Understanding Large Scale HPC Systems Through Scalable Monitoring and Analysis

Ann Gentile

EGI Technical Forum Sept 2010

OVIS Team: Jim Brandt, Frank Chen, Vincent DeSapio, Ann Gentile, Jackson Mayo, Philippe Pebay, Diana Roe, David Thompson, and Matthew Wong

http://ovis.ca.sandia.gov

Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Motivation

- **Large and complex systems**
  - Thousands of nodes
  - Tens of thousands of processing elements
    - Shared subsystems (memory, network, storage)
  - Both custom and commodity interconnects
    - Myrinet, Infiniband, 10 Gigabit Ethernet, Cray SeaStar, Cray Gemini,…
- **Difficult to understand and troubleshoot**
  - Hardware
  - System software
  - User applications
- **Resource-Aware computing could improve**
  - System throughput
  - Application performance

System understanding requires appropriate instrumentation and interpretation

ASC

Sandia National Laboratories

System Instrumentation

Wealth of information available:

- **Some typical platform related information and update frequencies**
  - Voltages, temperatures, fan speeds – LM sensors (query based but seems to have ~2 sec refresh period)
    - Kernel metrics such as CPU and memory utilization -- /proc (query based)
    - Could use more – example: power supply efficiency
    - Interfere with user applications
  - Log files – syslog, console logs (whenever appropriate event occurs + time lag)
  - Resource Manager information – user, app, resources, terminal state (beginning and end of job)
  - Network and storage infrastructure
- **Difficult to customize because typically systems are comprised of COTS servers**
  - Already instrumented
  - Can sometimes pay for access to advanced monitoring features (IPMI)
- **Environmental monitors**
  - Temp, humidity, power, cooling units, etc.

What is the relevant data? What data could be relevant given the appropriate analysis?

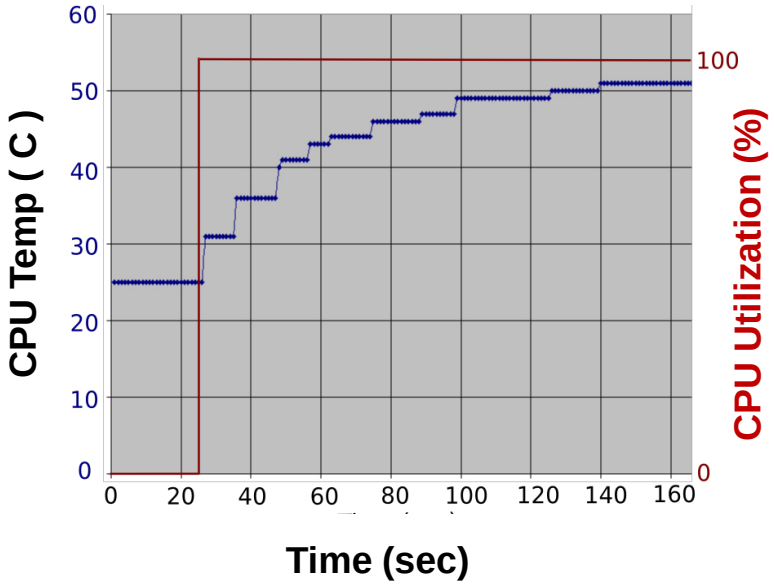
ASC™

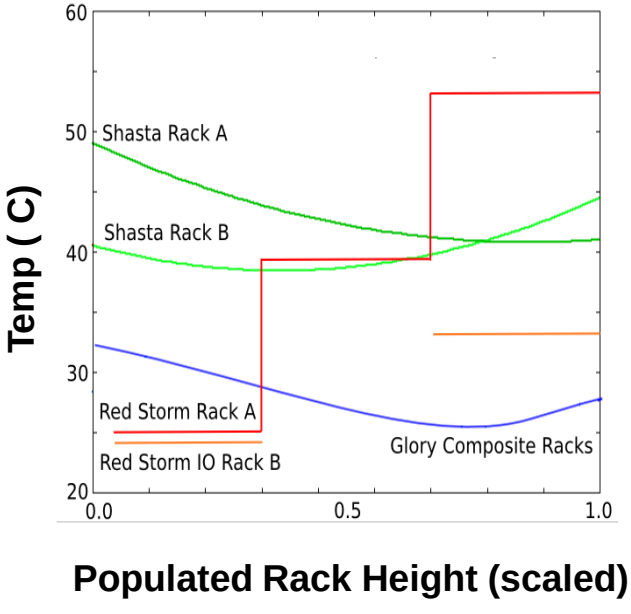
Sandia National Laboratories

Difficult to Extract Meaningful Information

Data Features and Analysis Complications

- Numeric and textual data
- Geographical and temporal variations (e.g. temperature, power)
- Multiple applications
- Components with both inertial and non inertial observables
- Meaningful window sizes depend on characteristics of observables
- Time skew (e.g. measurement, cause and effect)
- Large scale

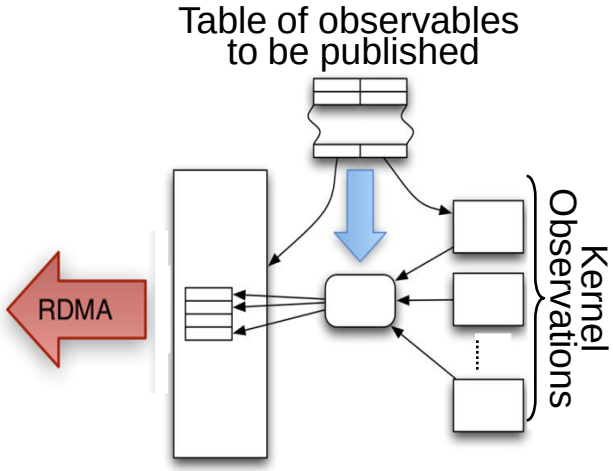


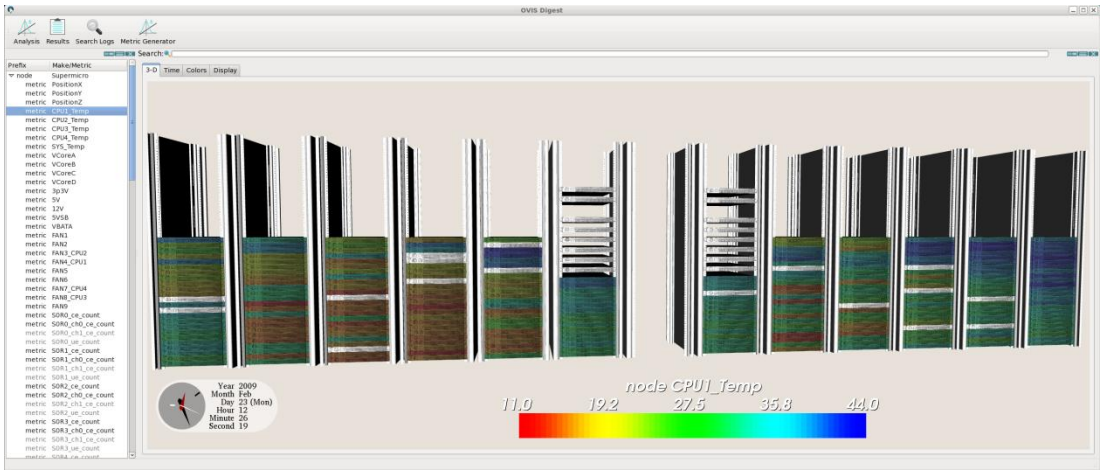


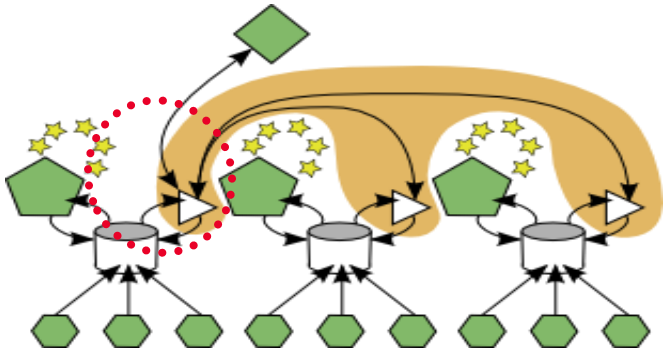
Sandia National Laboratories

OVIS: A Scalable Tool Targeted at Gaining System Understanding

- Lightweight information harvesting
- Information aggregation
- Information analysis
- Visualization



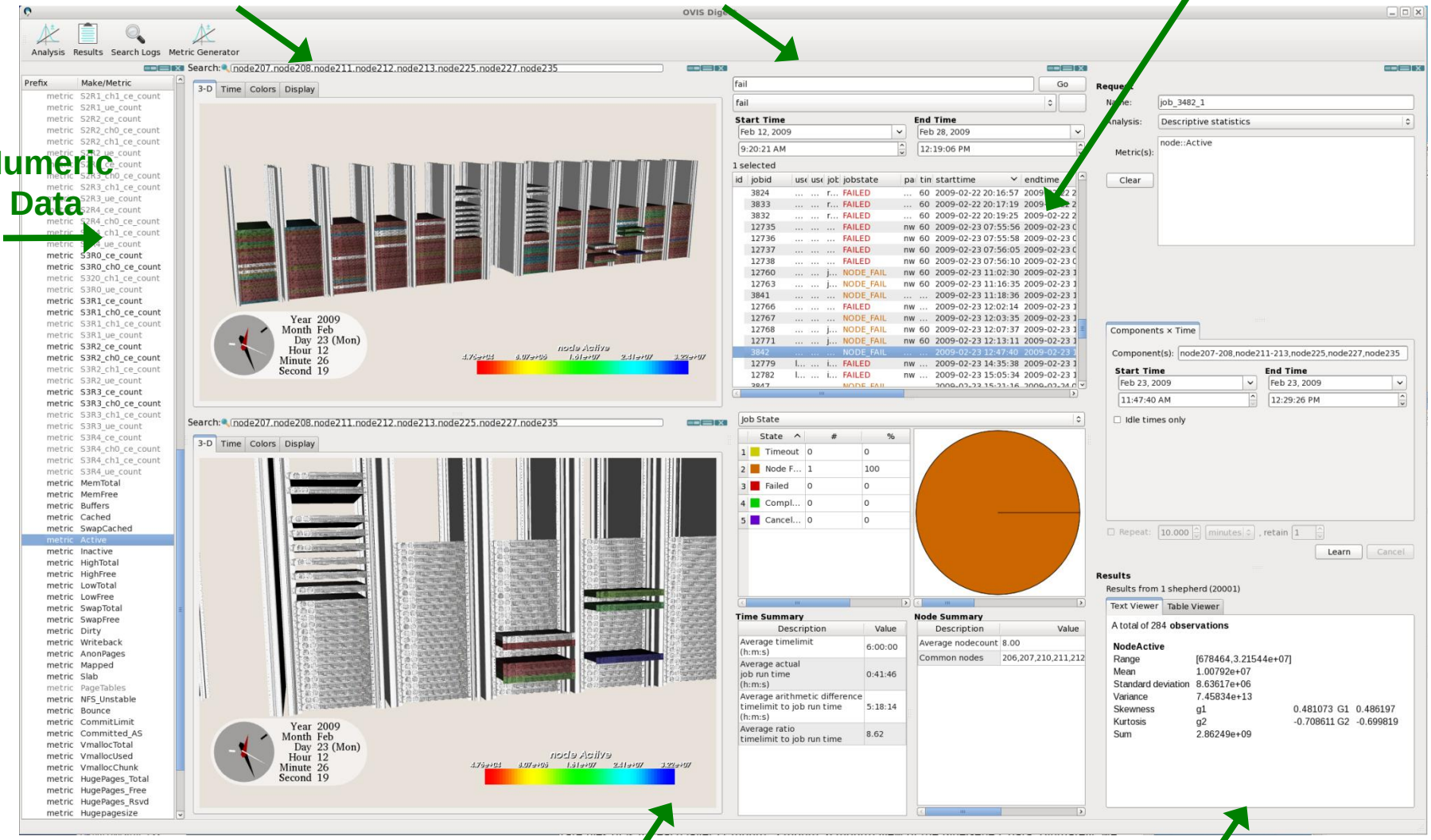




ASC

Sandia National Laboratories

Data Exploration Toward System Understanding

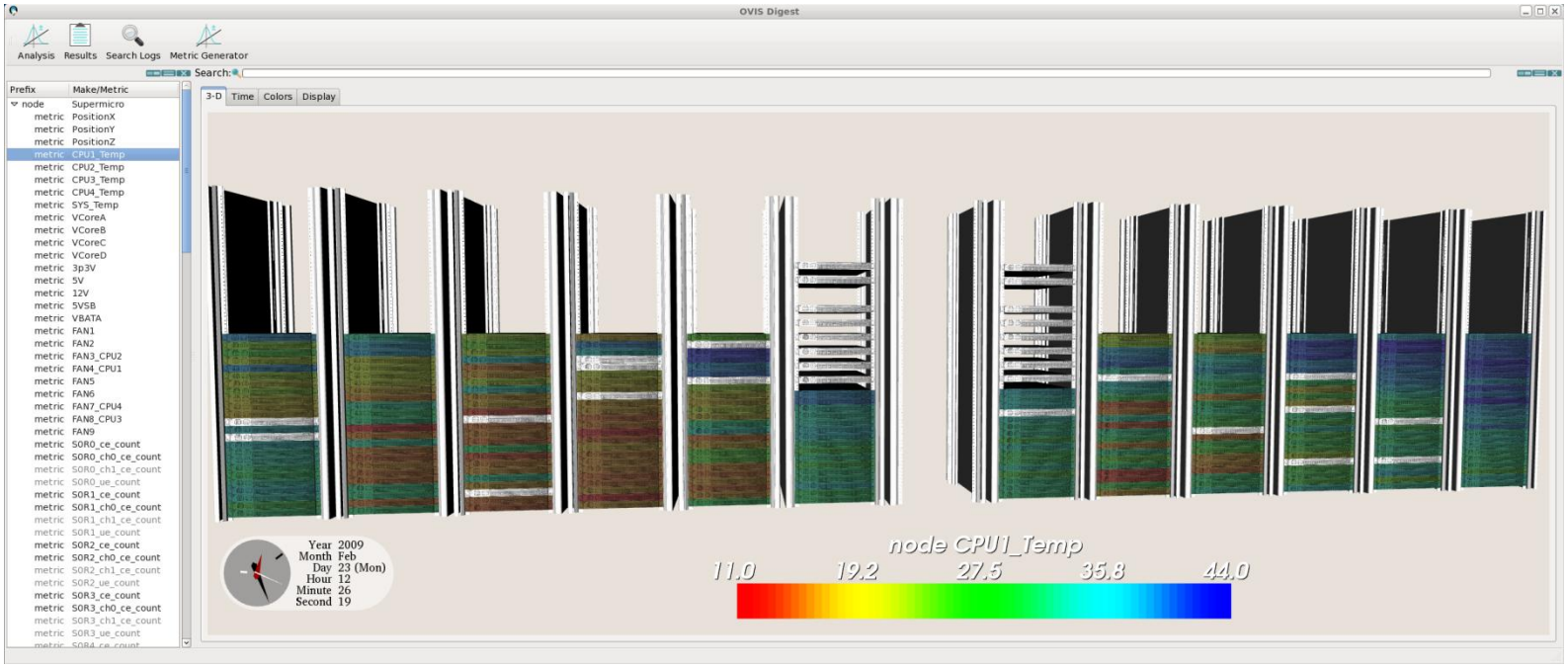


Sandia National Laboratories

Visualization

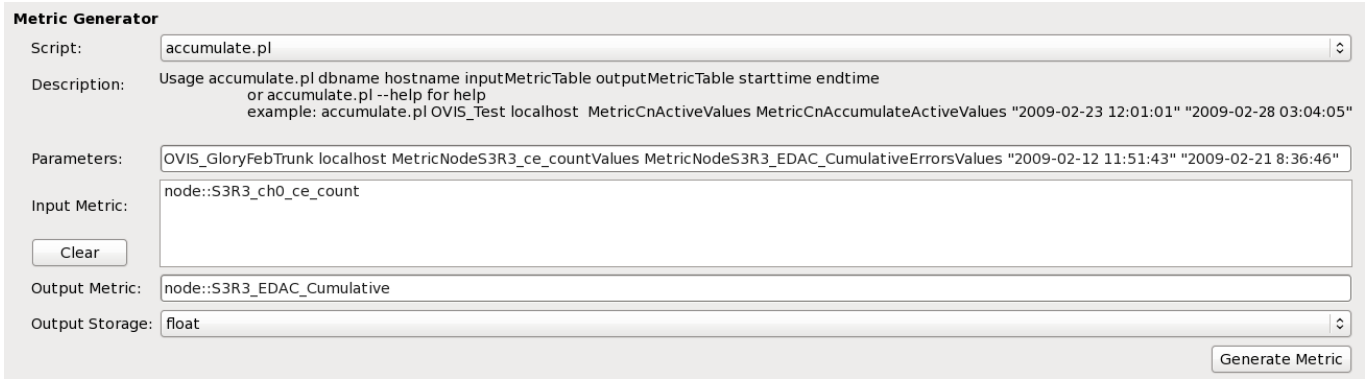
Spatio-temporal display can give intuition

- **Raw data**
  - Temperature data – exhaust recirculation leading to thermal issues
  - Voltage data – Power distribution/supply problems?
  - Network data – Bottlenecks and hot spots
- **Derived data and analysis results….**

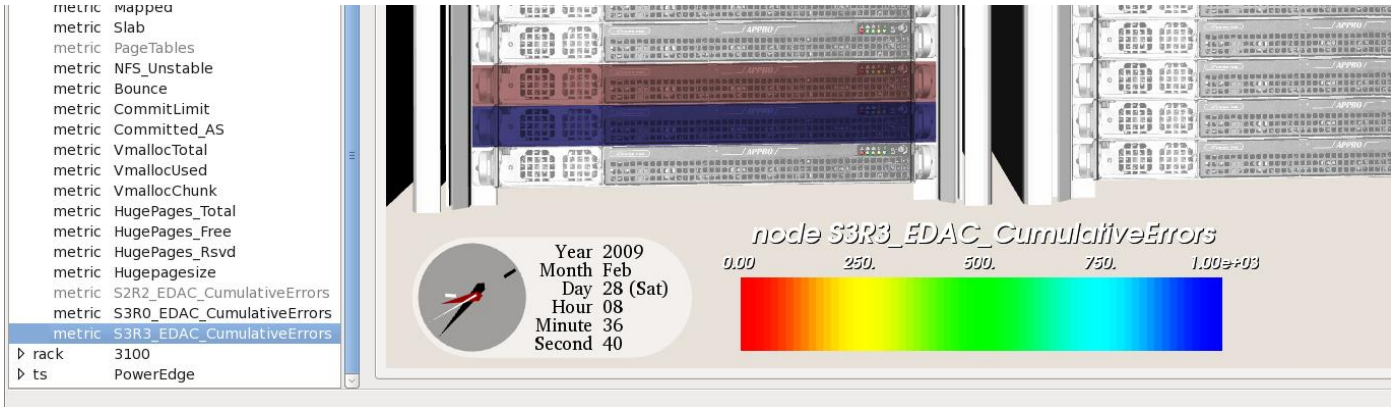


Derived Metrics via User Scripts

- **Custom Scripts:**
  - Read raw data from db (e.g., error counts)
  - Calculate derived quantity (accumulated error counts over a job)
  - Write derived data to db



- **Derived metric available for analysis, visualization**
- **Rapid prototyping**

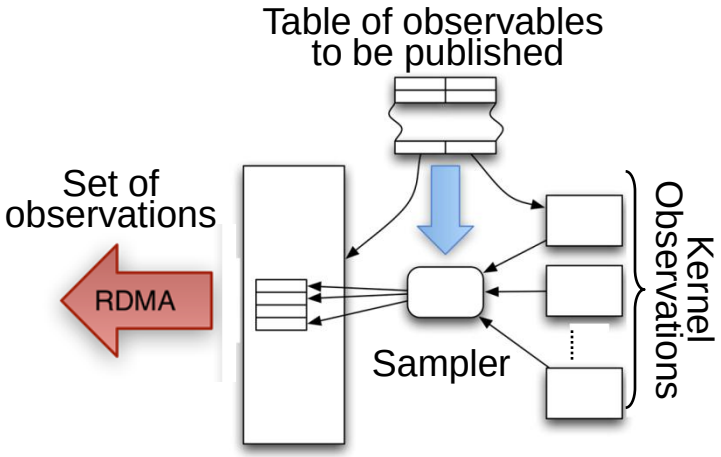


Data Collection Issues

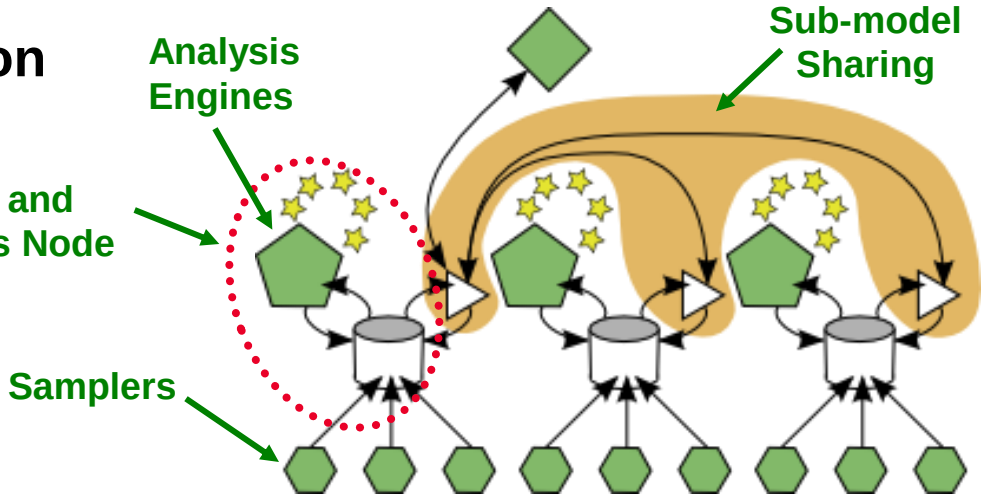
- **Sources and interference**
  - Out of band doesn't interfere with computational resources but may with communications – separate network
  - Collection of kernel metrics (CPU utilization, memory utilization, cache misses, etc.) can compete with application for computational and memory resources
  - Distributed
- **Scalability**
  - Volume
  - Bandwidth
  - Longevity
- **Performance**
  - Interference of collection and analysis
  - Currently write then read for analysis – will be changing to stream through analyses before write

Scalable Collection and Analysis of Observables

- **Data collection of kernel information for compute nodes**
  - Lightweight
  - Kernel code
    - Bound jitter/overhead by integrating with scheduler



- **Distributed data aggregation and analysis framework**
  - Parallel storage
    - Long term
    - High frequency data collection
  - Parallel analysis
    - Distributed data
    - Scalable

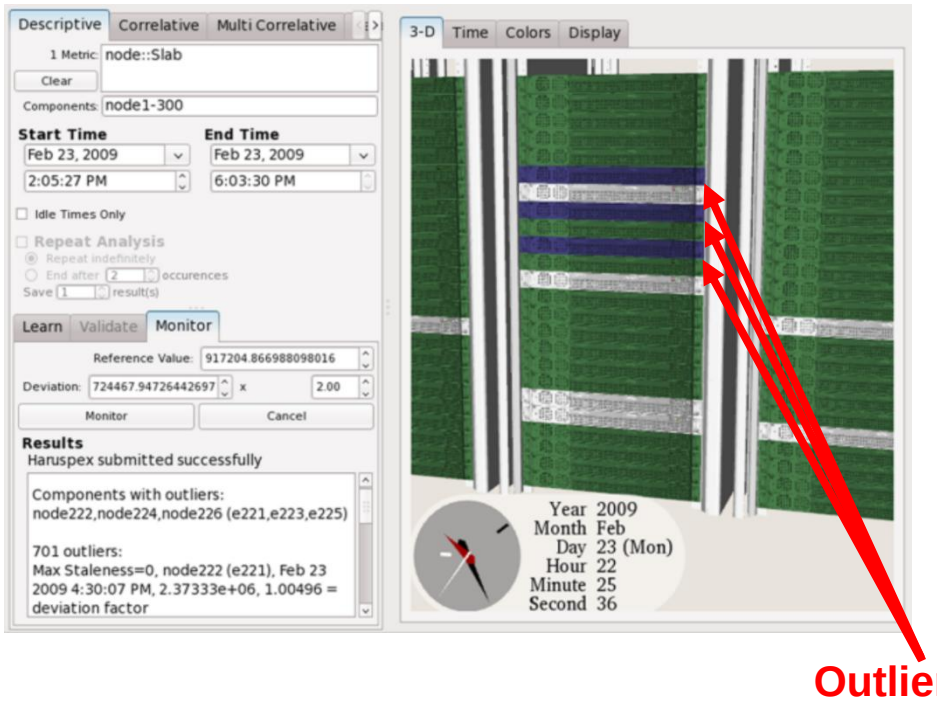


ASC

Sandia National Laboratories

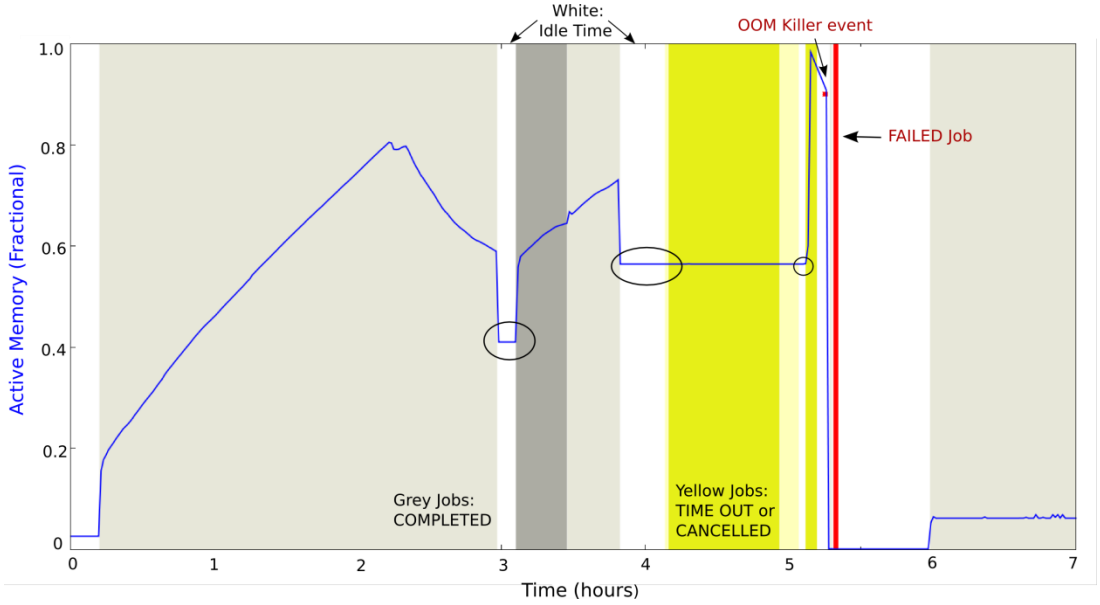
Data Analysis Beyond Custom Scripts

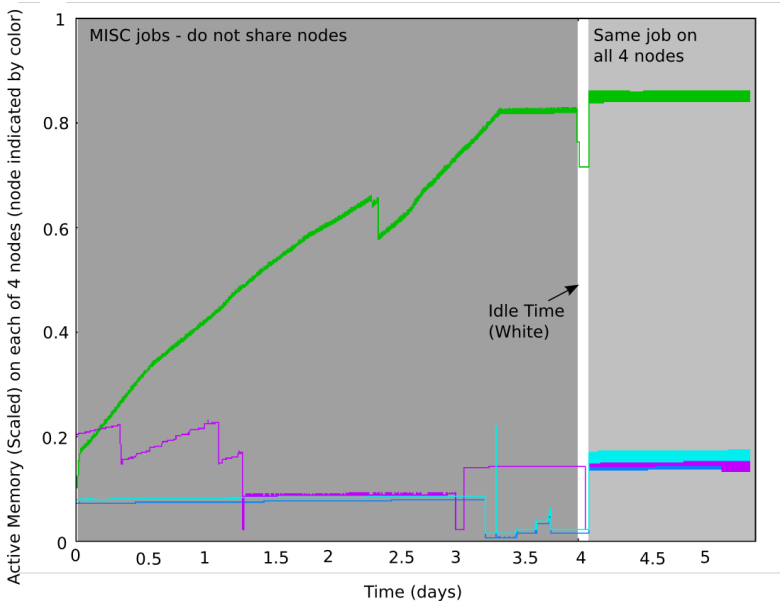
- **Analysis suite aimed at modeling and outlier detection**
  - Descriptive
  - Multi-variate correlation
  - Contingency
  - Time series (coming soon)
  - Wear rate based
  - Graph based
- **Raw and derived numeric data**
- **Text related – convert log file events into resource related numeric data (e.g. counting OOM events on a compute resource)**
- **Model drop (to be reintegrated)**
  - “Outliers” colored differently
  - Combined with pop-out-grey-out can show location of job related outliers



Troubleshooting and Research Areas

Detectable and Actionable Failure Indicator on TLCC





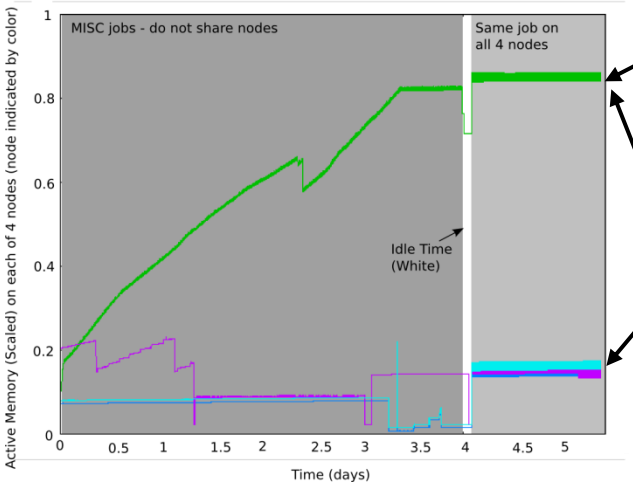
- Algorithmically detectable Failure indicator: Abnormally high memory utilization **during idle** or across nodes **sharing a job**
- Actionable: indicator detectable > 2 hours before actual failure (time is dependent on level and memory requirements of subsequent jobs)

Sandia National Laboratories

Detection of Both Dangerous and Anomalous Condition Invokes Mitigating Response

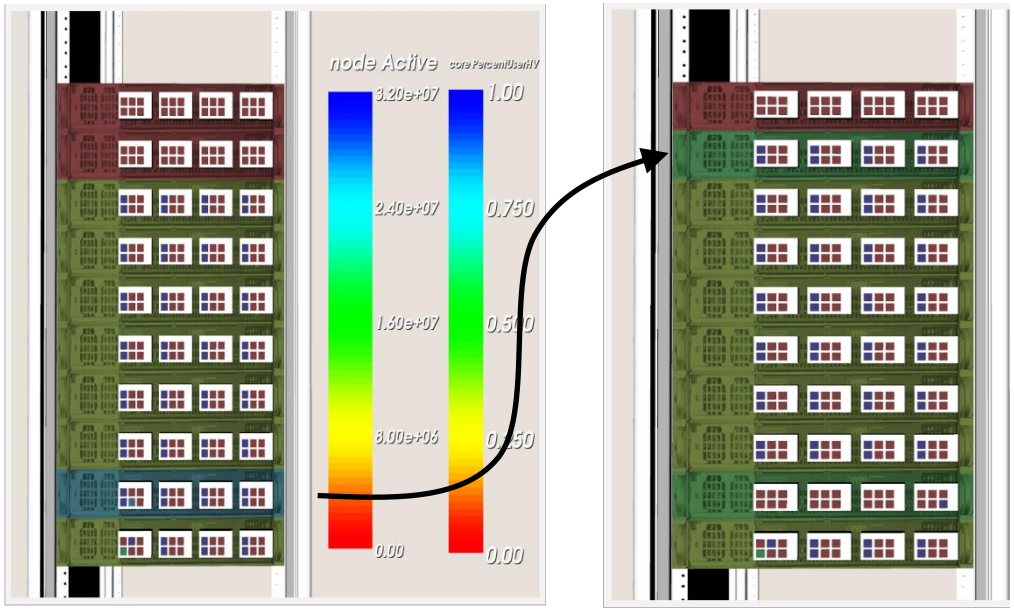
Impending Failure (Memory)

- Threshold checking -- positive
- Anomaly checking – positive
- Additional resource allocation in coordination with RM
- Notify application
- Migration
- Notify RM



Dangerous

Anomalous



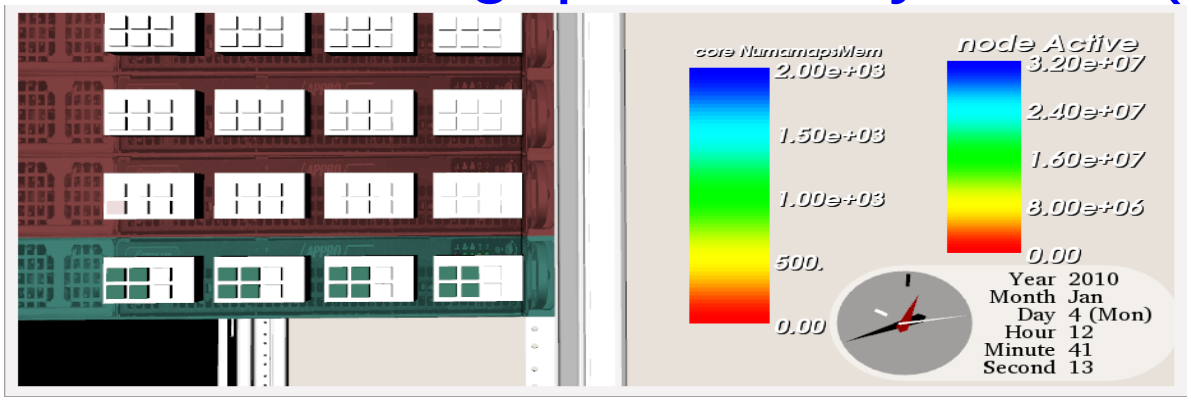
Anomalous condition detected

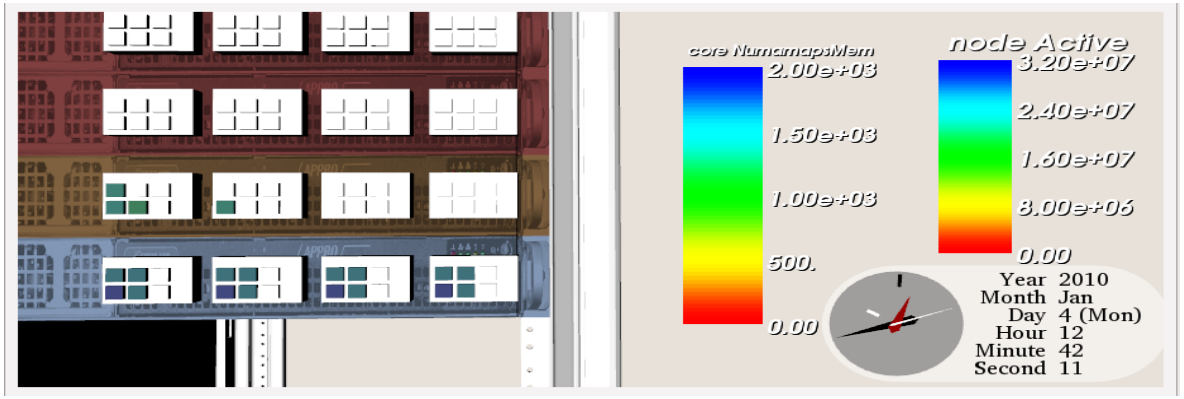
Anomalous condition mitigated via dynamic resource allocation and migration

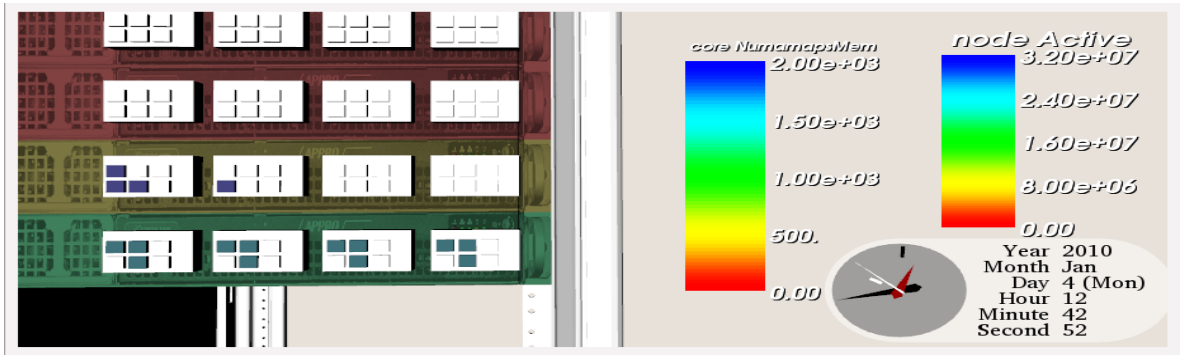
Combining Data for Increased System Understanding and Efficiency

- Combine job requirements, topology, and actual resource utilization information
  - Minimize resource contention
  - Maximize performance
  - Maximize system efficiency
    - Overlap of jobs on resources can minimize overall wall time for completion of a set of jobs (Mora, HPI-DC @ Cluster 2009)

Resource Analysis Used for Improving Thoughput and Dynamic (Re-)Allocation







- Intentionally oversubscribe based on known memory profile
  - Start job earlier than resource would otherwise allow
  - 1 process/core
- Trigger resource reallocation (migration) by run-time detection of oversubscription of memory

Top – triggering
Middle – during
Bottom – after

Sandia National Laboratories

- **OVIS Open source release at http://ovis.ca.sandia.gov**
- **Coming soon: OVIS 3.0 Official release with GIT repository**
- **Interested in partnerships and co-developers**

Questions???

ASC

Sandia National Laboratories

Abstract

- As HPC systems grow in size and complexity, diagnosing problems and understanding system behavior, including failure modes, becomes increasingly difficult and time consuming. At Sandia National Laboratories we have developed a tool, OVIS, to facilitate large scale HPC system understanding. OVIS incorporates an intuitive graphical user interface, an extensive and extendable data analysis suite, and a 3-D visualization engine that allows visual inspection of both raw and derived data on a geometrically correct representation of a HPC system. This talk will cover system instrumentation, data collection (including log files and the complications of meaningful parsing), analysis, visualization of both raw and derived information, and how data can be combined to increase system understanding and efficiency.

Data and Analysis

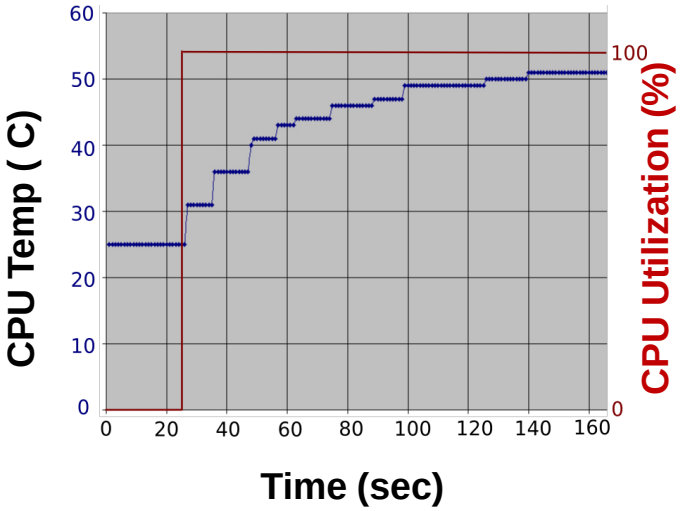
Data Features and Analysis Complications

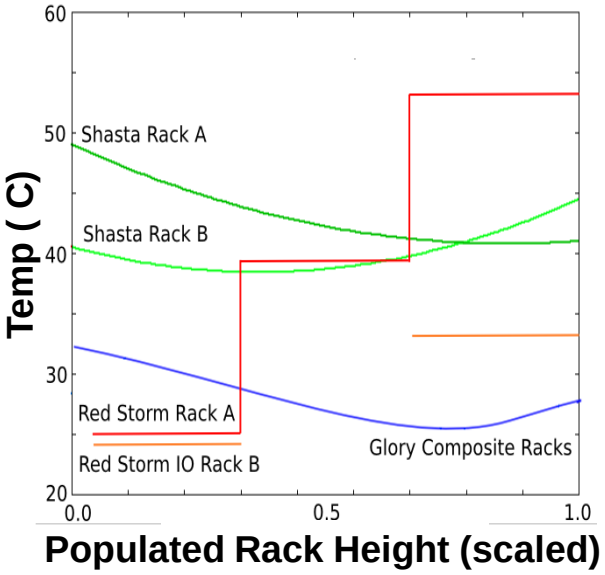
- Geographical and temporal variations (e.g. temperature, power)
- Multiple applications
- Components with both inertial and non inertial observables
- Time skew (e.g. measurement, cause and effect)
- Numeric and textual data
- Failure data is relatively sparse
- Large scale
- Meaningful window sizes depend on model type and observables

Parallel Analysis: Design Features

- Separate model learn from data compare for parallel scalability
- Design single pass numerically stable algorithms

Achieves optimal scalability for most analysis engines





Current analysis suite

- Descriptive
- Multi-variate correlation
- Bi-variate Bayesian
- Contingency and Information
- Principal Component Analysis
- Time series
- K-means
- Multi-variate Failure statistics (currently serial)
- Graph (currently serial)