

Continuous System Telemetry Harness

*Achieving True Autonomic Computing, Self
Healing Qualities, and Proactive Fault
Avoidance Through Telemetry*

Speaker: Kenny Gross Scalable Systems Group

Addn'l Team Developers:

Keith Whisnant, Aleksey Urmanov,
Kalyan Valdyanathan, Sajjit Thampy

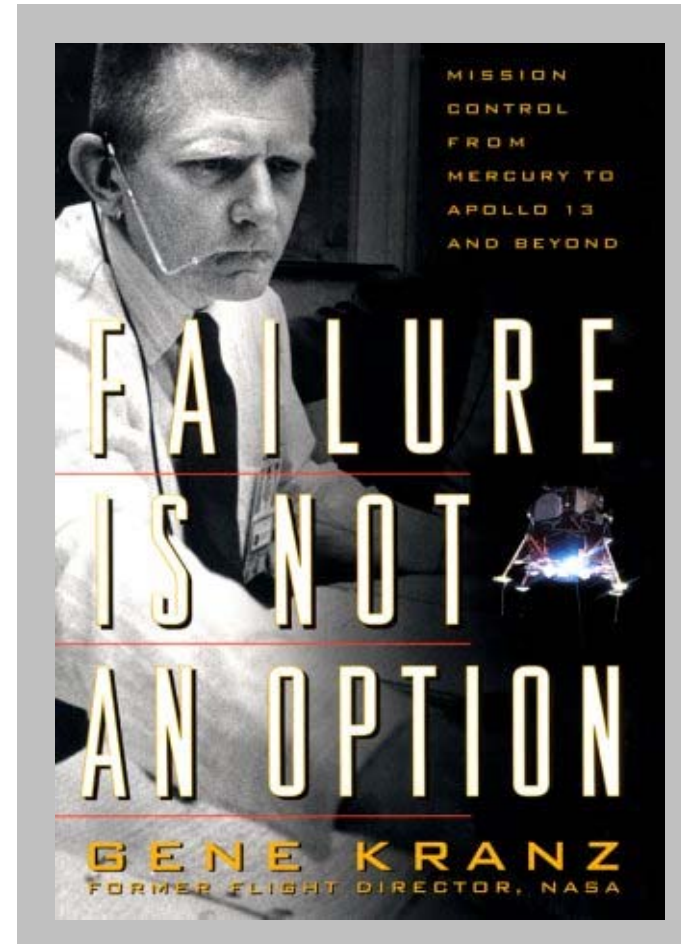


**2004
Sun Labs
Open House**

Why is Availability Important?

In today's internet-based computing model ...

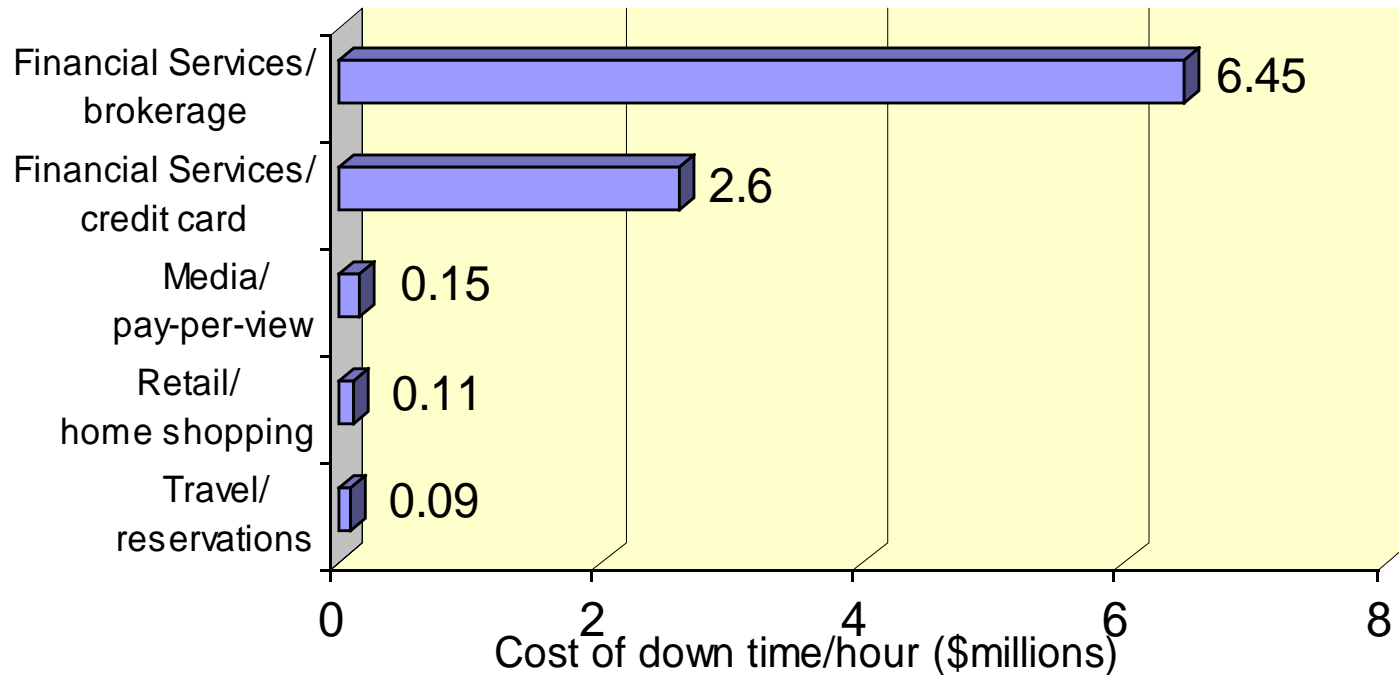
- To Customers
 - Downtime very costly
 - Measured by lost sales, reduced productivity, damaged business reputation, diminished customer loyalty
- To Sun Microsystems
 - Means of differentiation
 - Key corporate initiative
 - Goal to provide continuous application access with predictable performance



Mission director, Apollo 13

Costs of downtime at eCommerce datacenters can be substantial

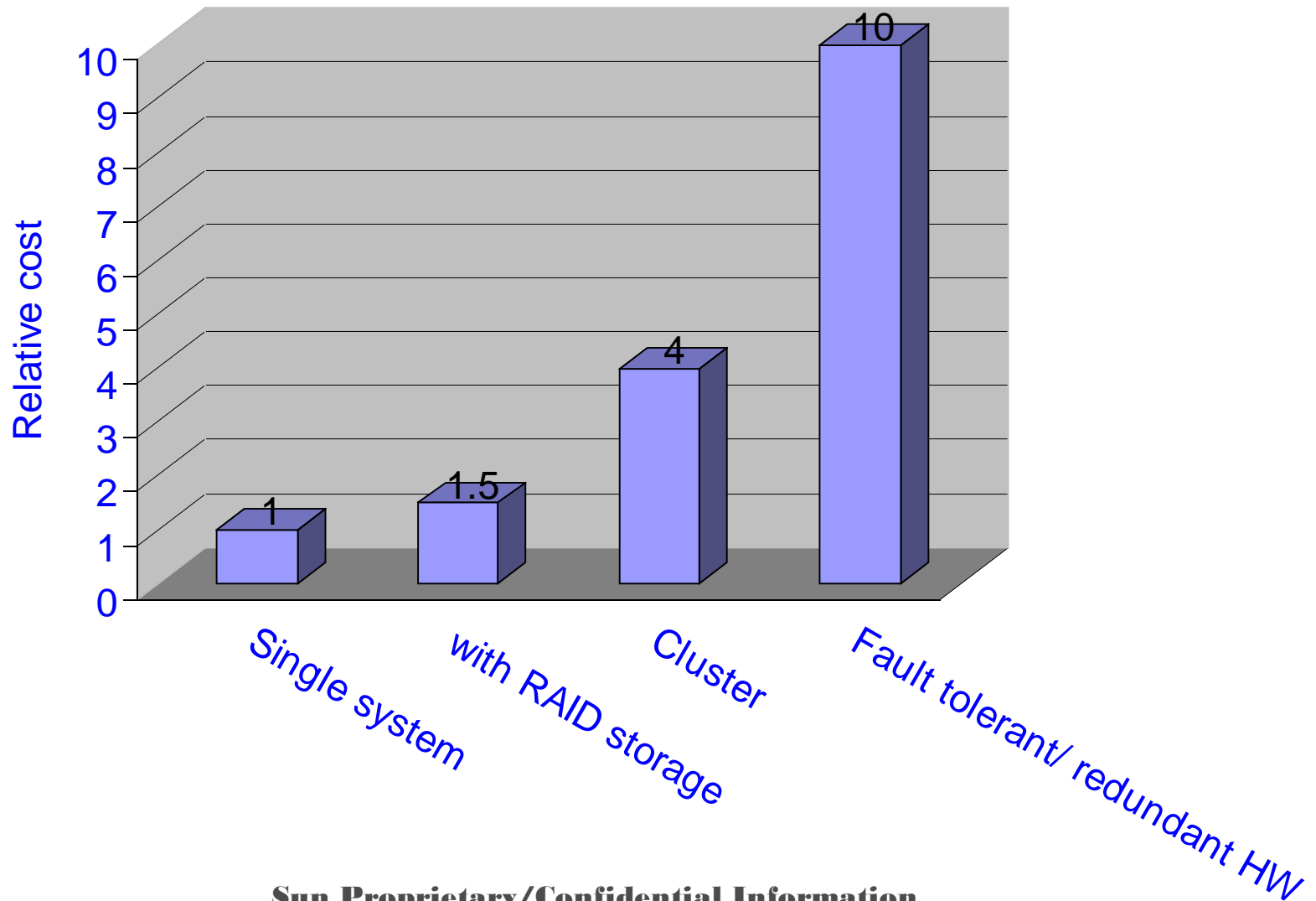
- Quantitative
 - Lost business, lost productivity









- Qualitative
 - Diminished reputation

Source: Hennessy and Patterson, Computer Architecture, Ch 1, Morgan & Kaufmann Publishers (2002).

Similarly, costs of avoiding downtime can be substantial



Evolution Vision

	Client-Server	3/N-Tier	Net Apps	Net Services	Next	After that
Catch Phrase	The Network Is the computer	Objects	Legacy to the Web	The Computer Is the Network	Network of embedded things	Network of Things
System Collections Components						
Scale	100s	1000s	1000000s	10000000s	100000000s	1000000000s
When/Peak	1984/1987	1990/1993	1996/1999	2001/2003	1998/2004	2004/2007
Leaf Protocol(s)	X	X	+HTTP (+JVM)	+XML, Portal	+RMI	Unknown
Directory(s)	NIS, NIS+	+CDS	+LDAP (*)	+UDDI	+Jini	+?
Session	RPC, XDR	+CORBA	+CORBA, RMI	+SOAP, XML	+RMI/Jini	+?
Schematic						

Correlation and Causality

- Correlation: Event A happens with Event B
 - Association
- Causality: Event A causes Event B
 - Association
 - Temporal Precedence
 - Non-spurious association
 - Mechanism

Key Enabler: Continuous System Telemetry

Advanced Pattern Recognition + Telemetry

Motivation:

Proactive fault monitoring (Detect incipient failures)

System Availability  Serviceability Costs 

Faster, more accurate Root Cause Analysis (RCA)...Each failure is used to build better future systems

Approach:

Adapt advanced statistical pattern recognition algorithm, MSET, that has been proven in a broad spectrum of safety-critical and mission-critical application domains.

BENEFITS:

- Enhanced end-to-end stack availability through predictive fault monitoring.
- Faster, more accurate RCA; Mitigation of No-Trouble-Founds (NTFs)
- Provides “Intelligent Agent” functionality when integrated with real time telemetry harness for monitored assets
- Provides “Software Aging and Rejuvenation” (SAR) capability for self-healing software systems
- Improved stability analysis, dynamic resource provisioning for networks of interacting dynamic elements
- Closed-loop autonomic control for future servers and networks

Advanced Pattern Recognition Tools for Ultrahigh-Reliability Surveillance

Sequential Probability Ratio Test (SPRT)

*For Stationary
Time Series*

- Advanced pattern recognition technique for high sensitivity, high reliability sensor and equipment operability surveillance.
- Developers proved in refereed journals that the SPRT provides the earliest mathematically possible annunciation of a subtle fault in noisy process variables.

*For Dynamic
Time Series*

Multivariate State Estimation Technique (MSET)

- Online model-based fault detection and identification.
- MSET predicts what each process should be on the basis of learned correlations among all process variables.
- MSET incorporates the SPRT to monitor the residuals between the actual observations and the estimates MSET predicts on the basis of the correlated variables.

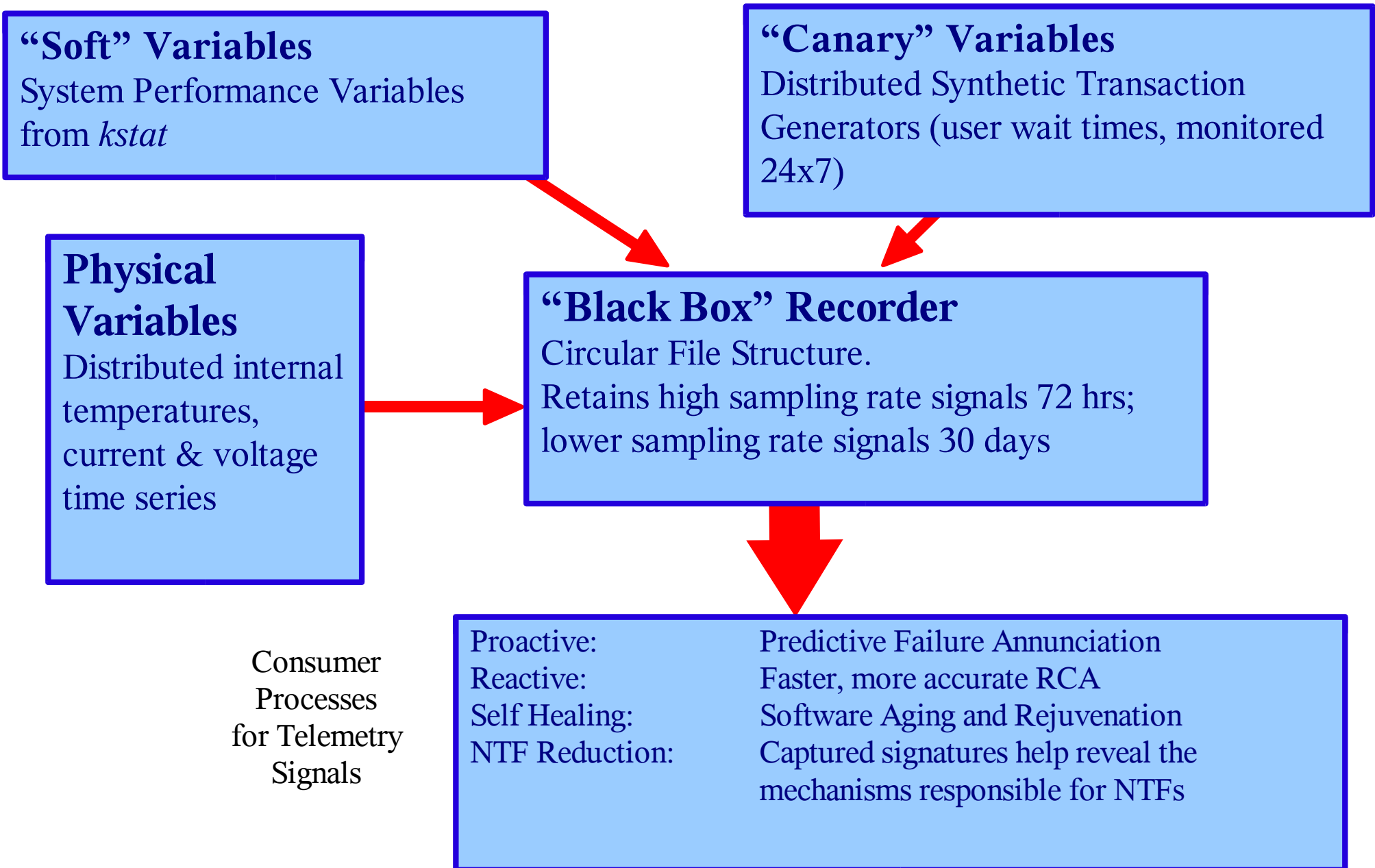
Continuous System Telemetry Harness (CSTH) Overview

- Software that works with the existing System Controller
- Collects valuable system data using existing physical sensors plus performance metrics
 - snapshots are available via showenv, CSTH provides continuous time series signals
- Signals provide both predictive and reactive failure information
 - preventative maintenance
 - reduce Failure Analysis time and costly NTFs (“Black Box Flight Recorder”)
 - enhance component reliability, system availability

CSTH Versions

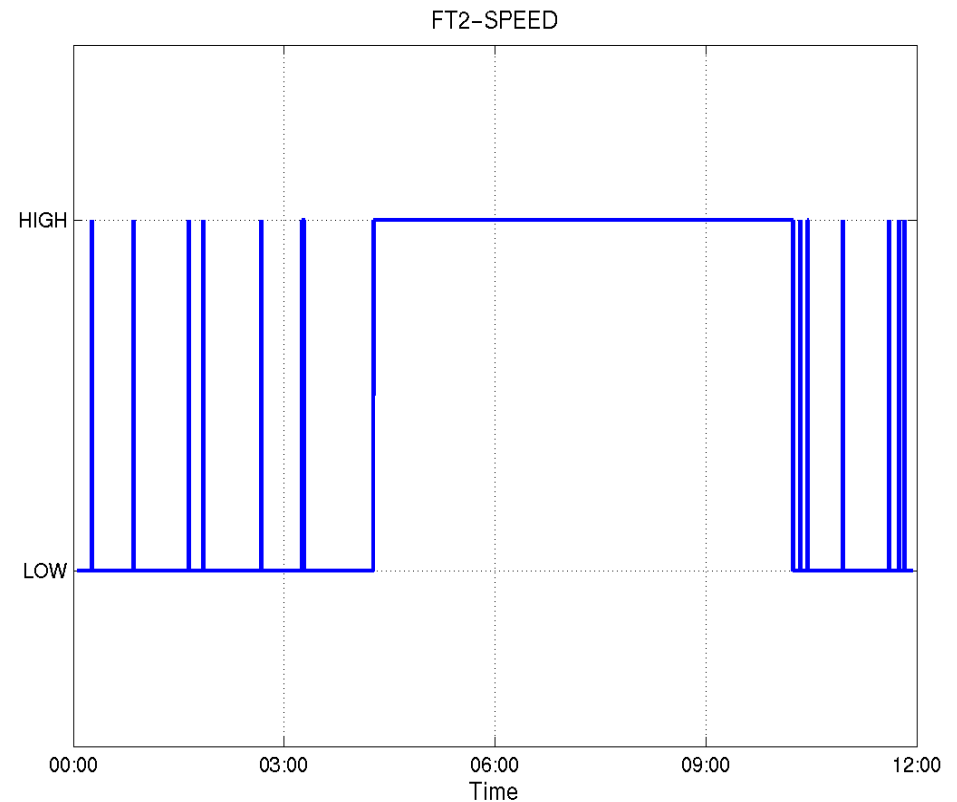
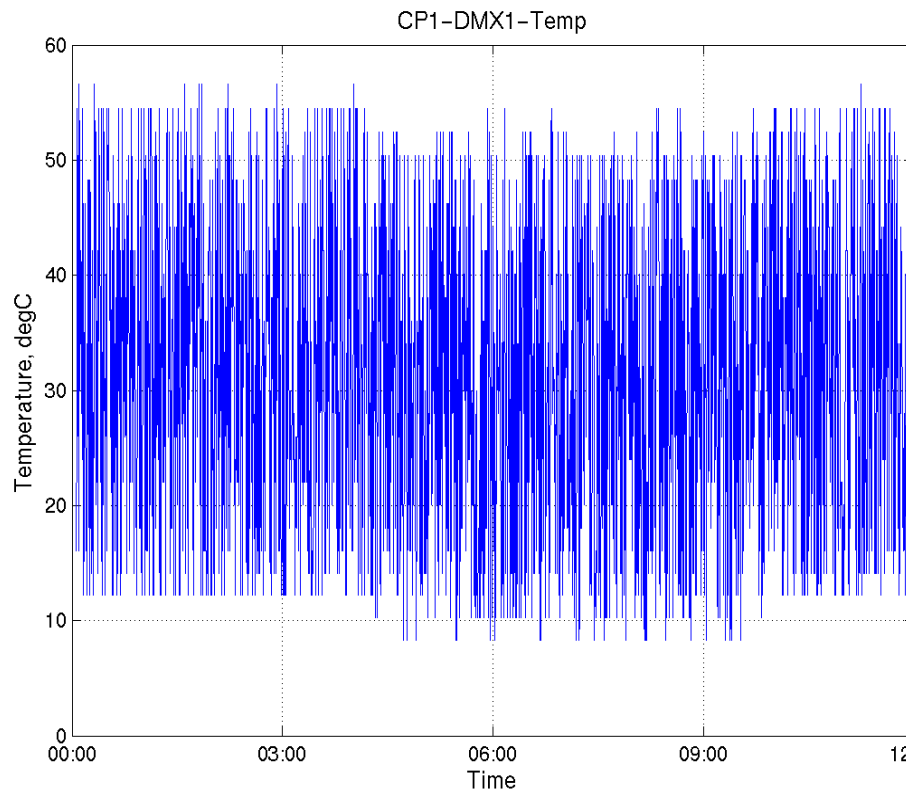
- Full CSTH
 - Over 1000 variables monitored
 - Successes at key customer sites, Manufacturing, Ongoing Reliability Testing, Sun's No-Trouble-Found Laboratory
- CSTH-Lite
 - Only processor voltage signals
 - Proactively detects symptoms of any types of socket/connector degradation
 - Enhances reliability of system boards in high-end and mid-range servers
 - Being productized throughout Sun's customer base for no charge (installs as a s/w patch; *requires no shutdown*)
- Real Time Power Harness (RTPH)
 - Accurately monitors total power consumption of machine during any types of load/memory/IO dynamics
 - Deployed in Mfg for system-board & server qualification
 - Helps achieve optimal energy utilization, reducing TCO for customer

Continuous System Telemetry Harness Structure



Early Telemetry Harness Success on StarCat Platform

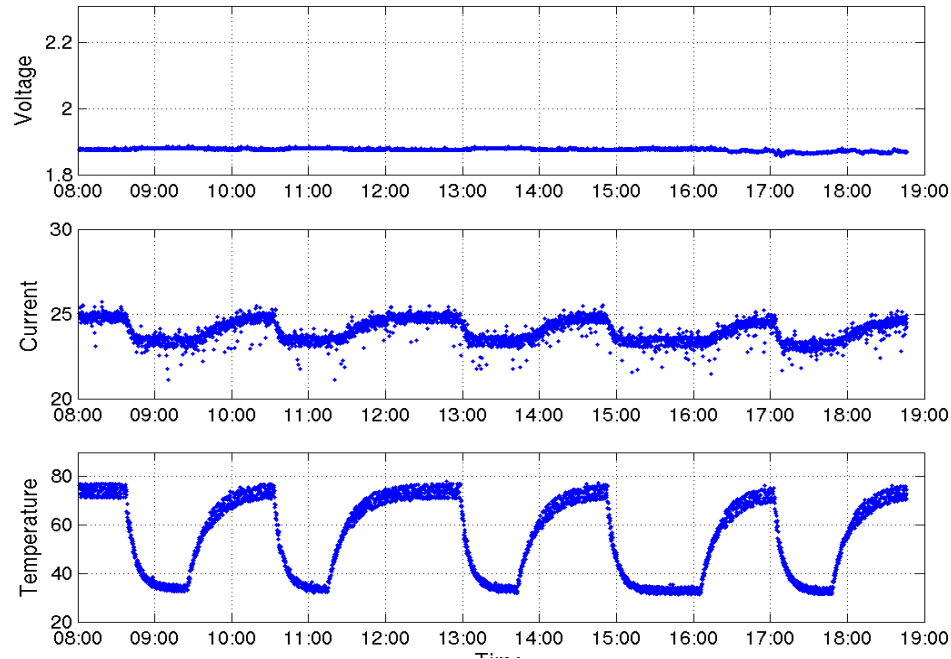
StarCat s/w bug identified during first week of telemetry testing of F15K.



StarCat environmental software bug reporting processor temperature improperly. Note wild swings between 12-50 deg C.

Faulty temperature values cause fans to continuously cycle between low and high (1 of 8 fans shown).

PS4 SN: 69061S000N05C Date: 16-Jan-2003

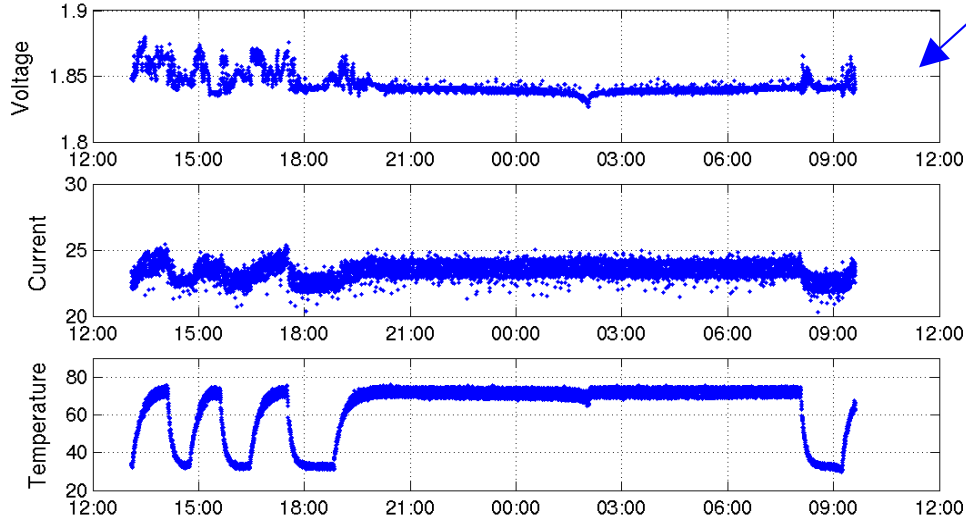


CSTH for NTF Mitigation
Ongoing thermal cycling experiments with NTF power supplies from E10Ks

Upper plots show flat voltage during temperature cycling with undegraded power supply.

Lower plots show voltage fluctuations from degrading power supply.

PS4 SN: 69061S000N05C Date: 22-Jan-2003



Voltage fluctuations from degrading power supplies are causing the system boards to throw out a number of failure messages, including:

- DTAG failures
- DTAG parity errors
- Ecache Failures
- Coherent processor errors
- UPA fatal error
- UPA parity error

What is MSET?

Multivariate State Estimation Technique

- Advanced pattern recognition system developed for 24x7 predictive fault monitoring in complex engineering systems like avionics and nuclear reactors
 - Continuous signal and sensor operability validation
 - Incipient fault annunciation on all monitored components
 - Extremely low probability of false alarms
- Award winning incipient fault surveillance system developed by Argonne National Laboratory
 - Capabilities surpass conventional pattern recognition approaches, including neural networks, in sensitivity, reliability, and computational efficiency

Telemetry + pattern recognition allow early prediction of failures

Traditional Monitoring: High/Low Thresholds

**Traditional monitoring techniques are limited in
their ability to identify failures proactively**



CSTH + MSET Pattern

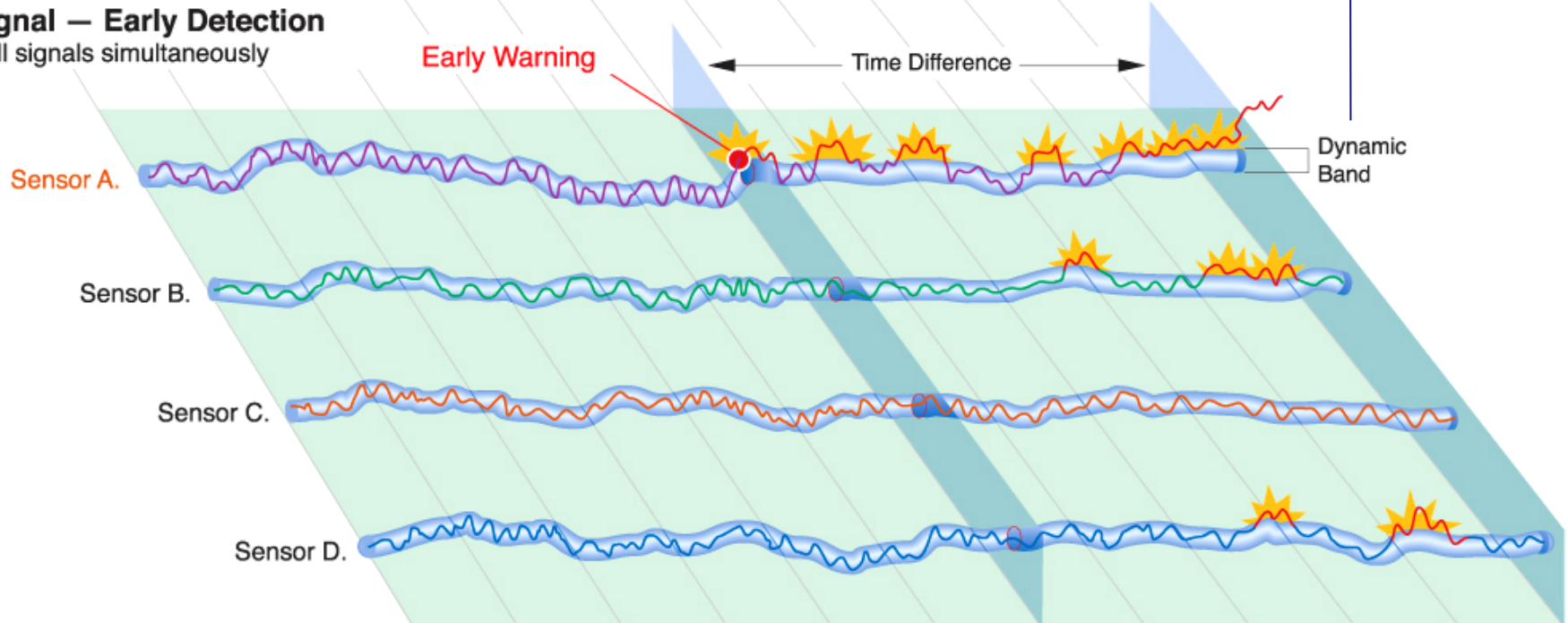
Recognition: Early Detection

Traditional Condition Monitoring
Monitors all signals separately



By creating a dynamic band around each sensor value in real time and correlating it to other sensor values, MSET is able to give early warning.

SmartSignal — Early Detection
Monitors all signals simultaneously



Thermal Anomalies Cause Customer Calls

Sun servers have high-temperature protection thresholds, but the thresholds are quite high (85-110 deg C)

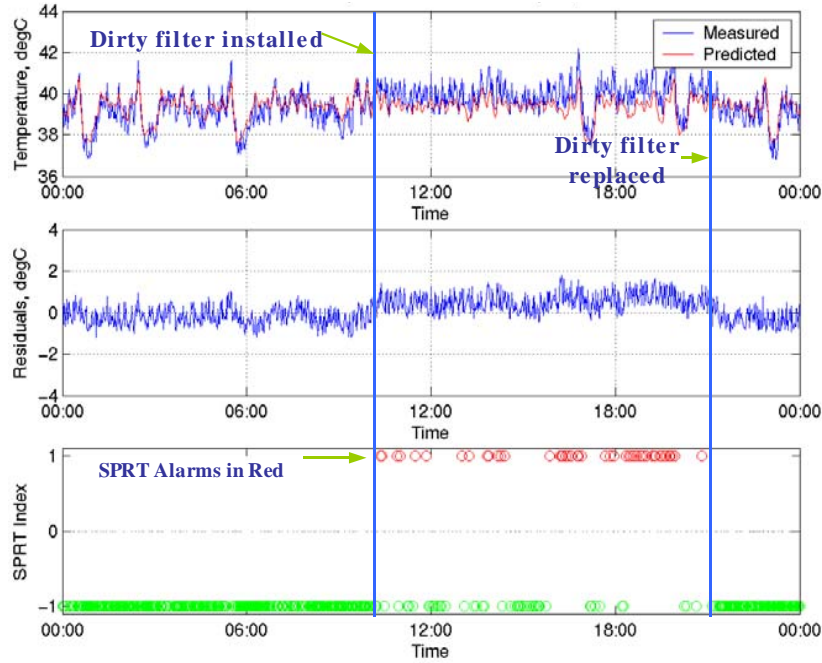
Many customer service calls originate from thermal problems that have much lower temperatures, but lead to long term reliability issues.

Examples:

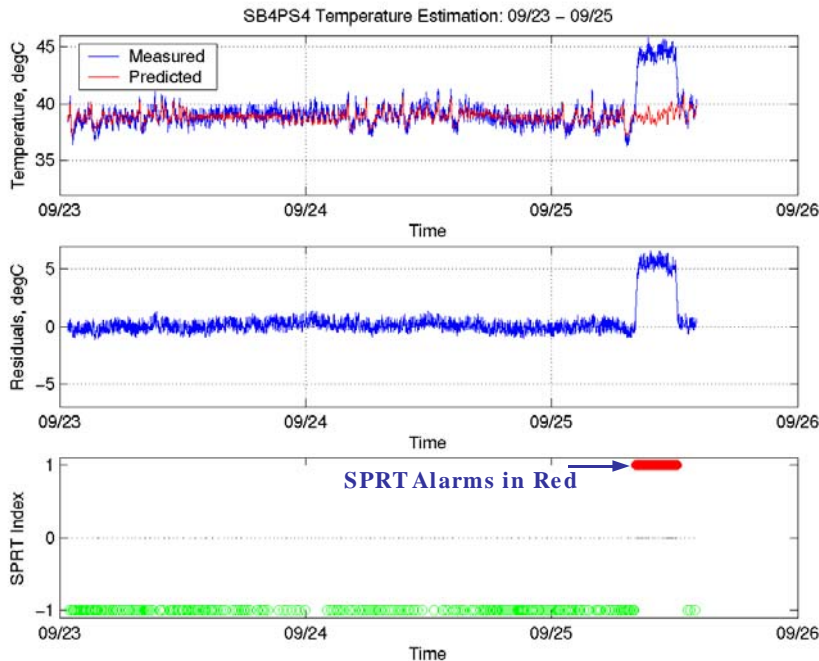
- Failing to change air filters
- Running cables in raised-floor cool-air channel
- Scrap papers get sucked onto bottom air inlet grill
- Inadvertently configuring hot-air exhaust from one machine into cold-air inlet of another

MSET detects all types of thermal perturbations with a high sensitivity and minimal false alarm probability.

SB4PS4 Temperature: Actual and MSET Estimate



MSET Detects Fouled Air Filters in Enterprise Servers

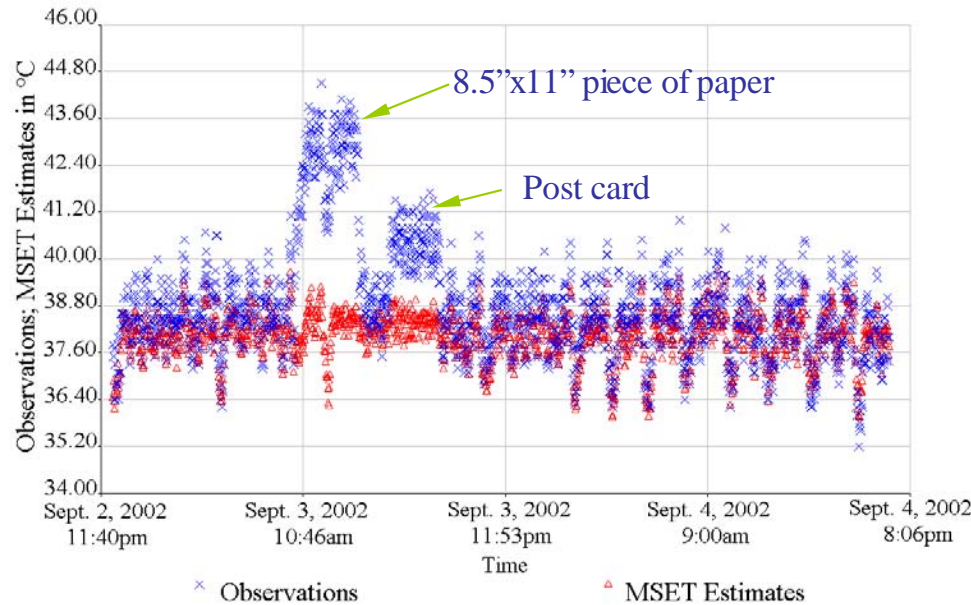


MSET Detects Degraded/Failed Fans (Eliminates Need for Hall-Effect RPM Sensors)

DATE: September 23 - 26 2002

Sun Proprietary/Confidential Information

Observations and MSET Estimates vs. Time
Variable: PS4 Temperature

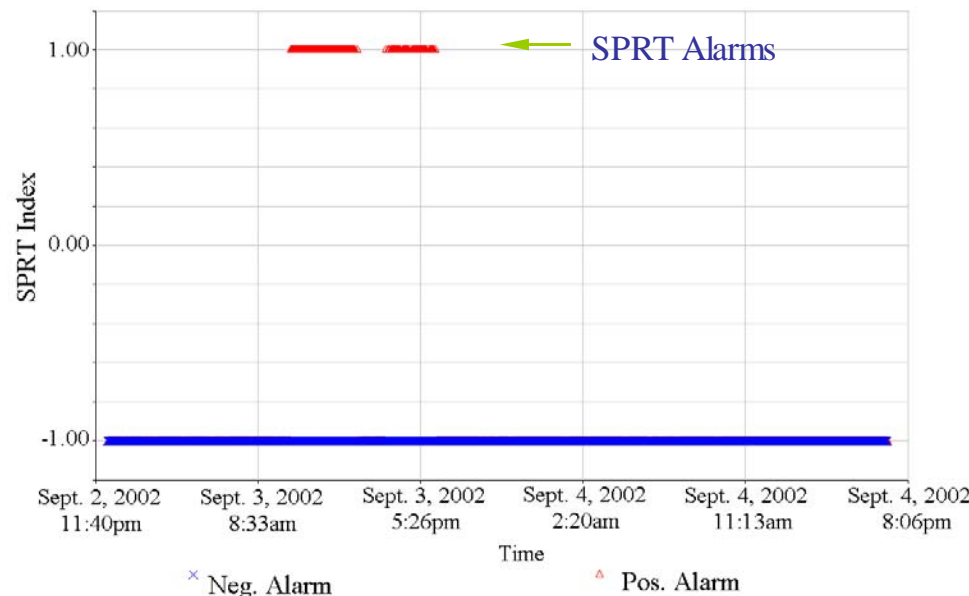


MSET Detects Coolant Air Flow Perturbations in Enterprise Servers

Occasional cause of service problems with high end servers: piece of paper falls from wall, notebook, etc. Works its way to bottom air inlet for server. Temps are not high enough to trip threshold; but over long term, can lead to accelerated reliability issues.

Experiments conducted with fully loaded E10K. MSET monitors dozens of performance variables. Piece of paper put on bottom air inlet.

SPRT Alarm vs. Time
Variable: PS4 Temperature



Immediate SPRT alarms observed.

2nd experiment conducted with 3x5 post card.

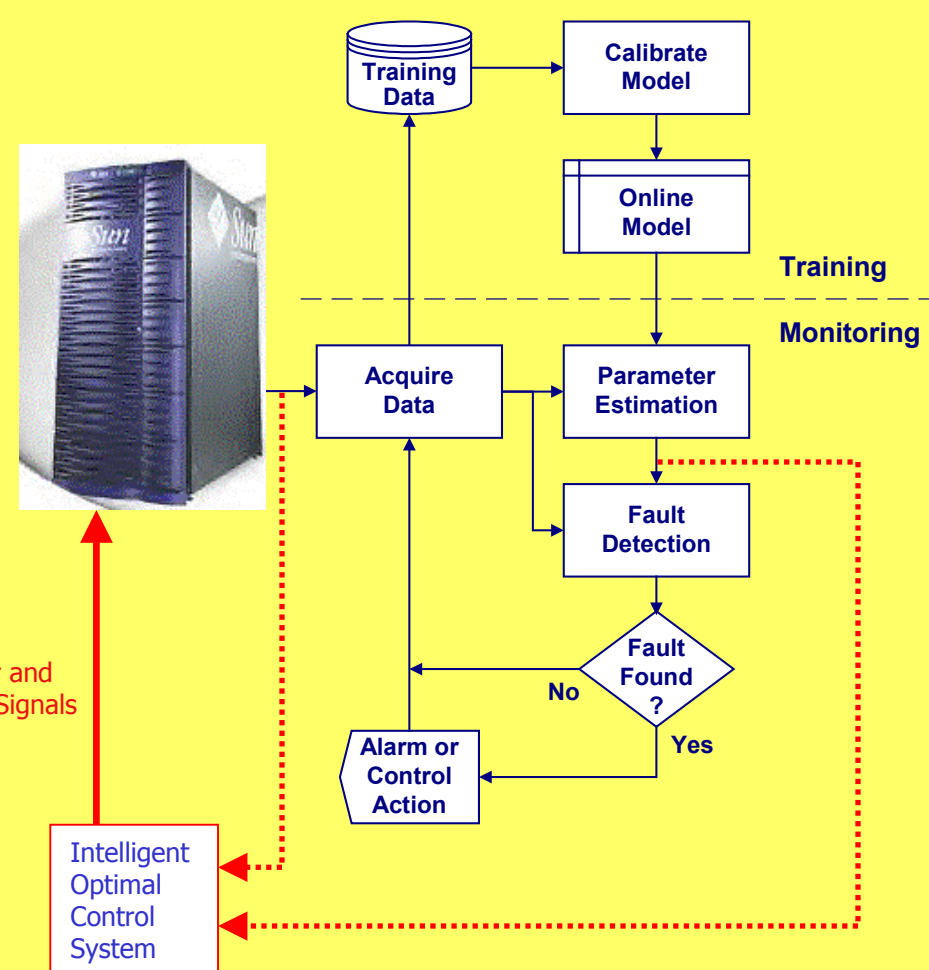
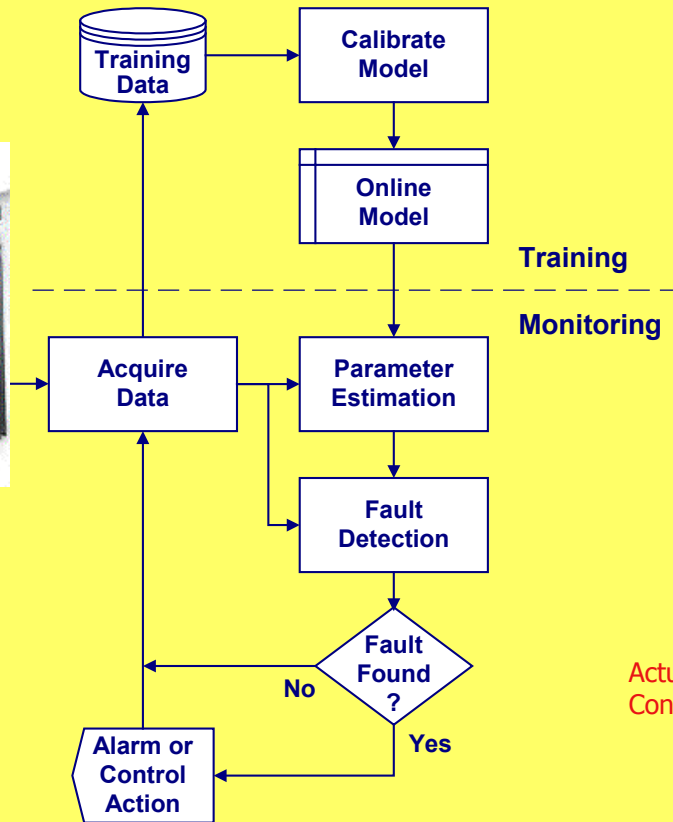
Sun System Dynamics Characterization and Control

MSET Now:

- Global System Telemetry
- Proactive Fault Monitoring
- Enhanced RCA
- Mitigate NTFs

MSET For the Future:

- Realtime Stability Assurance (N1)
- Dynamic Resource Provisioning
- Self Healing and Realtime Fault Avoidance
- Closed Loop Autonomic Control

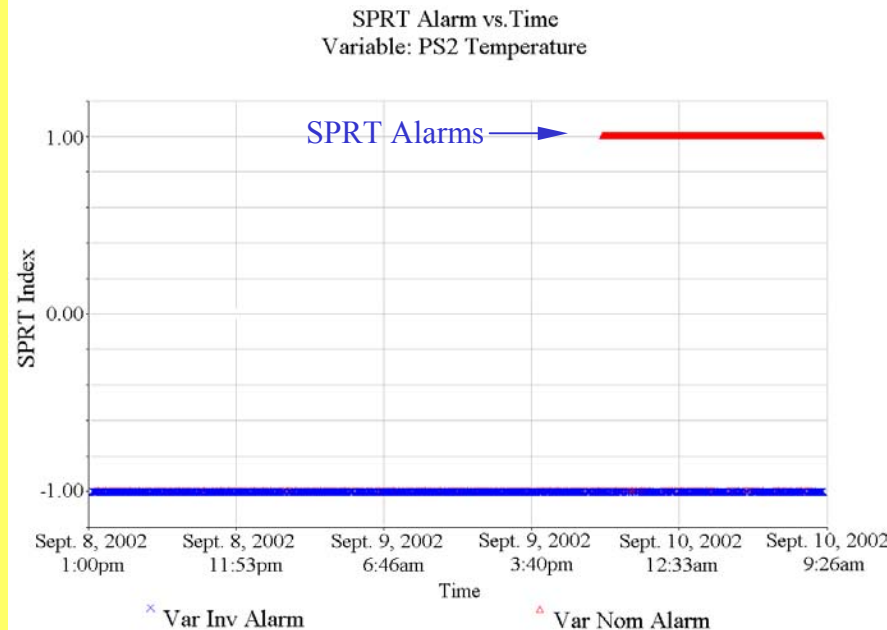
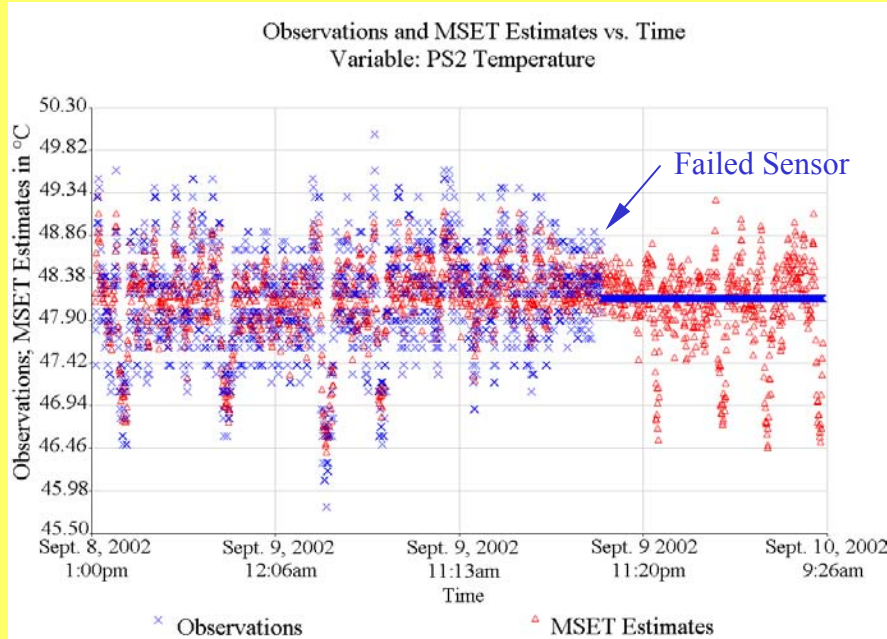


Inferential Sensing

Sun's high-end servers contain hundreds of physical sensors (distributed board, module, and ASIC temperature sensors, voltages, and currents) that protect the system by detecting when a parameter is out of bounds, and then shutting down a component, system board, domain, or entire system.

When a sensor failure is detected, a pattern recognition module swaps out the degraded sensor signal, and swaps in an "analytical estimate" of the physical variable. The analytical estimate is supplied by the pattern recognition algorithm and is called an "inferential sensor". This analytical estimate can be used indefinitely, or until the Field Replaceable Unit (FRU) containing the failed sensor needs to be replaced for other reasons.

Example: If a temp sensor fails on a Starcat system board, the pattern recognition module can replace the failed sensor signal with an analytical estimate (the "inferential sensor") until the FRU is replaced. (With IBM systems it is necessary to replace the entire system board when a sensor fails, or to multiply complexity by deploying redundant sensors).



Inferential Sensors via MSET

Physical sensors can fail. In many cases, the physical sensors have a shorter MTBF than the assets the sensors are supposed to protect.

With MSET, if a physical sensor fails or degrades in service, MSET can mask the sensor signal and swap in the MSET estimate (red variable in figure).

Immediate SPRT alarms observed.

Realtime Sensor Validation: Benefits

- **All control actuator functions now use fully validated signals**
- **For many (perhaps most) industrial systems, including Sun high-end servers, the sensors often have shorter MTBFs than the assets they are supposed to protect**
- **MSET has a unique capability, called inferential sensing, to detect the onset of sensor degradation and swap in a highly accurate analytical estimate. Sensor replacement can be postponed until the next scheduled outage.**

Software Rejuvenation

Software Aging Problems:

Resource contention phenomenon that can cause servers to hang or crash. Mechanisms can include:

Memory leaks; Unreleased file locks; Accumulation of unterminated threads; Data corruption/round off accrual; File space fragmentation; Shared memory pool latching; Thread stack bloating and overruns

Software Rejuvenation:

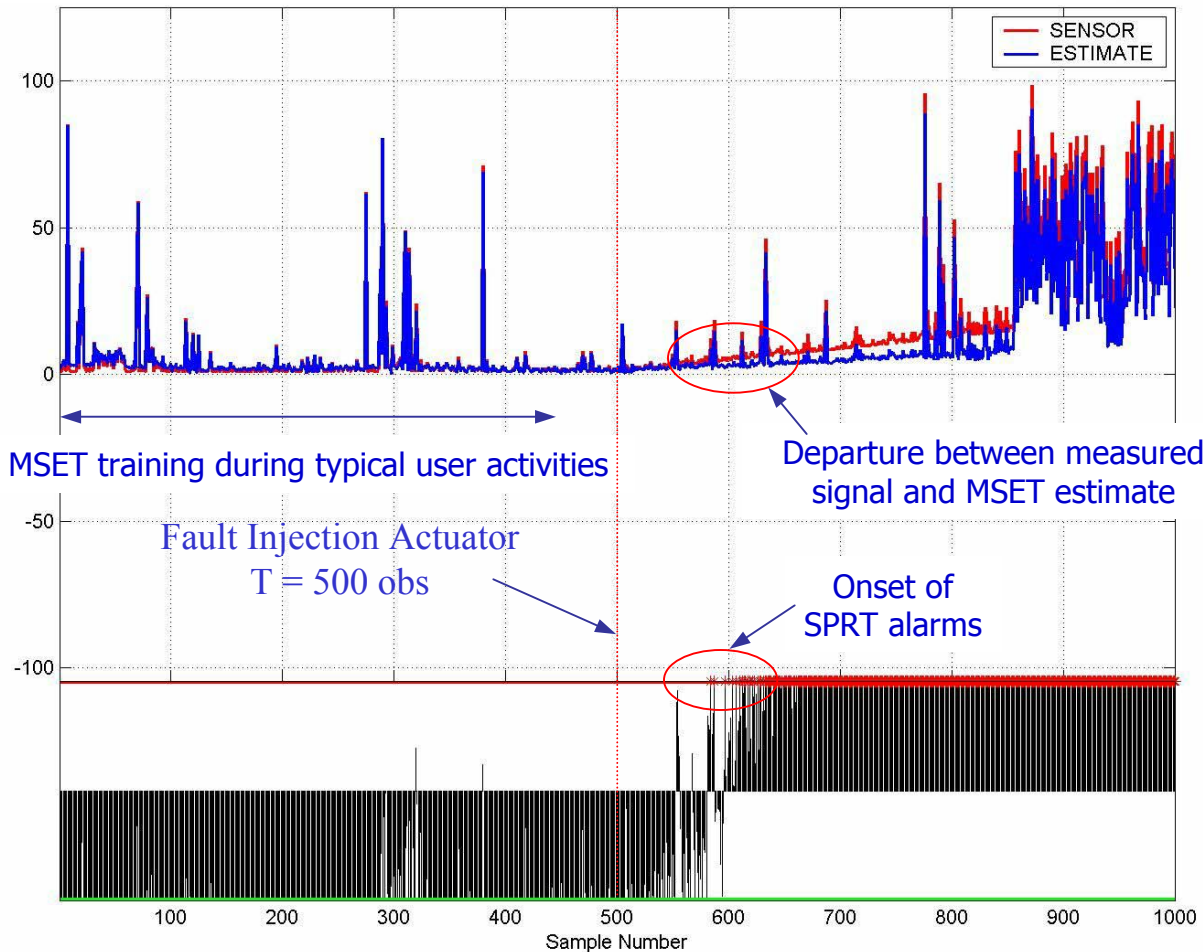
Proactive fault management technique to periodically “cleanse” the system internal state. Mechanisms can include:

Flushing stale locks; Reinitializing application components; Preemptive rollback; Memory defragmentation; Purging DB shared-pool latches; Node/application failover (cluster machines); Therapeutic reboots (primarily Wintel platforms)

MSET Experiments with Software Aging and Rejuvenation

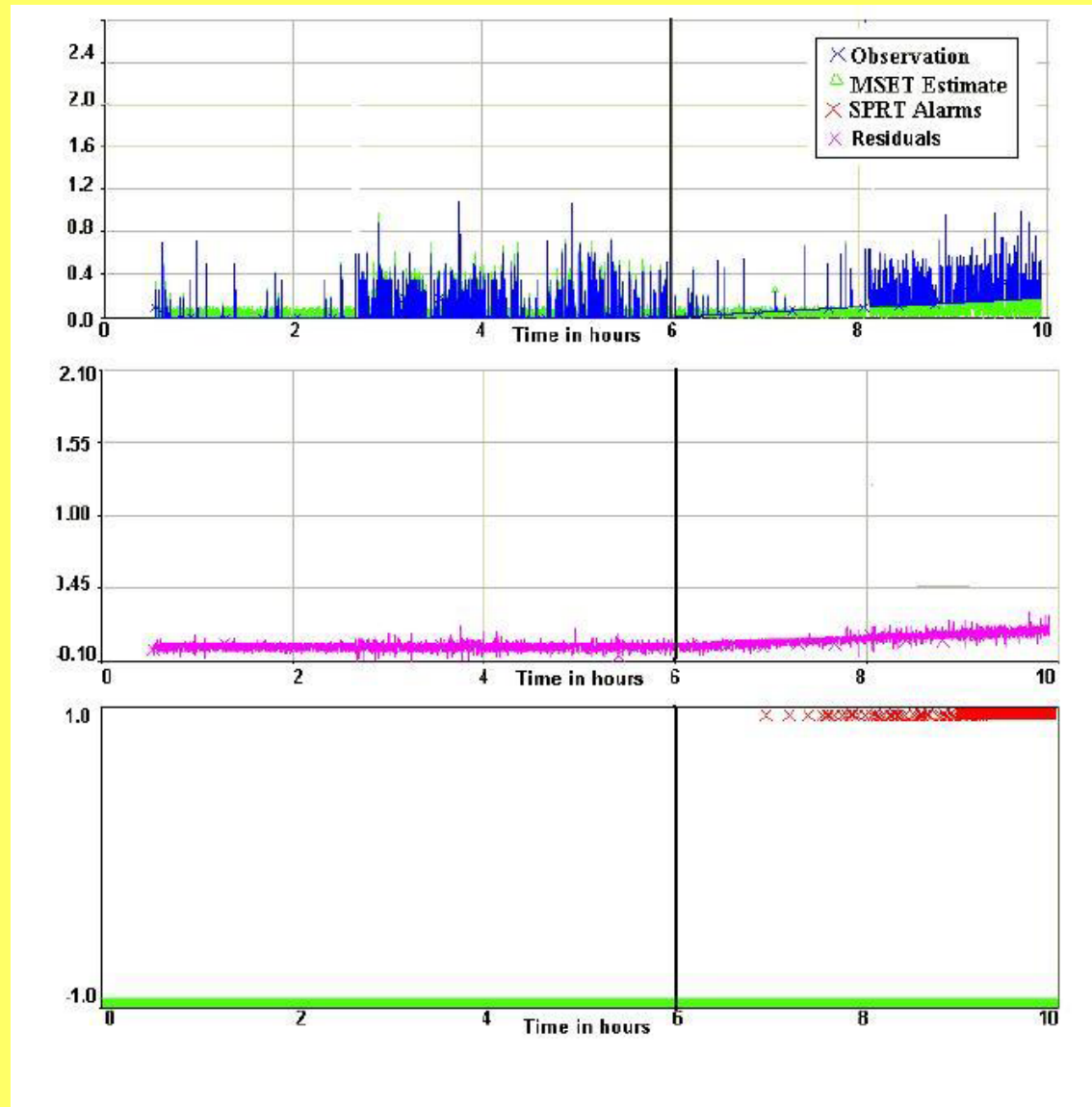
- MSET trained on 33 performance variables sampled in realtime
- Signals generated by standard Unix utilities (mpstat, iostat, cpustat)
- Subtle, linearly degrading memory leaks simulated
- MSET consistently demonstrated high alarm sensitivity with no false alarms

MPSTAT Response Variable (1 of 33 variables monitored by MSET)



MSET Experiments with Controlled Parasitic Resource Consumption

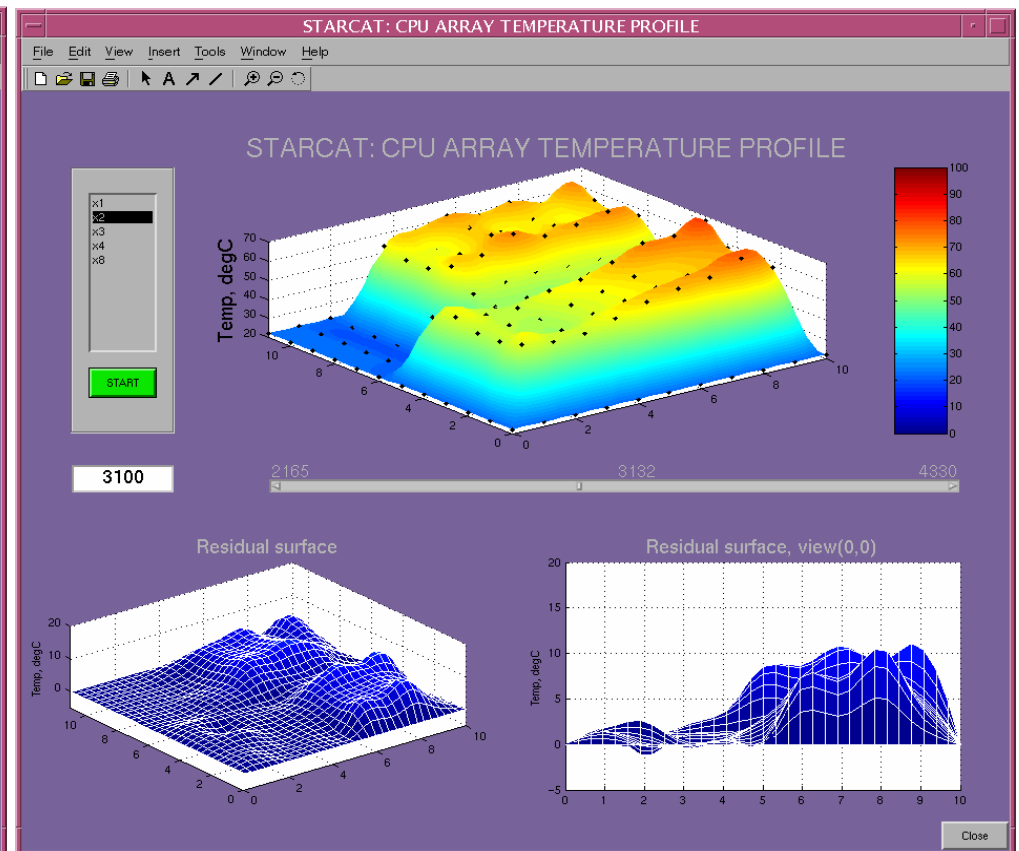
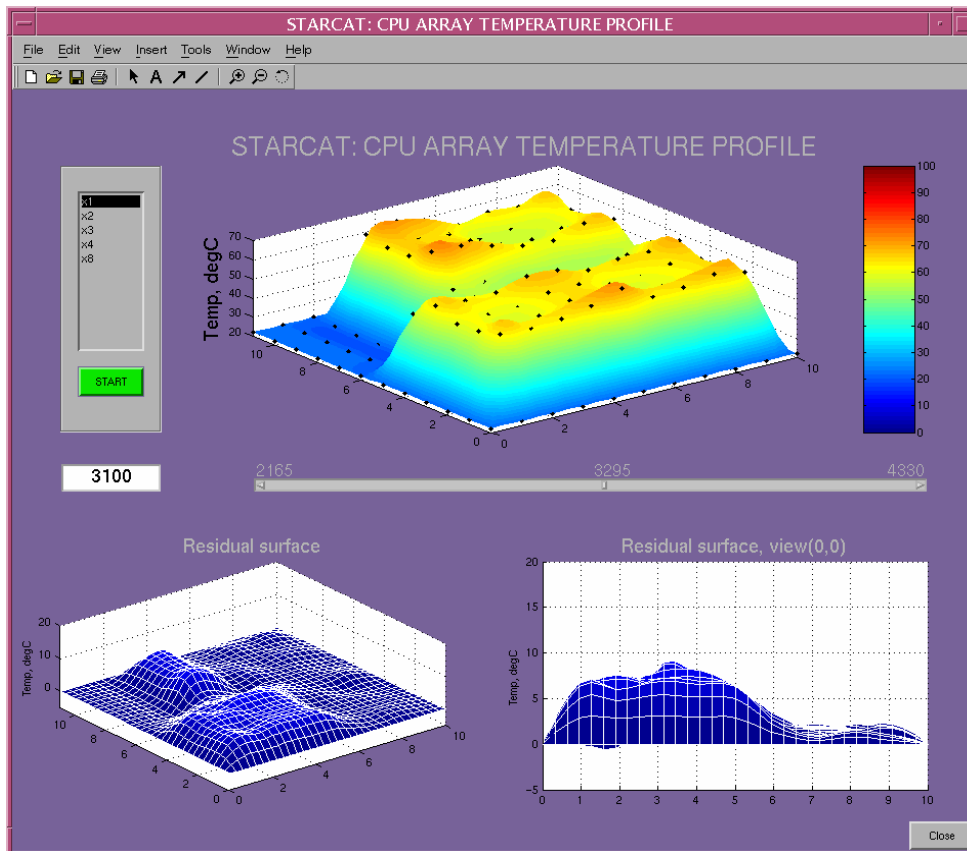
Top : Observation and MSET Estimate Middle : Residuals Bottom : SPRT Alarms



3D Dynamic Viewing Tool

CSTH enables dynamic visualization for thermal, power

- Displays 3D temperature profiles at horizontal cross sections of the server
- Note that blue dots on the 3D surfaces below are actual StarCat temp sensors
- Capabilities are being extended to view voltage and current profiles during dynamic load-perturbation experiments



kenny.gross@sun.com
keith.whisnant@sun.com



**2004
Sun Labs
Open House**