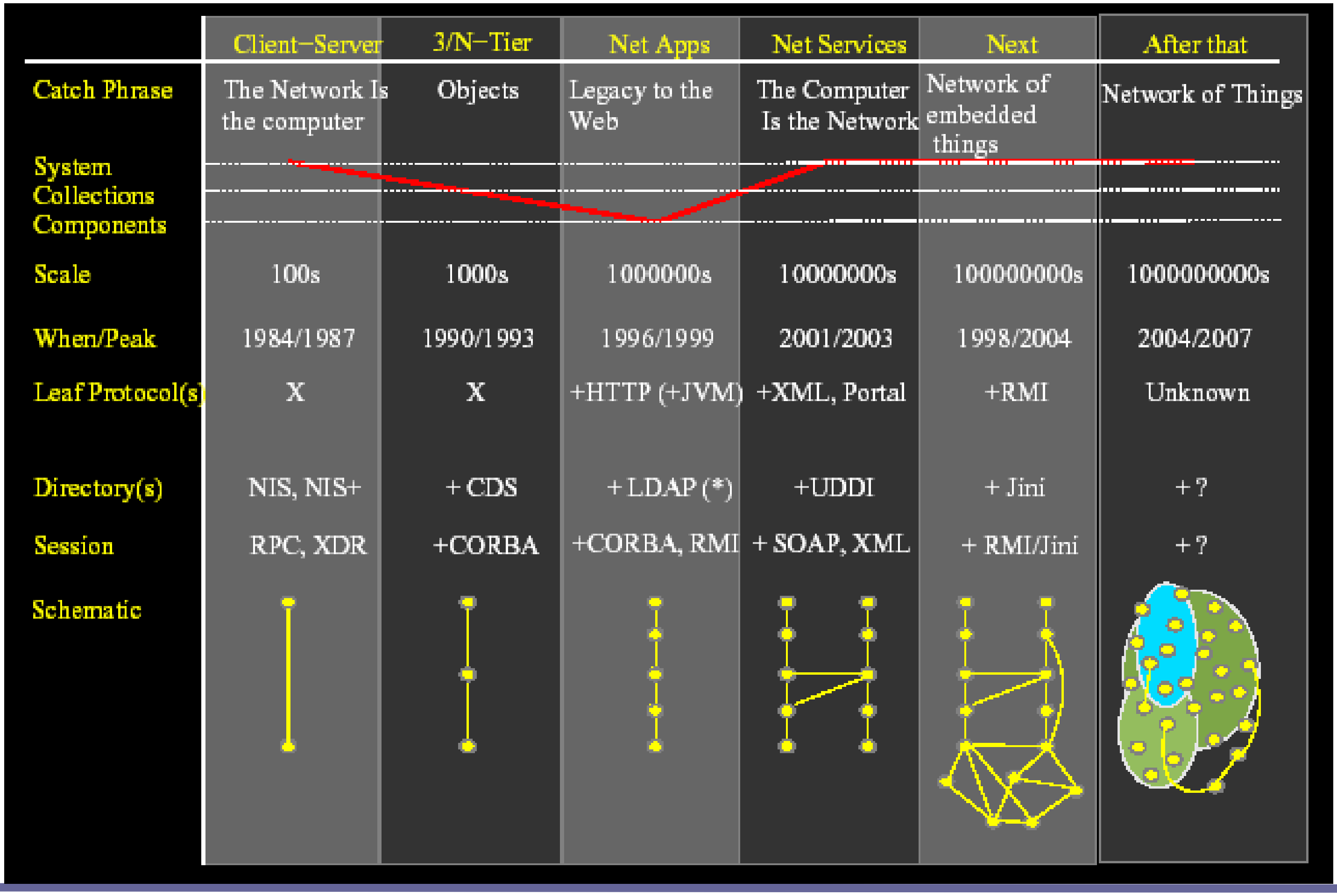
Reliability Through Telemetry

Kenny Gross

Sun Microsystems

Oct 14, 2004

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

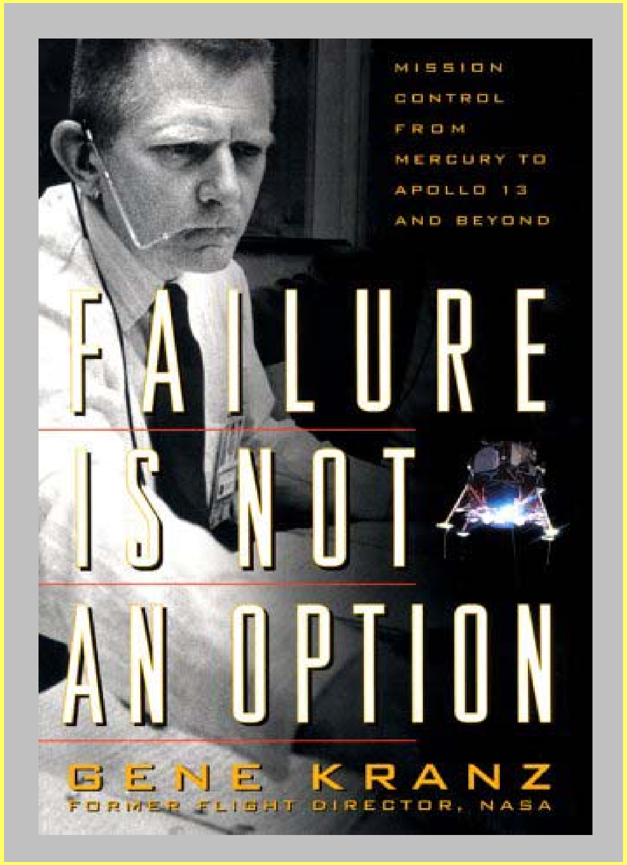
- Causality: Event A causes Event B
  - Association
  - Temporal Precedence
  - Non-spurious association
  - Mechanism

Key Enabler: Continuous System Telemetry

Why is Availability Important?

- 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

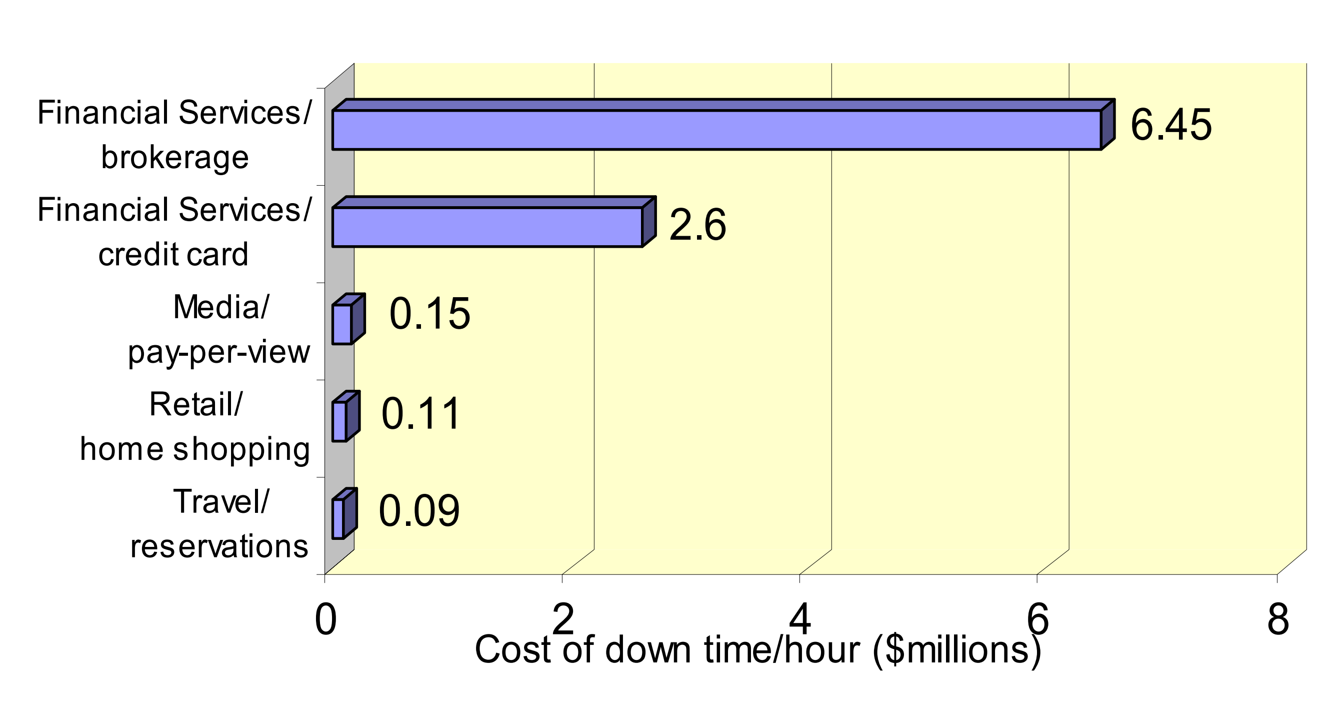
In today's eCommerce-based computing model …


MISSION CONTROL FROM MERCURY TO APOLLO 13 AND BEYOND
FAILURE IS NOT AN OPTION
GENE KRANZ
FORMER FLIGHT DIRECTOR, NASA

Mission director, Apollo 13

Costs of downtime can be substantial

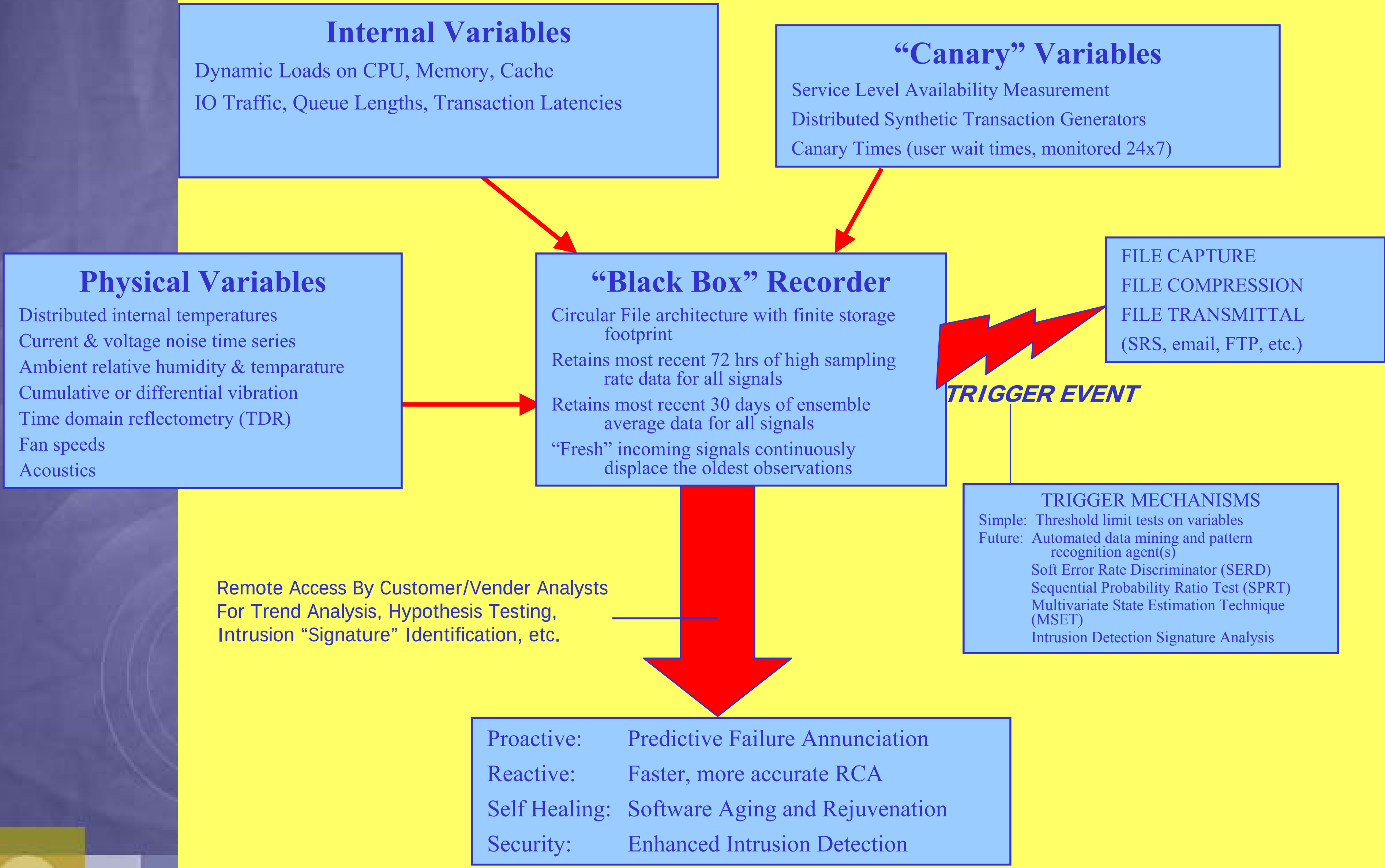
- Quantitative
  - – Lost business, lost productivity



- Qualitative
  - – Diminished reputation

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

Realtime Telemetry Harness for Enhanced Availability and Avoidance of “No-Trouble-Founds”



Internal Variables

Dynamic Loads on CPU, Memory, Cache

IO Traffic, Queue Lengths, Transaction Latencies

“Canary” Variables

Service Level Availability Measurement

Distributed Synthetic Transaction Generators

Canary Times (user wait times, monitored 24x7)

Physical Variables

Distributed internal temperatures

Current & voltage noise time series

Ambient relative humidity & temparature

Cumulative or differential vibration

Time domain reflectometry (TDR)

Fan speeds

Acoustics

“Black Box” Recorder

Circular File architecture with finite storage footprint

Retains most recent 72 hrs of high sampling rate data for all signals

Retains most recent 30 days of ensemble average data for all signals

“Fresh” incoming signals continuously displace the oldest observations

FILE CAPTURE

FILE COMPRESSION

FILE TRANSMITTAL

(SRS, email, FTP, etc.)

TRIGGER EVENT

TRIGGER MECHANISMS

Simple: Threshold limit tests on variables

Future: Automated data mining and pattern recognition agent(s)

Soft Error Rate Discriminator (SERD)

Sequential Probability Ratio Test (SPRT)

Multivariate State Estimation Technique (MSET)

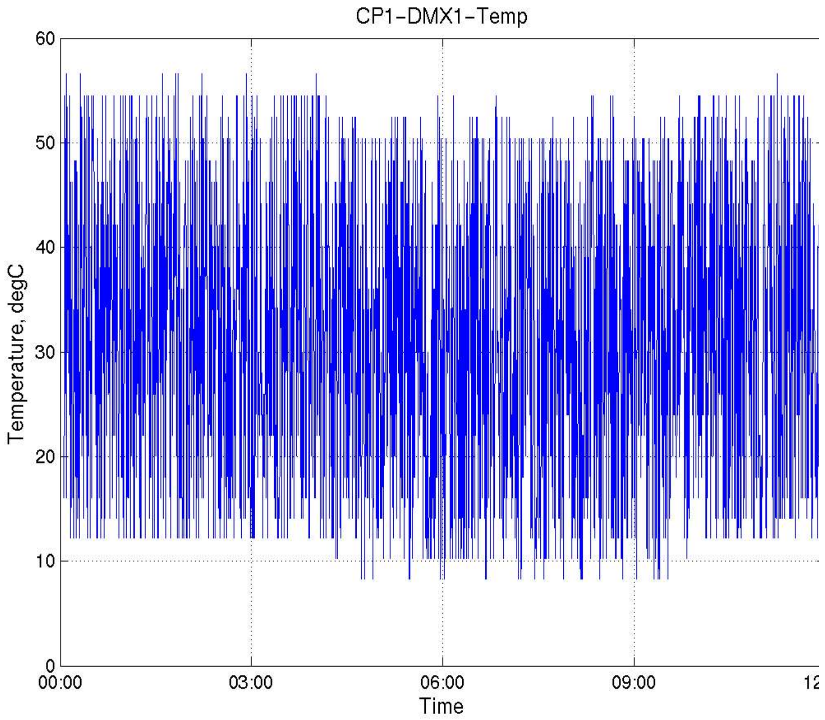
Intrusion Detection Signature Analysis

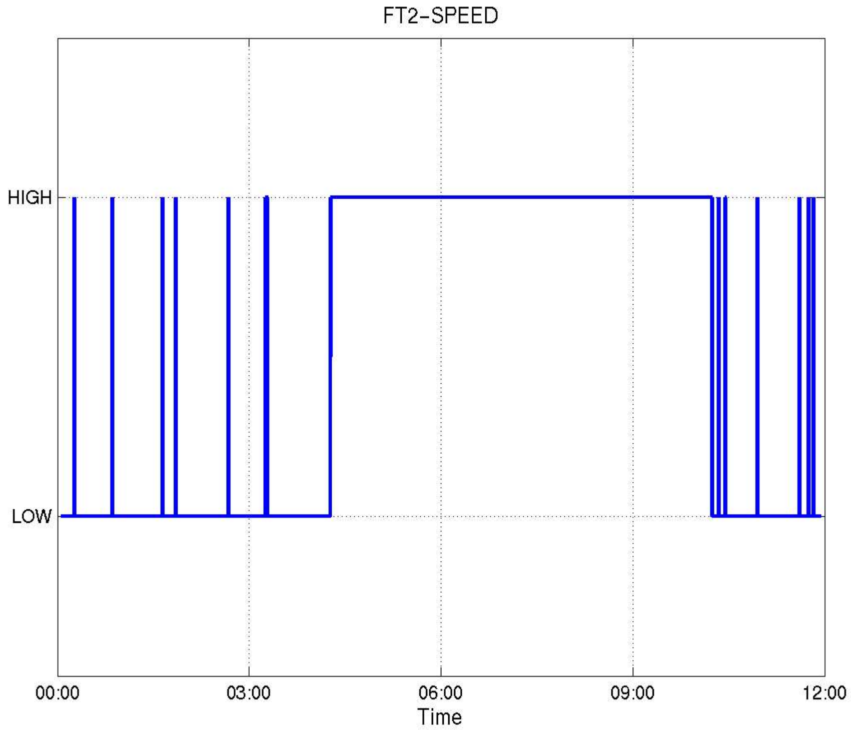
Remote Access By Customer/Vender Analysts For Trend Analysis, Hypothesis Testing, Intrusion “Signature” Identification, etc.

Proactive:	Predictive Failure Annunciation
Reactive:	Faster, more accurate RCA
Self Healing:	Software Aging and Rejuvenation
Security:	Enhanced Intrusion Detection

Early Telemetry Harness Success on StarCat Platform

Starcat EMS bug identified during first week of telemetry testing of F15K.



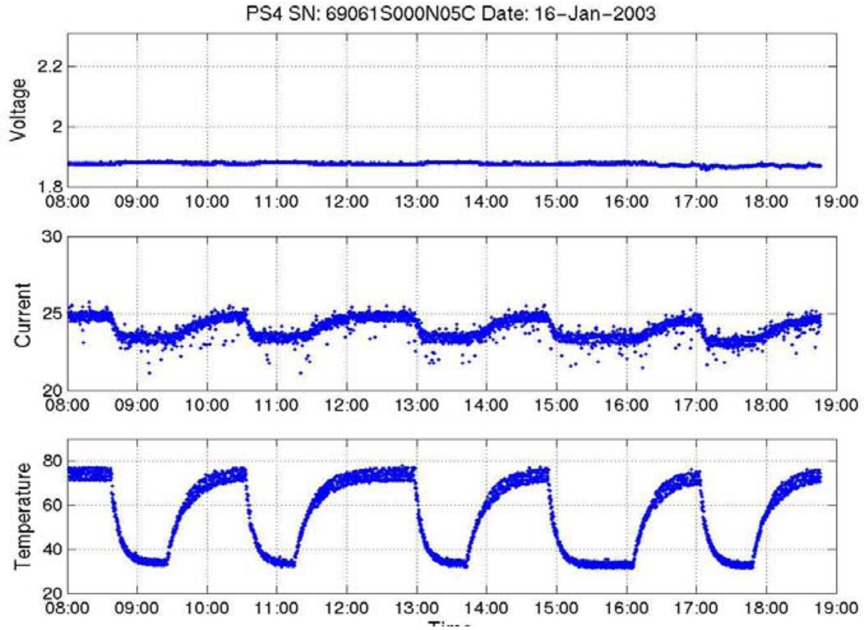


StarCat SMS reporting processor temperature improperly. Note wild swings between 12-50 deg C.

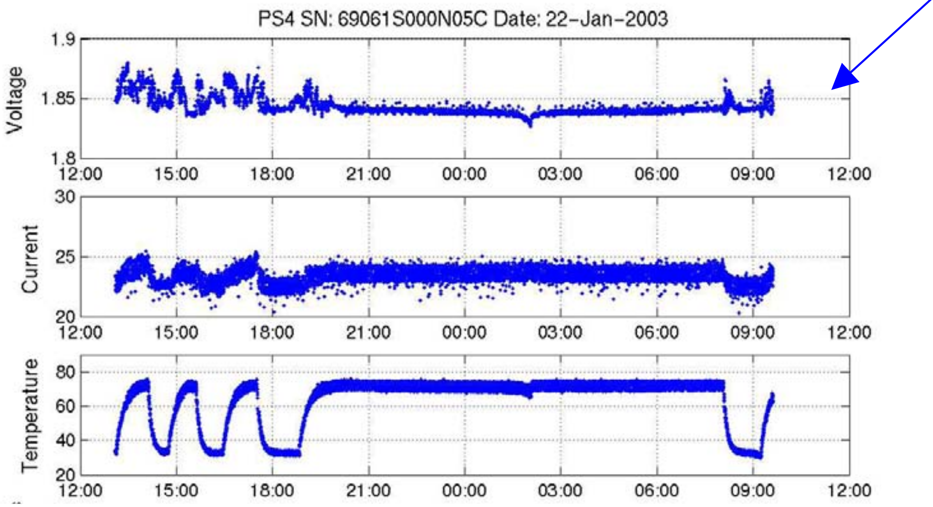
Faulty temperature values cause fans to continuously cycle between low and high (1 of 8 fans shown).

CSTH for NTF Mitagation

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.

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

Advanced Pattern Recognition

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
- Improved stability analysis, dynamic resource provisioning for networks of interacting dynamic elements
- Closed-loop autonomic control for future servers and networks

New and Ongoing Proactive Fault Monitoring Projects:

- Continuous signal validation, sensor operability validation in servers
- “Inferential Sensing”, allows analytical sensor replacement (i.e. temp sensor fails on StarCat system board, can swap in MSET estimate until the FRU needs to be replaced for other reasons).
- Dynamic resource provisioning for Eagle and Post-Eagle platforms
- Software aging problems at customer sites (memory leaks, resource contention issues)
  - Successful pilot project applied MSET to complex S/W aging problem on SunPeak business critical E10Ks; Recent additional successes with SunRay servers
- Improved environmental-variable proactive fault detection
  - Detects flow perturbations & partial blockages, dirty air filters
  - Detects degraded/failed fans without RPM sensors
- Closed-loop autonomic control for improved (and less human-intensive) performance management of servers and networks
- Proactive fault monitoring and more accurate RCA for mitigating NTFs and ELFs
- Stability assurance for N1 networks of dynamically interacting systems.

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.

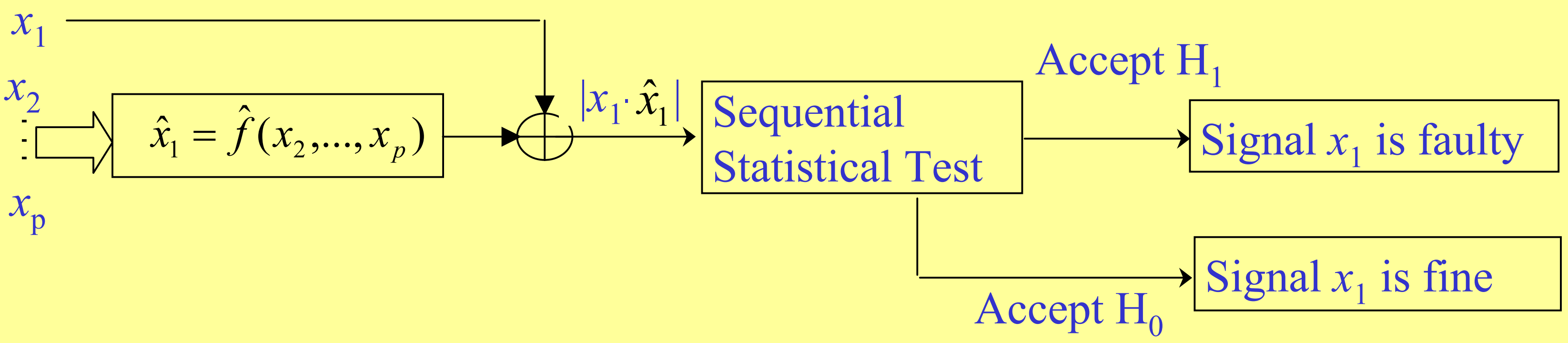
Monitoring Practice Today: Thresholds not adequate

Users typically set alert thresholds and then use tools to correlate multiple alerts and trigger alarms, but:

- Thresholds are difficult to set accurately;
- Static thresholds are either too tight or too loose;
- Correlations are not obvious or intuitive
- Tradeoff between false alarms vs. sensitivity leads users to ignore indicators til it’s too late

Sequential Probability Ratio Test (SPRT)

High sensitivity for subtle anomaly detection, but without increasing false alarm probability.



x_1

x_2

x_p

$\hat{x}_1 = \hat{f}(x_2,...,x_p)$

$|x_1 \cdot \hat{x}_1|$

Sequential Statistical Test

Accept H_1

Signal x_1 is faulty

Accept H_0

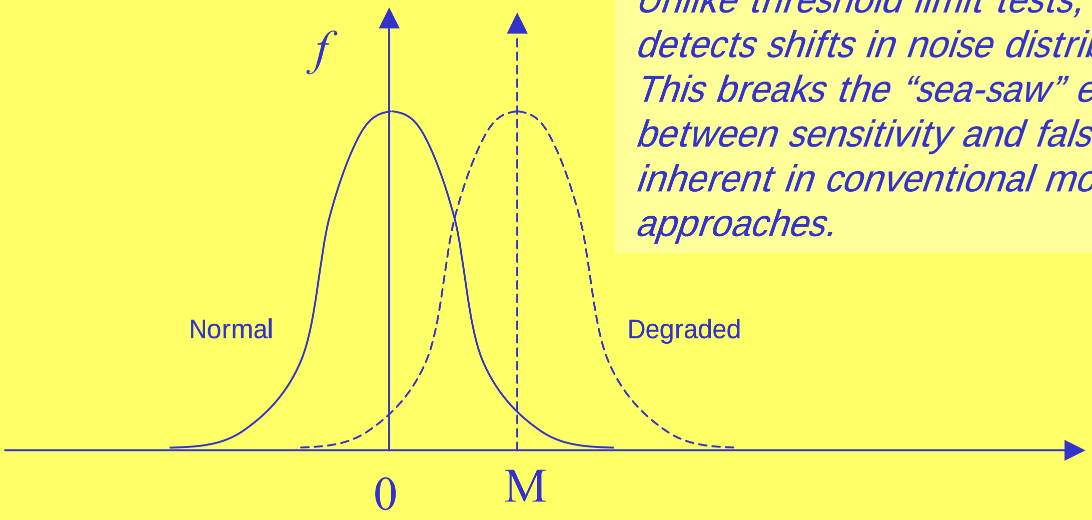
Signal x_1 is fine

Sequential test for residual signals is based upon two hypotheses:

$H_0 : \mu = \mu_0 = 0$

$H_1 : \mu = \mu_1 = M$

where μ is the mean of the distribution under test and M is a preset alarm threshold.



Unlike threshold limit tests, SPRT detects shifts in noise distributions. This breaks the "sea-saw" effect between sensitivity and false alarms inherent in conventional monitoring approaches.

MSET
Multivariate State Estimation Technique

Argonne National Laboratory developed an advanced pattern recognition system, MSET, for 24X7 predictive fault monitoring in complex engineering systems for extremely sensitive identification of the onset of component degradation or process anomalies.

MSET possesses unique surveillance capabilities that surpass any conventional approaches, including neural networks, in sensitivity, reliability, and computational efficiency.

MSET Provides:

- Continuous signal validation and sensor operability validation for safety-critical and mission-critical applications
- Incipient fault annunciation in all monitored components and process variables
- Extremely low probability of false alarms
- Improved dynamic feedback/control algorithms (MSET estimate can be used for control variable)

The MSET incipient fault surveillance system won a 1998 R&D-100 Award from R&D Journal for one of the top 100 technological breakthroughs for 1998.

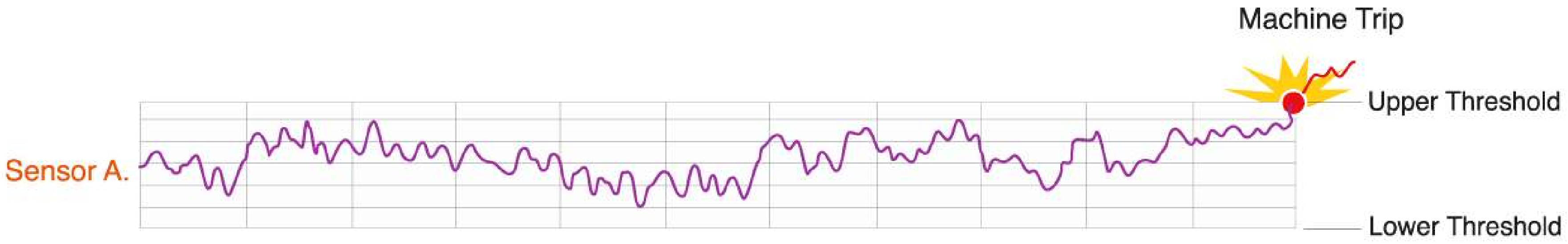
In 2000, MSET won Control Engineering Magazine's "Editor's Choice Top 40 Awards."

In 2001, MSET won a Best of Sensors Expo new product "Gold Award".

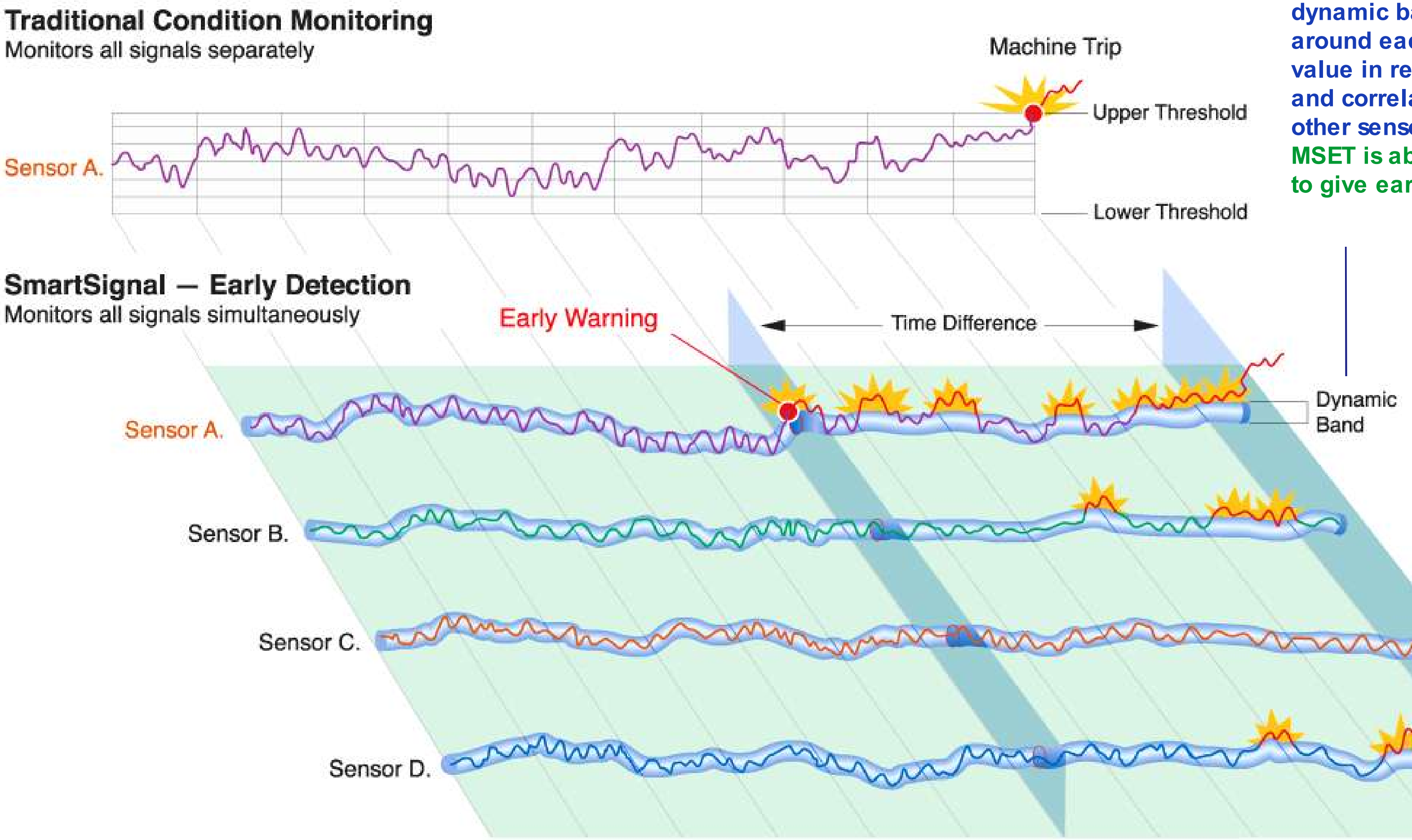
Telemetry + pattern recognition allow early prediction of failures

Traditional Monitoring: High/Low Thresholds

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

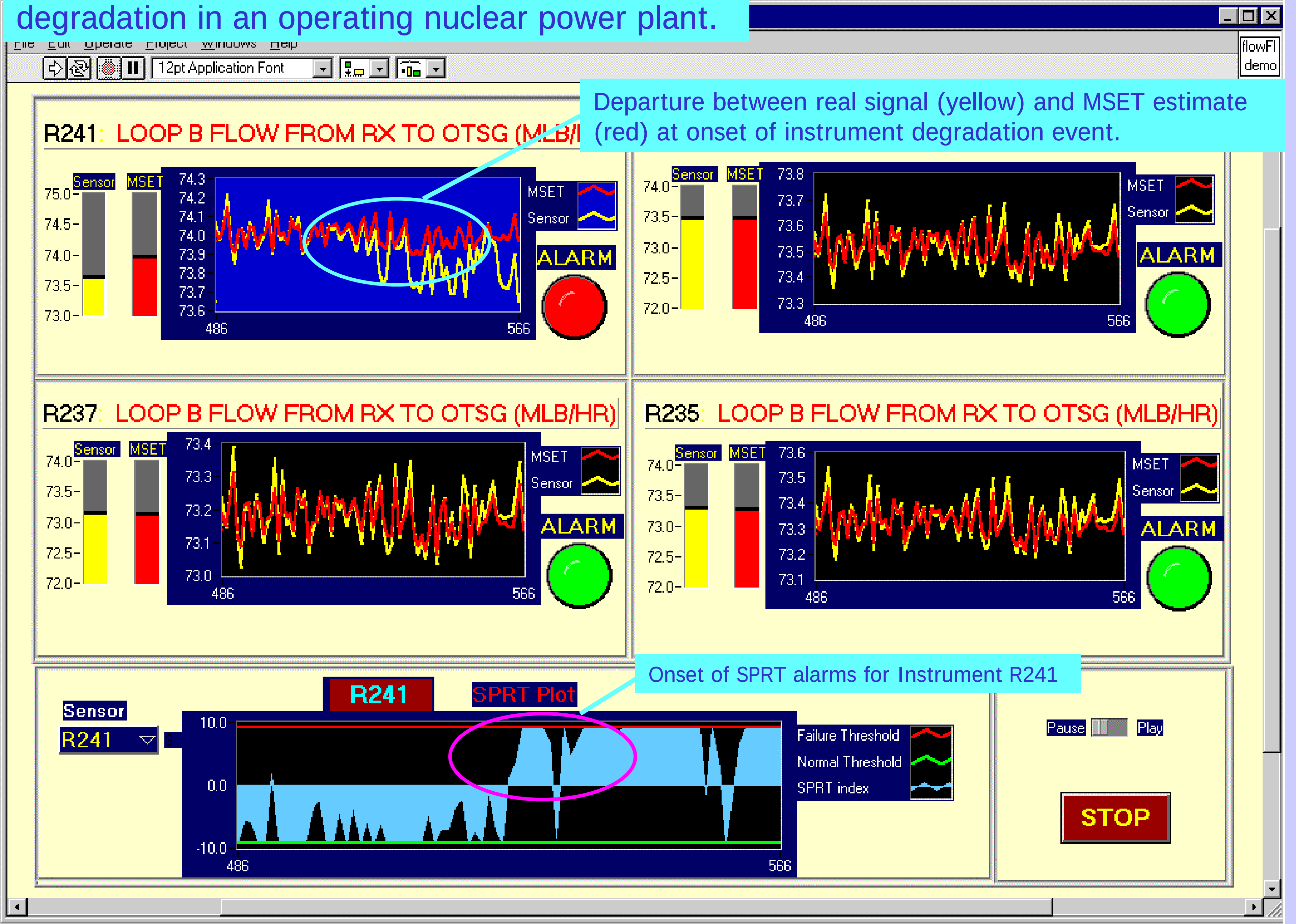


CSTH + MSET Pattern Recognition: Early Detection


Traditional Condition Monitoring
Monitors all signals separately
Machine Trip
Upper Threshold
Sensor A.
Lower Threshold
SmartSignal — Early Detection
Monitors all signals simultaneously
Early Warning
Time Difference
Sensor A.
Dynamic
Band
Sensor B.
Sensor C.
Sensor D.

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.

Legacy Viewgraph: MSET detects instrumentation degradation in an operating nuclear power plant.



Departure between real signal (yellow) and MSET estimate (red) at onset of instrument degradation event.

Onset of SPRT alarms for Instrument R241

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

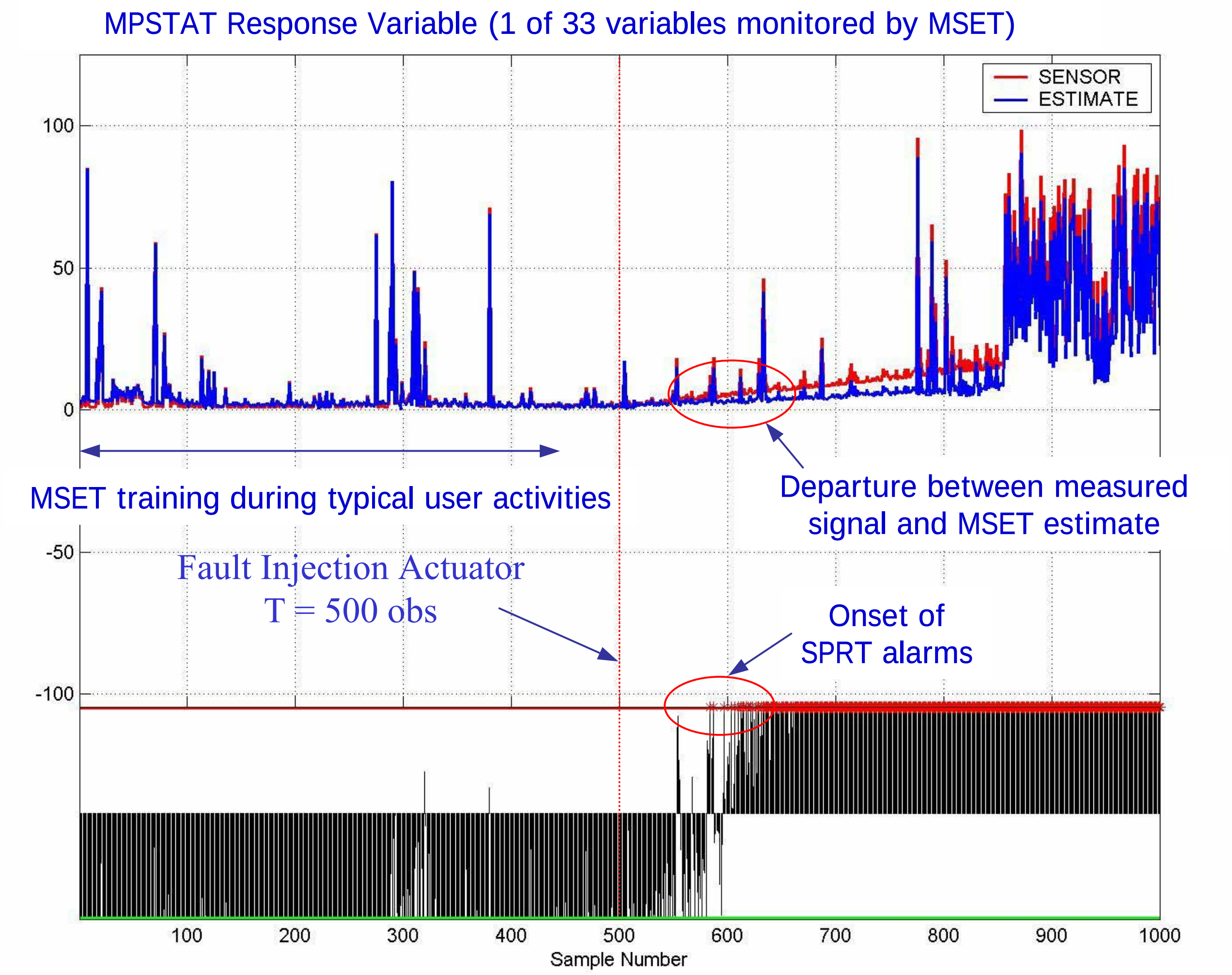
Software Rejuvenation:

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



MSET training during typical user activities

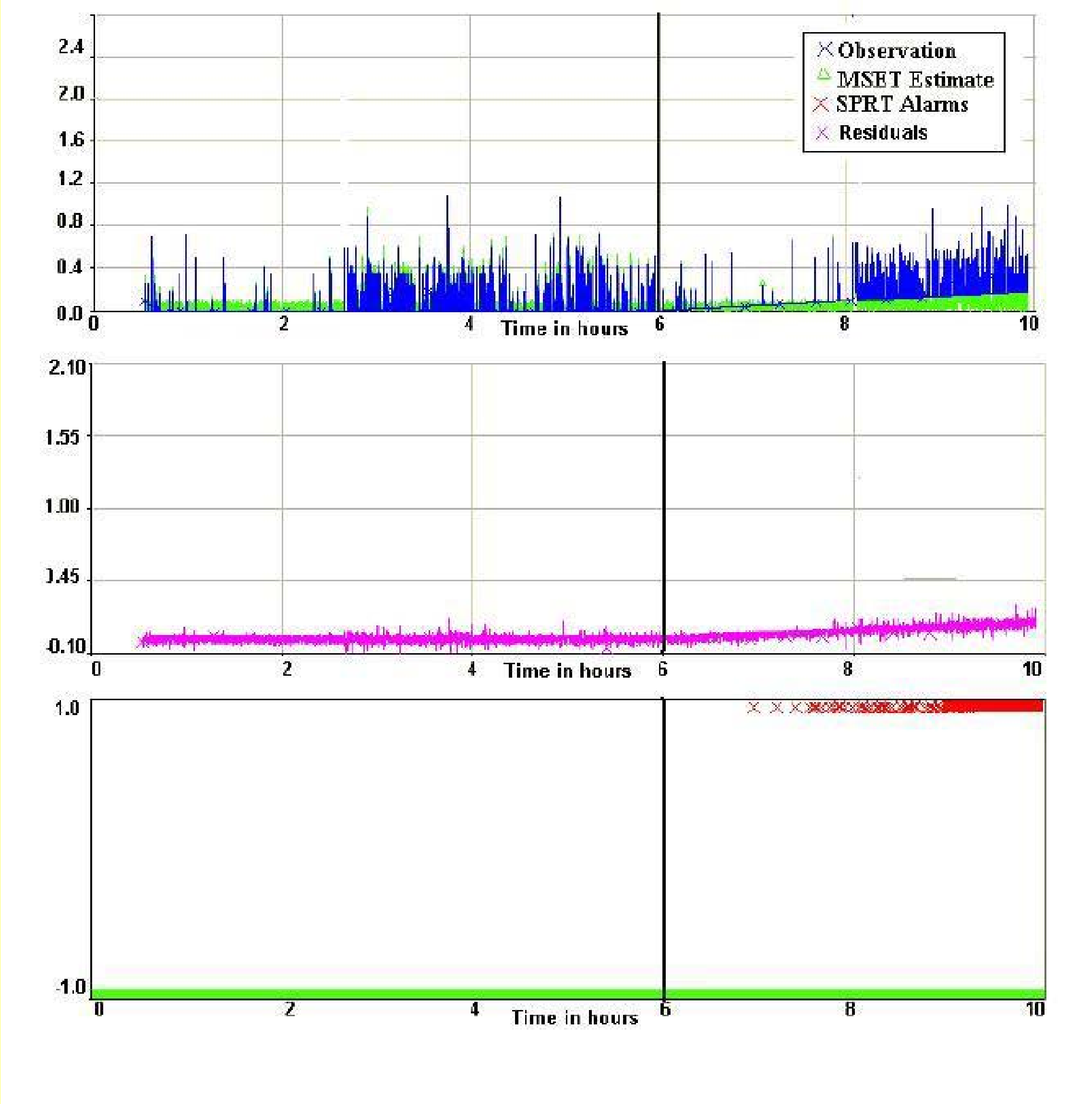
Departure between measured signal and MSET estimate

Fault Injection Actuator
T = 500 obs

Onset of SPRT alarms

MSET Experiments with Controlled Parasitic Resource Consumption

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



Thermal Anomalies Cause Customer Outages

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.

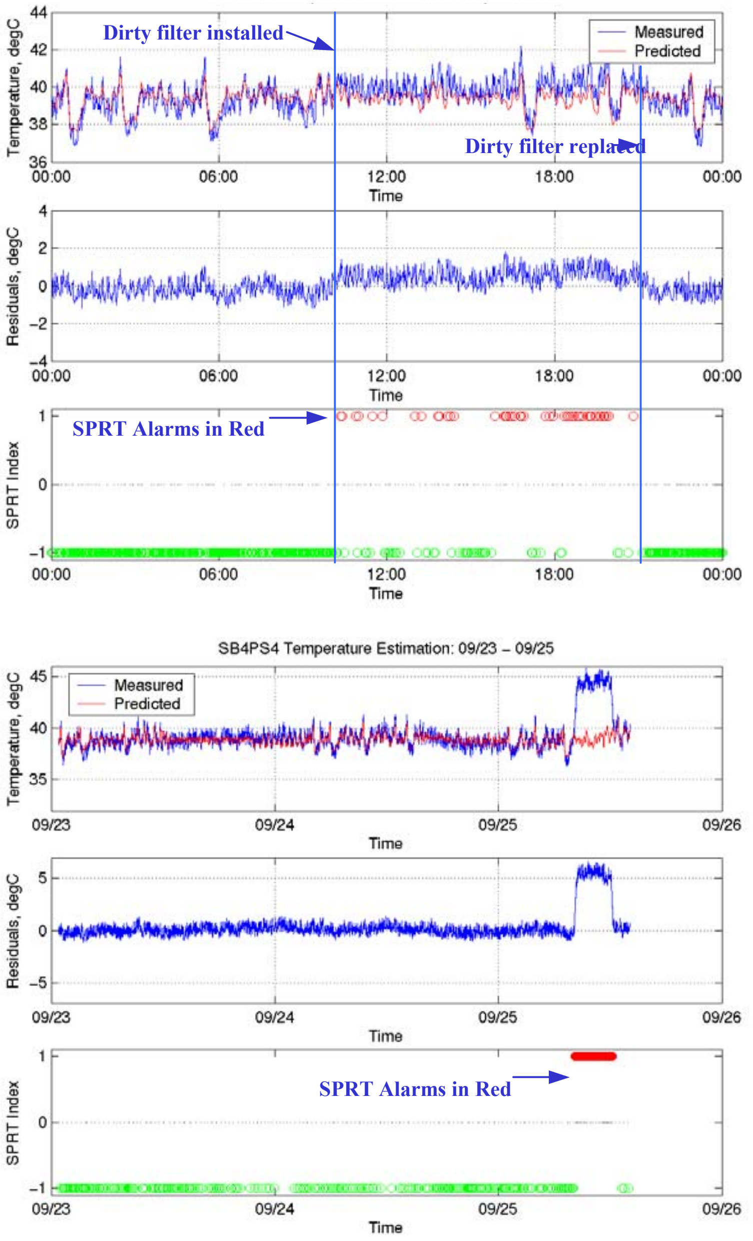
Examples:

- 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

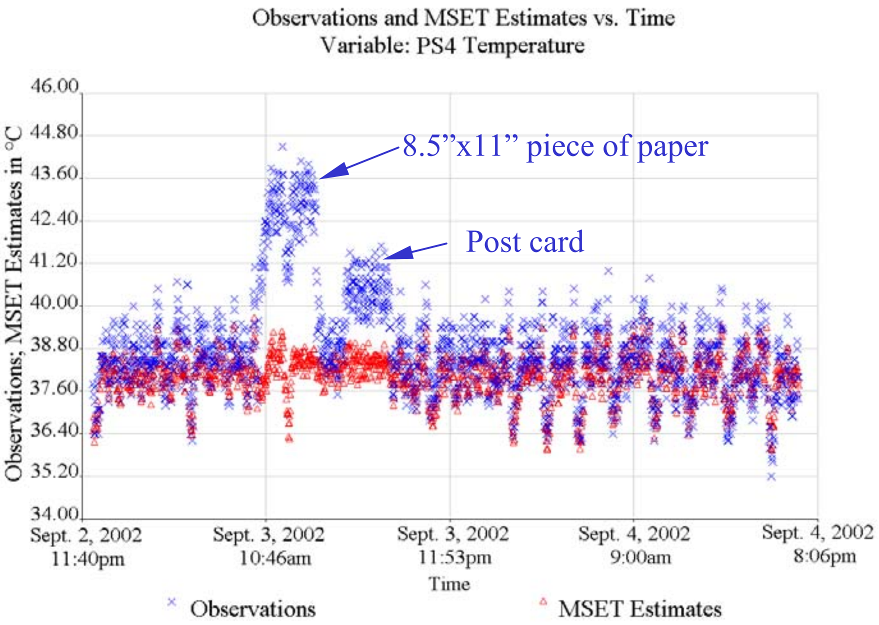


DATE: September 23 – 26 2002

MSET Detects Fouled Air Filters in Enterprise Servers

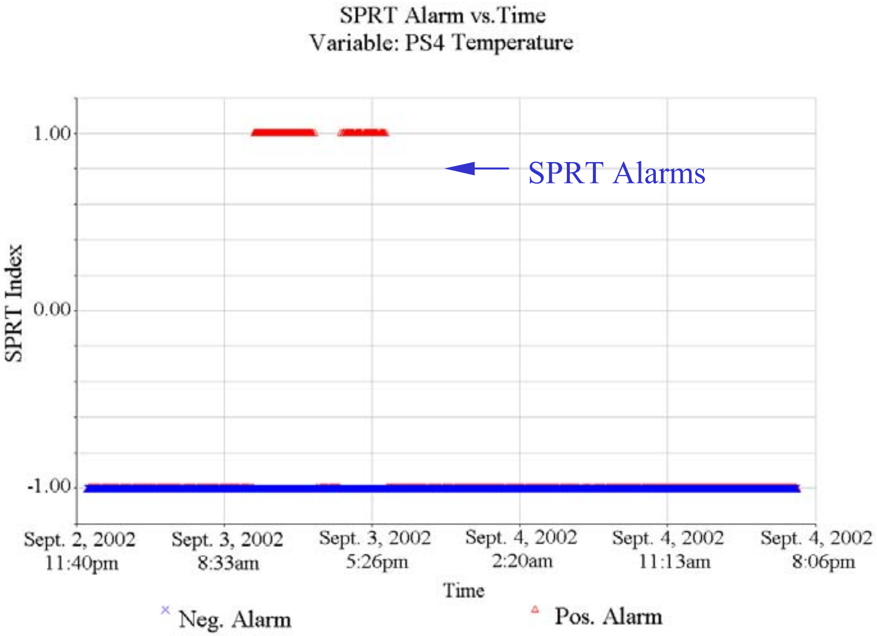
MSET Detects Degraded/Failed Fans

(Eliminates Need for Hall-Effect RPM Sensors)


Observations and MSET Estimates vs. Time
Variable: PS4 Temperature

8.5”x11” piece of paper

Post card


SPRT Alarm vs.Time
Variable: PS4 Temperature

SPRT Alarms

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.

Immediate SPRT alarms observed.

2nd experiment conducted with 3x5 post card.

Software Rejuvenation

Software Aging Problems:

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

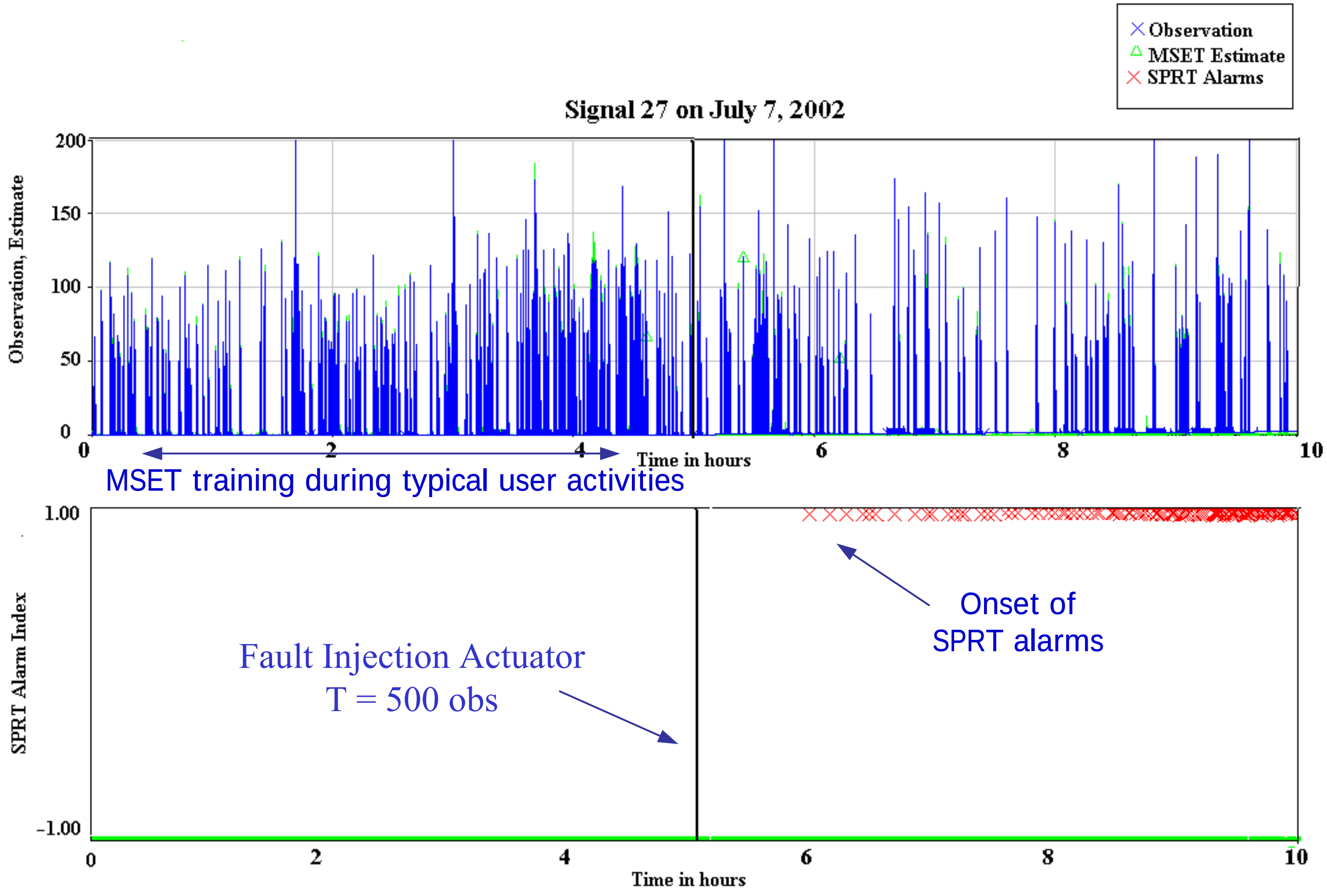
Memory leaks; Unreleased file locks; Accumulation of unterminated threads;   Data corruption/round off accrual;   File space fragmentation; Shared memory pool latching; Thread stack bloating and overruns

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 Detection of Onset of Software Aging Mechanism in Signal 27 Parasitic Resource Consumption Rate = 1% per 24 Hrs

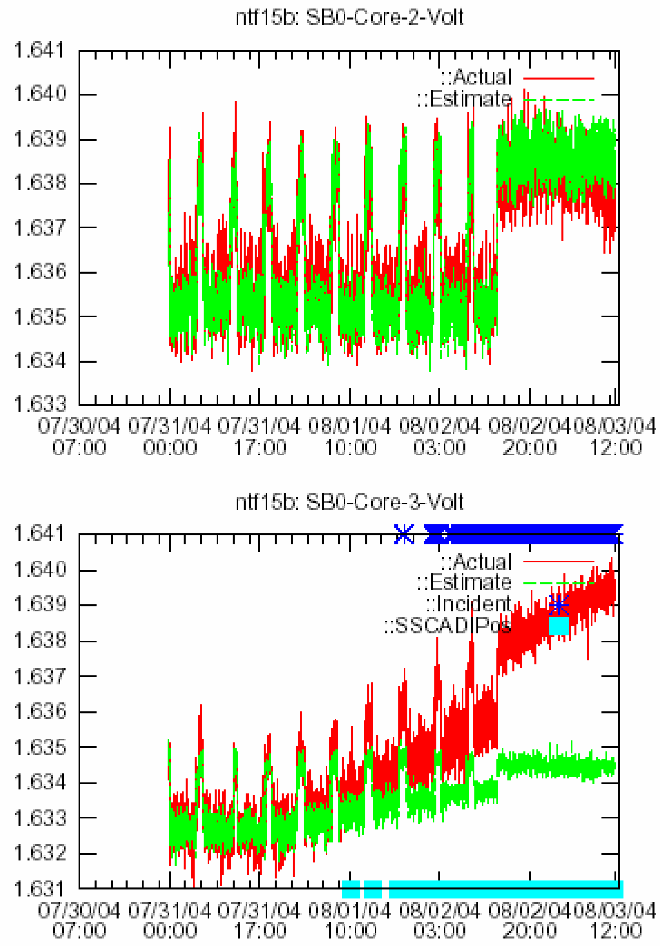


MSET training during typical user activities

Fault Injection Actuator
T = 500 obs

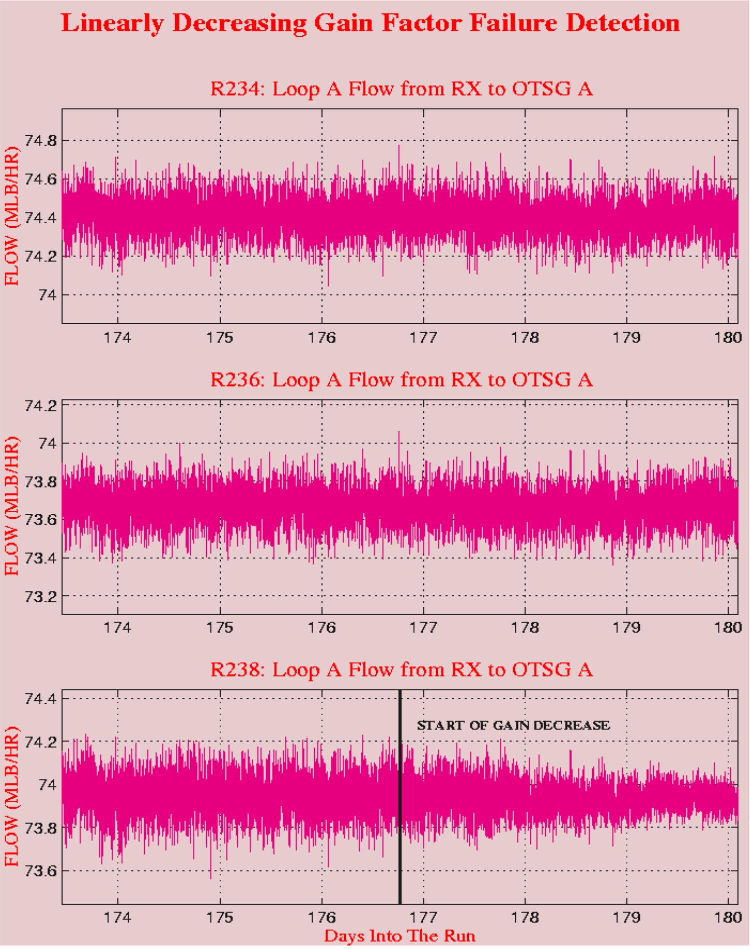
Onset of
SPRT alarms

New eCM MatLab Toolkit Detects Drifting Power Supply Voltages

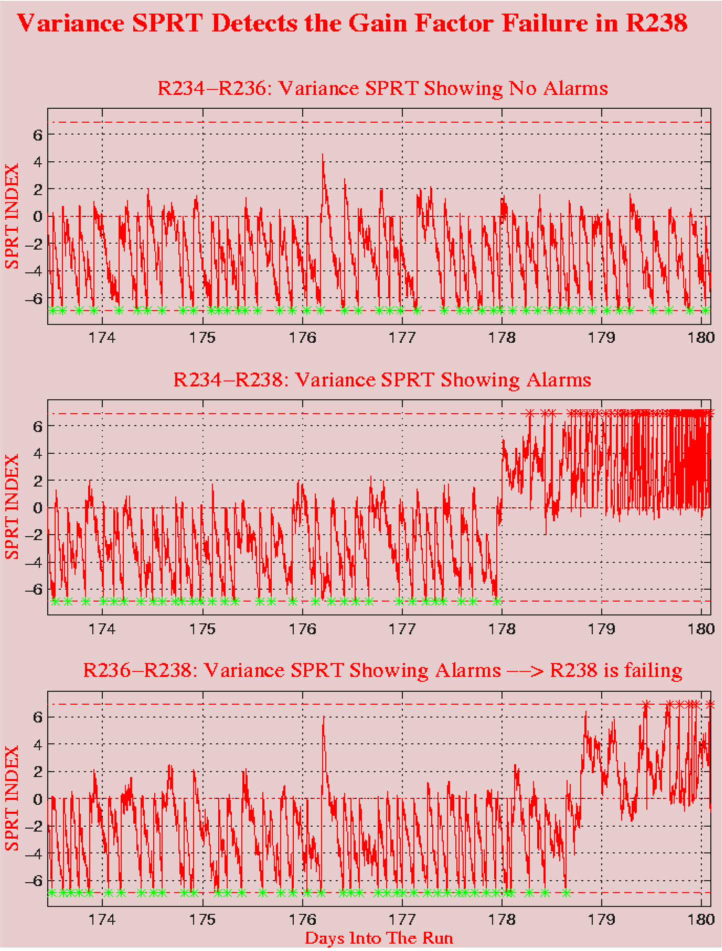


Can detect anomalies in nominally stationary signals.

eCM automatically takes into account load effects by monitoring correlated signals.



Linearly Decreasing Gain Factor Failure Detection



Variance SPRT Detects the Gain Factor Failure in R238

Use of Pattern Recognition with Six Sigma:

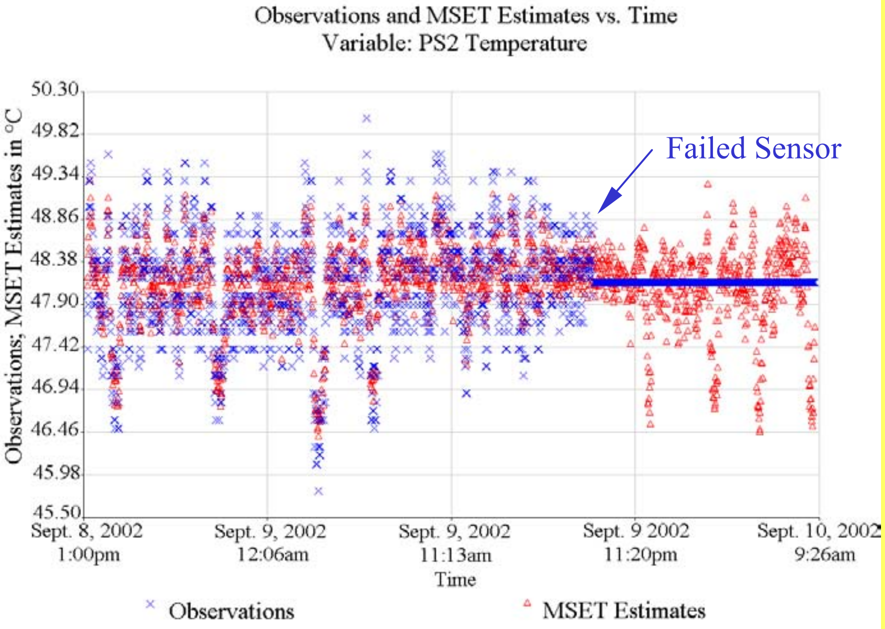
Note that advanced pattern recognition is useful not only for detecting the onset of degradation. MSET can also detect and quantitatively characterize the onset of process improvements, including the reduction in processes variability.

Inferential Sensing

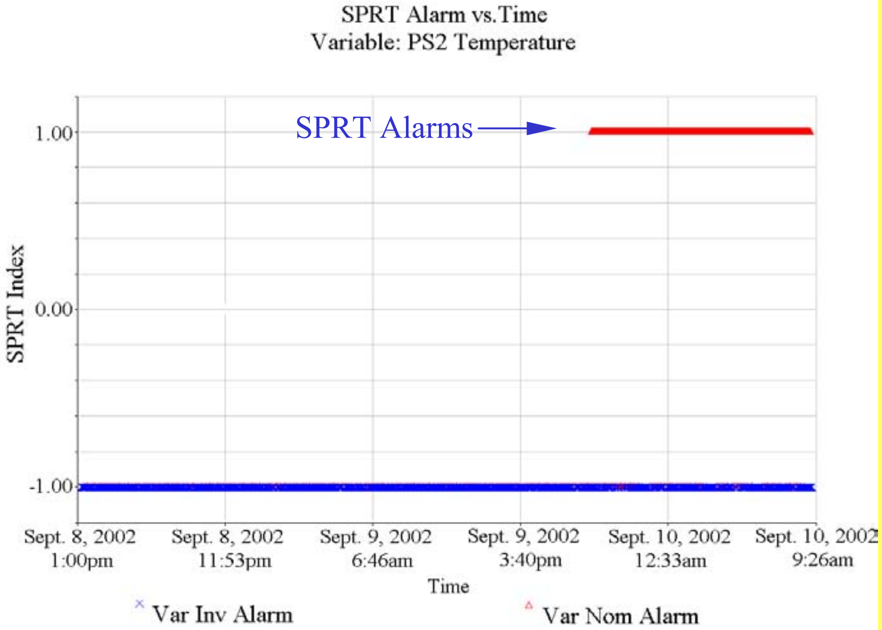
Sun's high-end servers contain hundreds of physical sensors (distributed board, module, and ASIC temperature sensors, voltages, and currents) that protect the system by detecting when a parameter is out of bounds, and then shutting down a component, system board, domain, or entire system.

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).



Observations and MSET Estimates vs. Time
Variable: PS2 Temperature



SPRT Alarm vs.Time
Variable: PS2 Temperature

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.

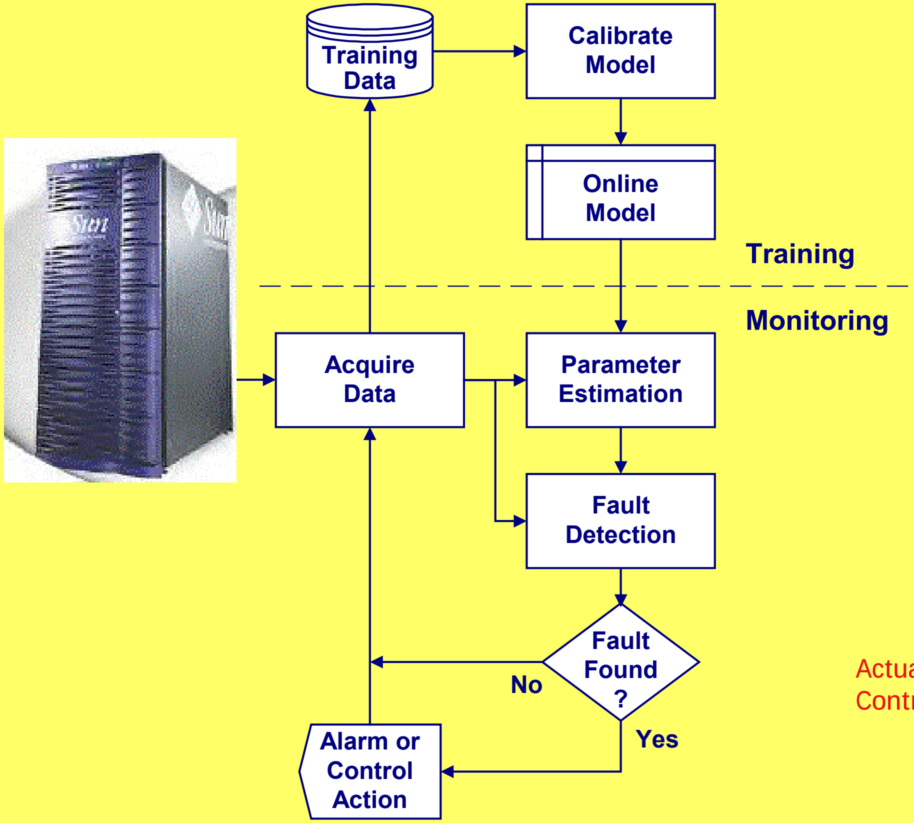
Realtime Sensor Validation: Benefits

- 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.

Sun System Dynamics Characterization and Control

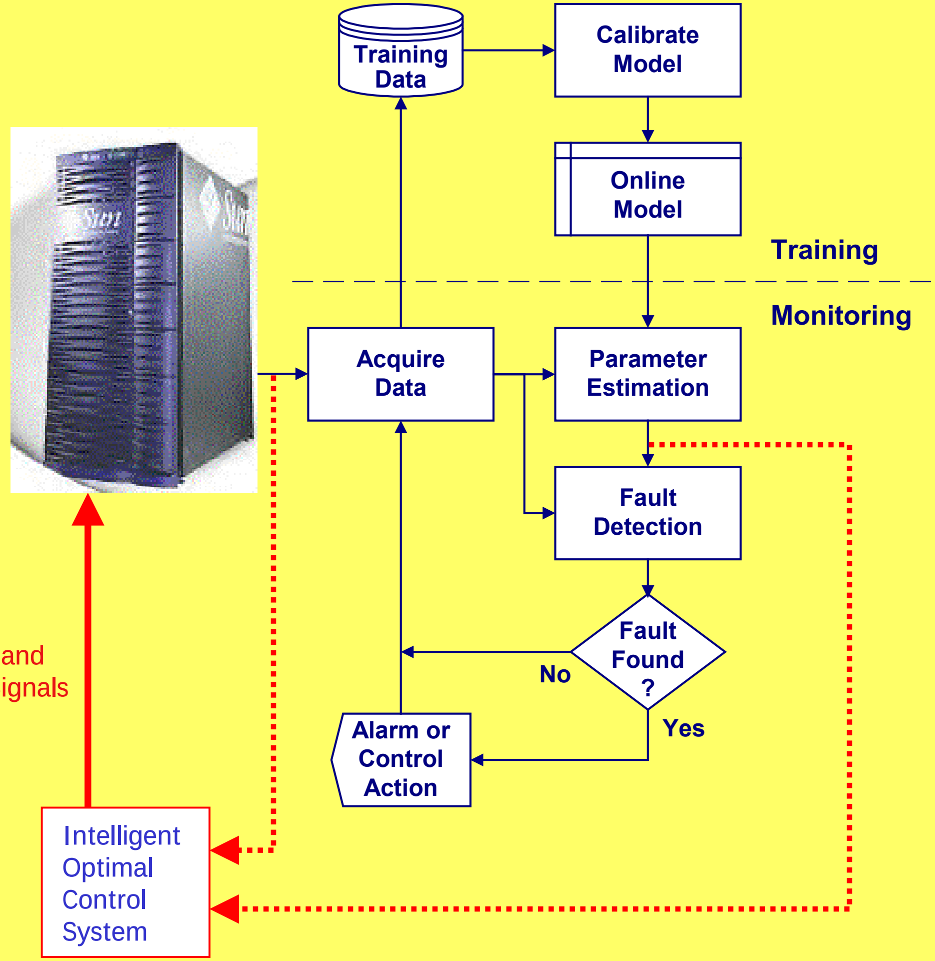
MSET Now:

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



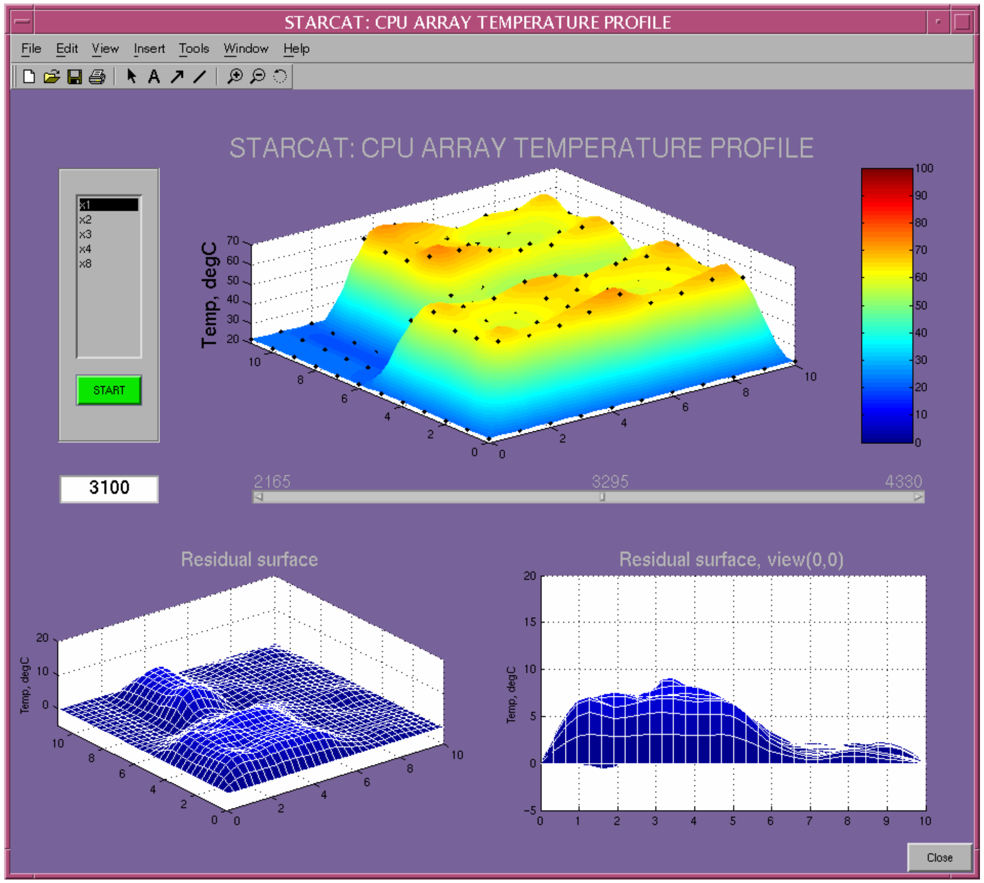
MSET For the Future:

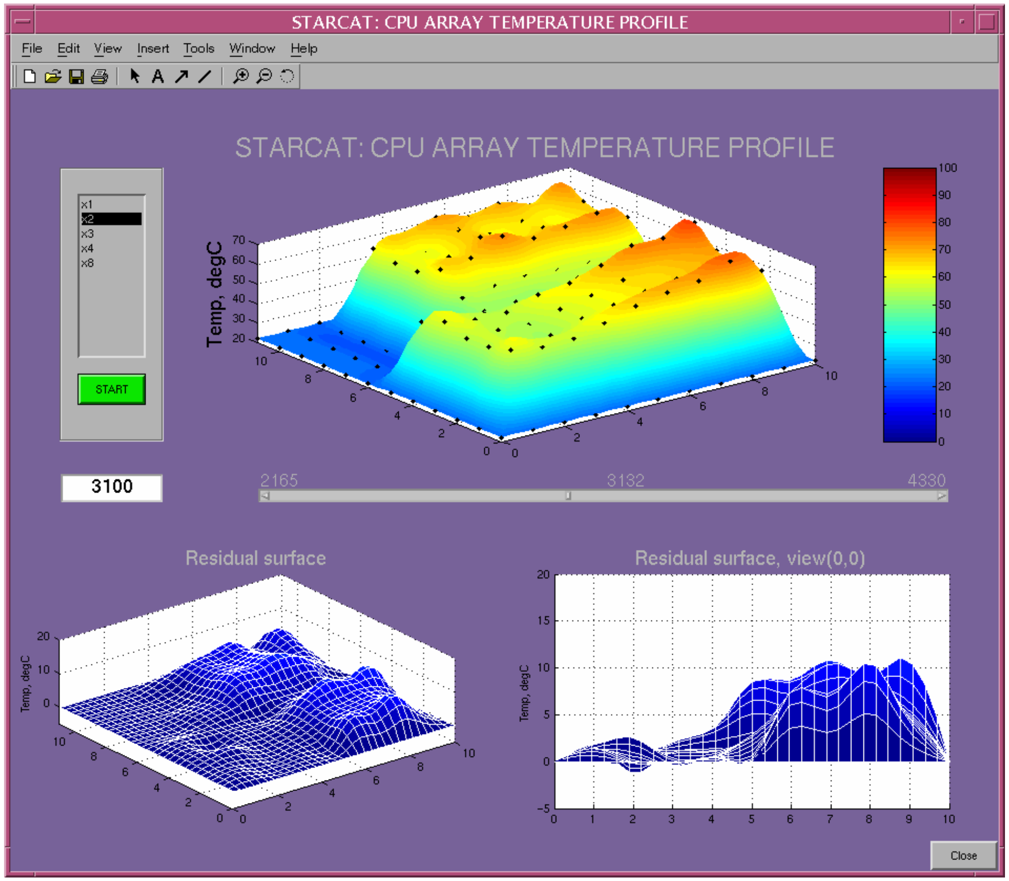
- ➢Realtime Stability Assurance (N1)
- ➢Dynamic Resource Provisioning
- ➢Self Healing and Realtime Fault Avoidance
- ➢Closed Loop Autonomic Control



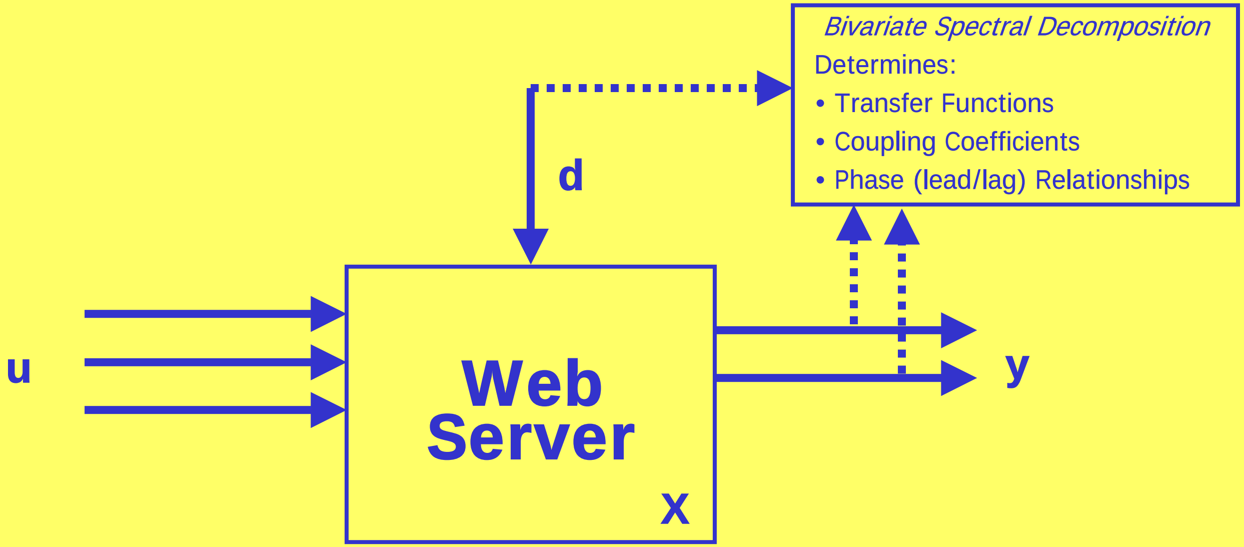
3D Dynamic Viewing Tool

- 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


STARCAT: CPU ARRAY TEMPERATURE PROFILE
File Edit View Insert Tools Window Help
STARCAT: CPU ARRAY TEMPERATURE PROFILE
x1
x2
x3
x4
x8
START
Temp, degC
3100
2165
3295
4330
Residual surface
Residual surface, view(0,0)
Temp, degC
Close


STARCAT: CPU ARRAY TEMPERATURE PROFILE
File Edit View Insert Tools Window Help
STARCAT: CPU ARRAY TEMPERATURE PROFILE
x1
x2
x3
x4
x8
START
Temp, degC
3100
2165
3132
4330
Residual surface
Residual surface, view(0,0)
Temp, degC
Close

New Approach for Dynamical System Characterization of Enterprise Servers



❖ Inputs, u	Typical user workload transactions
❖ Outputs, **y**	Performance metrics (e.g. loads, throughput, queue lengths, transaction latencies, etc.)
❖ States, **x**	Dynamical system states inferred via MSET
❖ Disturbances, **d**	Small-amplitude, multifrequency sinusoidal load perturbations
❖ **y = S[u,d]**	Bivariate spectral decomposition via the Normalized Cross Power Spectral Density (NCPSD) algorithm

CSTH Lab Console Displays For Coordination Of Dynamical System Characterization Experiments

Summary

- System telemetry harness provides “Black Box Flight Recorder”
- Can “roll back” signals to identify signatures that lead to crashes
- Observability mitigates complexity!
- Use MSET-based pattern recognition technology to interpret monitoring data on line, in real time (or can post-process archived telemetry data)
- Can enhance RAS and Quality of Service (QoS)
- Can mitigate No-Trouble-Founds (NTFs)
- Next generation system for N1 stability assurance, dynamic provisioning, & optimal control.

Bibliography

Recent RASCAL Pubs on Telemetry + Adv. Pattern Recognition for Enhanced Dependability of Complex Systems and Networks

"Frequency-Domain Pattern Recognition for Dynamic System Characterization of Enterprise Servers," K. Gross and K. Mishra, at the 2003 International Conference on Artificial Intelligence (IC-AI'03), Las Vegas, NV (June 23-26, 2003).

"Remote Analysis and Measurement of Software Systems," A. Porter, S. McMaster, A. Urmanov, L. Votta and K. Gross, 2003 Intn'l Conf. on Software Engineering (ICSE), Portland, OR (May 9, 2003)

"Spectral Decomposition of Performance Variables for Dynamic System Characterization of Web Servers," K. C. Gross, W. Lu, and K. Mishra, ***Proc. 2002 SAS Users Group International (SUGI 28)****,* Seattle, Wa. (Mar 30 - Apr 2, 2003).

"Proactive Detection of Software Aging Mechanisms in Performance-Critical Computers ," K. C. Gross, V. Bhardwaj, R. L. Bickford, ***Proc. 27th Annual IEEE/NASA Software Engineering Symp.***, Greenbelt, MD, (Dec 4-6, 2002.)

"Time-Series Investigation of Anomalous CRC Error Patterns in Fibre Channel Arbitrated Loops," K. C. Gross, W. Lu, and D. Huang, ***Proc. 2002 IEEE Int'l Conf. on Machine Learning and Applications (ICMLA),*** Las Vegas, NV (June 2002).

"Early Detection of Signal and Process Anomalies in Enterprise Computing Systems," K. C. Gross and W. Lu, ***Proc. 2002 IEEE Int'l Conf. on Machine Learning and Applications (ICMLA),*** Las Vegas, NV (June 2002).

"Advanced Pattern Recognition for Detection of Complex Software Aging Phenomena in Online Transaction Processing Servers," Karen J. Cassidy, Kenny C. Gross, and Amir Malekpour, ***Proc. Intnl. Performance and Dependability Symposium,*** Washington, DC, (June 23rd - 26th, 2002).

"Software Reliability and System Availability at Sun," K. C. Gross, ***Proc. IEEE 11th Intn'l Symp. On Software Reliability Eng. (ISSRE02)***, San Jose, CA (Oct 2000).

"Uncertainty Analysis for Multivariate State Estimation in Mission-Critical and Safety-Critical Applications," N. Zavaljevski and K. C. Gross, ***Proc. MARCON 2000,*** "Maintenance and Reliability in the 21st Century", Knoxville, TN, (May 7-10, 2000).

"Support Vector Machines for Multivariate State Estimation," N. Zavaljevski and K. C. Gross, ***Proc. ANS Intnl. Topical Mtg. On "Advances in Reactor Physics, Mathematics, and Computation into the Next Millennium,"*** Pittsburgh, PA, (May 7-11, 2000).