May 12, 2009

Stanford University / NASA

Prognostics in Data Centers

Aleksey Urmanov, Ph.D.
Anton Bougaev, Ph.D.

aleksey.urmanov@sun.com
anton.bougaev@sun.com

San Diego Physical Sciences Center

- **RASCAL** – Corporate resource for system reliability investigation and reliability technology development

> 30+ investigators; 12,000 sq. ft.
> Reliability Physics
>> Data driven diagnostics and root cause analysis
> Real Time Prognostics
>> Pattern recognition for signal analysis
> Technology Evaluation
> University and industry collaborations

Motivation: Failures Are Costly

- Customers suffer millions $/hr from downtime, depending on the industry
- Conventional approach for designing fault-tolerant and highly available systems: redundancy
- Redundancy has several drawbacks:
  - Expensive: replication of hardware components
  - Complex: additional hardware/software required to implement failover, voting, state replication, checkpointing, etc.
  - Difficult to validate: realistic failure scenarios often difficult to mimic, making fault tolerance functionality one of the more undertested pieces of the system

Anticipate Failures Before They Occur

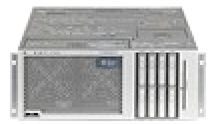
- Add observability to system
- Look for leading indicators of failure
- Take corrective actions to avoid failure
- Mitigate impact on executing applications

Computer Systems Observability

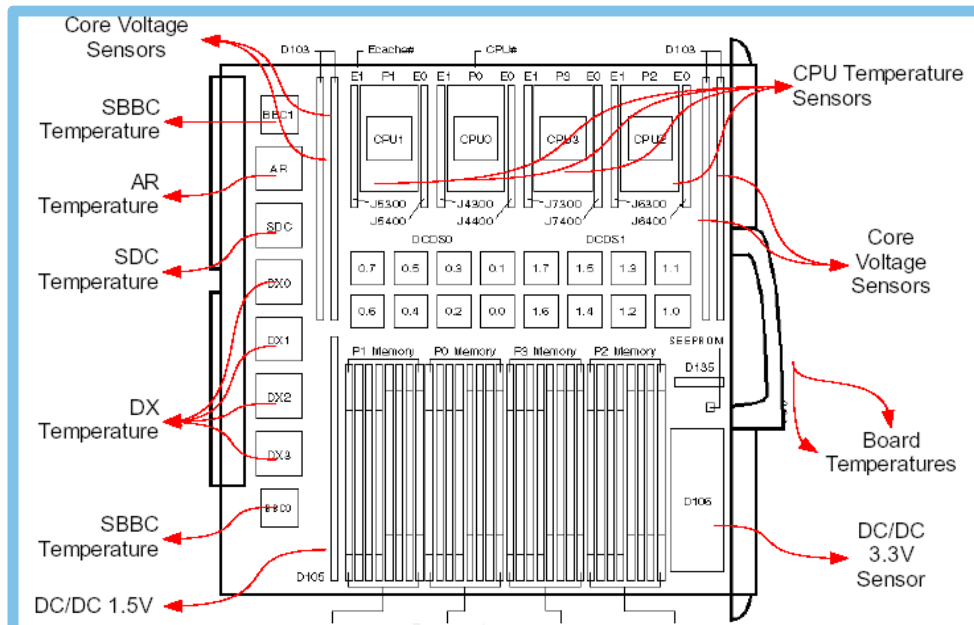
- Sensor data continuously sampled from sensors
  - > Data from physical sensors:
    - > Temperatures, voltages, currents, vibration, humidity, acoustics
  - > Data from logical sensors:
    - > CPU utilization, swap activity, queue lengths, HTTP GET response times, etc.
- Additional information about system state:
  - > Machine configuration
  - > Software patch levels
  - > Error messages or alerts
  - > SNMP traps

Sensor-Rich Computer Systems

- Entry Servers / CMT Servers / Desktop Systems
  - > Dozens of Sensors

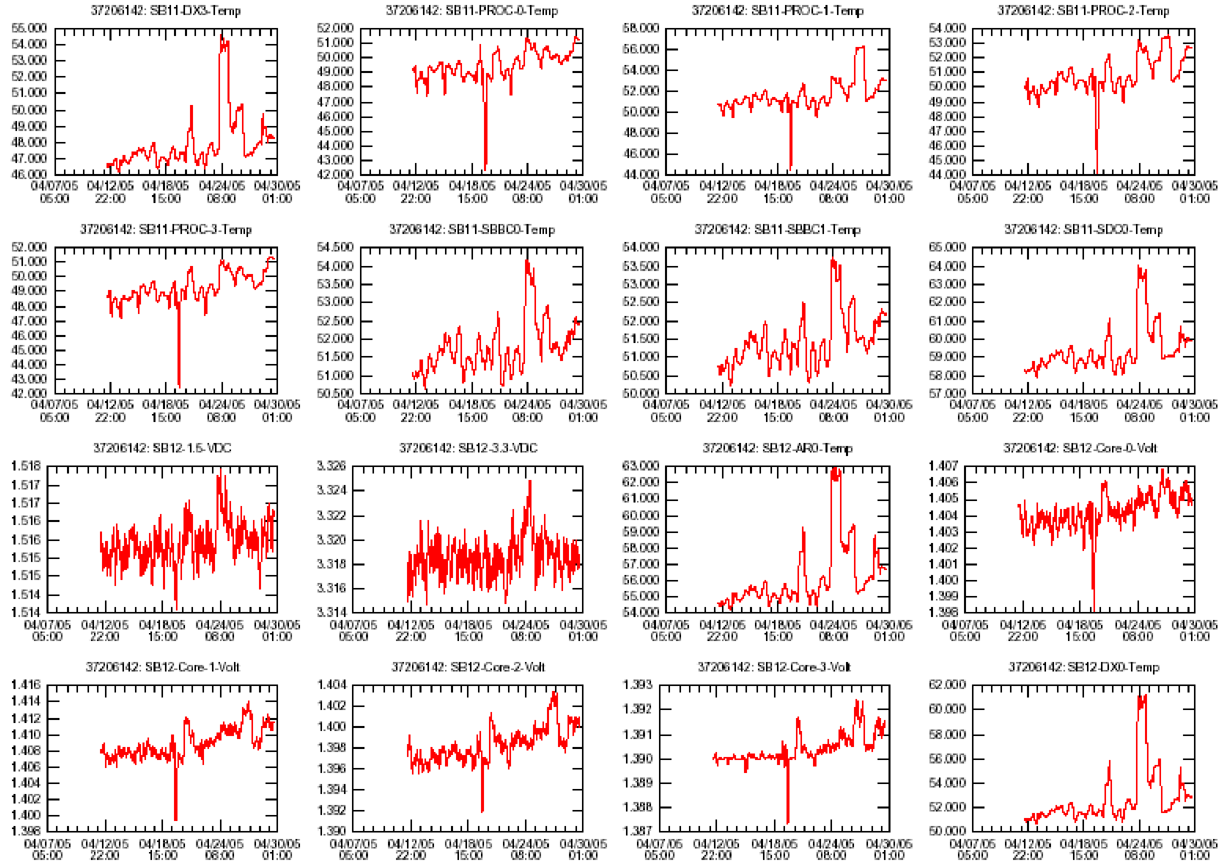


- Mid-Range Servers / Storage Systems
  - > Hundreds of Sensors

- High-End Servers
  - > Thousands of Sensors


Examples of Sensor Signals



Benefits of Monitoring Sensors

- Greater observability into system behavior:
  - > Insight into how system state evolves over time: not just a "point-in-time" snapshot
- Ability to characterize a correctly-behaving system:
  - > Distinguish between incorrectly-behaving system
- Foundation for proactive fault monitoring
  - > A.k.a. predictive analytics, electronic prognostics
- Preventive maintenance, improved RCA and reduced no-trouble-found (NTF) rates
  - > Analogous to airline black-box flight recorder
  - > Predictive and reactive failure information

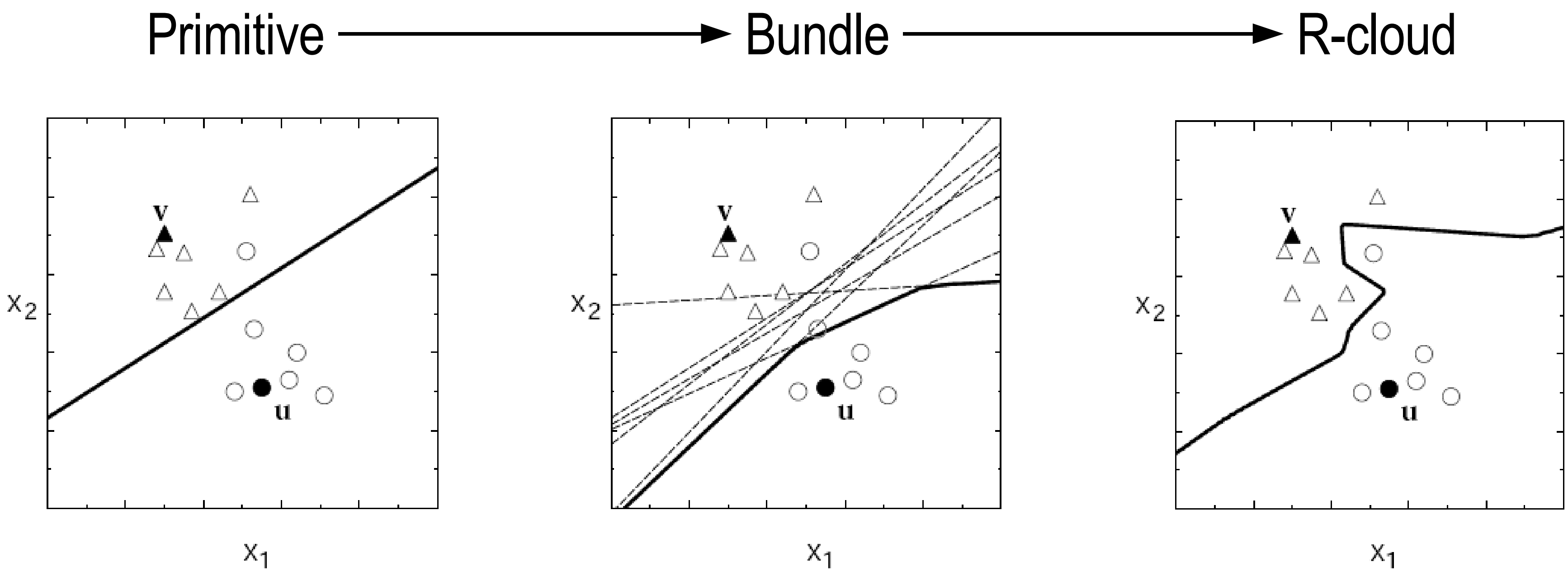
Analysis of Sensor Signals

- Save it to stable storage as “black-box flight recorder” for future root cause analysis
- Process the samples online for proactive fault monitoring
- Mine archived sensor data to discover trends in population and looming quality issues
- Infer system condition for RUL estimation

Software Aging and Rejuvenation

- Software aging: long-running applications degrade over time:
  - > Memory leaks, unreleased file locks, accumulation of unterminated threads, data corruption/roundoff accrual, file space fragmentation, shared memory pool latching, thread stack bloating and overrun
- Software rejuvenation: periodically “cleanse” internal system state to prevent software aging effects:
  - > Flush stale locks, reinitialize application components, preemptive rollback, defragment disk, therapeutic reboot (primarily Windows platform)

R-Cloud Classifier Construction for
Runaway Process Detection



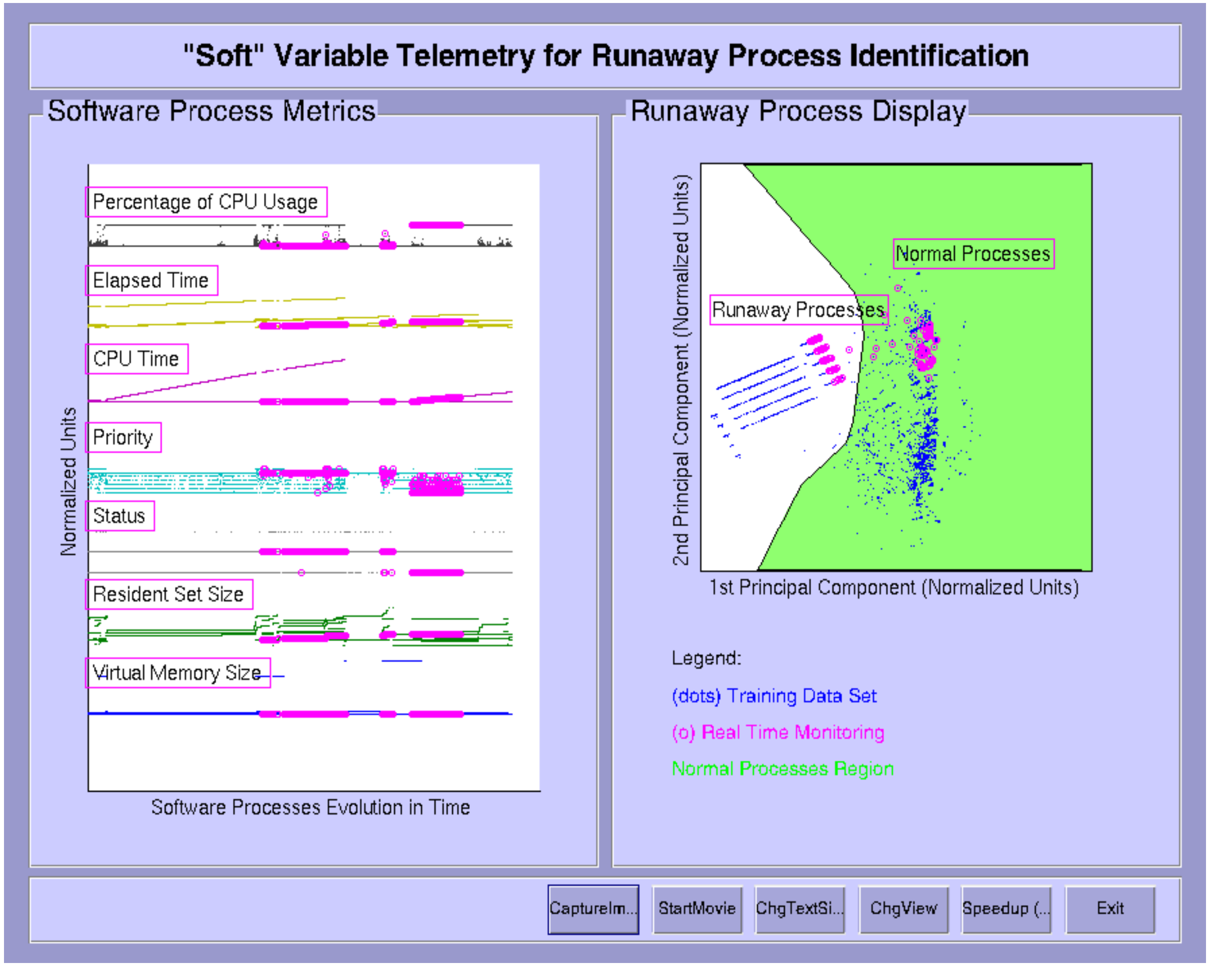
$$\rho_l(\mathbf{x},\mathbf{u},\mathbf{v},0) = d(\mathbf{x},\mathbf{v}) - d(\mathbf{x},\mathbf{u})$$

$$\beta^k(\mathbf{x},\mathbf{u}) = \bigwedge_{\mathbf{v}\notin\mathbf{C}_k} \rho(\mathbf{x},\mathbf{u},\mathbf{v})$$

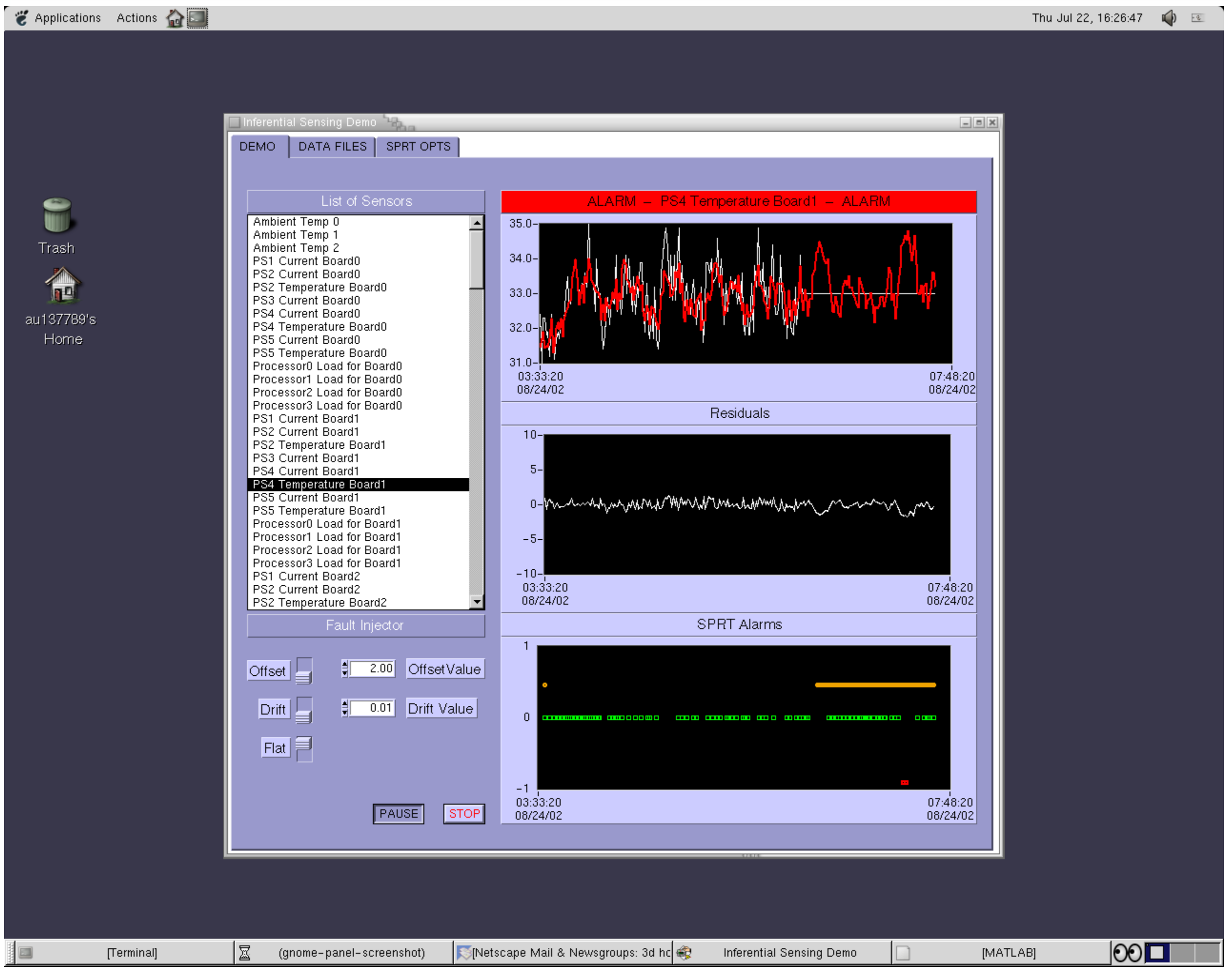
$$R_1(\mathbf{x}) = \bigvee_{\mathbf{u}\in\mathbf{C}_1} \bigwedge_{\mathbf{v}\in\mathbf{C}_2} \rho(\mathbf{x},\mathbf{u},\mathbf{v})$$

- R-cloud construction using a linear separating primitive in two dimensions

Runaway Process Detection



Inferential Sensors

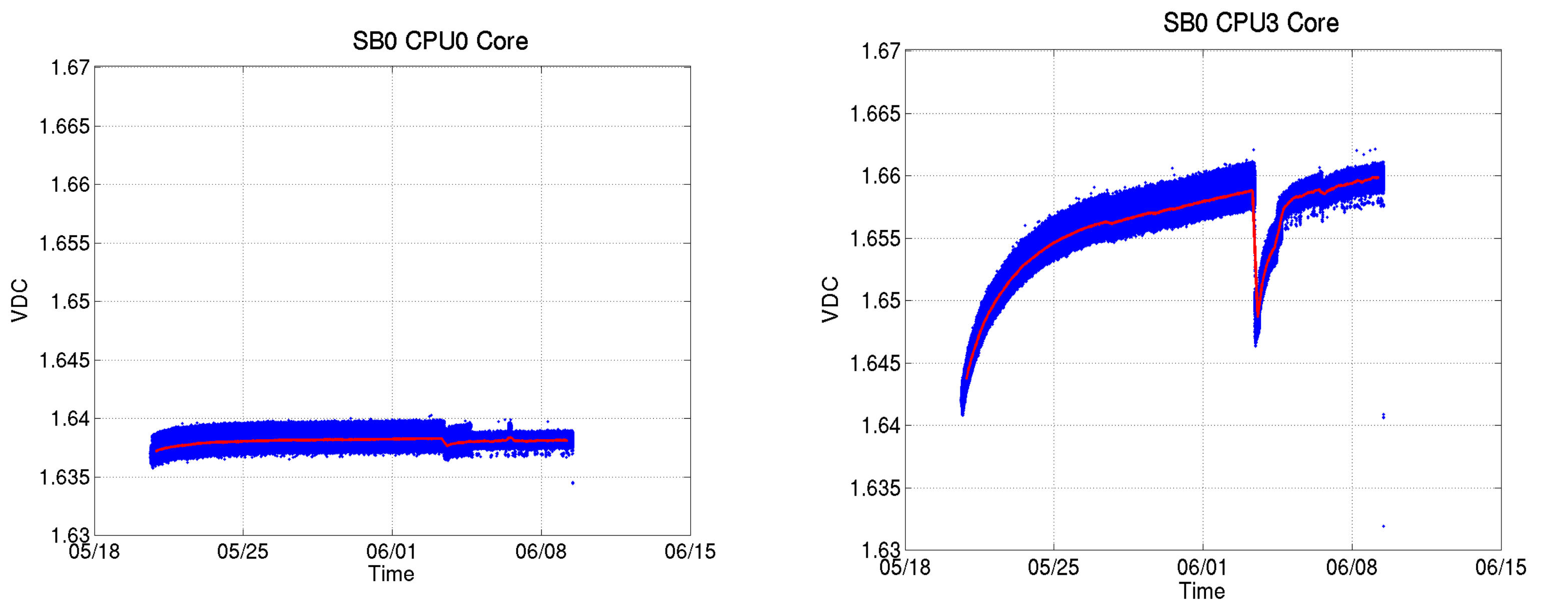


- Enables prognostics for legacy systems with limited # of physical sensors
- Analytical redundancy for fault detection
- Cost savings in low margin systems
- Improved RAS features

Connector Assembly

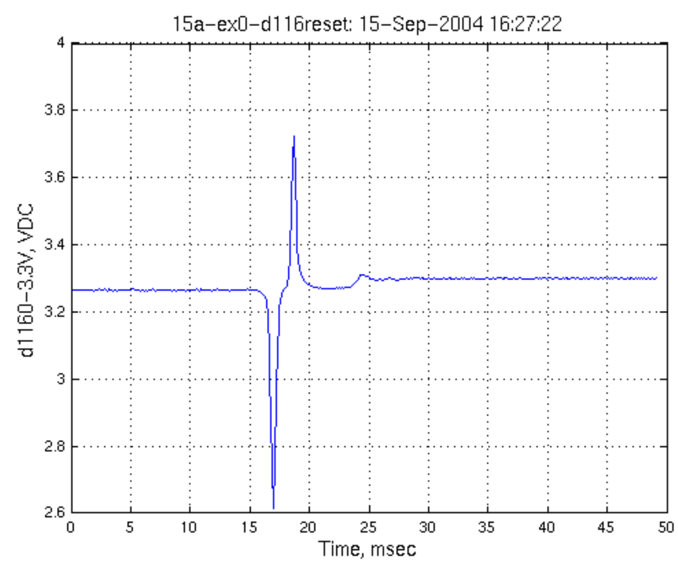
Faulty Connector Signature

Voltage is drifting upwards for faulty connectors.



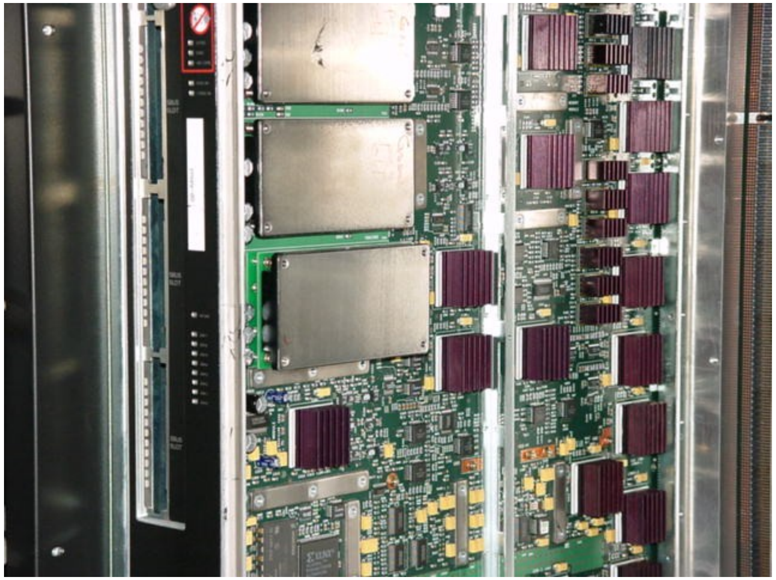
Cracking Solder Joints

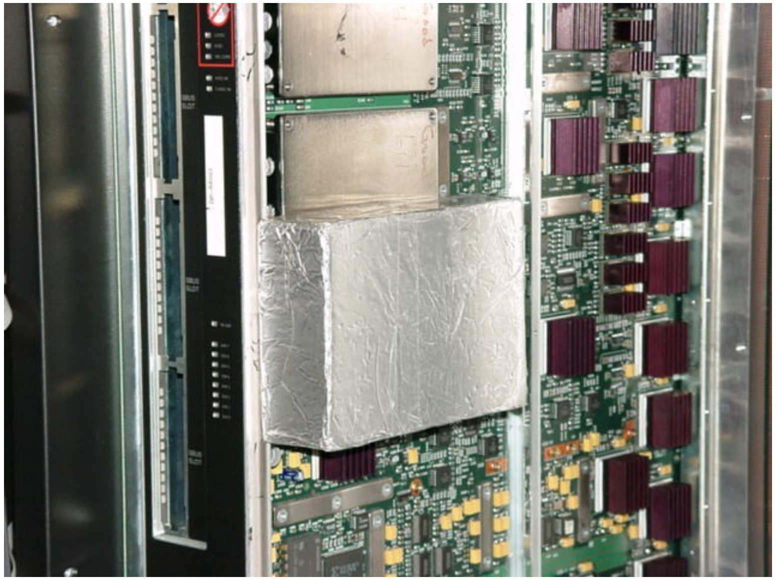
Cracks open and close with temperature and load swings, causing power supply resets. During a reset, a large transient may causes a system failure.



Power Supply Example

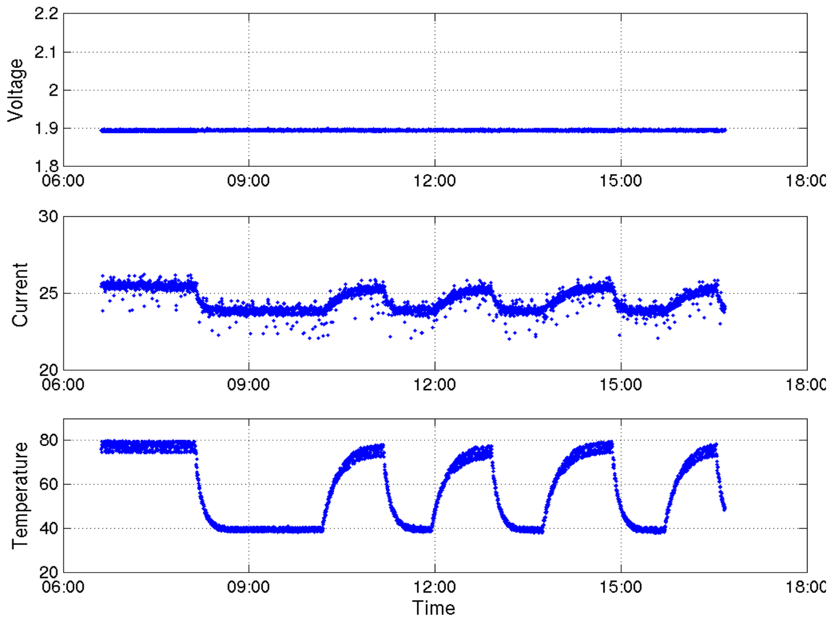
In-situ thermal cycling



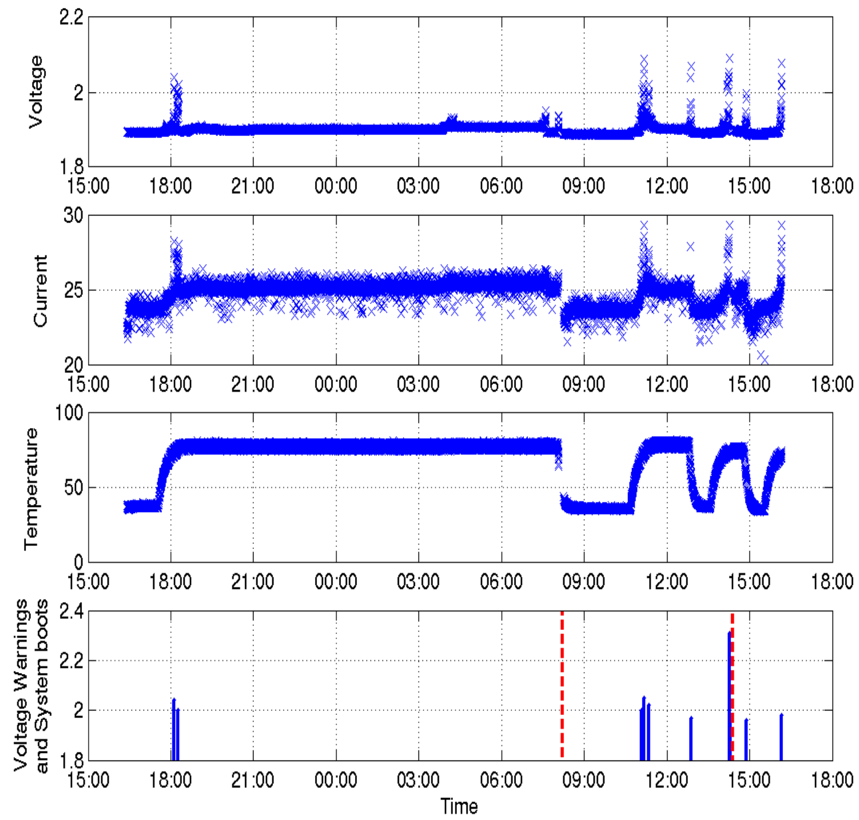


Signatures of a Good and Bad PSU's

Output voltage of a good power supply is stable during thermal cycling.



Output voltage of a bad power supply is unstable during thermal cycling.



Benefits to the Customer

- Improved availability:
  - > Isolate suspect hardware and software processes from application to avoid failures
- Condition-based maintenance:
  - > Schedule maintenance for system only when needed (not preventive, time-based maintenance that causes frequent interruptions)
  - > Know that you're servicing the right components
- Intelligent provisioning of resources:
  - > Telemetry gives you observability into how hardware is used, how power is consumed, etc.

Benefits to Sun

- Eliminating “No Trouble Found” events
- Improved product quality
- Lower cost of defects
- “Just in time” spares
- Data mining: early indications of quality issues in the field
- Higher customer satisfaction

References

- Bougaev, Urmanov, “ R-functions Based Classification for Abnormal Software Process Detection,” Lecture Notes in Computer Science, Volume 3801/2005, Computational Intelligence and Security, Springer Berlin / Heidelberg (PDF)
- Bougaev, Urmanov, “R-functions for Analytical Representation of Classification Boundary,” JPRR, To appear, (PDF)
- Urmanov, “Electronic Prognostics for Computer Servers,” Reliability and Maintainability Symposium, 2007. RAMS '07, 22-25 Jan. 2007 (PDF)
- Bougaev, “Pattern Recognition Based Tools Enabling Autonomic Computing,” Autonomic Computing, ICAC, 13-16 June 2005 (PDF)
- Usynin, “Model-Fitting Approaches to Reliability Assessment and Prognostic Problems,” JPRR, Vol 1, No. 1, 2006 p.32-36 (PDF)
- Hines, Garvey, “Development and Application of Fault Detectability Performance Metrics for Instrument Calibration Verification and Anomaly Detection,” JPRR, Vol 1, No. 1, 2006, p.2-15 (PDF)