SAND2006-7177C

OVIS:
A Tool for Intelligent, Real-time Monitoring of Computational Clusters

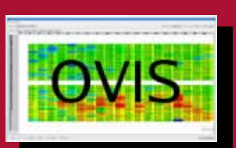
Matthew Wong, Jim Brandt, Ann Gentile,
José Ortega, Philippe Pébay, David Thompson
(with Youssef Marzouk)

Sandia National Laboratories
Livermore, CA

Sandia National Laboratories

Abstract

- OVIS is an open-source software tool for intelligent, real-time monitoring of computational clusters
- Visualization of deterministic information of cluster nodes:
  - CPU temperature
  - fan speed
  - memory error rate
  - etc.
- Built-in statistical tools for cluster analysis and prediction of future cluster health problems



Sandia National Laboratories

Traditional RAS Tools

Ganglia, Supermon and commonly-supplied vendor RAS systems

- Nodes within clusters treated in singleton
- Manufacturer determined extreme limits define thresholds used for failure detection/avoidance

OVIS

Statistical approach

– Enables earlier detection of abnormal behavior

– Can enable controlled failure avoidance

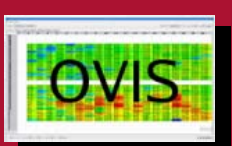
– Includes statistical and correlation tools

Spatial organization

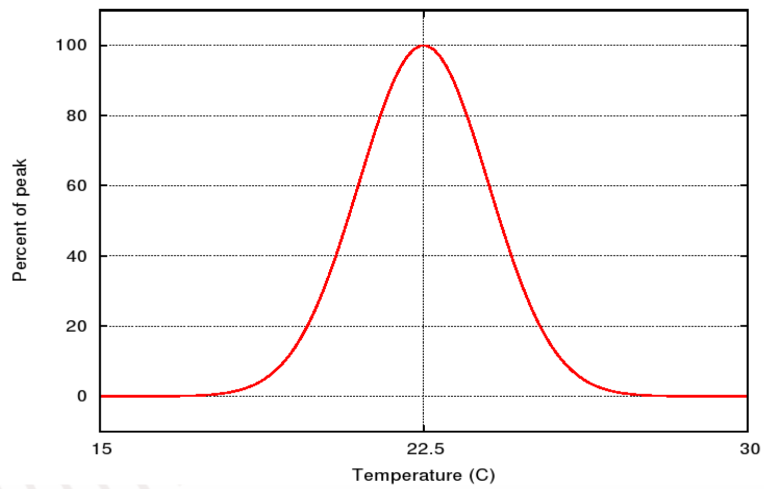
– Data is displayed in a meaningful realistic geographic layout

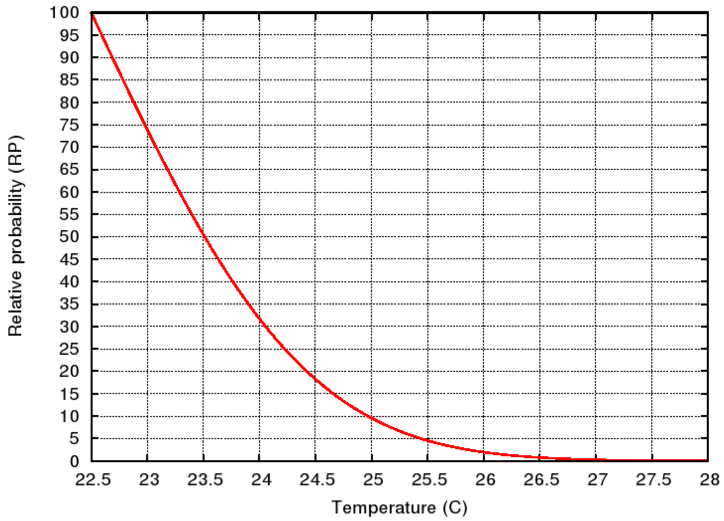
– Immediate visual feedback aids in environmental understanding and configuration

– Color map facilitates intuitive visualization of state

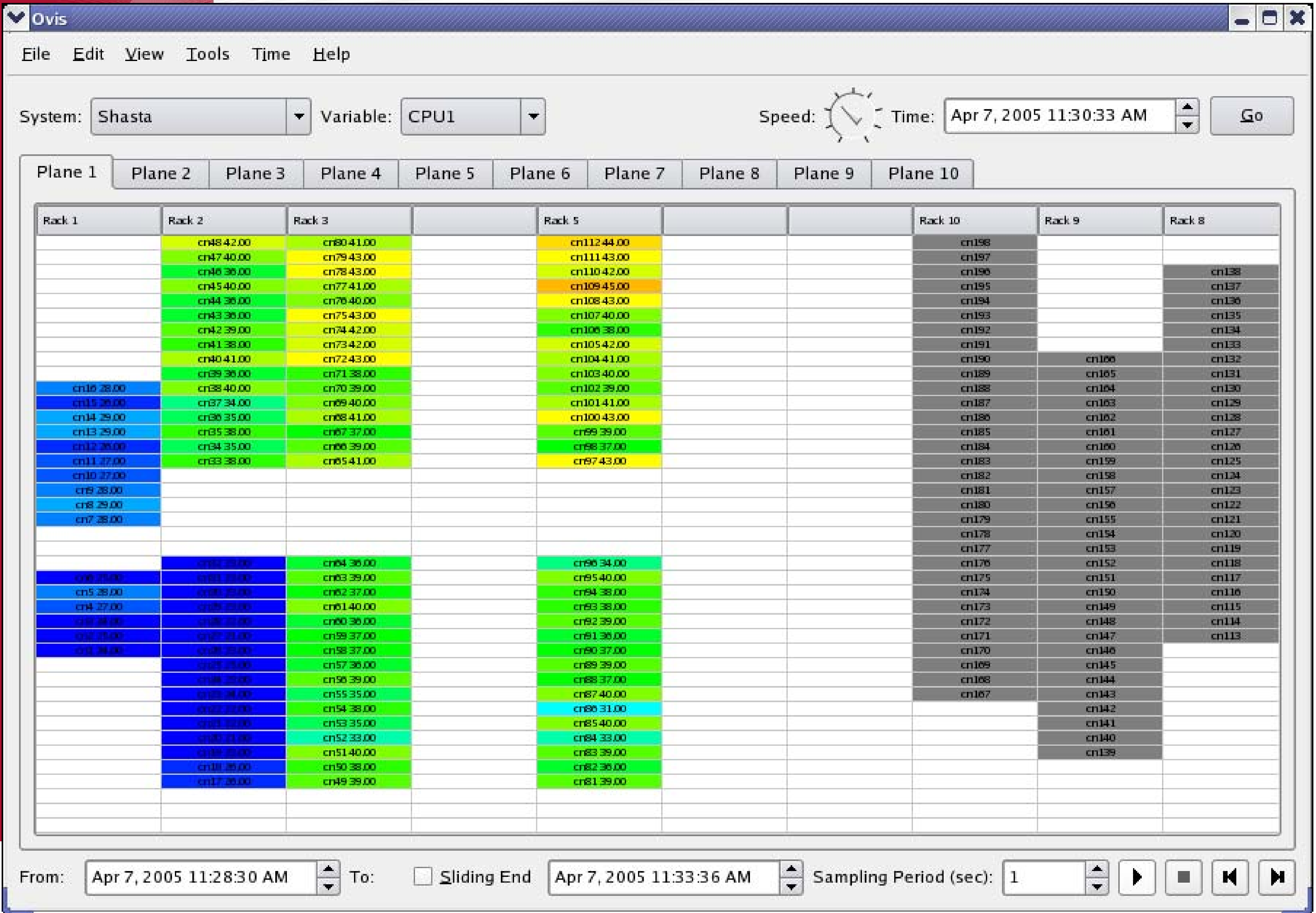


Statistical Approach: An example of using a PDF to determine a measure of the relative probability of a value occurring

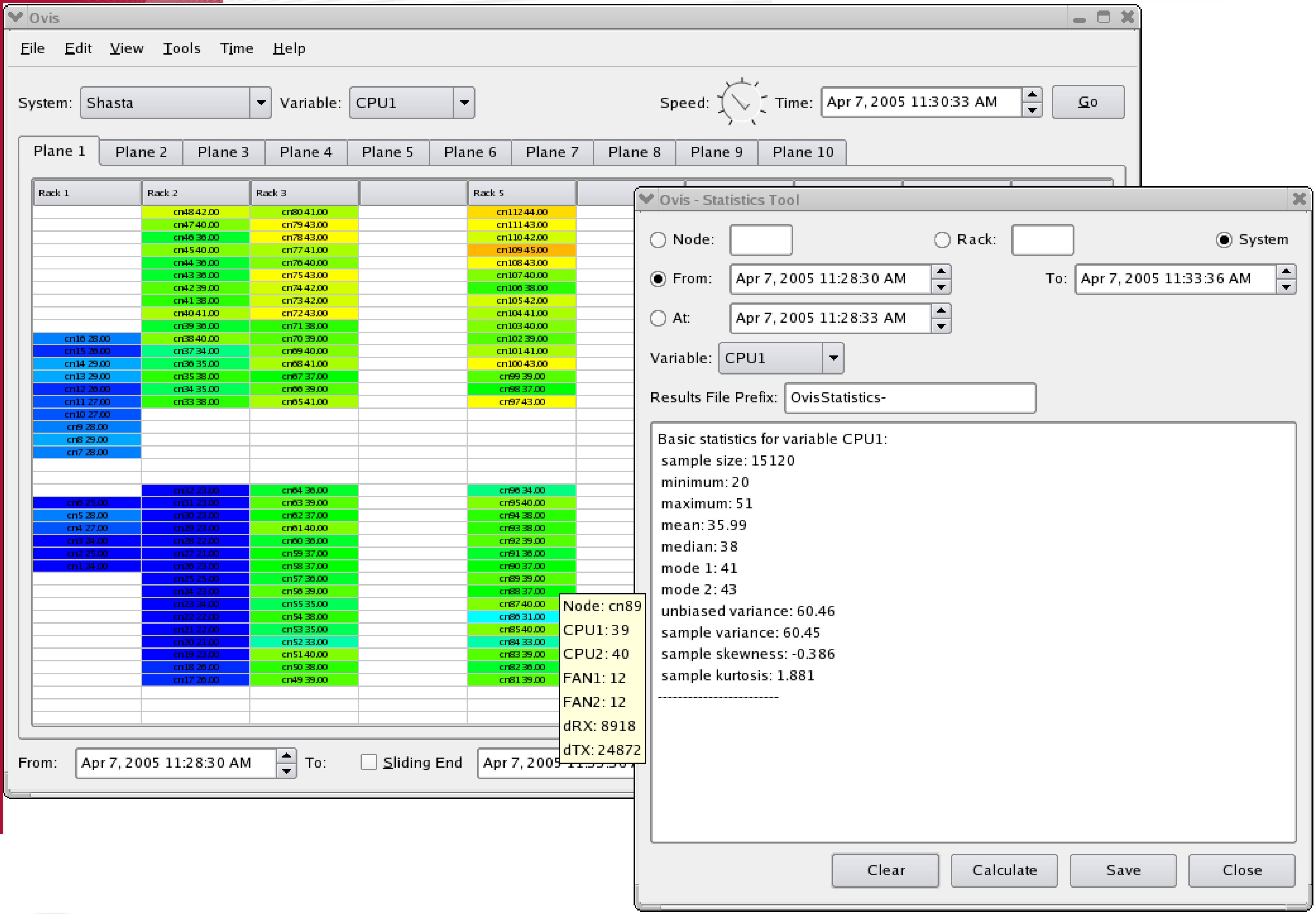




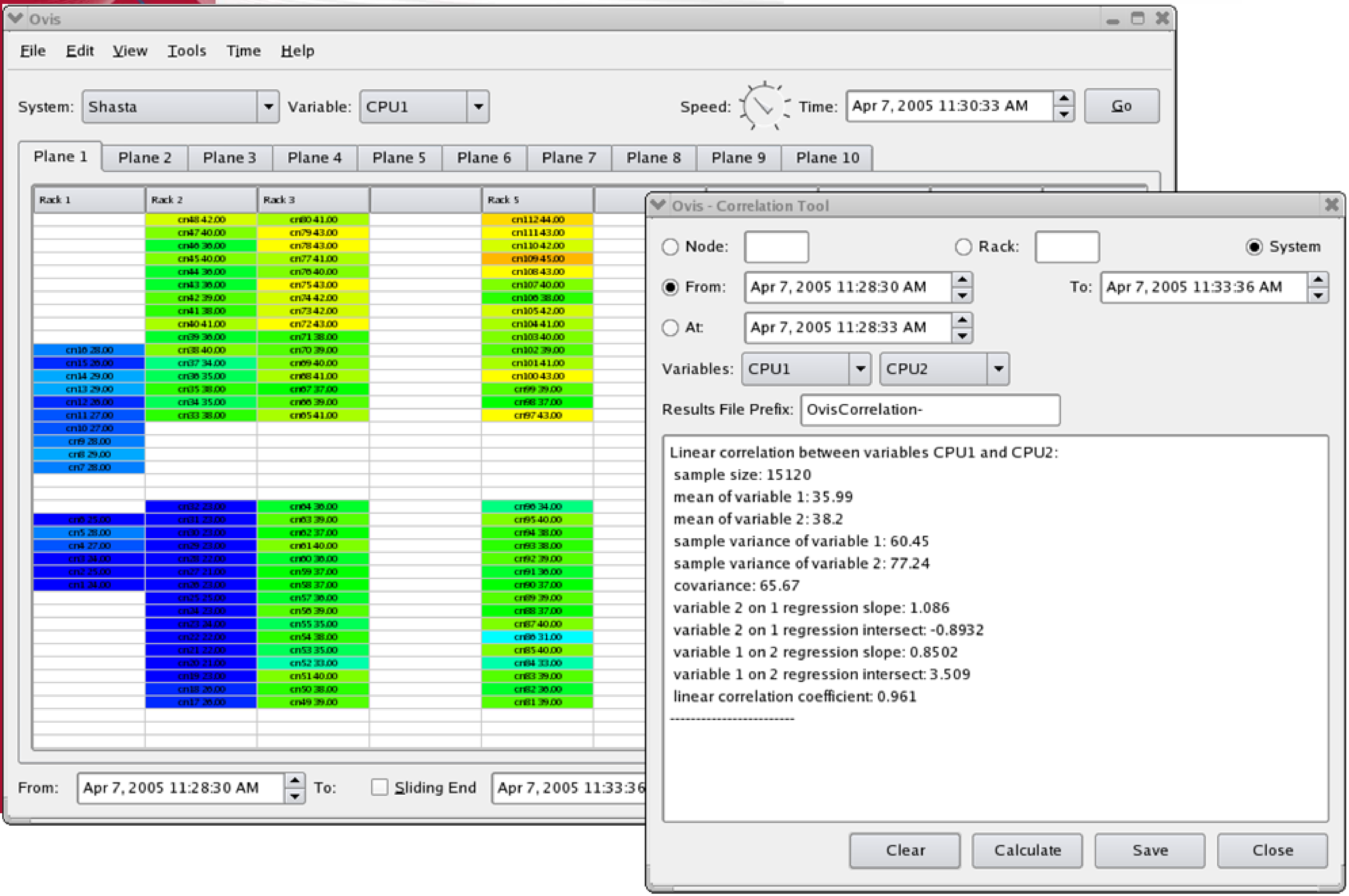
Spatial Approach



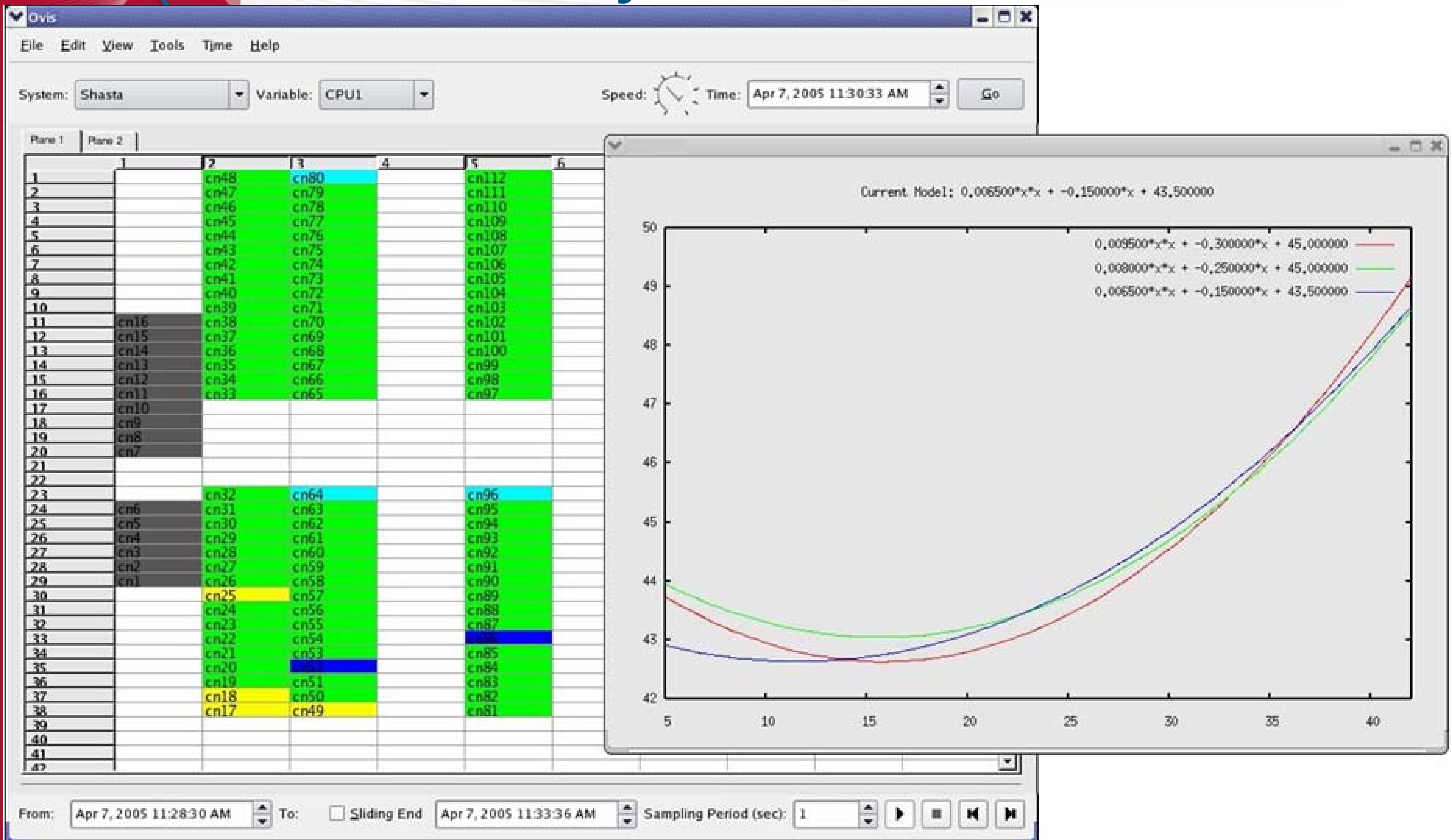
Statistics Tool



Correlation Tool

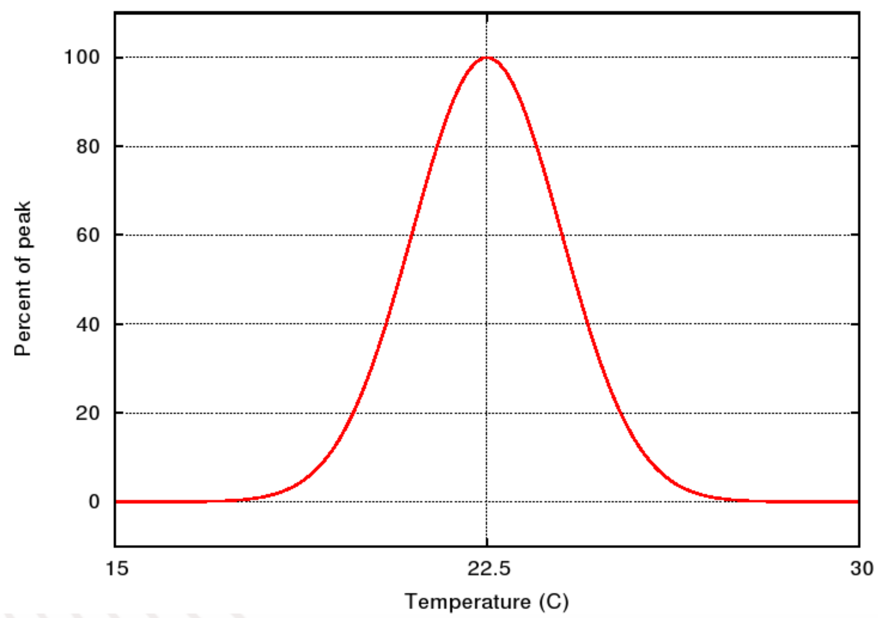


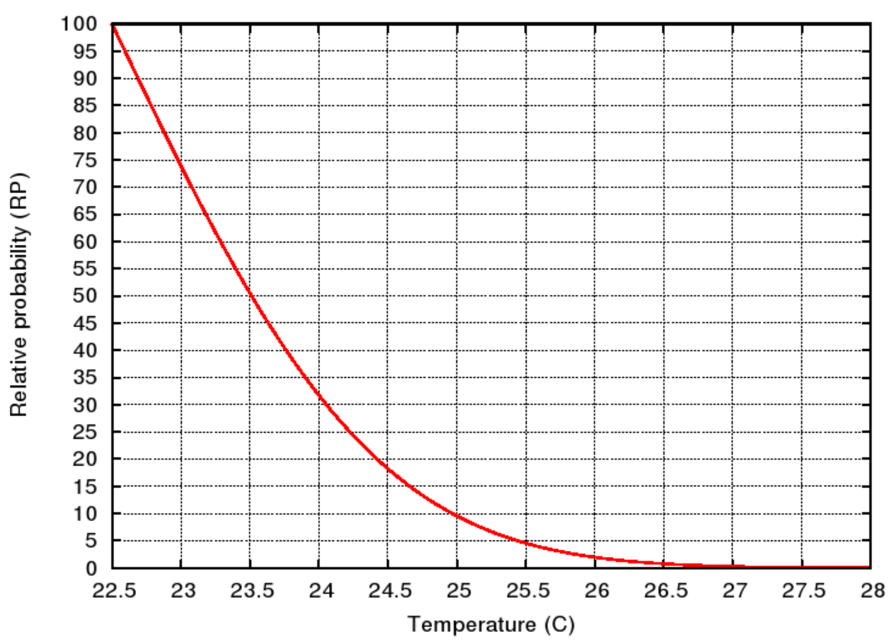
Environmental Modelling Using Bayesian Inference



Sandia National Laboratories

Statistical Approach: Plot of PDF of Idle Rack 3 at height 10



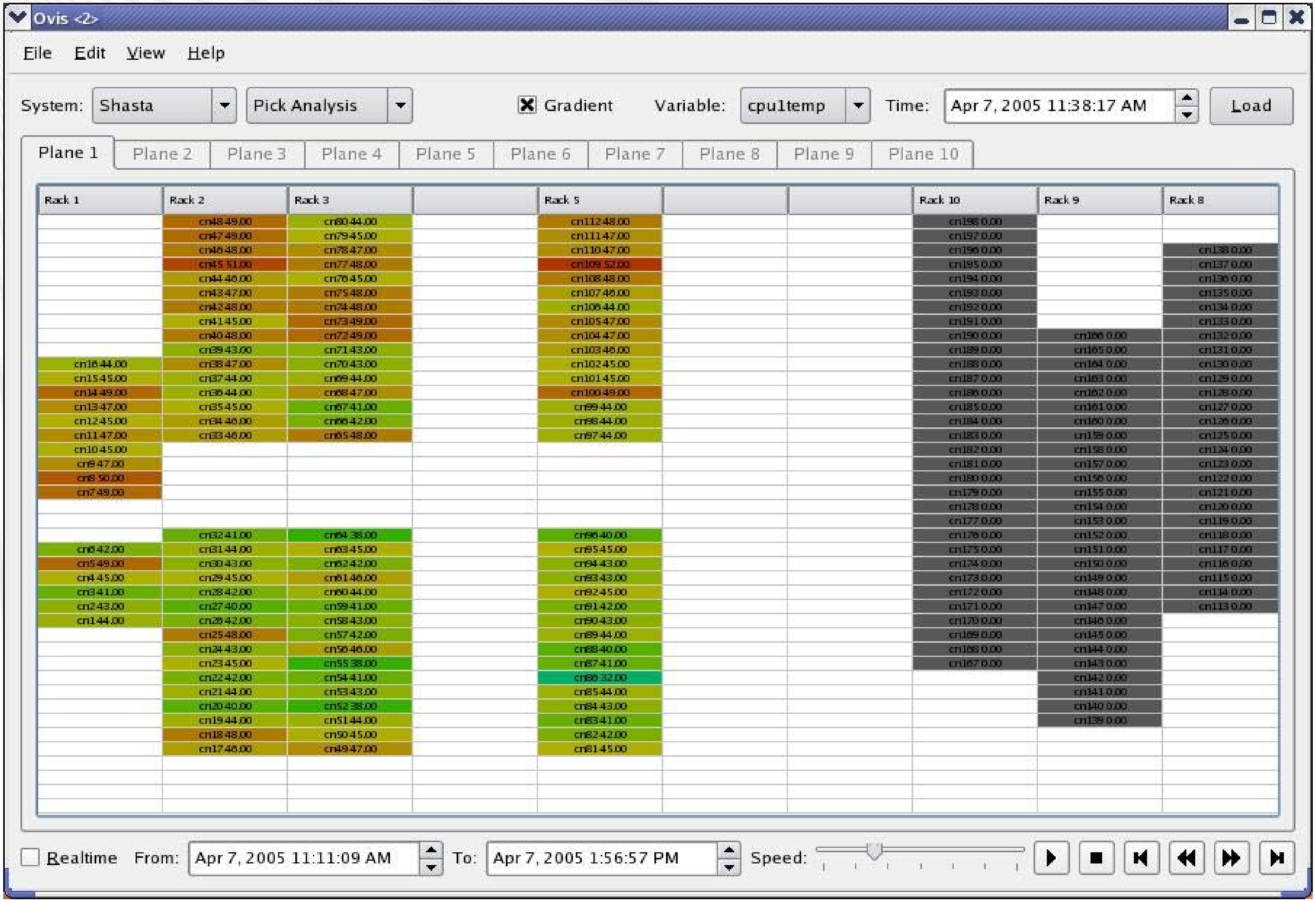


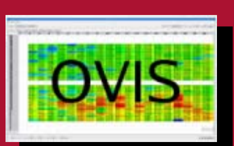
Sandia National Laboratories

OVIS – Other Features

- User-customizable color schema with automatic gradient shading
- Visualization of real-time data, or playback of historical data
- Easily adaptable to new systems
  - XML cluster description
  - Various readers to fetch collected data (e.g. Ganglia)

Demo


Ovis <2>
File Edit View Help
System: Shasta
Pick Analysis
Gradient
Variable: cpu1temp
Time: Apr 7, 2005 11:38:17 AM
Load
Plane 1 Plane 2 Plane 3 Plane 4 Plane 5 Plane 6 Plane 7 Plane 8 Plane 9 Plane 10
Rack 1 Rack 2 Rack 3 Rack 5 Rack 10 Rack 9 Rack 8
Realtime From: Apr 7, 2005 11:11:09 AM To: Apr 7, 2005 1:56:57 PM Speed:


OVIS

OVIS 2

- 3-D navigable cluster representation providing advanced visualization with drill-down functionality to component level
- Parallel architecture for real-time statistical characterization, modeling and analysis of large (10's of thousands of nodes) clusters
- Further data analysis tools including automated troubleshooting functionality

Contacts and Downloads

Contacts

ovis-help@sandia.gov

For OVIS downloads and more information:

http://ovis.ca.sandia.gov

Last Slide

Extras

Problems and Solutions

P **CPU utilization of different nodes could be different at any point in time → legitimately different CPU temperatures**

S **Model steady state temperature distribution(s) with CPU utilization dependence**

P **All nodes not co-located in space → non-homogeneous environment**

S **Model temperature distribution(s) with geographic location (e.g. distance off floor) dependence**

Problems and Solutions

P **Room temperature not fixed (but does have an acceptable range) → same CPU utilization rate yields different temperatures at different times (though change is slow)**

S **Use new data to continuously update model parameters and confidence in the model**

P **If sampling rate too slow --> Cannot look for stability**

S **Model upper and lower bounds**

OVIS

Sandia National Laboratories

Ganglia Cluster Toolkit
http://ganglia.sourceforge.net

Physical View for Thu, 13 Apr 2006 13:59:42 -0700

Get Fresh Data

Full View

Sandia ICC Shasta System Grid > Compute Partition > --Choose a Node

Compute Partition cluster - Physical View | Columns 4

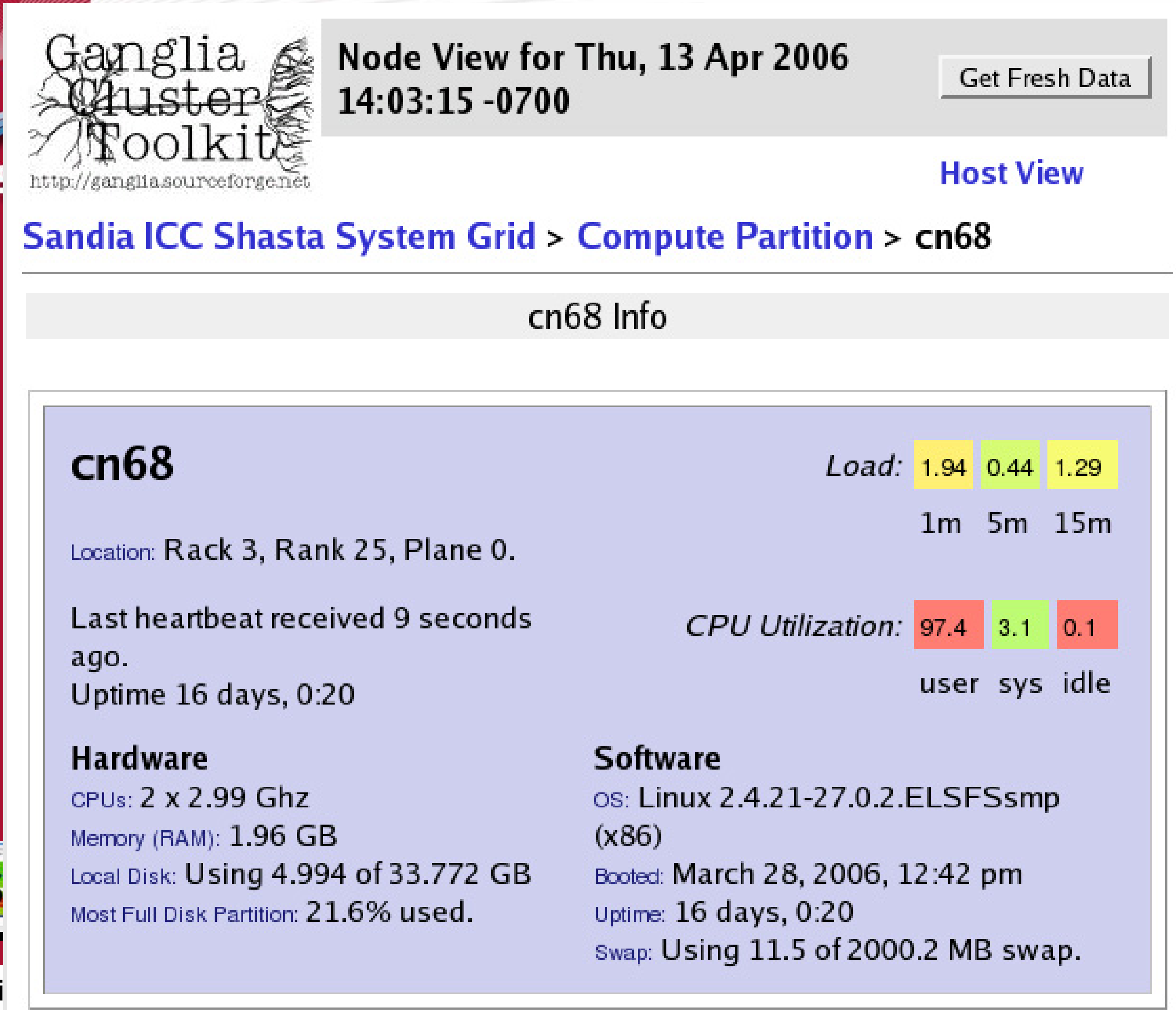
Verbosity level (Lower is more compact):
3 2 1

Total CPUs: **396**
Total Memory: **387.7 GB**

Total Disk: **6685.0 GB**
Most Full Disk: cn64 (38.3% Used)

Rack 1	Rack 2	Rack 3	Rack 5
cn16 0.20 cpu: 2.99G (2) mem: 1.96G	cn48 0.02 cpu: 2.99G (2) mem: 1.96G	cn80 0.03 cpu: 2.99G (2) mem: 1.96G	cn112 2.00 cpu: 2.99G (2) mem: 1.96G
cn15 0.00 cpu: 2.99G (2) mem: 1.96G	cn47 2.00 cpu: 2.99G (2) mem: 1.96G	cn79 0.02 cpu: 2.99G (2) mem: 1.96G	cn111 2.00 cpu: 2.99G (2) mem: 1.96G
cn14 0.00 cpu: 2.99G (2) mem: 1.96G	cn46 1.99 cpu: 2.99G (2) mem: 1.96G	cn78 0.01 cpu: 2.99G (2) mem: 1.96G	cn110 2.00 cpu: 2.99G (2) mem: 1.96G
cn13 2.00 cpu: 2.99G (2) mem: 1.96G	cn45 1.99 cpu: 2.99G (2) mem: 1.96G	cn77 0.02 cpu: 2.99G (2) mem: 1.96G	cn109 2.00 cpu: 2.99G (2) mem: 1.96G
cn12 2.09 cpu: 2.99G (2) mem: 1.96G	cn44 2.01 cpu: 2.99G (2) mem: 1.96G	cn76 0.00 cpu: 2.99G (2) mem: 1.96G	cn108 1.00 cpu: 2.99G (2) mem: 1.96G
cn11 2.00 cpu: 2.99G (2) mem: 1.96G	cn43 1.99 cpu: 2.99G (2) mem: 1.96G	cn75 0.01 cpu: 2.99G (2) mem: 1.96G	cn107 2.00 cpu: 2.99G (2) mem: 1.96G
cn10 2.00 cpu: 2.99G (2) mem: 1.96G	cn42 1.99 cpu: 2.99G (2) mem: 1.96G	cn74 0.00 cpu: 2.99G (2) mem: 1.96G	cn106 2.00 cpu: 2.99G (2) mem: 1.96G
cn9 2.00 cpu: 2.99G (2) mem: 1.96G	cn41 1.99 cpu: 2.99G (2) mem: 1.96G	cn73 0.00 cpu: 2.99G (2) mem: 1.96G	cn105 2.00 cpu: 2.99G (2) mem: 1.96G
cn8 2.00 cpu: 2.99G (2) mem: 1.96G	cn40 1.99 cpu: 2.99G (2) mem: 1.96G	cn72 0.00 cpu: 2.99G (2) mem: 1.96G	cn104 2.00 cpu: 2.99G (2) mem: 1.96G
cn7 2.01 cpu: 2.99G (2) mem: 1.96G	cn39 1.99 cpu: 2.99G (2) mem: 1.96G	cn71 0.00 cpu: 2.99G (2) mem: 1.96G	cn103 0.00 cpu: 2.99G (2) mem: 1.96G
cn6 2.00 cpu: 2.99G (2) mem: 1.96G	cn38 0.00 cpu: 2.99G (2) mem: 1.96G	cn70 0.00 cpu: 2.99G (2) mem: 1.96G	cn102 2.00 cpu: 2.99G (2) mem: 1.96G
cn5 2.00 cpu: 2.99G (2) mem: 1.96G	cn37 2.00 cpu: 2.99G (2) mem: 1.96G	cn69 0.00 cpu: 2.99G (2) mem: 1.96G	cn101 2.00 cpu: 2.99G (2) mem: 1.96G
cn4 2.00 cpu: 2.99G (2) mem: 1.96G	cn36 2.01 cpu: 2.99G (2) mem: 1.96G	cn68 0.02 cpu: 2.99G (2) mem: 1.96G	cn100 2.00 cpu: 2.99G (2) mem: 1.96G

Laboratories



Sandi
Natio...
Laboratories

cn68 Overview

This host is up and running.

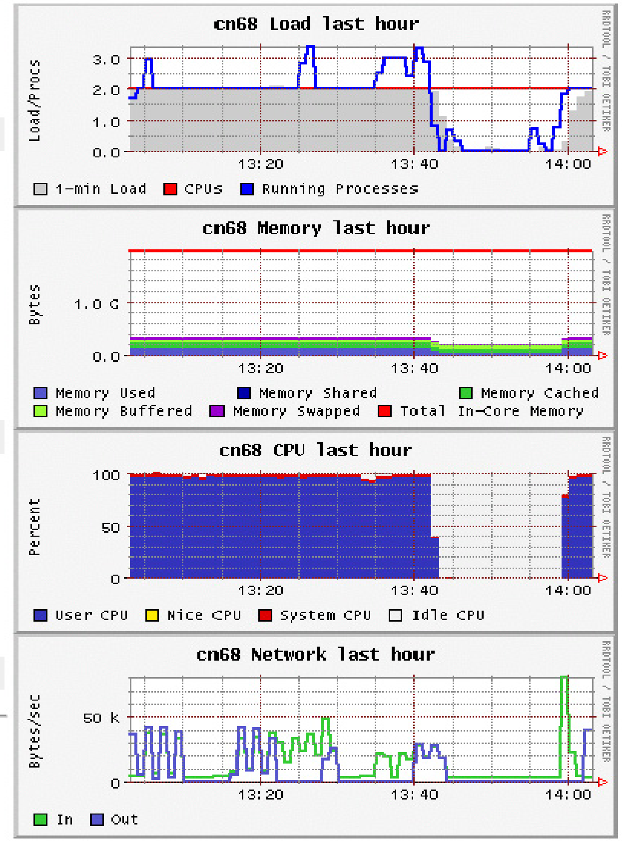
Time and String Metrics

boottime	Tue, 28 Mar 2006 12:42:10 -0800
gexec	OFF
machine_type	x86
os_name	Linux
os_release	2.4.21-27.0.2.ELSFSsmp
sys_clock	Mon, 3 Apr 2006 09:22:54 -0700
uptime	16 days, 0:20

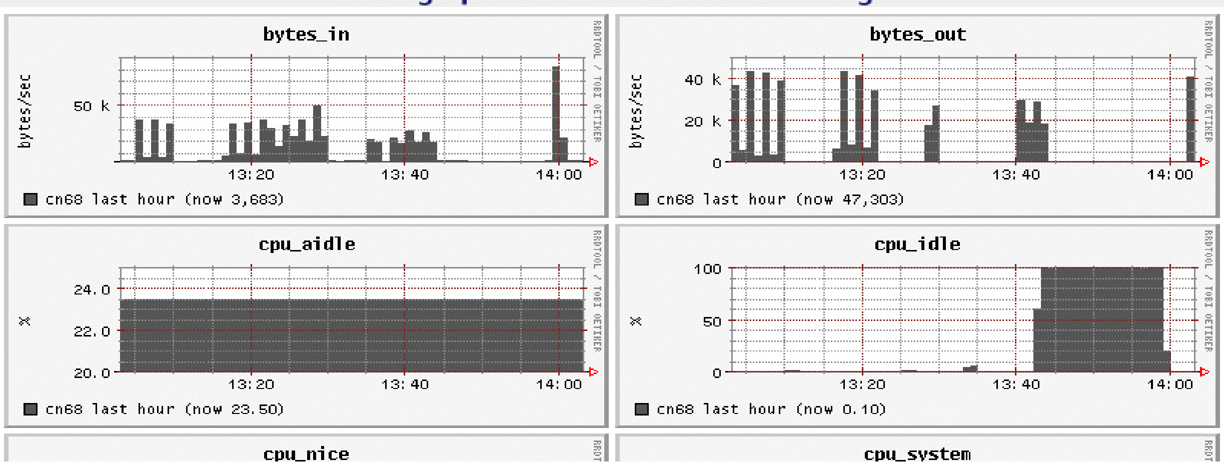
Constant Metrics

cpu_num	2
cpu_speed	3065 MHz
mem_total	2053420 KB
mtu	1500 B
swap_total	2048248 KB

Gmetrics



cn68 **graphs** last **hour** sorted **descending**



OVIS

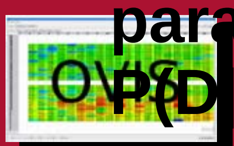
Sandia National Laboratories

Why simple statistics aren't enough

Need something about why we are using Bayesian inference.

Example of Bayesian Modeling: Dependence of Temperature on Height

- Bayesian learning allows us to incorporate expert knowledge in the model. e.g., it has been noted that on Shasta, the temperature T baseline varies with height h in the cluster.
- However, each individual node should behave similarly under similar conditions, within some variability range due to the manufacturing process.
- Thus, we model T as a Gaussian r.v. with mean Q(h) (Q is a quadratic => 3 degrees of freedom) and variance s. We can then infer the PDF of the 4 unknown parameters based on the data at hand using Bayes' formula.
- But in fact, we are mostly interested in the *most likely* parameters to characterize the model, using $P(X|D,M) \propto P(D|X,M) \times P(X|M)$
- Model training is automatically done until some user-defined maximum likelihood convergence ratio is reached



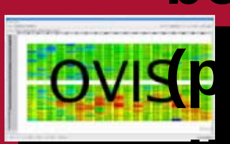
Sandia National Laboratories

Problems

- CPU utilization of different nodes could be different at any point in time → legitimately different CPU temperatures
- All nodes not co-located in space → non-homogeneous environment
- Room temperature not fixed (but does have an acceptable range) → same CPU utilization rate yields different temperatures at different times (though change is slow)
- Sampling rate too slow --> Cannot look for stability --> can only compare with upper and lower bounds

Solutions

- Model steady state temperature distribution(s) with CPU utilization dependence
- Model temperature distribution(s) with geographic location (e.g. distance off floor) dependence
- Use new data to continuously update model parameters and confidence in the model
- Model upper and lower bounds and call anything in between good though must detect shift in model (perhaps if many are falling out of bounds in same direction).



OVIS's Interface

OVIS can be used either in command-line mode or with a GUI Interface that:

- provides smooth transition between raw data and statistics and derived data and visualization,
- facilitates system administrator analyses for configuration and monitoring of a system,
- supports a variety of drop-in monitoring and analysis modules.

Current functionality

- Provide raw data visualization and archiving
- Provide statistical characterization, Bayesian modeling and analysis – not very fast and needs a homogeneous glob to work on

Posterior

- $P(X|D,M) = P(D|X,M) \times P(X|M) / P(D|M)$
- P(X|D,M) (posterior) is the probability distribution of model parameters given our data and choice of model.
- We chose the particular parameter set which yields the maximum of the PDF as most likely and use it and the model as the basis for comparison with any individual node

Likelihood

- $P(X|D,M) = P(D|X,M) \times P(X|M) / P(D|M)$

- P(D|X,M) (likelihood) is the probability distribution of actual data over all sets of parameters for the model

Prior

- $P(X|D,M) = P(D|X,M) \times P(X|M) / P(D|M)$

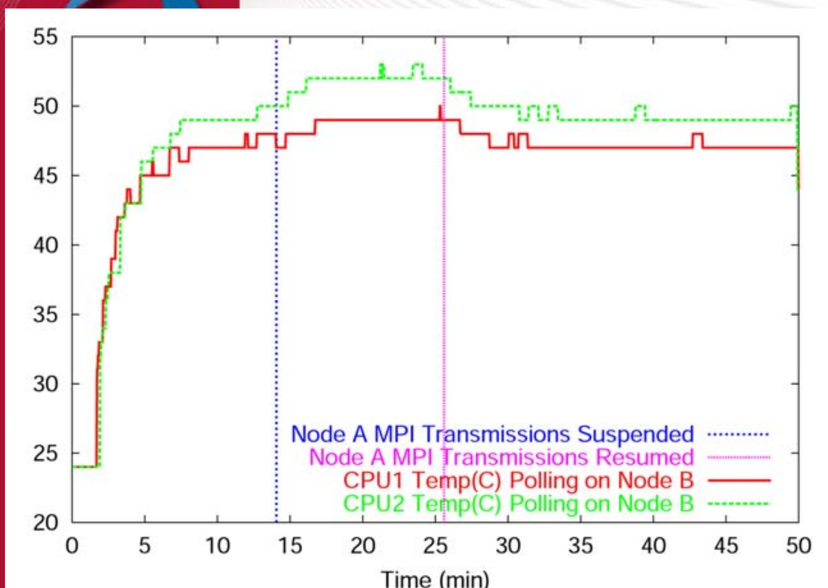
- $P(X|M)$ (prior) is the probability distribution of the parameter sets of the model
  - For initial calculation allows input of expert knowledge
    - Model selection
    - Knowledge of the actual distribution of parameter sets (we use a uniform distribution because of lack of knowledge)
  - Posterior from previous iteration is used on subsequent iterations

Evidence

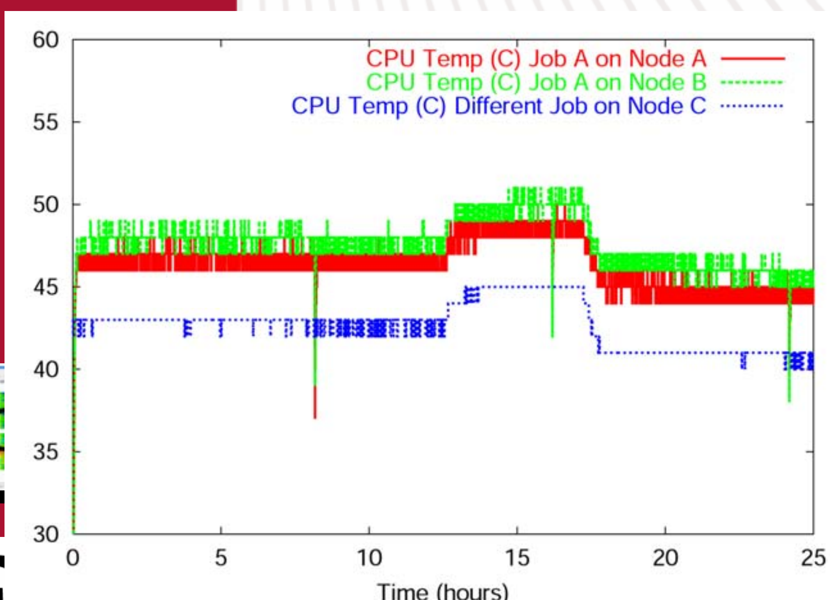
- $P(X|D,M) = P(D|X,M) \times P(X|M) / P(D|M)$

- P(D|M) (evidence) is a normalization term calculated by summing the posteriors before normalization has occurred (total CDF = 1)

Statistically Recognizable Abnormalities



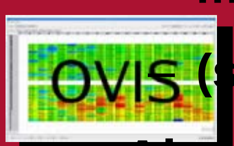
Emulate MPI failure – suspend one node, temperature rises in polling nodes



Room temperature changes – affect nodes in uncorrelated job groups

Demo: Example Analysis Sequence

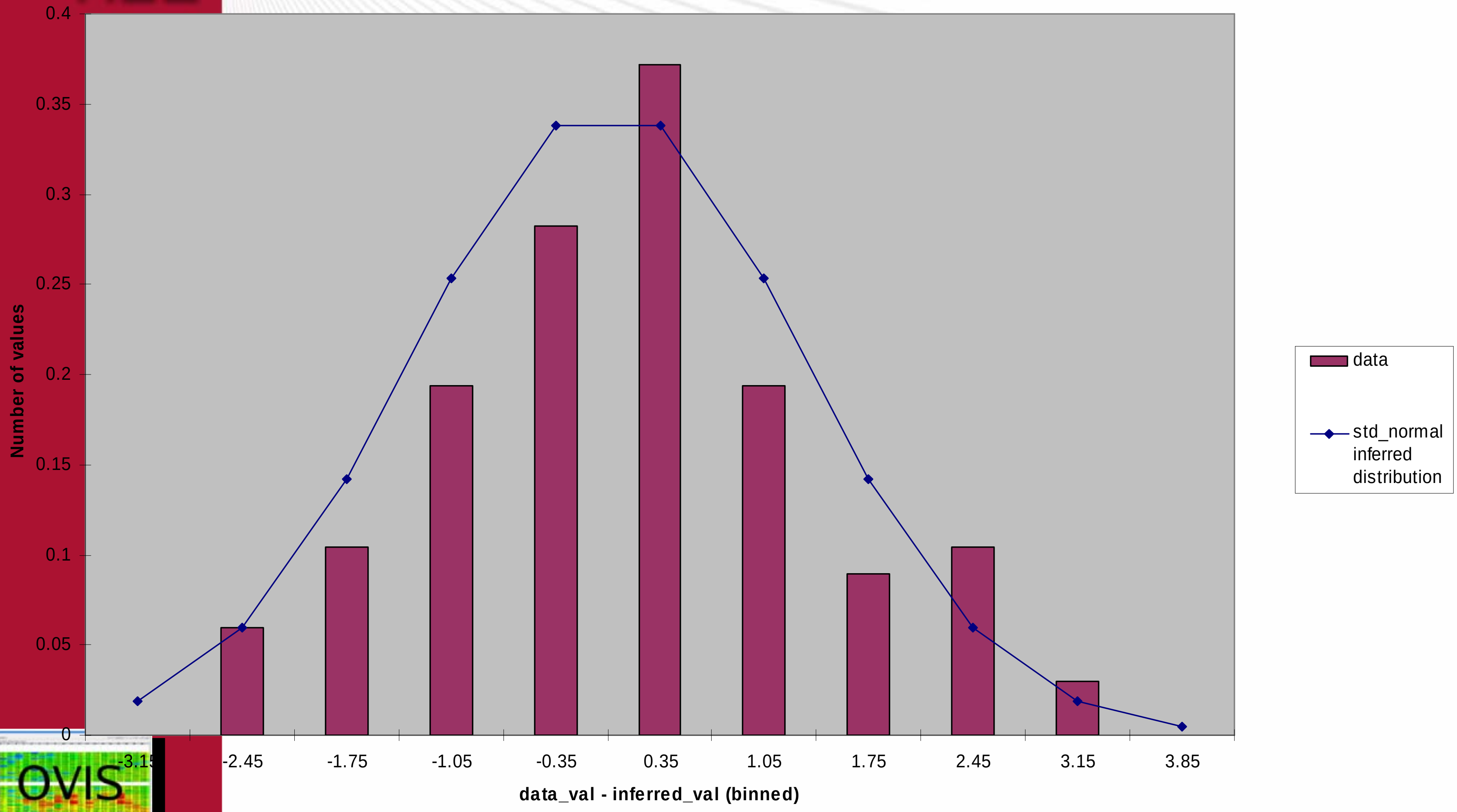
- General overview of variables:
  - Fan Initial
    - Faster around gaps
  - Altering Airflow - Fan
    - Problem solved when plugged racks
  - Temperature
    - Jobs come in and out (propagation of jobs onto the cluster starting at higher numbers)
    - Natural vertical gradient
    - Also hot near floor
    - Node 86
- Inference Model (Non-uniform and non-linear environmental effects)
  - Fit to Normal distribution with 2nd order polynomial (function of height) as its mean

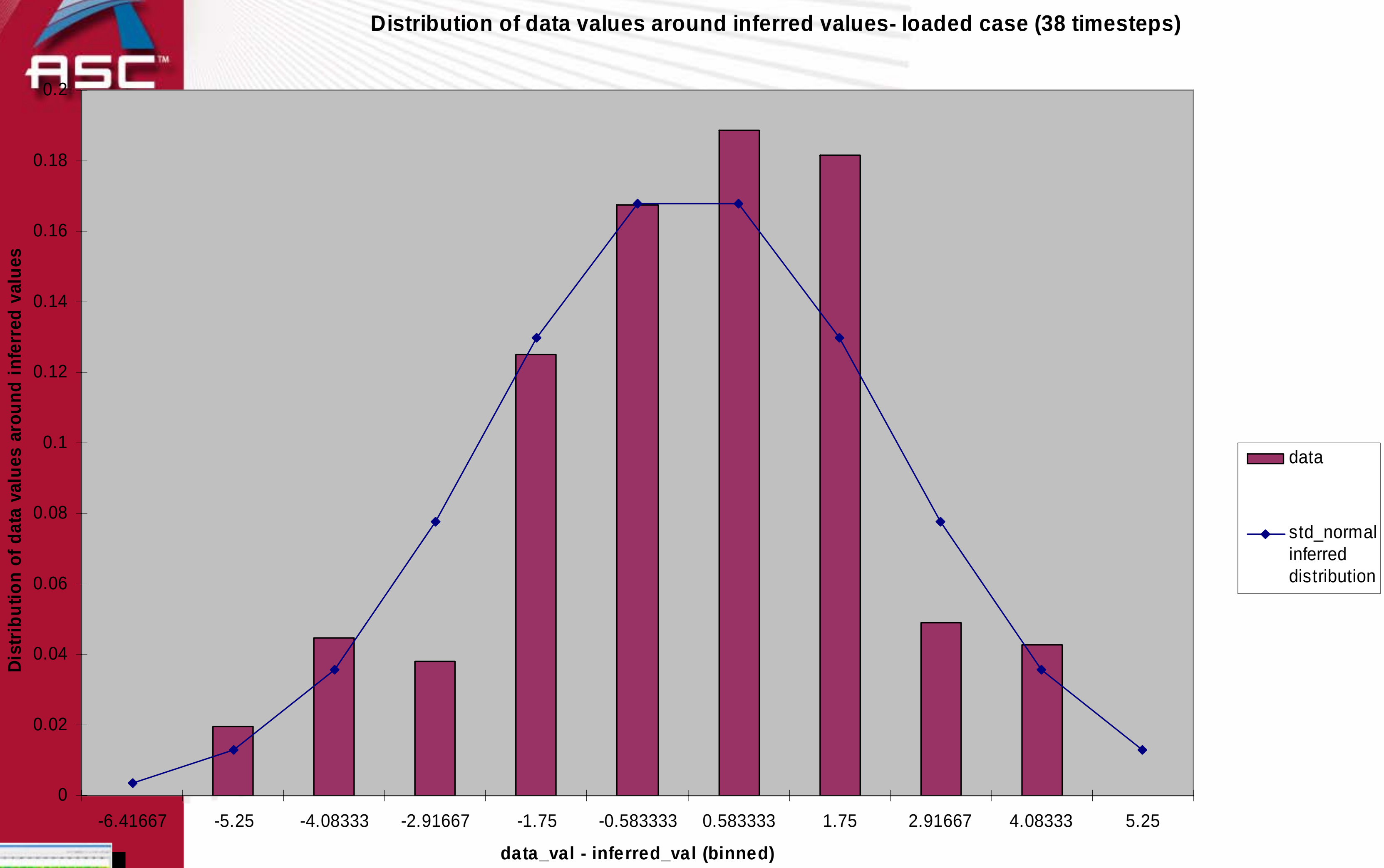


  - OVIS (screen data) converging likelihood, increasing confidence
- Abnormality Detection (Model comparison with the inferred curve)

Node 86

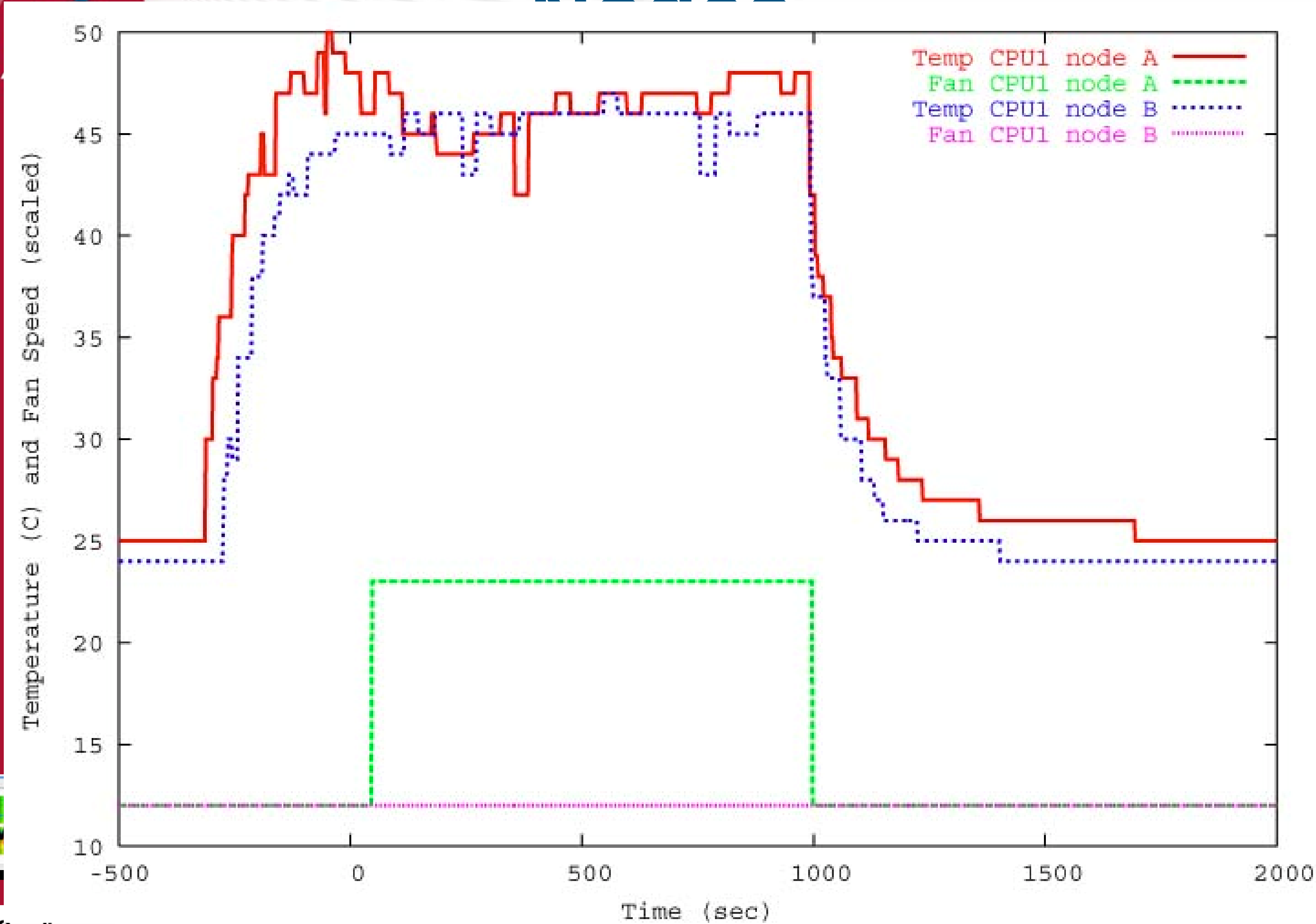
Distribution of data values around inferred values- idle case (1 timestep)





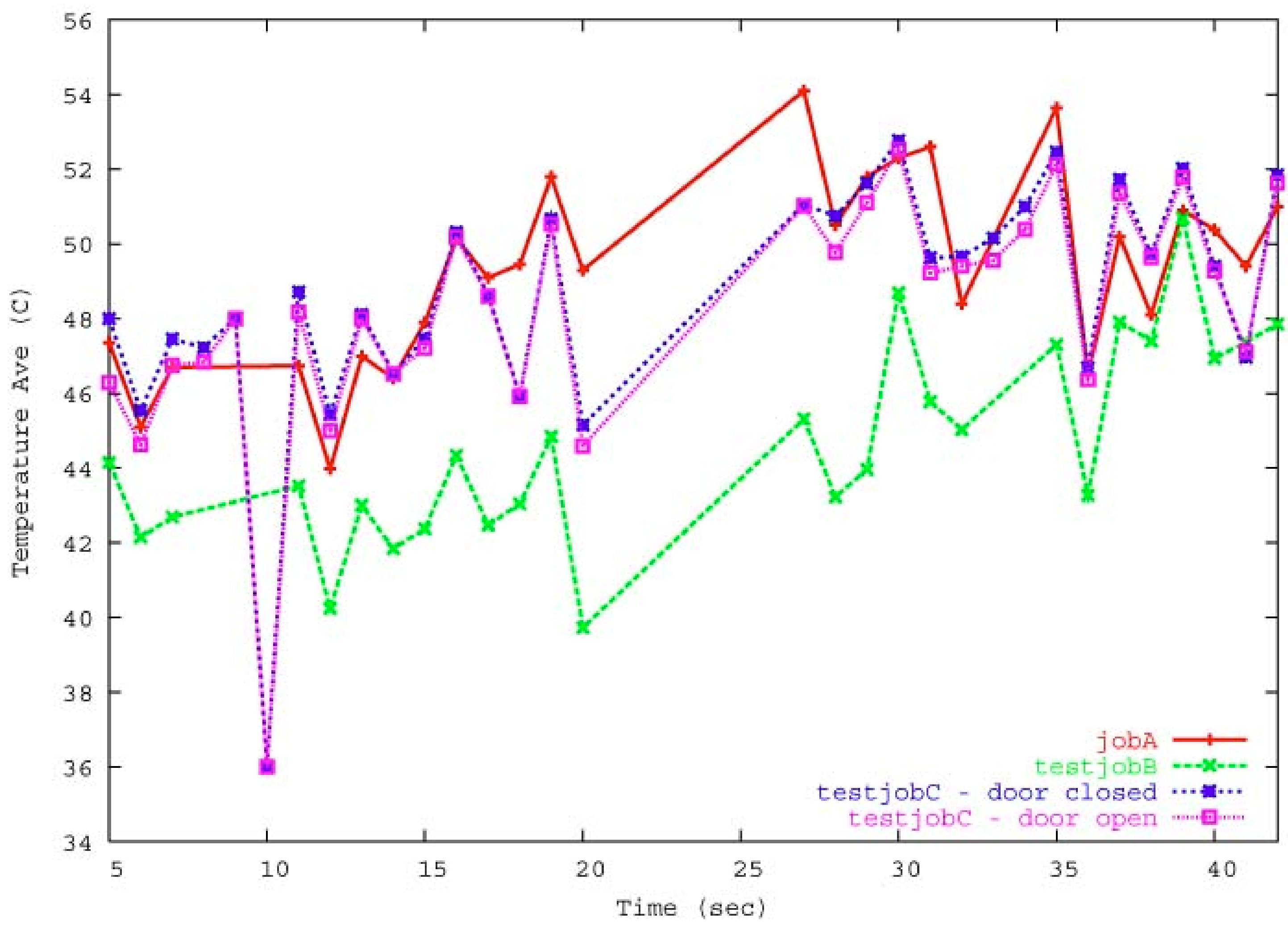
Distribution of data values around inferred values- loaded case (38 timesteps)
Distribution of data values around inferred values
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
-6.41667
-5.25
-4.08333
-2.91667
-1.75
-0.583333
0.583333
1.75
2.91667
4.08333
5.25
data_val - inferred_val (binned)
data
std_normal inferred distribution

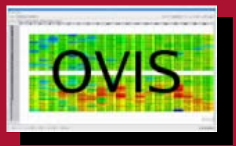
Natural Variations in Individual Nodes



Sandia National Laboratories

Characteristics Persist Across Different Job and Environmental Conditions





Sandia National Laboratories

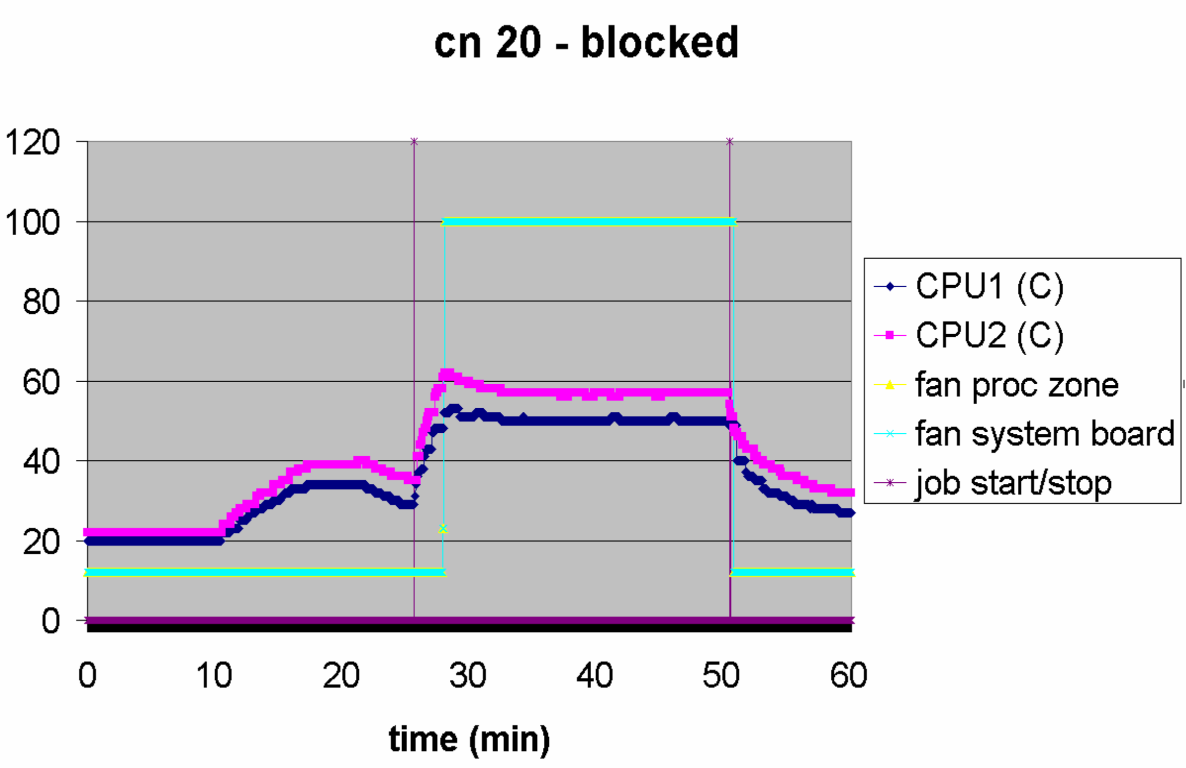
Notes

- Jags are the same in different circs, even if the curves (absolute and relative) arent. Individual variations overwhelm the conditions ?
- Research things:
  - how does model change as room temp change ? Does that differ from how model changes when nodes run harder (CPU utiliation)
  - Model – not over fitting

OVIS

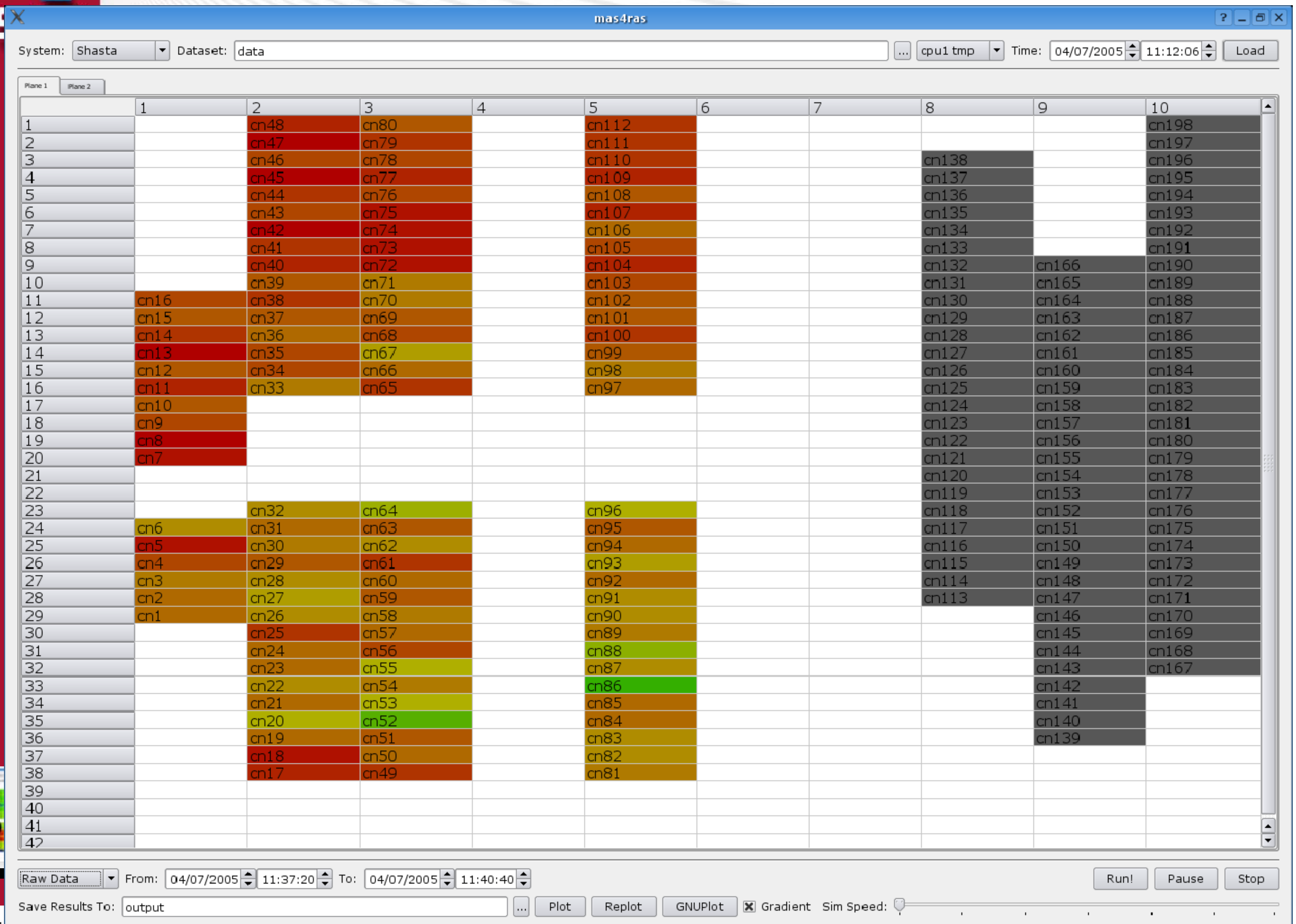
Sandia National Laboratories

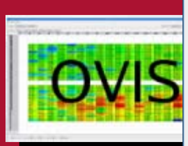
Recognizable Abnormalities (cont'd)



Partially blocked fan – fan speed increases in response to rising temperatures, while fans on other nodes remain at constant speed

Non-Uniform, Non-Linear Environmental Effects

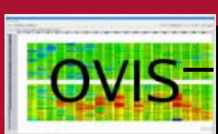




Sandia National Laboratories

Analysis and Monitoring

- **Analysis research**
  - **Understanding the system, changing likelihood's, other machine learning issues (clustering, adaptive methods etc.), time dependencies, dynamic issues, plotting capabilities, other variables (e.g., memory error rates).**
  - **Many different jobs – what is a statistically significant set?**
  - **Failure prediction and interactions with resource management**
- **Monitoring -- System administrators**
  - **Initial set up (HP) and effects of changes**
  - **Silently models, monitors and provides descriptive colour maps.**
  - OVIS– **Currently addressing a subset of the problem – thermal issues/airflow/cooling**



Sandia National Laboratories

Needs associated with RAS

- Advance warning of impending faults
- Fault detection
- Diagnostic help to identify actual problem(s)
- Interface to batch scheduler
- Interface to trigger check-point

Real-Time Requirements for Statistical Approach

- *Data sampling intervals short relative to change in monitored variables*
- *Data processing must keep pace with sampling*
- *Only do comparisons when data is stable*

Probabilistic Characterization Using Bayesian Inference

The keystone of our approach is Bayesian inference.

Reminder: Bayes' Theorem:

$$P(X|D,M) = P(D|X,M) \times P(X|M) / P(D|M)$$

or less formally: posterior = likelihood x prior / evidence

- X is a set of model parameters to be inferred (e.g., polynomial coefficients and variance in the model above);
- D is the data, i.e., measurements of the variables that are present in the model.
- M is the probabilistic model (e.g., temperature is distributed as a Gaussian r.v. whose parameters have a polynomial dependence on height in the cluster;

OVIS

Sandia National Laboratories

Abnormality Detection Using the Inferred Model

- **After model inference has been done (either with training data or "live" data), we have a stochastic model whose parameters optimally fit the data.**
  - **E.g., for idle rack 3:**
    - **$T \sim N(0.005\ h^2 - 0.1\ h + 23, 1.5)$**
  - Outliers can be defined automatically based on user-defined thresholds
    - **$RP(\{h=10, T=23\} \mid \{0.005, -0.1, 23, 1.5\}, M) \approx 95\%$**
    - **$RP(\{h=10, T=25\} \mid \{0.005, -0.1, 23, 1.5\}, M) \approx 25\%$**

OVIS

Sandia National Laboratories

Approach Summary

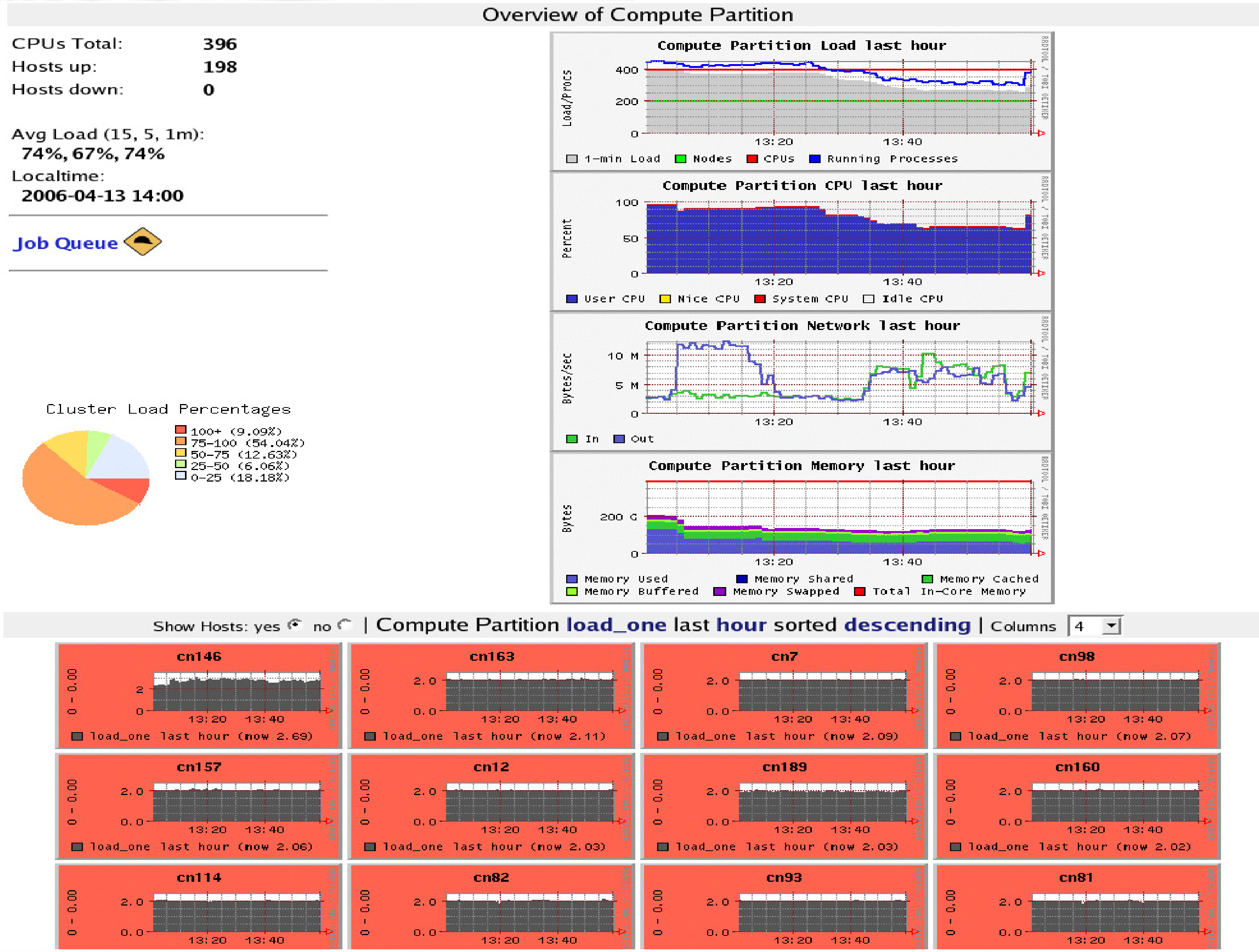
- Use statistical approach for probabilistic modeling using Bayesian inference
- Use these probabilistic characterizations to identify outliers, hot spots, etc.
- Use the constant influx of incoming data to update and improve the existing probabilistic models (machine learning).
- Scope the problem (thermal issues) for technique development then expand (memory errors, fans, voltages, cross-correlations)

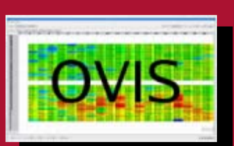
How Is This Different and Where Is the Intelligence?

- **Thresholds in terms of probability rather than raw data**
- **Determined by statistical processing on real data**
- **Numerical threshold values are *learned* and dynamic**
  – Will change in response to aging, environmental effects, etc.
- **Environmental modeling**

Sandia National Laboratories

Ganglia on Shasta





Temp vs. Time Plots on Shasta Cluster