Supercomputer Reliability and Mitigation

Ron T. Ogan  
Dept. of Electrical & Computer Engineering  
Jackson State University  
Jackson, MS. 39217 USA  
ron.t.ogan@jsums.edu

Paul E. Watson  
Dept. of Electrical & Computer Engineering  
Jackson State University  
Jackson, MS. 39217 USA  
paul.e.watson@jsums.edu

Marshall D. Boyette  
Dept. of Electrical & Computer Engineering  
Jackson State University  
Jackson, MS. 39217 USA  
marshall.d.boyette@jsums.edu

Khalid H. Abed  
Dept. of Electrical & Computer Engineering  
Jackson State University  
Jackson, MS. 39217 USA  
khalid.h.abed@jsums.edu

Abstract — Reliability, Availability and Serviceability (RAS) are key to high performance computing. Because of the relatively high costs of supercomputers and the required support needed to operate these systems, a holistic approach must be taken to assure system reliability as defined: Reliability is the probability that a material, component, or system will perform its intended function under defined operating conditions for a specific period of time. Extension of the computer system lifetime will be achieved through the implementation of built-in-test-equipment (BITE) monitoring that will provide the required feedback for optimum environmental controls. Reliability improvements of MTBF of 25% have been predicted based upon redesign of the microprocessor cooling system to reduce average case temperatures from 90 °C to 75 °C

Keywords—Supercomputer, Reliability, Availability, Serviceability (RAS), thermal analysis

I. INTRODUCTION

Supercomputers, the heart of High Performance Computing Centers, are designed with the latest processor technology to provide performance and scalable systems that integrate advanced interconnections to achieve the high-bandwidth performance to achieve incredible mathematical calculation rates. Advanced design methodology is utilized to network the rack assemblies with distributed memory to minimize system latency yet provide scalability and maintainability. Incredible performance has been achieved across computing network systems with distributed memory systems to provide programmers with global access to large memory arrays for parallel processing applications to support the most demanding global communication patterns. This paper addresses technological aspects of supercomputer technology that are predominantly being operated at U.S. Federal Laboratories to solve very large complex problems efficiently. The supercomputers are housed in a large environmentally controlled workspace with monitored support for cooling and power equipment to guarantee reliability, availability and serviceability (RAS) of >99%.

Table 1. Compute Module Faults and Mitigation

<table>
<thead>
<tr>
<th>Defect</th>
<th>Design Mitigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory soft failure</td>
<td>Provide electromagnetic shielding to extent possible</td>
</tr>
<tr>
<td>Memory hard failure</td>
<td>Provide fault tolerant systems</td>
</tr>
<tr>
<td>Capacitor drift</td>
<td>Design Center to accept drift</td>
</tr>
<tr>
<td>Power supply spikes</td>
<td>Power conditioning</td>
</tr>
<tr>
<td>Board impedance shifts</td>
<td>Design Center to accept shift</td>
</tr>
<tr>
<td>Microprocessor failure</td>
<td>Provide optimum cooling and monitoring</td>
</tr>
<tr>
<td>Microprocessor degradation</td>
<td>Provide optimum cooling and monitoring</td>
</tr>
</tbody>
</table>

A. Abbreviations and Acronyms

CPU—Computer Processing Unit

Reliability, Availability, Serviceability (RAS)

Reliability is a function of time that expresses the probability at time t+1 that a system is still working, given that it was working at time t. Availability is the measure of how often the system is available for use (such as a system’s up-time percentage). Availability and reliability may sound like the same thing, but it is worth noting that a system can have great Availability but no Reliability. An internet router is a good example of this; it stores no state data. It is one of the few systems wherein data loss is acceptable, as long as high availability is maintained. Availability is typically described in nines notation. For example, 3-nines means 99.9%. Obtaining 5 nines or 99.999% availability is an ambitious goal for many vendors when producing hardware and software modules as shown in Table 2 [8].

<table>
<thead>
<tr>
<th>Availability</th>
<th>9s</th>
<th>Downtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>One</td>
<td>36.5 days/year</td>
</tr>
<tr>
<td>99%</td>
<td>Two</td>
<td>3.65 days per year</td>
</tr>
<tr>
<td>99.9%</td>
<td>Three</td>
<td>8.76 hours/ year</td>
</tr>
<tr>
<td>99.99%</td>
<td>Four</td>
<td>52 minutes/ year</td>
</tr>
<tr>
<td>99.999%</td>
<td>Five</td>
<td>5 minutes/ year</td>
</tr>
<tr>
<td>99.9999%</td>
<td>Six</td>
<td>31 seconds/ year</td>
</tr>
</tbody>
</table>

Serviceability is a broad definition describing how easily the system is serviced or repaired. For example, a system with modular, hot- swappable components would have a good level of serviceability. Technology may limit unscheduled downtime by having systems with redundant dynamic reconfiguration, constant system monitoring and resource management. [5]

Mean Time between Failures (MTBF) is the average (expected) time between two successive failures of a component. It is a basic measure of a system’s reliability and availability and is usually represented as units of hours.

- Mean Time to Repair (MTTR) (or Recover) is the average (expected) time taken to repair a failed module. This time includes the time it takes to detect the defect, the time it takes to bring a repair man onsite, and the time it takes to physically repair the failed module. Just like MTBF, MTTR is usually stated in units of hours. The following equations illustrate the relationship of MTBF and MTTR with reliability and availability [4] and [5].

\[
\text{Reliability} = e^{-\frac{\text{Time}}{\text{MTBF}}}
\]

\[
\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}
\]

The following conclusions can be reached based on these formulas [4]:

* The higher the MTBF value is, the higher the reliability and availability of the system.
* MTTR affects availability. This means if it takes a long time to recover a system from a failure, the system is going to have a low availability.
* High availability can be achieved if MTBF is very large compared to MTTR.

Sandia Laboratories has suggested new Energy-Reliability (EneRel) metrics that are noted as reliability-aware by the use of a subscript r, for example $E_{\text{rD}}$ and $E_{\text{rD2}}$. At a very high level, EneRel can be thought of as: [9]

\[
\text{EneRel} = E_{\text{rt}} + (\text{Efail recov} \cdot (p(\text{fail}) + p(\text{failadd} \cdot \text{rt})))
\] (1)

The Efail for an energy saving technique under test differs from the Efail of the baseline implementation since the baseline technique’s $p(\text{failadd} \cdot \text{rt})$ is zero. Figures 1a, 1b and 1c show the impact that failures can have on energy consumption given different failure rates and increases in runtime for evenly distributed failures over the runtime. For large node/socket counts the amount of energy that can be lost due to failures in the increased runtime period is non-negligible as HPC component counts keep rising and will continue to rise for Exascale computing and beyond.
Reliability must improve along with rising component counts to make future systems viable. Despite improvements in component reliability, reductions in whole system reliability are still likely to occur. The forecasts in Figure 1 a,b,c show that reliability mechanisms will no longer play a minor role in energy consumption. EneRel proves useful for illustrating that sometimes spending slightly more energy to finish a job faster, thereby reducing the probability that a failure occurs during the job’s runtime, may in some cases actually result in lower energy consumption. The exploration of the effect of reliability on energy consumption is made all the more important by the fact that reliability can be reduced when using energy saving techniques due to thermal cycling [10] and lowered operating voltages result in an increase in soft faults [11].

Figure 1 (a) Energy overhead due to reliability for varying percentage increases in runtime for 3 year, MTBF

Figure 1 (b) Energy overhead due to reliability for varying percentage increases in runtime for 5 year, MTBF

III. THE US ARMY ENGINEER RESEARCH & DEVELOPMENT CENTER (ERDC) INFORMATION TECHNOLOGY LABORATORY (ITL) FACILITY

The US Army Engineer Research & Development Center (ERDC) Information Technology Laboratory (ITL) facility has 10,000 square feet of air conditioned space with four feet deep false flooring for utilities. The computer systems are Cray Sonexion (12 processors) and experimental LEO. There is a section of UNCLASSIFIED servers/processors and a much larger set of CLASSIFIED servers/processors. There is a section of Silicon Graphics Inc. (SGI) computers with speeds of 1.2 or 4.6 Penta Flops. One petaflops is equal to 1,000 teraflops, or 1,000,000,000,000,000 FLOPS. Computer systems cost $25M and are operated for 4 years before replacement. ERDC HPCC has HPC computers that are ranked 16th fastest in the world. User computers have access to 12 Penta-Bytes of memory. HPCC’s do not charge for use of the computer 1.9 Billion core hours per year, but applications must be for US government or non-profit (University) entities. Online access is available at www.uit.hpc.mil Backup generators supply 2.4 Mega Watts of power and there are extensive cooling systems to support operations. Emerson power systems and Data Direct 3.5” disk drives were observed. Cray uses a proprietary interconnect circuitry but Silicon Graphics Inc. (SGI) uses open architecture. Both Cray and SGI compute module have large copper cooling heat transfer plates. Cooling water flow rates provided in channels to maintain the required exchange rate and return the heated water to cooling towers before the return cycle to the HPC system.

IV. PERFORMANCE DEGRADATION

The typical reliability curve is shown below in Figure 2. The initial failure region depends upon the screening and quality of the components and workmanship. The initial failures may be high due to latent defects that were introduced during the manufacturing or assembly process and not caught during the testing phase. The center useful life region is where random, contact hazard occurs. The end
of life region is where mechanical devices such as air cooling fans fail and cause over heating of the electronic and electro migration of device junctions which degrade performance or result in catastrophic failures.

![Reliability Curve – "typical bathtub shape"](image)

Figure 2 – Failure rate as a function of time at a single environmental condition (temperature, humidity, salt atmosphere, etc.) curve [1]

V. THERMAL CONTROL AND HEAT DISSIPATION IS CRITICAL FOR MICROPROCESSOR PERFORMANCE AND LIFETIME

One dimensional conduction requires that the junction and case nodes be isothermal, however, neither the die nor the lid are isothermal; therefore, this thermal model requires corrections for heat spreading from the chip to the heat sink. Theta jc is the temperature rise from the device junction to the case.

\[
\theta_{jc} = \frac{(T_j - T_c)}{\text{Power}} \tag{2}
\]

Intel® Xeon® microprocessors have a specified Theta jc of 0.3 °C per Watt for maximum applied power of 90 Watts. Rearranging the equation and substituting the specification values results in \( T_j = T_c + 27 \) °C; therefore, the junction temperature is a direct function of the cooling water temperature with these results predicted: \( T_j = 62 \) °C for 35 °C Tc, \( T_j = 57 \) °C for 30 °C Tc, and \( T_j = 52 \) °C for 25 °C Tc, neglecting losses across the Thermal Interface material (TIM).

A simplified one dimensional heat flow model was proposed by Bar Cohen as shown in Figure 5 [6] where Theta junction-to-case jc, Theta case-to-sink cs and Theta sink-to-ambient sa or to a reservoir when liquid cooling is used.

![Figure 3 Single axis heat flow model [6]](image)

Figure 3 Single axis heat flow model [6]

Figure 4 shows the actual heat flow from the junction to case then lateral through the case lid and spreading into the external heat sink for dissipation through the combination of heat conduction into the cooling liquid or convection to the surrounding air flow created by cooling fans. [7]

![Figure 4 heat flow from flip chip ball grid array microprocessor to the heat sink](image)

It is clear that control of the microprocessor junction temperature is critical to performance and minimizing degradation due to device operation and elevated temperatures near the maximum limits.

VI. TECHNOLOGICAL AND PROCEDURAL METHODS TO EXTEND SUPERCOMPUTER USEFUL LIFE

The microprocessor is the heart of the high performance computing system that generates thermal energy during operation. There would be performance degradation unless adequate heat transfer cooling is provided which is normally a case temperature of 0° to 70°C. Thermal case temperature regulation is provided by thermal conduction through the heat sink and transfer to external cooling water or other exchangers to an external support facility. Cooling air is provided by keeping the room temperature in the 65°-70°F range and providing fan air flow rates to change the typical 72 cubic foot equipment cabinet air out every 5 minutes or 14.4 CFM flow rate.

A standard model used by the industry is the Sum-of-failure-rates (SOFR) model, which makes two assumptions to address this problem: (1) the processor is a series failure system – in other words, the first instance of any structure failing due to any failure mechanism causes the entire processor to fail; and (2) each individual failure mechanism has a constant failure rate (equivalently, every failure mechanism has an exponential lifetime distribution). Microprocessor failures are accelerated by temperature, duty cycles and electro and stress migration effects resulting from the AL and Cu metal interconnections and junctions.
The FIT value targeted by reliability qualification, FIT\textsubscript{target}, is a standard. Currently, processors are expected to have an MTTF of around 30 years—this implies FIT\textsubscript{target} is around 4000 [12].

From MIL-STD-217Fn2, paragraph 5.1 the failure rate $\lambda_p$ for microprocessors is estimated by

$$\lambda_p = (C_1 \pi_T + C_2 \pi_E) \pi_0 \pi_L \text{ Failures/ } 10^6 \text{ Hours per million hours of operation.}$$

The most critical parameter is the $\pi_T$ the effects of elevated temperature operation. From the MIL-STD-217Fn2, Section 5.2 and 5.8 tables for microprocessor CMOS technology shows values of $\pi_T$ at 90 °C = 1.1 and $\pi_T$ at 75 °C 0.71 which will directly result in an approximately +25% increase in MTTF.

The lower temperature operation may be achieved by implementing these changes in the High Performance Computing cooling systems:

- Install temperature monitoring and coolant flow rate sensors at the input of rack assembly
- Redesign the parallel and series routing of coolant to achieve a maximum exchanger temperature at the microprocessor heat sink of 70 °C
- Monitor and control processor work flow to avoid peak duty cycles without providing additional cooling.

Performance modeling and testing has shown that Intel® microprocessors (at the very least Sandy Bridge, Ivy Bridge, and Haswell) can run at their maximum Turbo Boost frequency all the way up to 100 °C. While there may be a tiny performance difference between a microprocessors running at 30 °C and one running at 95 °C, our testing has found that the difference is miniscule. In fact, even after running benchmarks dozens of times the difference is so small that it is essentially nonexistent.

Puget Systems recently performed tests to measure the effects on operating frequency when a microprocessor core temperature approaches 100 °C. The Intel Core i7 4790 microprocessor was cooled it with a Gelid Silent Spirit Rev. 2 cooler that was connected to a manual PWM fan speed controller. By running Linpack (a benchmark widely used in the scientific community) and slowly dialing the fan speed down in careful increments, to allow the microprocessor to overheat by incremental amounts. At each cooling increment a log of the Linpack benchmark results were monitored as well as using CoreTemp to record the microprocessor core temperature and frequency, as shown in Figure 5 [1]. Since the Intel microprocessor thermal limit is 100 °C, the amount of overheating occurs when the core temperature was running at > 99 °C.

The testing showed that while the minimum core load frequency started to drop as soon as the core hit 100 °C, the average microprocessors frequency didn't drop by more than 0.1GHz until the core was overheating more than 30% of the time. In fact, Intel microprocessors are surprisingly good at being able to handle large thermal changes with only a small reduction in the average frequency. However, for high performance computing centers, the frequency changes and subsequent performance degradation could extend run times significantly and also reduce the component lifetimes similar to degradation results from accelerated life testing at elevated temperatures that shortens the useful lifetimes.

Intel® server-board-s2600jf with Xenon® processors, are used in high performance computing applications and as shown in Figure 6, may have two or more microprocessors with heat sinks to dissipate the thermal energy. The Mean Time Between Failures (MTBF) is 25,000 hours, 100,000 hours or 250,000 hours depending upon the operating environment and physical configuration.
Figure 6 – Intel® server-board-s2600jf with Xenon® processors that are used in High Performance Compute Modules

VII. CONCLUSIONS AND FUTURE PLANS

Reliability, Availability, Serviceability (RAS) are key factors to consider in improving the performance and extending the lifetime of high performance computer center systems. The Mean Time Between Failures (MTBF) and Mean Time To Repair (MTTR) are critical to system availability. Modular hot swappable compute modules, cooling fans and even redundant systems may be required to meet program operating objectives.

Dynamic system monitoring similar to an electrical power generation facility may be required to detect and correct support equipment out-of-specification operation to minimize performance degradation of the computing system.

Further work with the temperature control facility and the compute module supplier RAS may be needed to measure and characterize the relationship of performance degradation as a function of cooling systems under several operating loads.

Extension of the computer system lifetime will be achieved through the implementation of built-in-test-equipment (BITE) monitoring that will provide the required feedback for optimum environmental controls.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Defense High Performance Computing Modernization Program under the U.S. Army Engineer Research and Development Center (ERDC) contract number W912HZ-15-2-0001 entitled: “Strategic Cyber Science, Warfare, Security, Application Development and High Performance Computing Research and Development,” and the research project is entitled: “Investigating High Performance Heterogeneous Computing with Advanced Architectures,” and in part by Army Research Office HBCU/MSI contract number W911NF-13-1-0133 entitled: “Exploring High Performance Heterogeneous Computing via Hardware/Software Co-Design.”

REFERENCES