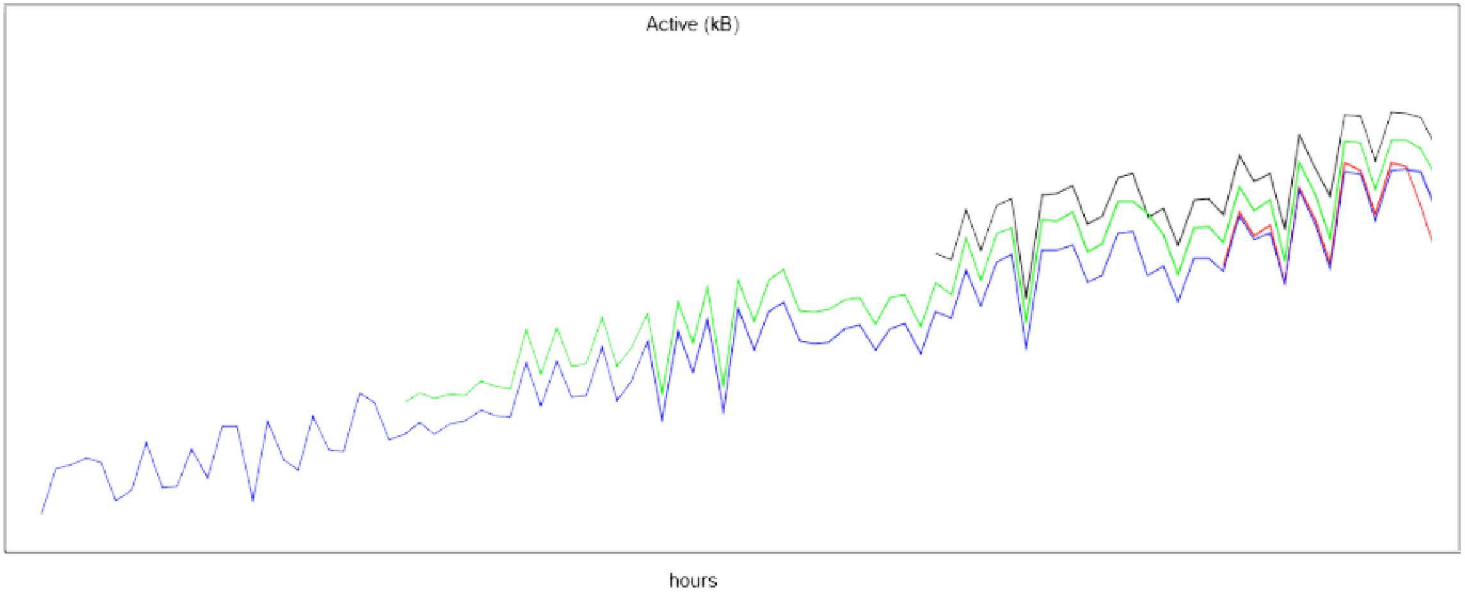
Aggregated data is pushed up toward root collector. Server can either write the data to log files or store the data in a database.



My participation (2013 – early 2019)

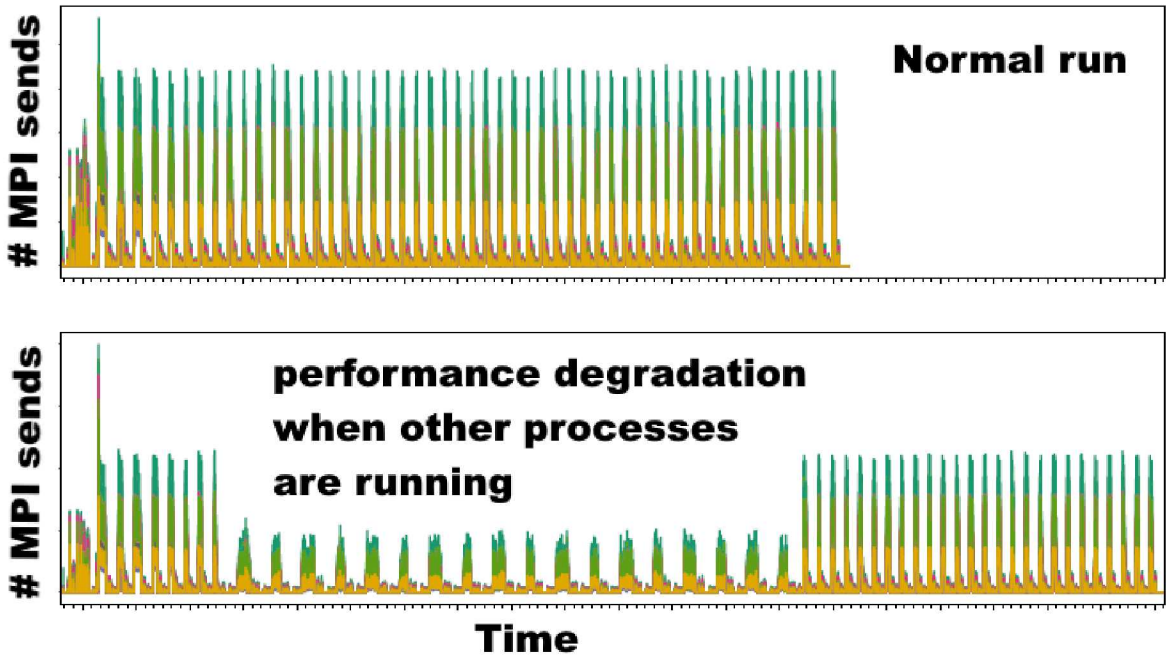
Designed and implemented visualization clients.

Example:
Real time XY plot of Active Memory (recently used memory) for 4 nodes



Can view one parameter on one node

Visualization makes it easy to see problems.

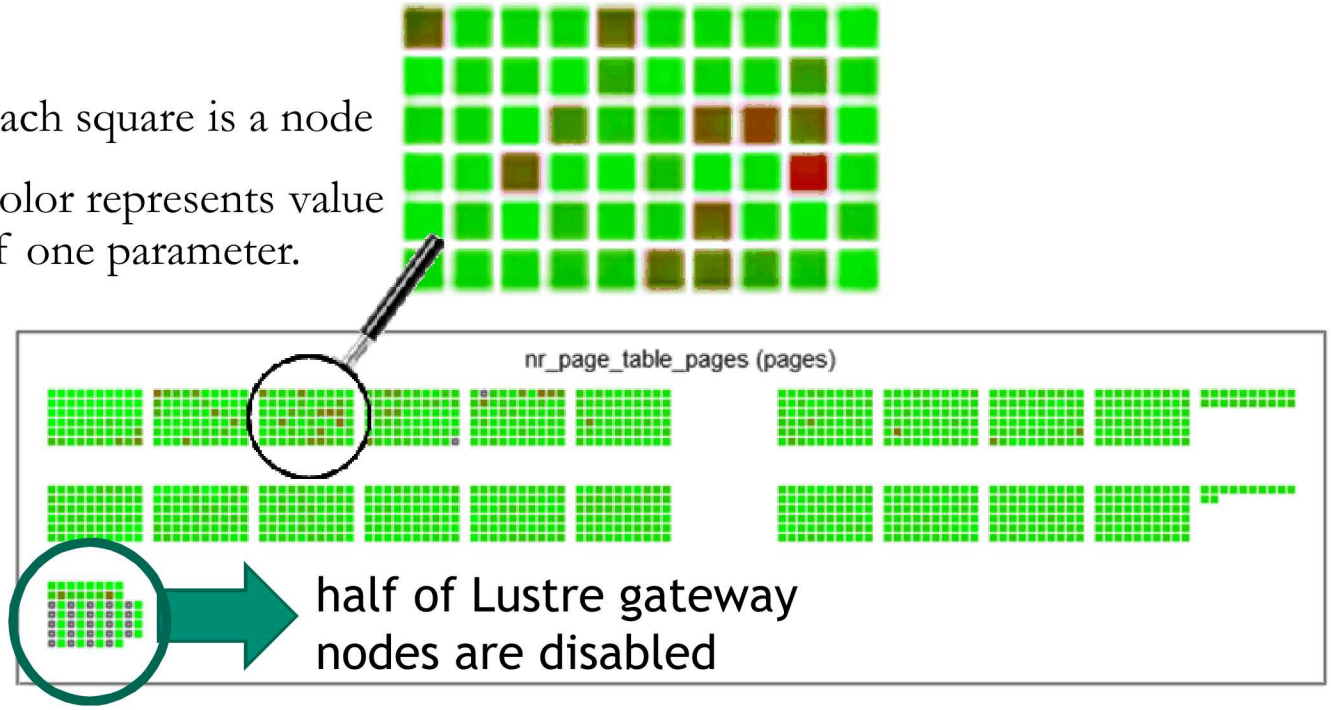


Can view one parameter on all nodes

OVIS client heat map of one parameter for every node. Easy to see outliers.

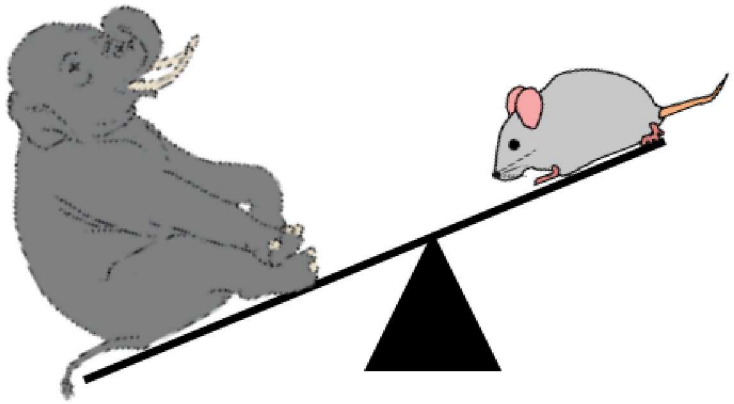
Each square is a node

Color represents value of one parameter.



Is load balanced?

Is job load balanced (job sends equal amounts of work to all nodes)?
Can look at CPU utilization, CPU temperature, CPU fan speed for all nodes in job.

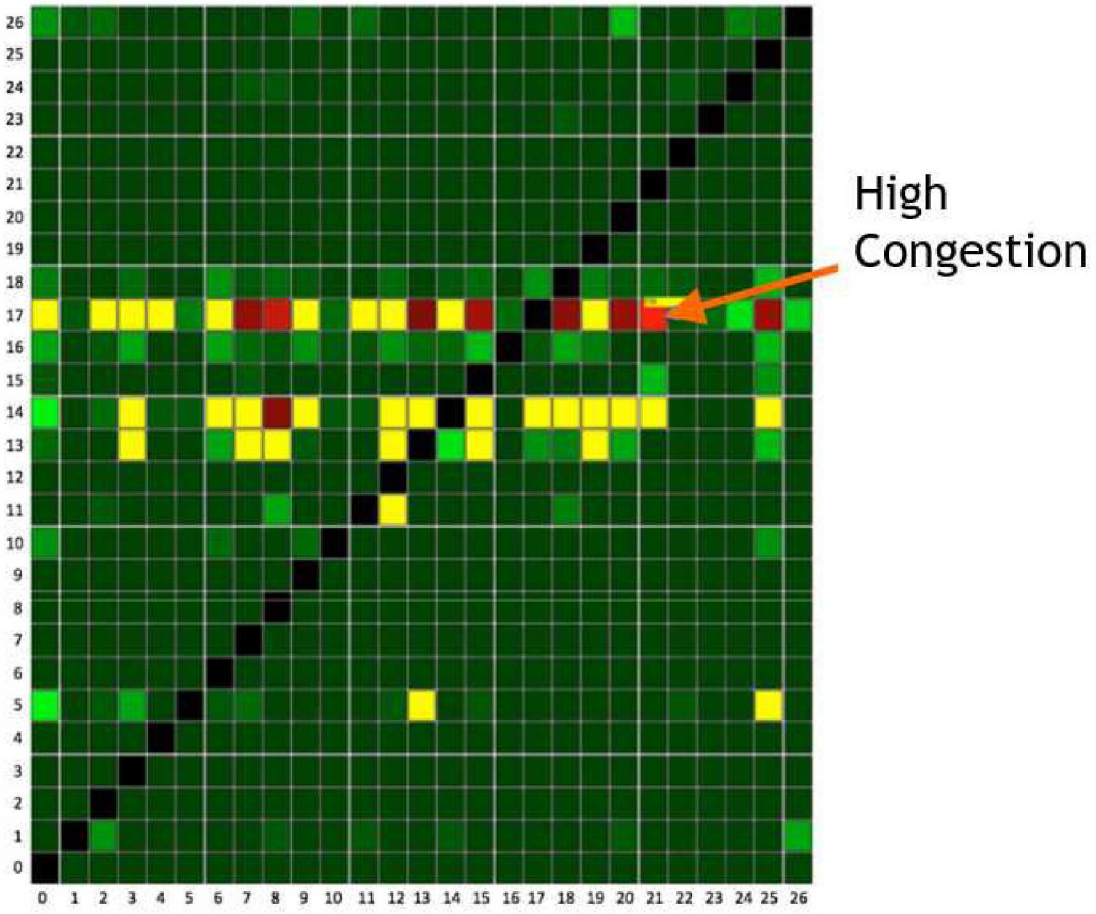


Derived from https://publicdomainvectors.org

Network congestion inside a Cray

Interior network allows nodes to communicate with each other.

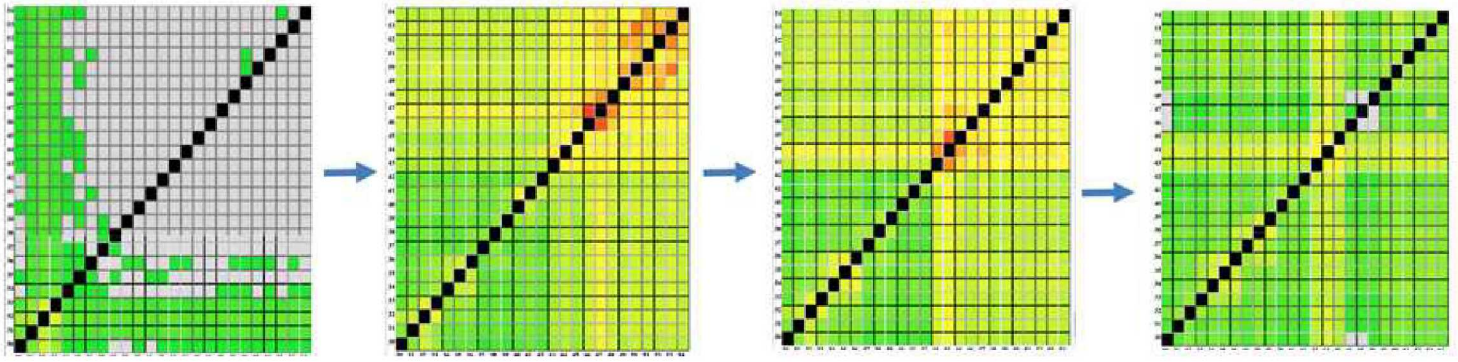
Client visualization of network congestion.



communications density

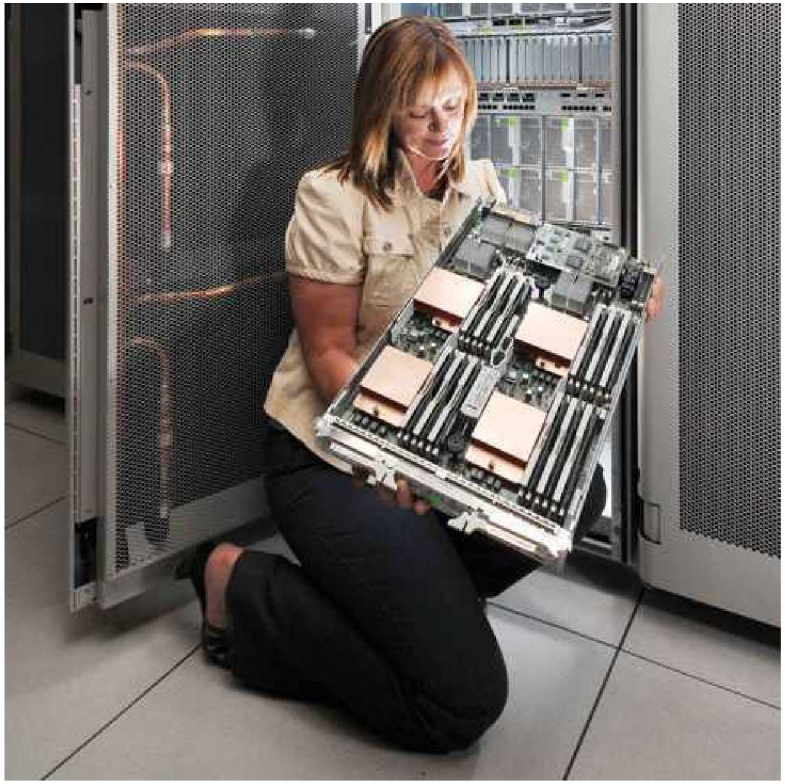
Can see real-time changes in communication traffic.

Timeseries stills



Published Results

Some results from various papers, written by numerous people, over 13 years.



https://www.sandia.gov/news/publications/research_magazine/articles/2017/03/beyond_moore.html

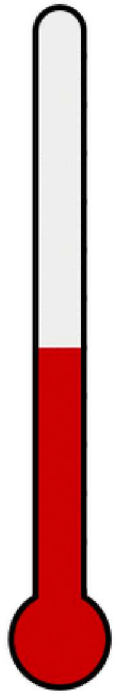
Real time monitoring of CPU temperature

Several case studies found nodes whose fans were running too fast.
Causes:

1) Hot exhaust air from one node enters the air intake of another node.
--SAND2005-4558

2) Different nodes have different fan grills. Different amounts of air enter/exit nodes.
--https://ovis.ca.sandia.gov/images/d/d9/Ipdps-2006.pdf

3) Faulty fan controller ran fan at wrong speed and sent wrong fan speed to OVIS.
--https://ovis.ca.sandia.gov/images/d/d9/Ipdps-2006.pdf



https://www.pdclipart.org/displayimage.php?album=search&cat=0&pos=4

of failed read/write on a hard drive

As hard drive gets older, expect number of failed reads/writes to increase.

For critical data, want to use a hard drive with very low read/write failures.

Paper makes this suggestion:
Monitor number of read/write failures. If number is high, then may want to consider replacing the hard drive.

Can't always see problem on simple plots

Plot of one parameter on one node may not reveal a problem.

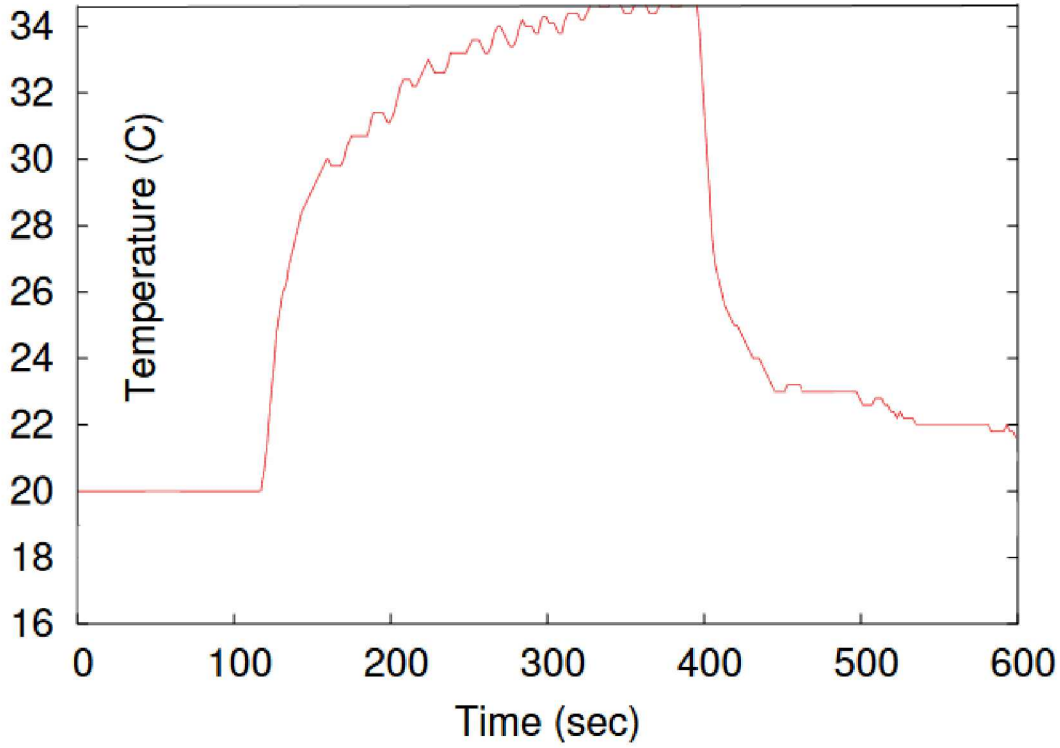
Temperature of one node.

Node is cool before job starts.

Node heats up during run.

Cools down when job ends

This looks normal.
…BUT…
this node has a serious problem



Derived from SAND2005-4558

Statistical analysis of identical nodes

A supercomputer has thousands of identical nodes. Nodes, when placed under identical workloads, should have nearly identical performance (i.e. normal distribution). OVIS uses a statistical approach to characterize behaviors of nodes.

Can not compare different cars

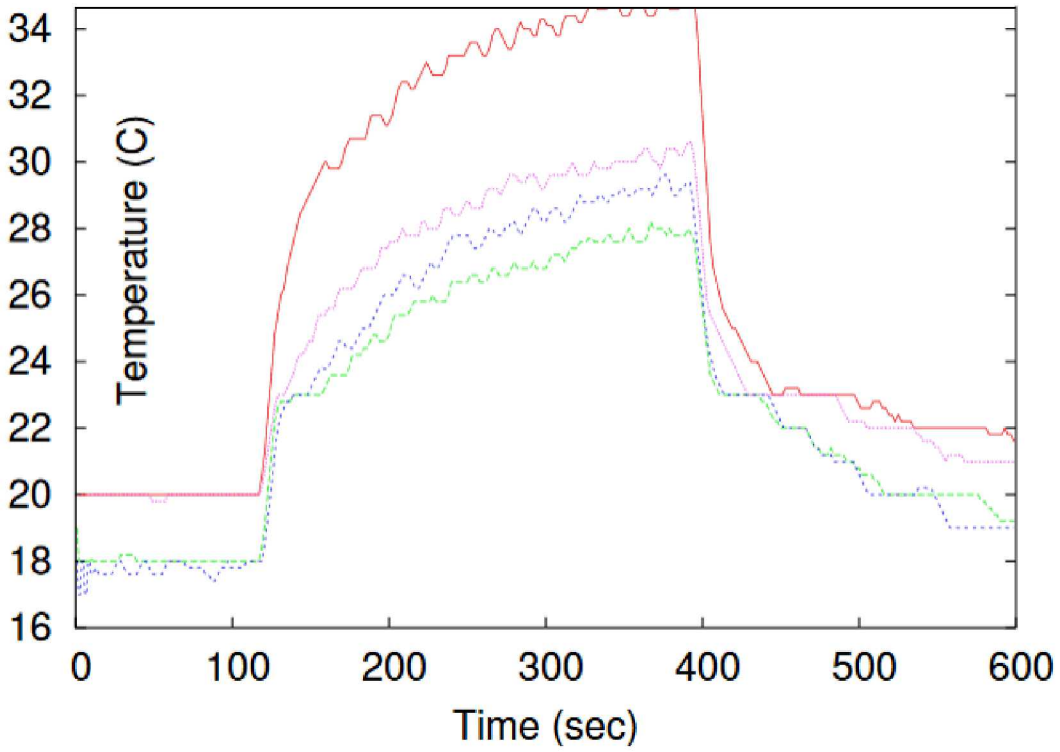
Can compare identical cars

Clipart is from pixabay

Plot of one parameter on several nodes.

All 4 notes are running similar parallel tasks in same job.

The top node is too hot.
Cause:
The CPU is not firmly attached to its heatsink.



Derived from SAND2005-4558

OVIS looks at ensemble of nodes

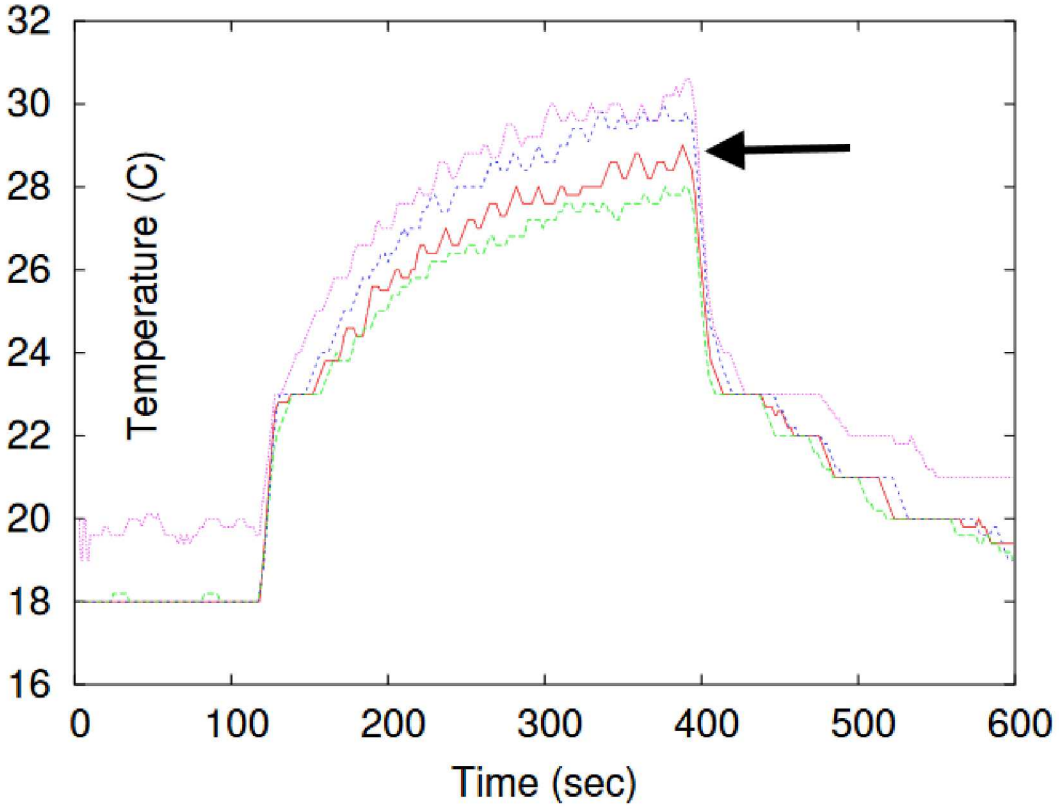
Prior to OVIS:
Monitoring tools looked at raw data values, one node at a time.

OVIS client plot of "# of processes" in 8 nodes.
X axis represents time.
Each row is one node.
Color represents "number of processes."

Collect statistical data for a baseline

Gather data when all 4 nodes are operating normally. Arrow points to the node that in no longer running hot.

All nodes, in this ensemble (identical nodes with similar work loads) have similar temperatures.

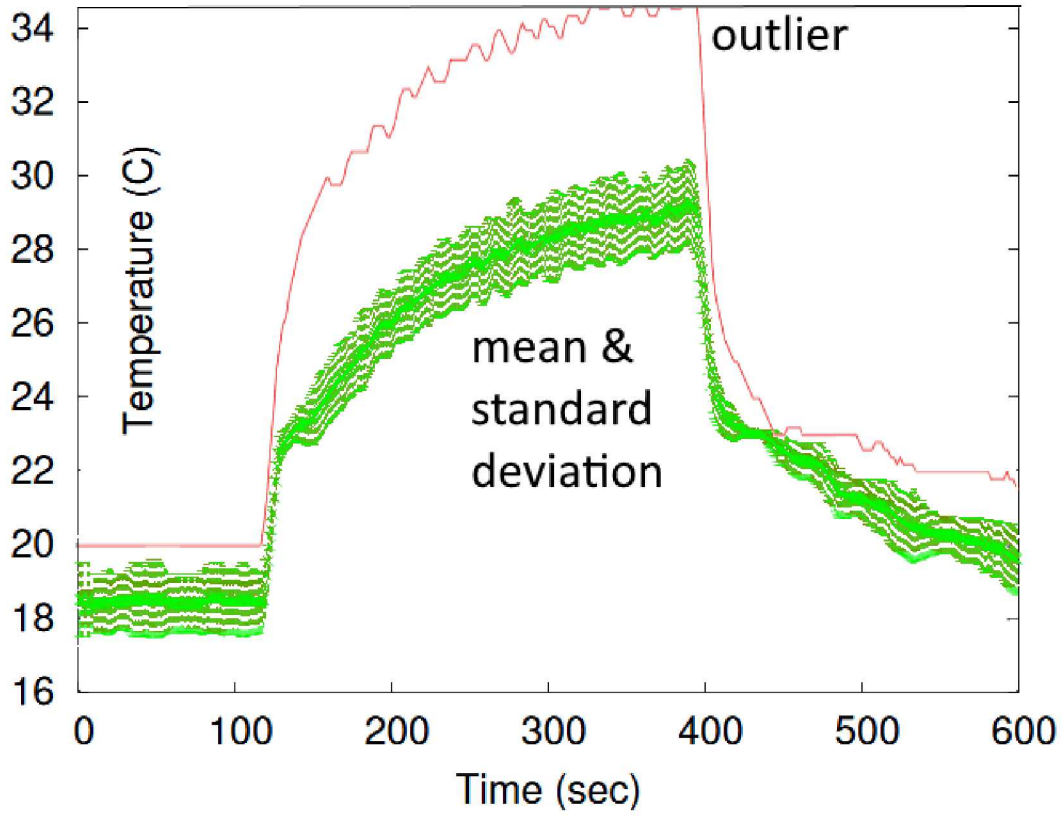


Derived from SAND2005-4558

Easy to find outliers

Green: mean and standard deviation of nodes when there is no problem.

Red: node is running too hot. Outlier is several standard deviations outside of mean.



Derived from SAND2005-4558

Changes impact nodes

Example:

Temperature in a node goes up when:
- workload goes up
- temperature in the room goes up
- fan speed decreases

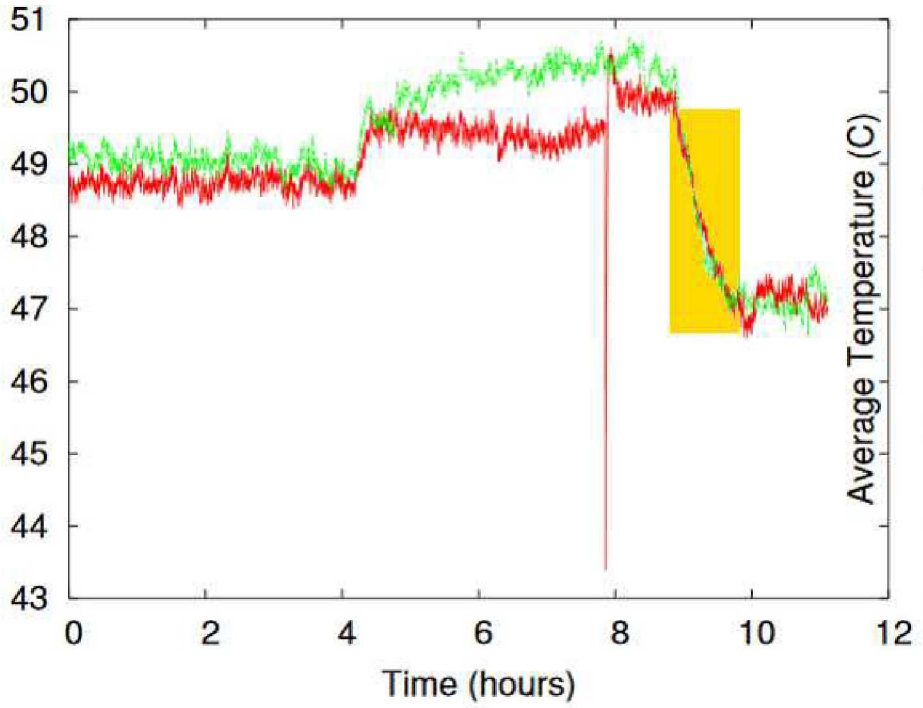
Average behavior of jobs

Look at average behavior of jobs to see impact from environment.

Two independent jobs.
Each job has several nodes.

Average temperature
of nodes in each job

Temperature drop in yellow
is NOT due to job nor node.
Both jobs see same drop.
Cause:
The room temperature dropped.



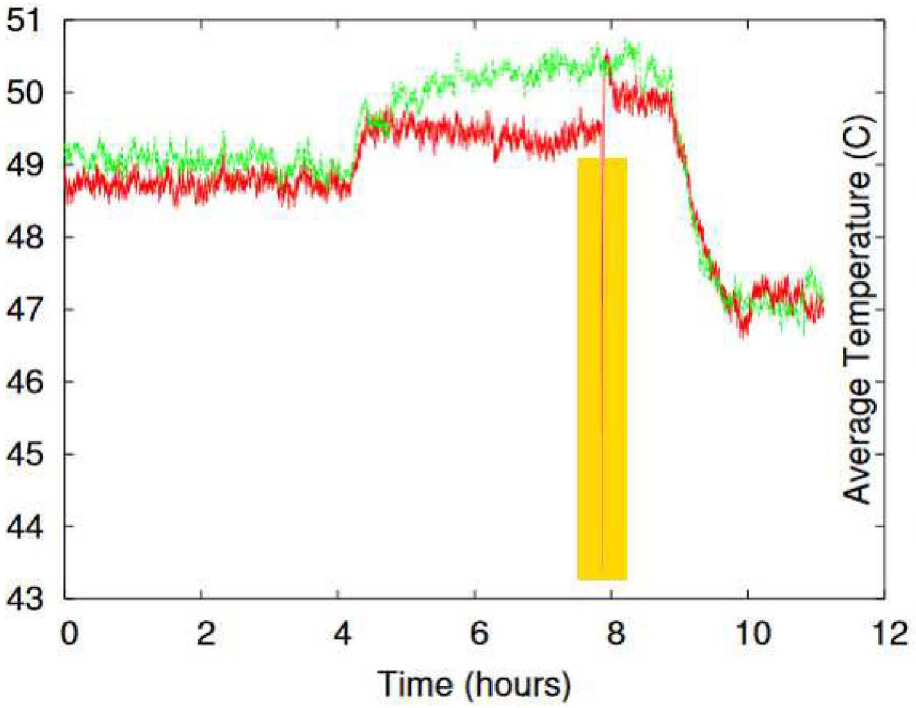
Derived from SAND2005-4558

Non-environmental changes

What about the temperature drop highlighted in yellow?

Only one job sees the drop.
Thus, NOT due to environment.

Cause:
A disk write occurred at this time on the red job. CPU stall.
Less work = lower temperature



Derived from SAND2005-4558

Server room

Cold air is pushed up from the floor and exits through ceiling. Fans, in the nodes, bring in cold air and expels hot air. The expelled hot air rises to the ceiling. Nodes, up high, are pulling in warmer air.

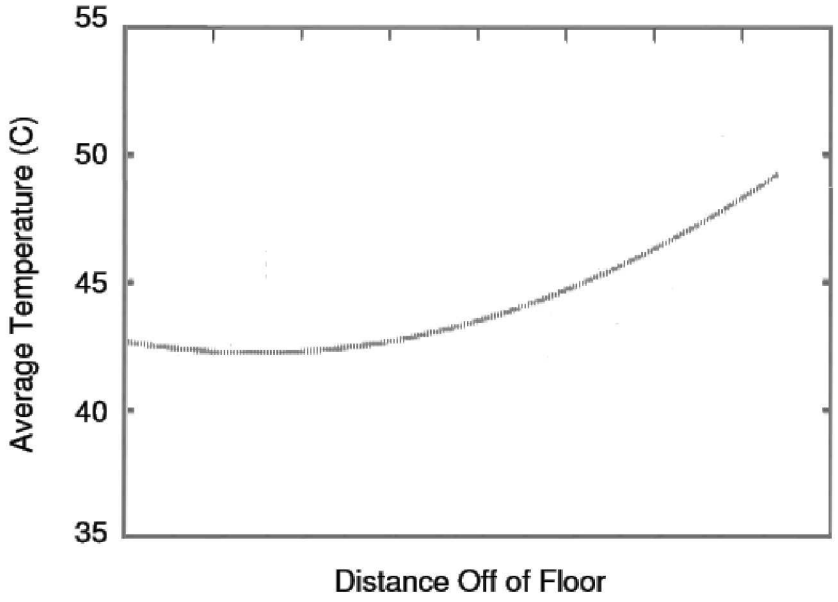
Expected behavior of the nodes that are under similar workloads is:
As distance from floor increases, node temperature increases.

https://share-ng.sandia.gov/news/resources/news_releases/images/2009/redsky_kc.jpg

Case study: coldest nodes above the floor

Average temperature of all nodes in a job.

For this particular cluster, coldest node is NOT on the floor.
Cause:
Bottom nodes pulling in warm exhaust air from nodes above it.



Derived from SAND2005-4558

Why track environment?

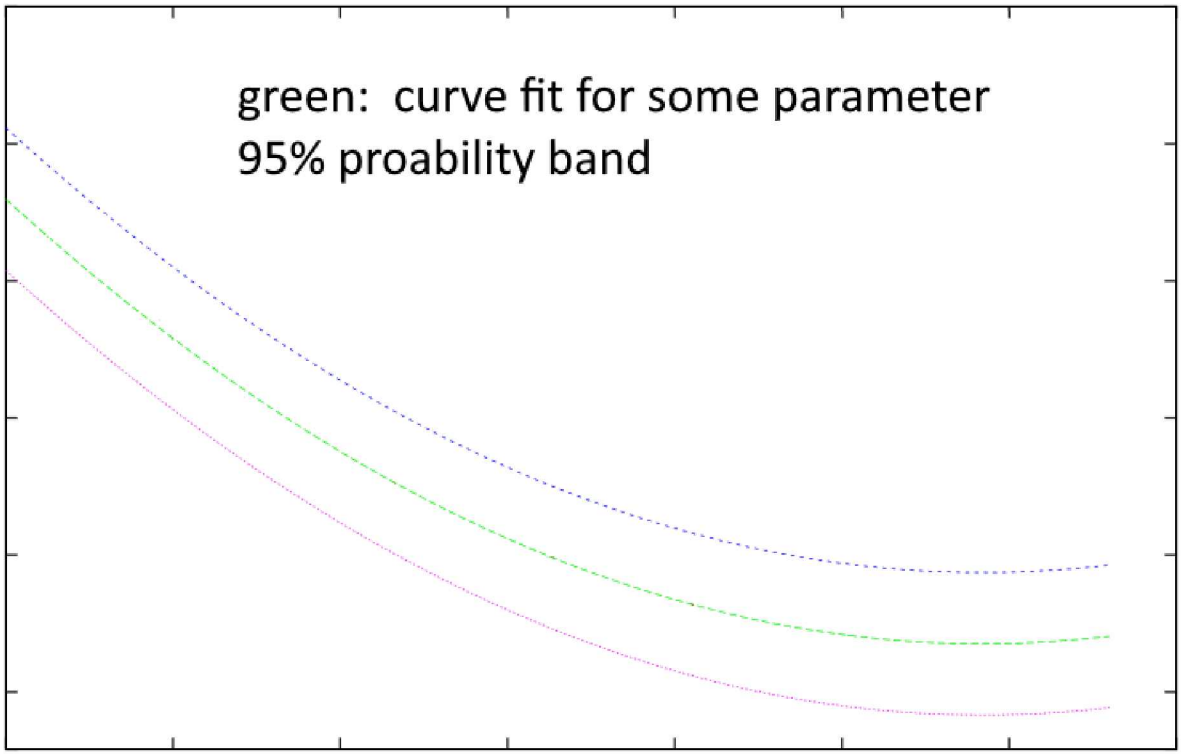
To determine if a node has a temperature issue (e.g. running too hot), must take-into-account the ambient temperature

temperature due to workload and fan speed = observed temperature − Tempurature of room

Once node temperatures are normalized (i.e. effects of ambient temperature are factored out), then the temperature behavior are identical for all nodes with similar work loads. This clears the path to do statistical analysis.

Bayesian Inference used to model a parameter

If acquired measurement for a node is outside 95% band, then node may have a problem.



https://ovis.ca.sandia.gov/images/d/d9/lpdps-2006.pdf

Memory errors due to cosmic radiation

One study finds cosmic radiation, when hitting clusters located at high altitudes (such as in Albuquerque), causes memory errors.
--SAND2006-7939.pdf

Fault Tolerance proposal

Bigger job = more nodes
More nodes = higher chance that job won't complete because a node will fail

Before OVIS:
Monitoring notified system administrator of a component failure AFTER the component failed.

OVIS uses a statistical approach to create a model of a typical node. Can compare observations of one node against the model to evaluate the health of the node. i.e. monitor degradations BEFORE a failure occurs.

Paper suggests reprograming a job scheduler to pick healthy nodes.

https://ovis.ca.sandia.gov/images/9/99/Resilience08.pdf

Parameters that determine health of node

voltage
temperature
fan speed
NIC (network)
SMART disk controller for hard drives

https://ovis.ca.sandia.gov/images/9/99/Resilience08.pdf

Active Memory (Recently Used Memory)

If a previous job exited a node with a high amount of Active Memory
and if a new job takes over the node
then higher probability that the new job won't run to completion because of an out-of-memory error.

https://ovis.ca.sandia.gov/images/4/44/OVISHPDC09-rev.pdf

Summary

Thousands of identical nodes allow for statistical analysis.

Generate models for ensembles of similar nodes operating under similar conditions.

Use the model to find degradations BEFORE a failure occurs.

Visualization makes it easy to see degradations.