PNNL-16271

Energy Smart Data Center (ESDC) Phase I Reliability and Uptime Report

Isothermal Systems Research, Inc.

Tahir Cader
Harley J. McAllister
Levi Westra

September 20, 2005

Prepared for the National Nuclear Security Administration (NNSA)
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Page Break

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

**Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,**

P.O. Box 62, Oak Ridge, TN 37831-0062;ph: (865) 576-8401fax: (865) 576-5728email: reports@adonis.osti.gov

**Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161 ph: (800) 553-6847 fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm**

Table of Contents

ESDC PHASE I RELIABILITY AND UPTIME REPORT

1. EXECUTIVE SUMMARY

Current trends in computing are leading to more powerful and larger HPC cluster systems. These trends are driving the need for larger data center facilities with increased power consumption and lacing increased stress on facility power and cooling systems. To investigate ways to help reverse these trends, Pacific Northwest National Laboratory (PNNL) and Isothermal Systems Research, Inc. (ISR) initiated a muli-phased program titled the Energy Smart Data Center Program with the focus on three key areas:

1. Driving adoption of energy efficient cooling solutions
2. Demonstrate computing architectures that enable dense computing solutions
3. Investigate and improve system reliability and uptime

The primary focus for this report is item number 3, system reliability and uptime. Phase 1 of the program converted a rack of HP Rx2600 servers utilizing ISR's SprayCool technology. The rack level solution was deployed in the EMSL production facility at PNNL and underwent a 1 year reliability and uptime study. During the 1 year study the system was monitored continuously and uptime hours were calculated for both the CPU availability and SprayCool cooling system availability. For the period from July 1st, 2005 to August 10th, 2006, the rack uptime based upon CPU availability was 95.48%. For the same period, the cooling system uptime was 96.9%. The reliability of the system was tested by inducing monthly robustness tests targeted at determining the weak points in the system design; several areas were identified and are discussed in this report. The lessons learned from the robustness testing has resulted in improvements in the ISR system design and reliability models that are now predicting system MTBF hours of over 21,000 hours.

2. BACKGROUND

The following report will detail the investigation and lessons learned from the Phase 1 effort of the Energy Smart Data Center program. For the demonstration, a rack of 18 HP RX2600 Itanium 2U servers was converted from air cooling to evaporative spray cooling. The system was comprised of a Thermal Management Unit (TMU) that contains the working fluid pumps with control system and a Heat Exchanger Unit (HXU) that is connected to facility chilled water. The TMU is also connected to a rack manifold that distributes the working fluid to individual SprayModules in each of the servers which capture the heat from the CPU in place of the traditional heat sink.

This system was first benchmarked to show performance enhancements when compared to the original air-cooled state, and following this testing the system was installed at PNNL's Environmental Molecular Sciences Laboratory to operate in an actual HPC Datacenter facility for a full year. The intent was to monitor the system over this period to gain insight into the performance of the system in terms of reliability,

serviceability, maintenance, and uptime when deployed in a real world environment. Results of this study are presented below.

Glossary

PNNL – Pacific Northwest National Lab
EMSL – Environmental Molecular Sciences Lab (at PNNL)
MSCF – Molecular Sciences Computing Facility
ISR – Isothermal Systems Research, Inc.
TMU – Thermal Management Unit
HXU – Heat Exchanger Unit
HPC – High Performance Computing
CPU – Central Processing Unit
FEA – Finite Element Analysis
PCA – Printed Circuit Assembly
OEM – Original Equipment Manufacturer

3. RESULTS

3.2 ROBUSTNESS TESTING

The robustness testing was designed to subject the cooling solution to an extraordinary level of stress on a monthly basis, and to expose inherent weaknesses in the system. The results from the testing were used to improve the system design as it moved from the test/alpha state to beta, and eventually productization. The Test Methodology is presented in the Robustness Testing Methodology section with results presented in the Robustness Test Results section.

3.1.2 Robustness Testing Methodology

The test methodology requires the computer and cooling systems to operate under extreme operating conditions. Electrical and mechanical components are stressed under these conditions with the intent of causing failures. The test focuses on the cooling system's pumps, pump controllers, power supplies, electrical cabling, coolant plumbing, facility water plumbing, and fail safe protection systems. The computing system is exercised at full power, which stresses the computers as well as the cooling system.

The primary operational test focus on how the system responds to higher than normal current draws and pumping pressures. These conditions are generated by increasing the discharge pressure of the TMU's pump(s). This stresses the pump, the electrical cabling, the pump controller, system pressure transducers, the system's power supply, and the system plumbing due to higher operating pressure. A secondary test includes cycling the pump speed. This stresses the system by introducing inrush current and mechanically stressing the system's mechanical power switch.

A key feature of the SprayCool modular rack solution is the ability of the system to detect and respond to a water leak. The system uses facility water in its liquid-to-liquid heat exchanger, so it must respond 100% of the time to protect the computing system and the facility. The robustness test for the water detection system included testing the water detection sensor, controller, fail safe shut-off valves, the TMU, and the

MSCF facility operator's response to the leak. The water detection sensor and controller were triggered using a damp rag which then triggered the fail safe shut off valves to close. This event also creates an alarm that is sent to the MSCF facility operator.

In addition to testing the mechanical and electrical systems, a battery of tests were run on the TMU control system to exercise the firmware and verify correct operation. This testing mainly consisted of executing commands and verifying the outputs from the firmware in response to these commands.

3.1.2 Robustness Test Results

A selection of results from the robustness testing are shown in Tables 1-4. The robustness testing has been fully executed three times to-date. The testing on 1/10/06 was only a partial test since certain fixtures were not yet ready. The intent of the robustness testing has been to exposes weaknesses in the cooling system design. The results in Tables 1-4 indicate that since the first round of testing the system has performed without any major issues.

		1/10/2006	3/10/2006	4/14/2006	5/12/2006
Mechanical Test		**PASS/FAIL**			
Component	**Test Criteria**				
Spray cooled cold plate	Fluid flow rate (and pressure drop)	pass	NA	NA	NA
	Sealing against leaks	pass	pass	pass	pass
Fluid Tubing	Tubing has not collapsed under vacuum	NA	pass	pass	pass
	Tubing not deformed at high pressure	NA	pass	pass	pass
Rack Manifold	Manifold is not leaking	NA	pass	pass	pass
	Manifold not deformed at high pressure	NA	pass	pass	pass
Pump	Pumps respond to speed changes	NA	pass	pass	pass
Mechanical Switches	Power switch cycled 10 times	pass	pass	pass	pass
TMU Sealing	No detectable leaks	NA	pass	pass	pass
Water Throttling Valve	Valve responds to control inputs	NA	pass	pass	pass
Water Shut Off Valve	Valve responds to control inputs	NA	pass	pass	pass
*"NA" Indicates that test was not conducted					

Table 1 Tabulated mechanical robustness test results

		1/10/2006	3/10/2006	4/14/2006	5/12/2006
Electrical Test		**PASS/FAIL**			
Component	**Test Criteria**				
Pump Motor Controllers	Pumps respond to speed changes	NA	pass	pass	pass
	Pumps respond fail over command	NA	pass	pass	pass
Fluid Level Sensor	Level sensor responds to input commands	pass	pass	pass	pass
Pressure Transducers	Power supply normal while cycling valve	NA	pass	pass	pass
	Operational when load applied to 12VDC	NA	pass	pass	pass
Internal Cabling	Power supply normal while cycling valve	NA	pass	pass	pass
	Operational when load applied to 12VDC	NA	pass	pass	pass
	Pumps attain/maintain specified pressure	NA	pass	pass	pass
TMU Power Supplies	Power supply normal while cycling valve	NA	pass	pass	pass
	Operational when load applied to 12VDC	NA	pass	pass	pass
	Operational when load applied to 12VDC	NA	pass	pass	pass
Cooling System Controller	Cooling system responds to test inputs	pass	pass	pass	pass
	Pumps respond fail over command	pass	pass	pass	pass
	Pumps attain/maintain specified pressure	NA	pass	pass	pass
External Cabling	Pumps attain/maintain specified pressure	NA	pass	pass	pass
	Power supply normal while cycling valve	NA	pass	pass	pass
	Operational when load applied to 12VDC	NA	pass	pass	pass
Water Detection	Facility detects alarm, valves respond	NA	pass	pass	pass
*"NA" Indicates that test was not conducted					

Table 2 Tabulated electrical robustness test results

		1/10/2006	3/10/2006	4/14/2006	5/12/2006
		PASS/FAIL			
Firmware Test					
Component	**Test Criteria**				
Cooling System Controller	Cooling system responds to test inputs; issues warnings	NA	pass	pass	pass
Cooling System Failsafe	All nodes shutdown when failsafe is triggered	NA	pass	pass	pass

Table 3 Tabulated computing and thermal robustness test results

		1/10/2006	3/10/2006	4/14/2006	5/12/2006
		PASS/FAIL			
Computing/Thermal Test					
Component	**Test Criteria**				
Node CPU Temperatures	CPU temperatures are within ±2C of DCT data values	pass	pass	pass	pass
Node benchmark performance	Entire rack Linpack benchmark results are normal	NA	NA	pass	pass
*"NA" Indicates that test was not conducted					

Table 4 Tabulated firmware robustness test results

3.2 UPTIME

A key aspect of any high performance computer is its availability to users for running production jobs. This is typically referred to as “uptime”, and this metric was tracked for the liquid-cooled rack Rx2600s for a full year.

An average uptime value across the entire high performance computing market is 92%, with certain high availability mission critical systems achieving 95% or even 98% uptime. Typically these higher values are achieved by having an increased number of spares on hand, contracting with the hardware vendor for higher level service agreements, and so forth. However, even these values are relatively low compared to other markets such as telecommunications, a result of the fact that HPC systems are typically customized solutions using leading edge components in specialized applications which can put greater stress on the overall system.

3.2.1 Definition of Uptime

For this study, the uptime is determined in two different manners. The first calculation considers the total number of hours of availability for all 36 CPUs in the spray cooled rack. For this case, the uptime is calculated by considering whether each individual CPU is either operational or being serviced/repaired, and then comparing the operational time against the total amount of time it could have been available. The second calculation considers the total number of hours during which the cooling system is online and providing sufficient cooling capacity to the 18 servers in the rack. This is referred to as the cooling system uptime. In this case the uptime is calculated by comparing the time the cooling system was actually operational to the total amount of time it could have been available. This second calculation ignores the amount of time it took to bring up individual servers that experienced unique events. For example, a particular event involved the replacement of a power pod in an individual server. While the cooling system was available to provide cooling during this service, the CPUs in that server were not available during this time.

It is worth noting that PNNL typically performs system maintenance once a month, at which time the system may be brought down for routine service elements. Since this is typical operation for the system and is planned for, these times were not considered downtime for the purposes of the calculations.

TMU Availability	96.9%
Total Hours	9,229

ESDC SprayCool Rack Uptime

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
July	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
August	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
September	1	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
October	1	1	1	1	1	1	1	1	1	1	1	1	1	3	2	1	2	2	3	2	3	1	1	1	1	1	1	1	1	1	1
November	1	1	3	1	1	1	1	1	3	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
December	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
January	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1
February	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1			
March	1	1	1	1	1	1	1	1	1	4	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
April	1	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
May	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	3	2	2	2	1	1	1	1	1	1	1	1	1
June	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
July	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
August	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

if 1 = up	Green cells indicate normal operation.
if 1 = up	Bold text - rack is running applications (NWChem), benchmarks (NASA parallel, HPL, etc.)
if 2 = precautionary shutdown	The September shut down was precautionary, pending installation of water monitoring.
if 2 = down	1. October event resulted from poor workmanship during water monitoring install. 2. November event was to replace a leaking HXU. > Root cause was a bad glue joint design, new units use brazed joints. 3. The March stress test induced the downtime. Root cause was an intermittent thermistor failure. 4. The May downtime was caused by a false positive leak detection from a facility ball valve.
if 3 = down<10 hrs	Service events: Minor periods of down due to testing and facility maintenance interrupts
if 4 = robustness testing	Starting in 2006, monthly testing to "stress" the system to look for potential weak points.

Figure 3.1 Cooling uptime table for the spray cooled rack.

3.2.2 Results

Figure 3.1 shows a graphical representation of the uptime for the rack. For the period from July 1st, 2005 to August 10th, 2006, the rack uptime based upon CPU availability was 95.48%. For the same period, the cooling system uptime was 96.9%. Below we will discuss each of the major downtime periods in greater detail.

In October it was decided jointly between PNNL and ISR to install a water monitoring system to guard against external water leaks that might impact the operation of the datacenter. During the installation of this system several new electrical cables were built and crimped in the field, and poor workmanship while building this cable subsequently caused a loss of power to the water shut-off valve, which starved the TMU from its cooling source and caused the system shut down. So the failure was not product related but rather caused by human error.

The November shutdown was caused by a failed HXU as detailed in the following reliability section. The design issue that caused the leak has been corrected by going from a design with glued joints to a new HXU design that uses welded and brazed joints to eliminate the risk of a serious internal water leak. An FEA on the new HXU has an expected life of 5 million cycles, with a cycle being a significant change in pressure. The anticipation is that the active venting system will activate roughly twice per week, or 100 times per year. The anticipated life of the HXU is then on the order of tens of thousands of years, which is to say it should be more than sufficient for any anticipated application.

As discussed in an earlier section, in 2006 the team decided to initiate robustness testing on the system which was designed to put an extraordinary level of stress on the cooling system in order to provide data to be used to improve system design. The robustness testing of March 10th, 2006 did, in fact, cause a shutdown of the rack from anomalies induced on a single temperature sensor. This shutdown lasted 4 days because it happened late on a Friday and could not be addressed until the following week.

The final shutdown of significance occurred in May of 2006 when a small drip from a facility ball valve triggered the water detection system to activate and shut off the chilled water supply to the TMU. This facility leak was so minor that it could not have caused any damage to equipment under the floor, but it only takes a single droplet to activate the system. And once again the event occurred on a Friday which delayed the response until the following week.

In summary, of the four major downtime periods, two can be attributed to actual product design, with one of the failures being induced by the robustness testing. Likewise, several of the small shut downs were incurred as a result of tests that were undertaken to gather data, but that were unrelated to actual system performance. In consideration of the fact that the spray cooled rack being monitored is an initial prototype, the results achieved to-date are encouraging and point to the fact that a highly reliable liquid-cooling solution is achievable for production systems.

3.3 RELIABILITY

Reliability is commonly defined as "the probability that an item will perform a required function without failure under stated conditions for a stated period of time." In practice it is often measured in a variety of ways depending on what is most pertinent to the user, with different criteria for components vs. systems. For instance, a common measure for components is the Mean Time Between Failure (MTBF), which then correlates to system level measures such as Mean Time To Interrupt (MTTI), Availability/Uptime, etc. As stated in the original definition, typically these results are then presented as the probability of that component or system operating for a given time, say 1000 hours, 1 year, etc.

At the system level these measures become more complex because elements such as service, maintenance, and repair can impact the results. These considerations are reflected in the Mean Time to Repair (MTTR), which is a function of the service and maintenance level being supplied by the system vendor. For example, if a system has built in redundancy and an ample supply of spare parts on site with a trained technician close at hand to perform the work, a component failure will not necessarily have any impact to the system's availability, even though a component failure has occurred.

3.3.1 Reliability Methodology

The exponential model was assumed as the baseline for the individual components and sub assemblies. The model is a widely used in the electronics industry but has limited application in the mechanical reliability because of its lack of modeling for wear-out. For the total system rollup, the standard series predictions model was used. Redundant parts, such as the pumps and power supplies, were modeled using the appropriate parallel model. A more detailed explanation of the techniques used is in Mil-Std-756B.

Reliability modeling mechanical components for wear-out effects is quite difficult. A common methodology is to evaluate similar components and use failure information to define the appropriate reliability model. If similar component failure data is not available, using the exponential distribution as an initial reliability model is the more conservative approach until failure information becomes available. When actual test data and failure information becomes available, the mechanical reliability models can be refined. More details of plans in this regard are found in the following section.

3.3.2 Results

The initial reliability modeling tools used for the Phase I analysis resulted in an MTBF prediction of only 601 hours, actual performance was 3.3 times that at 1,984 hours over the one year period, revealing several areas for improvements in both the modeling methods and system design improvements. The lessons learned from Phase I were incorporated into the reliability analysis models and the system design for Phase II, resulting in a predicted MTBF of 21,031 hours.

Because the model is predictive and based largely on standard values from the MIL handbook and vendor datasheets, the initial confidence interval is 50%. The 'parts count' methodology is an inherently conservative approach, which is supported by the fact that the Phase I system performance was 3.3 times better than predicted. To address this with the Phase II model, some of the standard failure rates were factored down to what are believed to be more accurate values based on engineering judgment and experience gained from the Phase I system, and this is reflected in the model. Accordingly, it is not expected that the actual performance of the Phase II system will be 3+ times above the predicted value. To start gathering actual data, ISR will begin long term reliability testing of 10 pumps and 4 TMU's starting in Q4 2006. Assuming no failures, the 4 TMU's will need to run for 12,106 hours to achieve 90% confidence interval.

The main contributors to predicted failures for the Phase I system were the heat exchanger, pumps, o-rings, and throttle valve controller. The initial heat exchanger design used glued joints that proved unequal to the system pressures and temperatures, and it has since been replaced in future designs with a unit that utilizes brazed and welded joints for substantial gains in reliability. Similarly, the pump used in the Phase I system was an early prototype design that has since been replaced by a commercially available model from an established company with expertise in designing and manufacturing cost effective and reliable pumps. These two changes in components represent a shift in approach for these systems that leverages components from experts in the respective fields rather than custom in-house designs where possible.

Any liquid based system that is modular will rely heavily on o-rings, and these two systems are no exception. In the Phase I system a number of component failures were seen due to the use of radial o-rings on the quick disconnects. It was all too common for an o-ring to become nicked or scratched when mated, resulting in a small leak and resulting charged failure. In the Phase II system a change was made to quick disconnects that use an o-ring as a face seal, thus resulting in a seal that is not subjected to significant wear during mating cycles. In addition to the improvement in the robustness of the o-ring seals used, ISR has incorporated its active venting system in Phase II design. With active venting the system operates in a vacuum so that o-ring failures will typically result in air ingress to the system, which will simply cause the venting system to actuate more frequently and with no impact to the ability of the system to provide cooling to the components.

The final major contributor to failures was the PCA controller for the throttling valve. This component was added late in the design due to condensation concerns, and a commercially available component was chosen for the application. After witnessing a failure of this component, the vendor was contacted and admitted to a failure rate of 25% for these devices. This came as quite a surprise, and has been addressed in the Phase II system. It is worth noting that a failure of this kind did not result in a lack of cooling to the system, but rather an increased likelihood of condensate on the under floor plumbing.

4. CONCLUDING REMARKS AND RECOMMENDATIONS

The PNNL Phase I and II systems are quite complex and are comprised of many components. Accordingly, while each individual component has a high degree of reliability, in aggregate the result is a predicted MTBF on the order of 2.4 years for the Phase II system. However, this rate applies to the likelihood of an individual component failure and not the overall system availability or uptime, which is more interesting.

The primary concerns for the SprayCool system reliability continue to be the o-rings, tubing, and valves. However, as noted before, the addition of an active venting has resulted in a system that is quite fault tolerant. Tubing and o-ring leaks are not catastrophic in nature but tend to be incremental changes that exceed a given threshold. Furthermore, with a system operating below atmospheric pressure (such as the Phase II rack), these failures will result in air ingress into the system rather than a coolant leak out, which will have no impact on system cooling since it will simply be periodically purged, perhaps more frequently, until the next service period when repairs can be made. Accordingly, the active venting system is quite effective in mitigating the impacts of this type of component failures. Clearly, adding this function has added an additional component that can fail, but a failure of this kind would result in a gradual increase in system pressure and temperature that could be repaired on the order of days, not hours, with minimal impact to the system. Furthermore, because the active venting system is anticipated to activate for only 60-90 seconds 2-3 times per week, its energy consumption is minimal.

One of the few components whose failure would have immediate and direct impact to the cooling system is the pump. In this case, the system has built in redundancy with up to 3 pump slots available. In addition, the pumps are designed to be field replaceable so that failed components can be quickly replaced in the field.

For these reasons, availability or uptime is the more useful measure for system performance. It is encouraging to note that the Phase I prototype, with all its issues, was able to demonstrate an uptime of 97%. The anticipation for the Phase II system is that a higher level of uptime will be achieved with substantially less intervention, reflecting a maturity of design that is suitable for OEM adoption and customer deployment.