PNNL-17281

Energy Smart Data Center (ESDC)
Phase II Final Report

Isothermal Systems Research, Inc.
Tahir Cader
Rich Maes
Harley J. McAllister
Levi Westra

Pacific Northwest Nation Laboratory
Andrés Márquez

December 5, 2006

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

Pacific Northwest National Laboratory
Richland, Washington 99352

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**

ESDC Phase II
Final Report

Presented to PNNL

December 5, 2006

Isothermal Systems Research, Inc.
Tahir Cader
Rich Maes
Harley J. McAllister
Levi Westra

Pacific Northwest Nation Laboratory
Andrés Márquez

TABLE OF CONTENTS

Glossary

1U—The standard unit of measure for designating the vertical usable space, or height of racks. 1U is equal to 1.75 inches.

AC—Alternating Current

ASIC—Application Specific Integrated Circuit

ATX—Advanced Technology Extended

BMC—Board Management Controller

BurnI2—Program that uses a calculation loop to exercise an Itanium II processor

COP—Coefficient of Performance, ratio of cooling power to compute power

cPCI— compact Peripheral Component Interconnect

CPU—Central Processing Unit

DAQ--data acquisition system

ESDC—Energy Smart Data Center

DC—Direct Current

DIMM— Dual In-line Memory Module

FET—Field Effect Transistor

GB—Gigabyte, 2^{30} bytes when discussing memory

GFLOP—Billion (giga, 10^{9}) Floating point Operations Per second

HP—Hewlett-Packard Company

HPC—High Performance Computing

HPCS—High Performance Computing System

HXU—Heat Exchanger Unit, component of the TMU

IHS—Integrated Heat Sink, also referred to as the processor lid

IPMI--Intelligent Platform Management Interface

ISR—Isothermal Systems Research

MCH—Memory Control Hub, also known as Northbridge

MPB—Multiple Processor Board

MTBF—Mean Time Between Failure

nbench—A benchmark program originally developed by BYTE Magazine to expose the capabilities of a system's CPU, FPU and memory system.

OEM—Original Equipment Manufacturer

Pa-S—Pascal-second, the SI unit of dynamic viscosity equal to 1 kg/m/s

PF5060—One of the Fluorinert™ brand dielectric fluids made by 3M Corporation.

PNNL—Pacific Northwest National Laboratory

PSIA—Pounds per Square Inch Absolute

PSID—Pounds per Square Inch Differential

RAM—Random Access Memory

RAS—Reliability, Availability, and Serviceability

RMS—Root Mean Square

RPM—Revolutions per Minute

SMK—Spray Module Kit

TC—Thermocouple

TDP—Thermal Design Power

TMU—Thermal Management Unit, manages heat generated by a computer, usually composed of a pump, reservoir, heat exchanger, and logic controller

UDP—User Datagram Protocol

W—Watt, 1 joule per second

Executive Summary

Thermal problems have existed in data centers for years. But today, thermal problems have transcended into real business problems that have a direct effect on a data center operator's ability to control operating costs and install the highest performing system with reduced acquisition budgets. Some of the key drivers to the challenges are: significant increases in the cost of power, growing server power requirements, facility infrastructure incapable of delivering adequate air to individual racks and rising costs in both capital and time of data center construction.

Advanced cooling technologies present the most promising method for reducing the impacts of today's data center thermal-business problems. Advanced cooling technologies provide the ability to:

- Capture heat in a way that enables computer companies to grow server performance
- Transport heat in a consolidated medium that reduces facility requirements
- Reject heat utilizing energy-efficient equipment.

The Energy Smart Data Center (ESDC) program has been exploring and demonstrating the benefits of alternative cooling technologies, such as evaporative spray cooling. Isothermal Systems Research (ISR) has developed a line of SprayCool™ data center products, optimized for 1U and 2U clusters that has demonstrated the ability to revolutionize how new data centers are built and operated by delivering the following value to data center operators:

- Extending the life of existing data centers
- Enabling payback periods of less than a year by deferring the capital cost associated with facility construction
- Reducing the power required to operate a facility by 20-30%
- Reducing the energy cost per server by 30-50% per year
- Enabling a 30-60% increase in the number of servers deployed in a facility.

The Pacific Northwest National Laboratory (PNNL) teamed with ISR for the first phase of a project to explore the efficacy of using the SprayCool product to address the cooling and power issues at the server, rack, and facility levels of a supercomputing data center facility. This phase included demonstrating a SprayCool rack of Hewlett-Packard Company (HP) rx2600 servers, similar to what PNNL employs in its High Performance Computing System (HPCS). Phase II of the program has been focused on understanding and improving the products Reliability, Availability, and Serviceability (RAS) performance when deployed into a production data center environment. In addition, the phase II effort further developed next generation products with a 1U reference design optimized for liquid cooling and system updates and expansion to the performance of the SprayCool optimized global system.

1.0 Program Description

PNNL's ESDC is a multi-phase program. Phase I demonstrated the feasibility of SprayCool technology to effectively cool a rack of HPC servers in a data center environment. This phase included thermal performance of the system 1) to maintain CPU temperatures below that encountered with air cooling, 2) to efficiently remove the rejected heat through the facilities water system and reduce the load on the air handlers, thereby enabling densification, and 3) to achieve all this with a high degree of system reliability and uptime.

Phase II is an outgrowth of Phase I wherein the insights gained from operation of the initial system at the PNNL facility would be implemented into a next generation design that is ready for market adoption. The focus of the development was to produce a product that had improved reliability and serviceability while still making gains in performance. In addition, claims of densification in Phase I were to be demonstrated by cooling a rack of 1U rather than 2U servers and to include in this rack a chassis of ultra high density SprayCool server blades in a globally spray cooled chassis. The entire system would then have almost double the heat load in a single rack while still being cooled by a single Thermal Management Unit (TMU), thus demonstrating both near-term and long-term product adoption strategies. Finally, a reference design for a 1U server optimized for SprayCool technology would be developed to enable easier adoption by Original Equipment Manufacturers (OEMs) interested in integrating SprayCool technology into their product portfolios.

In Phase III, verification of performance gains in terms of energy efficiency, reduced air flow requirements, a higher coefficient of performance (COP), and the resulting improvements in total cost of ownership will be verified by installing an actual SprayCool supercomputer in a fully instrumented facility. By monitoring all system performance parameters of the supercomputer and facility under a variety of facility cooling implementations (rejecting to chilled water, cooling tower water, free cooling, and an air cooled baseline), the actual effects of SprayCool technology on the facility will be demonstrated and supported with real-time data.

2.0 Phase I Lessons Learned and Resulting Improvements to Phase II

2.1 Lessons Learned

Following the installation of the ESDC Phase I hardware onsite at PNNL, a team of project contributors met and generated a comprehensive list of design enhancements and improvements for the Phase I deliverable. This list identified a wide range of design improvements covering reliability, serviceability, system performance, and cost reduction. This list was the basis of the system updates for operational methodology, design guidelines, and specifications. The intent was to capture all of the knowledge gained from observed strengths and weaknesses in this first SprayCool system and improve the next generation by embracing the strengths and addressing areas for improvement. All lessons learned from the Phase I system would be used to improve the system delivered in Phase II and have been grouped into three categories, functionality, reliability, and serviceability. Functionality is defined as the SprayCool system's ability to effectively cool the electronics in various operating conditions. The reliability category includes issues that were revealed in the Phase 1 system that minimize the system's Mean Time Between Failure (MTBF). ISR has defined a failure as an event which causes the computing system to experience an unplanned shutdown. Serviceability incorporates features that simplify servicing the system during scheduled maintenance and repairs. It also includes features that allow service organizations to easily monitor system status and predict impending system failures.

Functionality of the SprayCool system includes features and traits that allow the system to operate both in normal conditions and during startup and that react appropriately to aberrant temperature and pressure conditions. The key component of the SprayCool system is the TMU. For the ESDC Phase I project, a first generation TMU was created and characterized. This TMU, shown in Figure 1, was the first system design driven by requirements and specifications for a datacenter facility. Many best judgments were used during this phase to address all possible issues the SprayCool system would encounter while operating in a data center. During the lesson's learned review, it was obvious that a number of the requirements for data center environments were addressed well by the first system prototype. However, several very important weaknesses in system functionality became evident after the system had been characterized and operated for a year at PNNL.

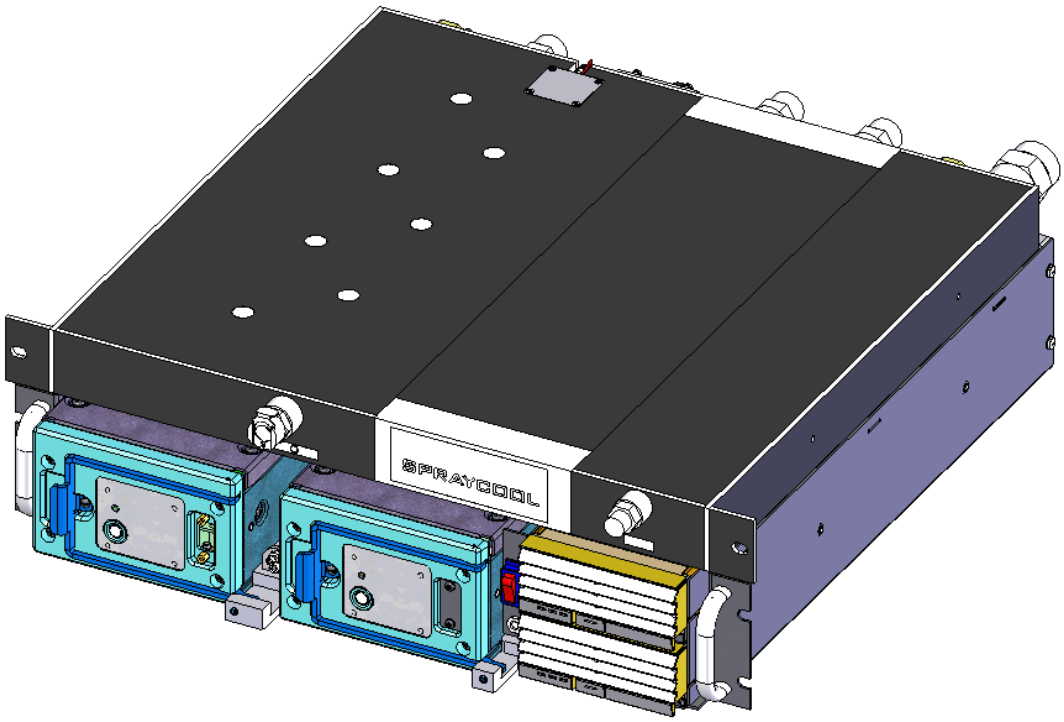


Figure 1: Generation 1 Thermal Management Unit (TMU)

The primary design weaknesses are reviewed in detail in the following section. First, the heat exchanger was designed with glue joints that were subsequently determined to be unreliable when the system experienced an internal leak a few months into operation. This failure led to contamination of the system fluid by a propylene glycol and water mixture that proved problematic to the system. The filters in the system were able to remove the water, but the remaining propylene glycol then crystallized and caused excessive wear in the pumps and fouling of one of the spray modules. The presence of water also caused issues for the pump because it utilizes a fluid “flow through” design. In this type of pump design, system fluid is in contact with the control electronics in order to cool more effectively, improving power efficiency of the pump, but the presence of water obviously has an adverse impact on the operation of the control electronics of the pump. In fact, the majority of the issues that were experienced could be traced back to this single point of failure. As a result of this, the failed heat exchanger was replaced with a bar and plate heat exchanger design where the entire component relies on either welded or brazed joints, a much more robust design, resulting in a higher predicted system reliability. Furthermore, a replacement chemical filter was implemented that included activated carbon to remove any propylene glycol that might be present, before it crystallizes.

A second issue that arose during initial characterization of the TMU is that the turbine pump used in the TMU had challenges with priming the SprayCool system during the initial startup. Extraordinary measures were required to prime the system, and this was not acceptable. Two changes were implemented in the

next generation TMU to address this issue. First, the orientation of the reservoir to the pump inlet was changed. In the initial TMU, the pump inlets were connected to the reservoir via a tube that could entrap air bubbles and impair priming or at the very least restricted flow to the pumps. In the new TMU design, the pump inlets mount directly to the side of the reservoir such that the pump inlets are always flooded. Second, a new centrifugal style pump is being used that leverages a more robust, commercially available pump design. The new pump design, in conjunction with the modification in the pump-reservoir interface, has eliminated the potential for priming problems. In addition, the new pump does not utilize a flow through design for the control electronics, thus removing the risk of water exposure to the electronics of the pump and the problem noted in the previous paragraph.

The final major system design improvement is the presence of an active venting system. In the first iteration, the system was designed to be as leak-tight as possible, but this is a very challenging design problem, and it was observed during the yearlong operation that air tended to leak into the system over time, which would change system pressures and therefore performance, necessitating the need to periodically purge the system of excess air.

Accordingly, in the Phase II TMU, a small pump was implemented to periodically remove the excess air, together with a vertical chamber adjacent to the rack manifold to house these gasses. Pressure transducers are located in several key areas of the system, and the TMU monitors these devices constantly. When the pressures rise to a predetermined level, the TMU signals the pump to activate until the desired vacuum is reached, and the pump shuts off, typically requiring only a minute or two. Because it is likely that this process will also capture some Fluorinert™ dielectric coolant vapor, the vertical chamber is large enough to allow separation of the coolant vapor from the air; and because it is heavier, it will collect at the bottom. All that is left then is for a service technician to periodically drain the active venting chamber back into the TMU reservoir.

The active venting system has several key advantages. First, it eliminates the need for human involvement on a regular basis. Second, it allows the system to operate in a vacuum within a specified range where the thermal properties of the performance fluid are optimized for maximum cooling of the components; and this can vary depending on the source of the water being used for heat rejection, be it chilled water or facility cooling tower water. Furthermore, by reducing the stringent requirements to manufacture a system that is “leak free”, it eases the burden on component manufacturers and system manufacturing parameters such that build costs are reduced. Finally, the active venting system changes the system balance from above atmospheric pressure to below, meaning that air tends to come in more than fluid tends to leak out. This greatly reduces the operational expense of having large inventories of performance fluid on hand as consumables and essentially creates a self-healing system that is much more robust to both long and short term system perturbations.

2.2 Phase II Software Architecture Enhancements

Phase II also presented the opportunity to advance the software architecture of the ISR TMU and to create a system that was more robust with the appropriate hooks to allow for monitoring and maintenance. The original software implementation of the TMU was code leveraged from previous generations of TMU type equipment that focused on the very specific application of regulating pumps to a given pressure.

At the time, coding for those products was very direct and lacked layers of abstraction that are considered to be standard practice today in embedded applications. As a result, modification of the TMU code was tedious and time consuming.

For PNNL's Phase II program, code developed for the new TMU was re-engineered from the ground up with extensibility in mind. Development started by using an off the shelf, high reliability operating system (OS) from Micriμm, Inc. Additionally, software architecture was created that allowed for easy modification and maintenance. Software tools were identified that could aid in debugging and testing.

The end result is a software architecture/toolset that now includes layers of abstraction appropriate to an embedded system and a mechanism for field troubleshooting and testing that will improve overall software quality and software defect response time.

The Abstraction layer now allows us to modify hardware and create an appropriate interface to the application program interface (API) without having to redesign the application. Additionally, it also allows for the same application to be reused on multiple hardware platforms in the future (see Figure 2).

2.2.1 Customer Usability Driven Software

Phase II was an opportunity to productize the TMU. Getting customer acceptance of SprayCool™ technology is the first step in becoming a major player in the liquid cooled computing business, but becoming the leader means providing the tools and quality to make your customers lives better (see Figure 3).

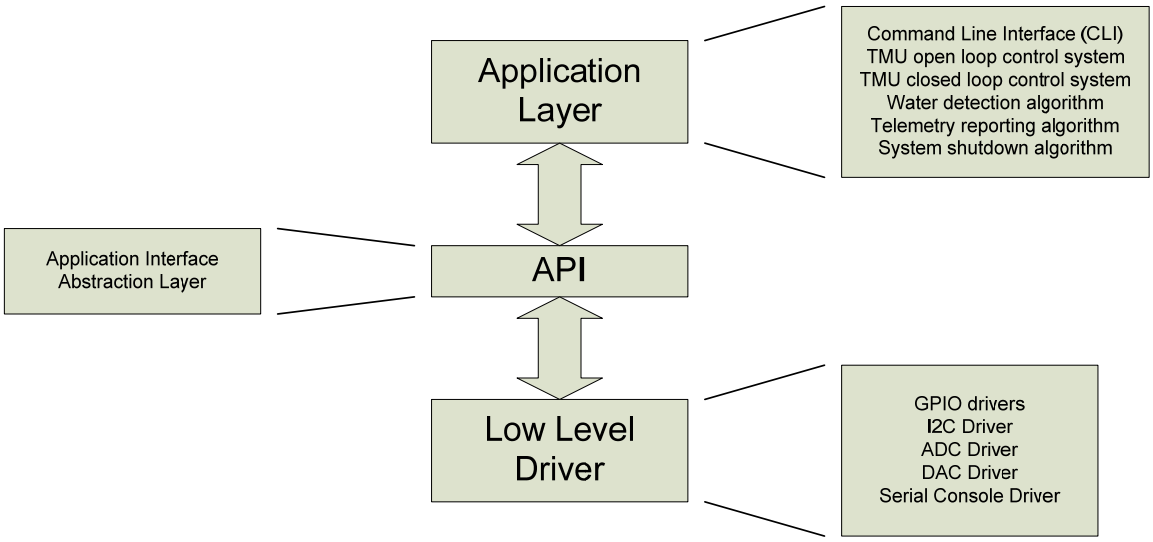


Figure 2: Phase II Software Architecture Enhancements

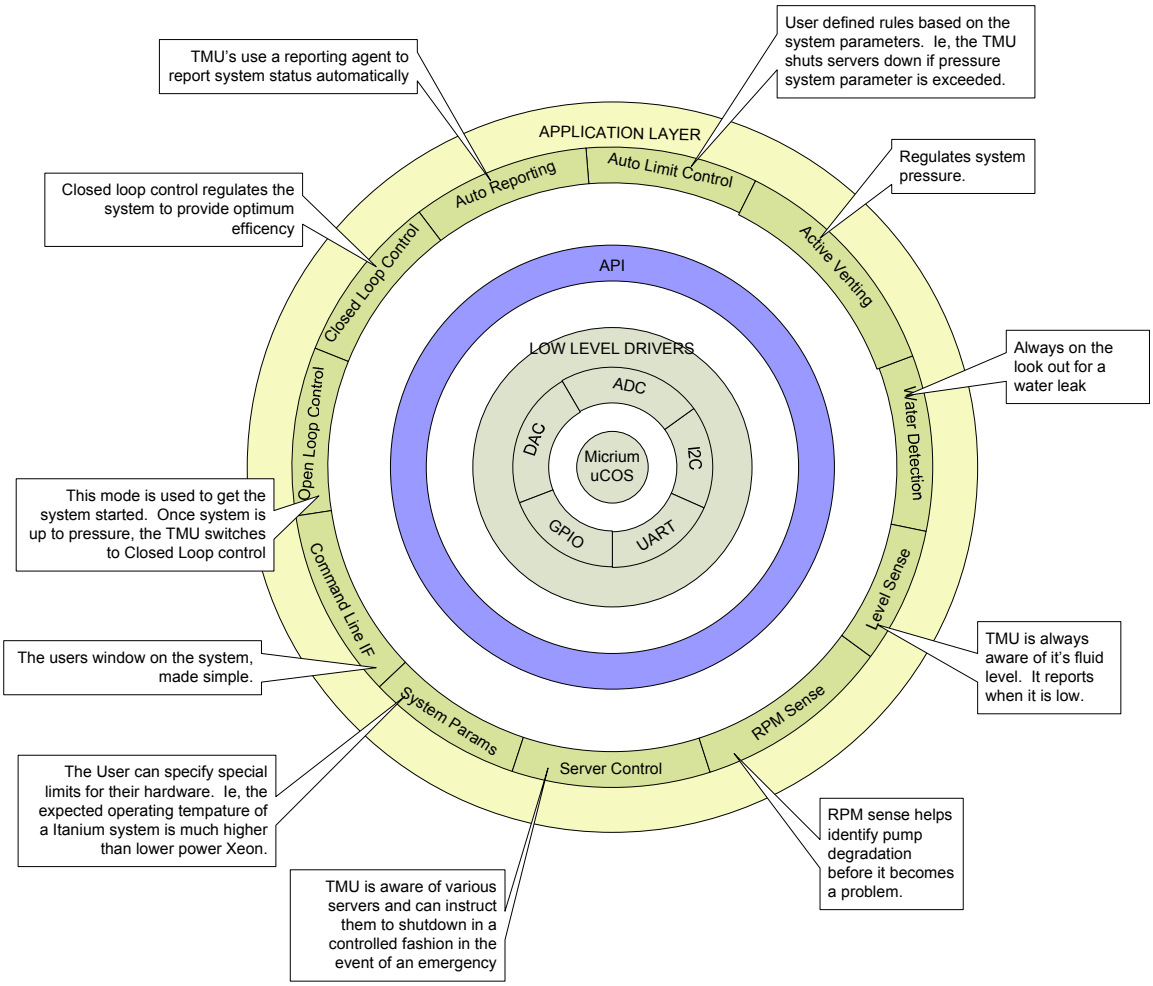


Figure 3: Customer Usability Driven Software

2.2.2 Command Line Interface

In Phase I, the software command sets were based on acronyms that were difficult to remember. Online help was also underdeveloped. For the new TMU design, new command sets where chosen that could extend to multiple platforms in the future. Command syntax was chosen to be very self explanatory. For example, deactivating or activating the SprayCool TMU is accomplished by typing “spraycool on” & “spraycool off” respectively. Activating 1U servers is accomplished by typing “server on all”. Activating the global server is done by typing “global on”.

2.2.3 System Parameters

The Phase I TMU software was written in such a way that system operational parameters had to be re-compiled into the software. This effectively meant that customer management of the TMU platform was impossible. The Phase II TMU allows for users to alter the setting for the TMU from the console port. In particular, this will be very useful for OEM’s who desire to set special warning limits or equipment shutdown limits with their customer shipments.

2.2.4 Server Control

PNNL asked that the SprayCool system should have a pro-active method for shutting down servers in lieu of the OEM integrated self protect method. This solution is only currently designed to work with HP rx1620’s, ISR SprayCool Blades™ and ISR SprayCool Global solutions. For the system to work with HP rx1620’s an external serial switch is required together with the HP rx1620 serial console port.

2.2.5 Water Detection System

The Phase II TMU has two methods for water detection.

1. An external third party water detection system with a dry contact relay output.
2. Internal water detection system.

The first method is the external water detection device. In the current implementation, any water detect system with a normally closed dry contact can be interfaced to the TMU. On water detect, the contact must open which will notify the TMU that a water leak has been detected. By default the TMU is configured to respond to the water detection by initiating a system shutdown procedure via the server control algorithm.

The integrated water detection system can support up to four zones. Support for the integrated water detection system is currently disabled pending product planning on the implementation. The current thinking is that one or two of the

zones may be deployed within the rack and TMU to aid in determining leak locations.

2.2.6 RPM Sense and Level Sense

The new TMU provides high accuracy measurements for both level sense and RPM sense. This is an improvement over the older TMU because we can now use this high accuracy data to provide service alerts, failure prediction and remote trouble shooting support.

2.2.7 Open Loop Control

Open loop control mode in the new TMU provides the same basic feature that was provided for in the original TMU, plus support for establishing active venting and water detection synchronization.

2.2.8 Closed Loop Control

The closed loop control in the new TMU regulates pressure and is robust enough to support flow rates as low as one 1U server or as high as sixteen 1U servers and a Spraycool™ Global Chassis simultaneously.

2.2.9 Auto Reporting

The PNNL Phase I deployment required extensive monitoring as part of a reliability demonstration. The method for obtaining the data was implemented through a variety of systems that proved to be problematic because the system was not streamlined. This feature, when functional, proved to be very effective for remote troubleshooting. The new TMU implements several features that make data collection, reporting, and analysis much simpler. This feature is called Auto reporting and can be enabled and disabled by the user. When enabled, the TMU periodically transmits a packet containing a slice of operational data. The data packet is described Figure 4.

1	2	3	4	5	6
Serial Number	Date	Time	Run Time (sec)	Application State	Active Vent State

7	8	9	10	11
Level Sensor Reading (raw)	Differential Pressure (PSID)	Reservoir Pressure (PSIA)	Global Vapor Pressure (PSIA)	Global Discharge Pressure (PSID)

12	13	14	15	16
Global Temp 1 (C)	Global Temp 2 (C)	Return Temperature (C)	Reservoir Temperature (C)	Water Proportional Valve (% Open)

17	18	19	20	21	22	23	24
Pump 1 Control Voltage (%)	Pump 1 Speed (RPM)	Pump 1 Current (A)	Pump 1 Voltage (V)	Pump 2 Control Voltage (%)	Pump 2 Speed (RPM)	Pump 2 Current (A)	Pump 2 Voltage (V)

25	26	27	28
Lower Shutdown Flags	Lower Warning Flags	Upper Warning Flags	Upper Shutdown Flags

29	30	31 Thru 60	61	62
Server 1 Temp 1	Server 1 Temp 2	Servers 2 thru Server 15	Server 16 Temp 1	Server 16 Temp 2

Figure 4: Data Packets

Data is transmitted via User Datagram Protocol (UDP) and can be collected using standards based methods and relayed to multiple sources.

2.2.10 Auto Limit Control

Auto limit control is a feature that allows the OEM and user to define special rules and actions to take in the event that performance rules are broken. The rules themselves are controlled by the limit value that the user has set in the TMU. As an example, the reservoir pressure (Field 9 above) may drop to an abnormally low level but not so low as to be considered critical. The system would then make the decision to send a warning message via the auto reporting mechanism. If the system reservoir pressure continues to drop past the critical level, the auto limit control mechanism would then implement the system shutdown procedure.

2.2.11 Active Venting

The active venting algorithm in the new TMU overcomes some very particular problems related to two-phase liquid cooled computing. The first issue is that the boiling point of the fluid is regulated by pressure. This algorithm is designed to bring the system pressure down and thereby decrease the boiling point of the 3M PF5060 fluid.

The algorithm must also take into account the fact that the system pressure will rise or fall with temperature. This keeps the TMU from shutting down when the compute hardware shuts down or slows enough that the resulting temperature drop causes the internal pressure to drop below the critical point.

3.0 System 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 MTBF, which then correlates to system level measures such as Mean Time To Interrupt, Availability/Uptime, etc. As stated in the original definition, typically these results are then presented as the probability of that a component or system operating for a given time, say 1000 hours.

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, 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.

3.1 Reliability Methodology

The exponential model was assumed as the baseline for the individual components and sub assemblies. In the electronics industry, the model is widely used, but it 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.

3.2 Results

The initial reliability modeling tools used for the Phase I analysis resulted in an MTBF prediction of only 601 hours; the 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 1 were incorporated into the reliability analysis models and the system design for Phase II, resulting in a predicted MTBF of 21,031 hours.

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 later 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 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.

3.3 Reliability, Availability, and Uptime

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 active venting has resulted in a system that is in many ways “self healing.” 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 because it will simply be periodically purged, perhaps more frequently, until the next service period when repairs can be made.

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 three 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.

4.0 ESDC PHASE II DELIVERABLES

Three different pieces of technology were delivered for ESDC Phase II. The first was a rack of 16 HP rx1620s. This rack of servers would have CPUs spot cooled using the second generation TMU. The HP rx1620 is a 1U server that has similar computing capabilities as the HP rx2600 2U server which was part of the ESDC Phase I. Using the 1U rx1620s, demonstrates a 2x increase in computing density (flops/rack) when compared to the 2U rx2600s. The second deliverable is an upgrade of the Global system delivered in Phase I. This upgrade doubled the number of Multi-Processor Boards (MPB) from four to eight and increased the amount of memory on each MPB. The third major deliverable includes two 1U SprayCool ATX servers. The motherboard used by these servers was designed specifically for SprayCool.

Each of the three deliverables was appropriately tested prior to delivery to PNNL. The test methodology and results are presented in the following sections.

4.1 SprayCool Overview

The heart of the SprayCool system is the TMU. The TMU contains the pump, (to circulate the dielectric coolant), the control electronics, and the liquid to liquid heat exchanger. The SprayCool system block diagram, Figure 5, shows how the TMU provides the rack supply manifold with cool single-phase fluid to distribute to the various spray modules mounted in the rack. In the spray module, the cool liquid is sprayed, heated, and then vaporized creating a two-phase fluid which is returned back to the TMU via the rack return manifold. The heat exchanger unit within the TMU condenses and sub-cools the fluid readying it for redistribution to the rack supply manifold.

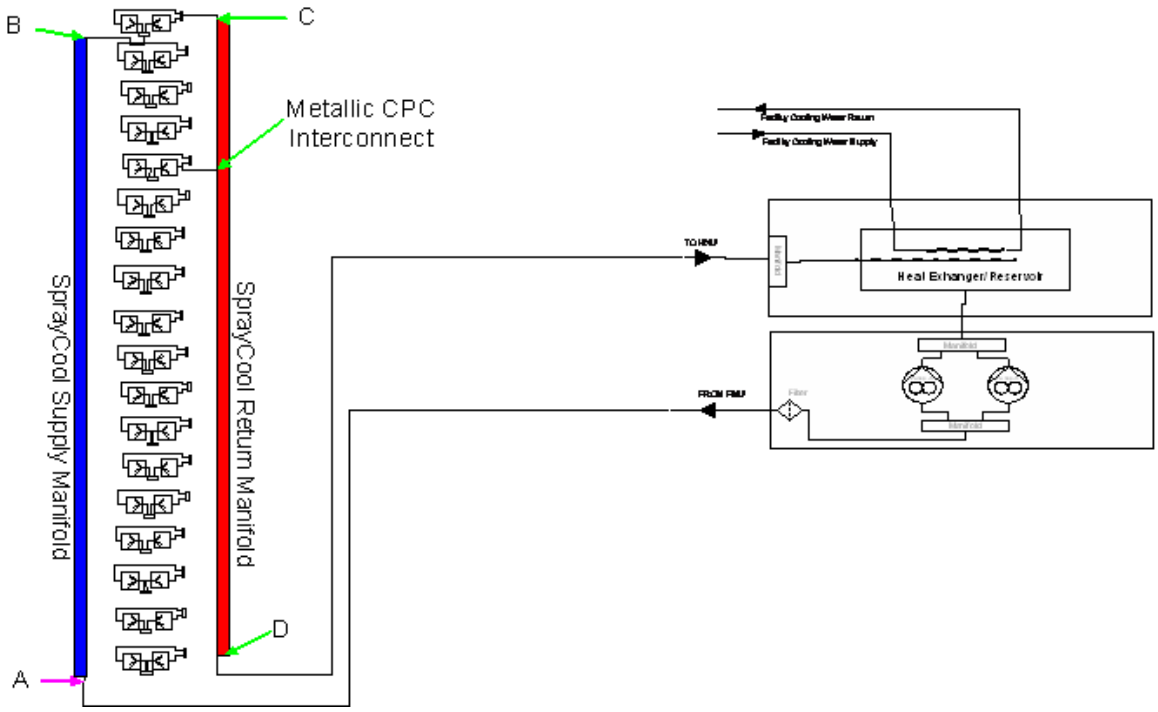


Figure 5: SprayCool System Block Diagram

For ESDC Phase II, the TMU is the second generation TMU, shown in Figure 6, incorporating many of the lessons learned from ESDC Phase I, as discussed in Section 2.0, Phase I Lessons Learned and Resulting Improvements to Phase II.

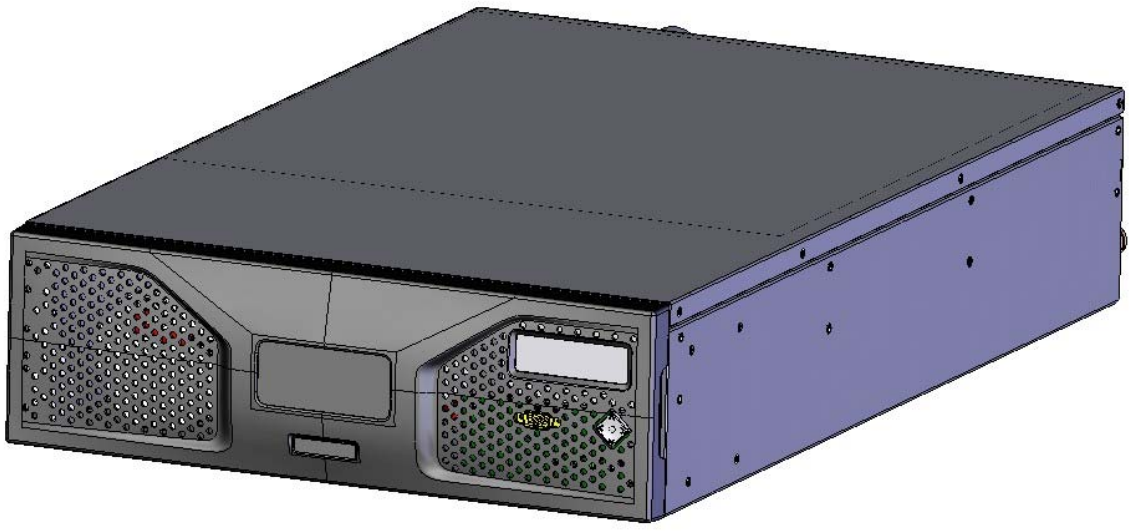


Figure 6: Phase II TMU

4.2 HP rx1620 – Technology Description

The HP rx1620, see Figure 7, is a 1U dual processor server configured with two 64 bit 1.6 GHz Itanium II processors with 2 GB of memory. The Itanium II processors operate with 3.0 MB of L3 cache at a system bus speed of 267 MHz. The server itself is rated to have a maximum input power of 585 W with a typical input power of 440 W. The 1.6 GHz Itanium II processor has a thermal design power (TDP) of 99 W with a maximum case temperature of 83° C. The processors are cooled using standard passive copper air-cooled heat sinks. Air is delivered to the heat sinks via the six 40 mm fans (note two of the fans are not shown in Figure 7, they are located in front of the power supplies, lower right of the picture).

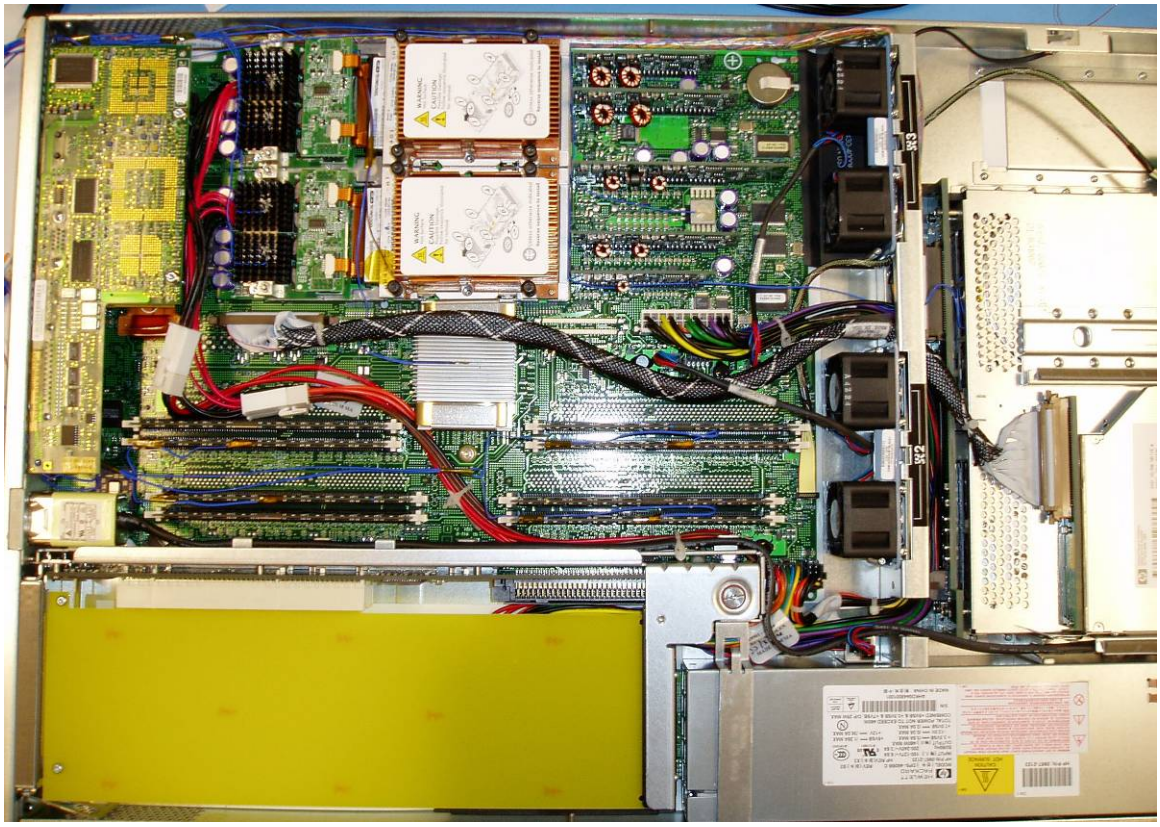


Figure 7: HP rx1620 dual Itanium II server, with (4) of (6) 40 mm chassis fans visible

Figure 8 shows the 16 units racked in a 19 inch 42U HP Proliant 10000 series rack. A single HP ProCurve 24 port GB Ethernet switch was used to communicate with the servers.

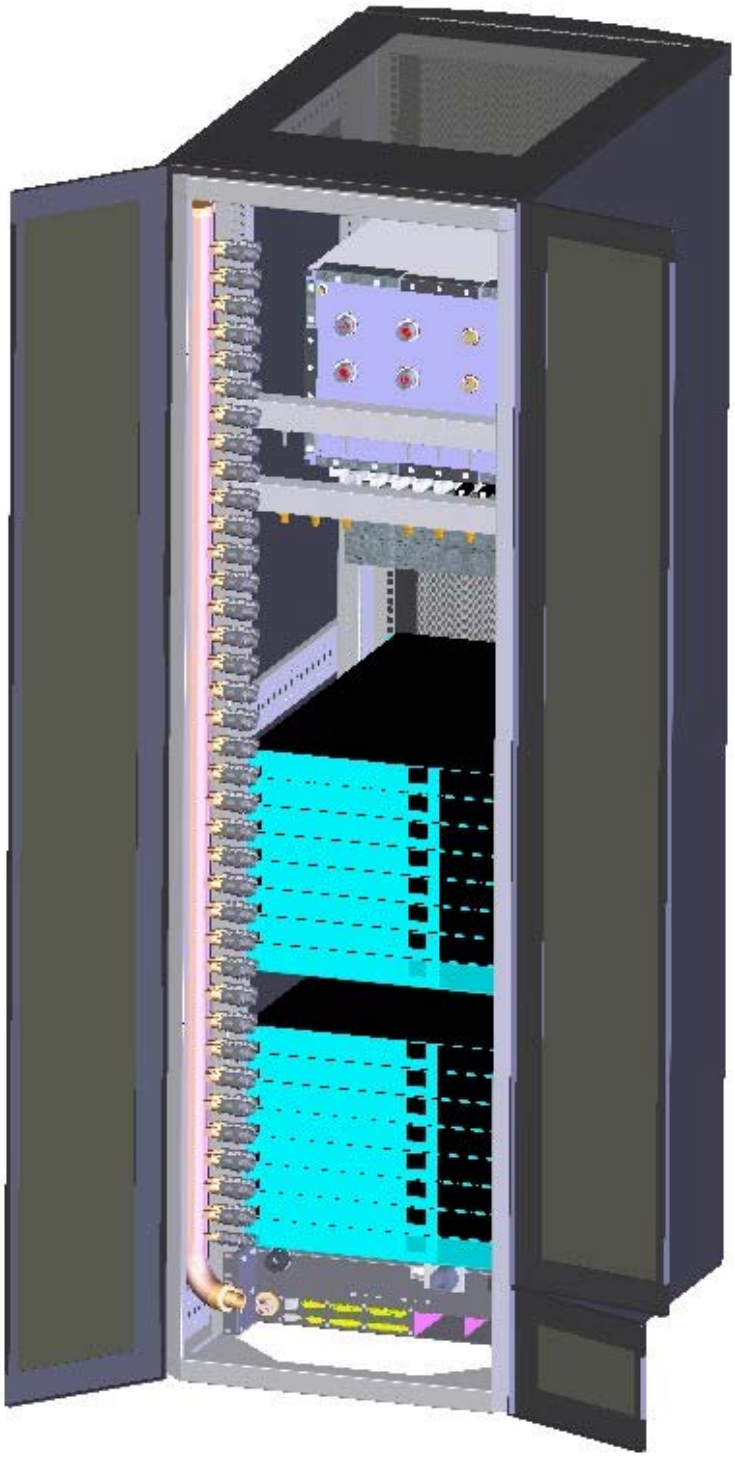


Figure 8: Rear of the Phase II SprayCool rack with (16) HP rx1620s.
A SprayCool Global system and typical SprayCool hardware

When converted to SprayCool, the two copper air-cooled heat sinks are removed and replaced with a Spray Module Kit (SMK), see Figures 9 and 10. For this retrofit, no server fans were removed or replaced. The SMK includes two spray

modules, supply and return plumbing, particle filter, and a server manifold. The server manifold is connected to the SprayCool system using flexible tubing and quick disconnect fittings with interconnect with the rack manifold, as shown in Figure 8. The rack manifold runs the height of the rack and supplies the SMKs with coolant and returns the two phase heated coolant from the SMKs back to the TMU.

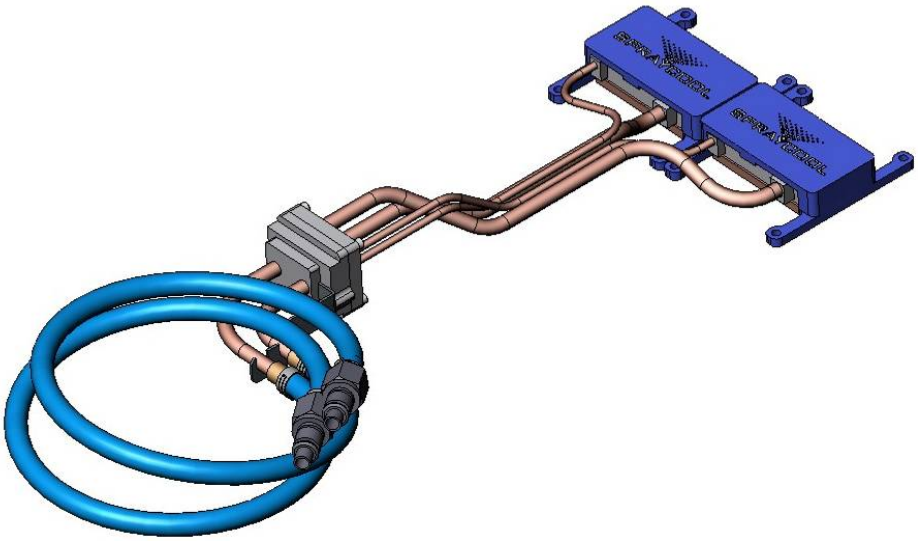


Figure 9: Typical Spray Module Kit (SMK)

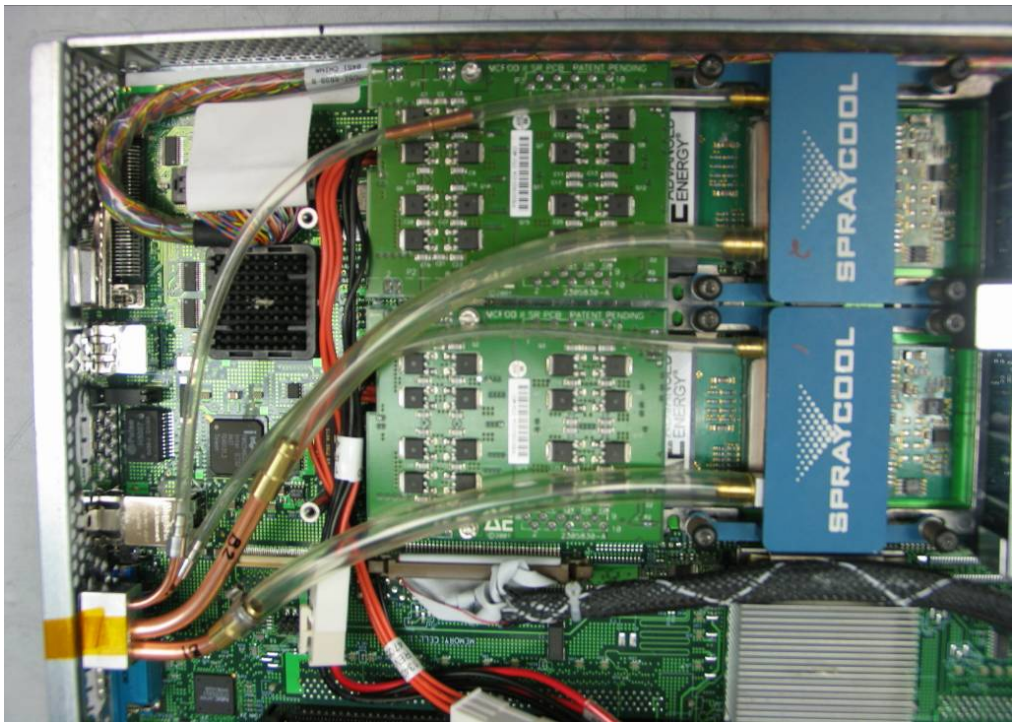


Figure 10: HP rx1620 with air-cooled heat sinks removed and SMK installed

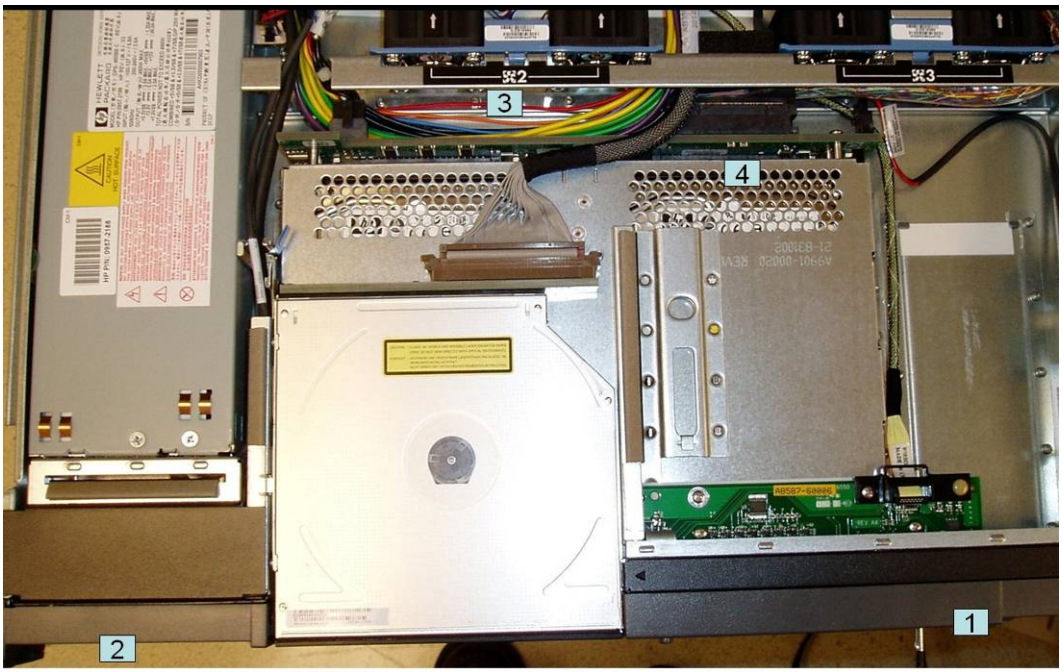
4.2.1 Test Methodology

Server Power Ranking

Once delivered to ISR the 16 rx1620s were characterized on the bench, see Appendix D, TMU Test Plan for test procedure details. Server AC inlet voltage, server AC current draw, CPU DC voltage, CPU DC current draw, CPU diode temperatures, and motherboard temperature were gathered for each server in an average ambient temperature of 23° C. Testing included running the servers in two states, while at idle and fully exercised. Eight instances of BurnI2 and a single instance of nbench were used to fully exercise the servers. Server inlet power and CPU power was calculated for each server. Servers were then ranked from 1-16 based on highest power draw. CPU power draw was used to discriminate servers having similar total inlet power draws. The three highest ranked servers (#1-#3) were then fully instrumented per the rx1620 Thermal Characterization Plan located in Appendix B. To summarize, memory, Memory Control Hub (MCH, also referred to as the Northbridge), various power Field Effect Transistors (FETs), air inlet, and air exhaust temperatures were monitored. Data from the server while air-cooled and cooled via SprayCool will be compared.

Server Air-Cooled Baseline

The three most powerful servers were fully instrumented per Appendix B. The instrumented components included all of the memory DIMMs, DC to DC CPU power conversion boards, the MCH, a hard drive, and the Ethernet controller. The air inlets and exhausts for the CPUs, the DIMMs, and the power supply were instrumented. Figures 11, 12, and 13 show the instrumented locations of the server. The remaining 13 servers (#4-#16) were instrumented to measure a single air inlet and the CPU and DIMM exhaust temperatures.



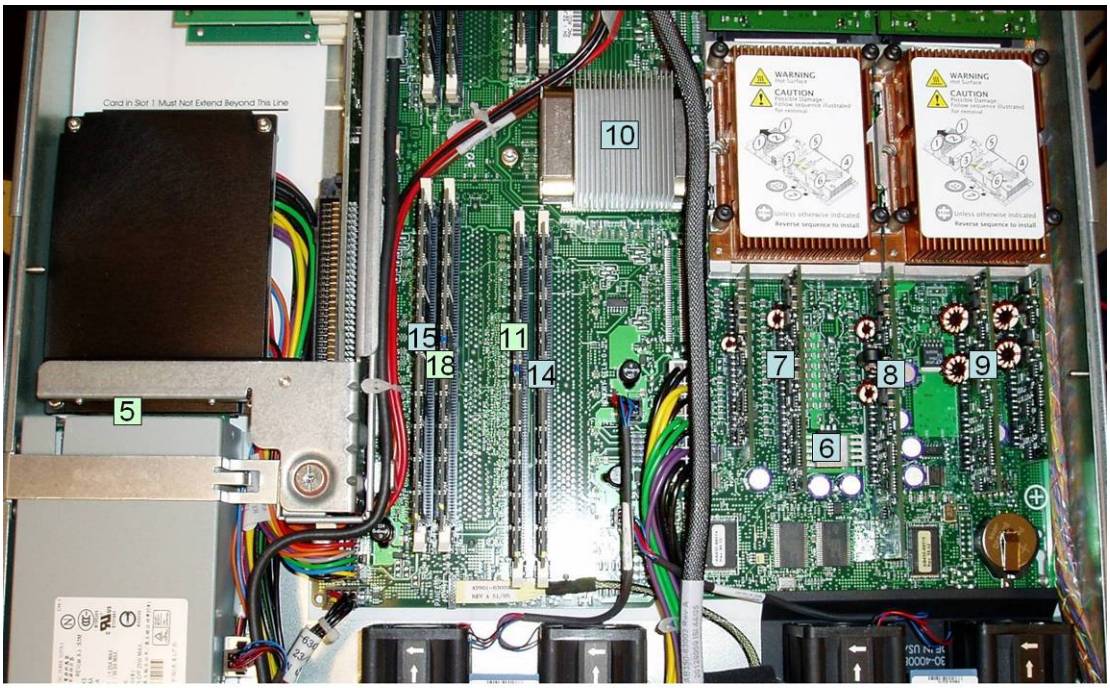
1 Air intake CPU, front of chassis

2 Air intake, power supply

3 Air intake, RAM side

4 Rear of HD

Figure 11: Fully instrumented server, air inlet, and hard drive instrumentation



5 Power supply exhaust

6-9 FETs in front of CPU

10 Northbridge

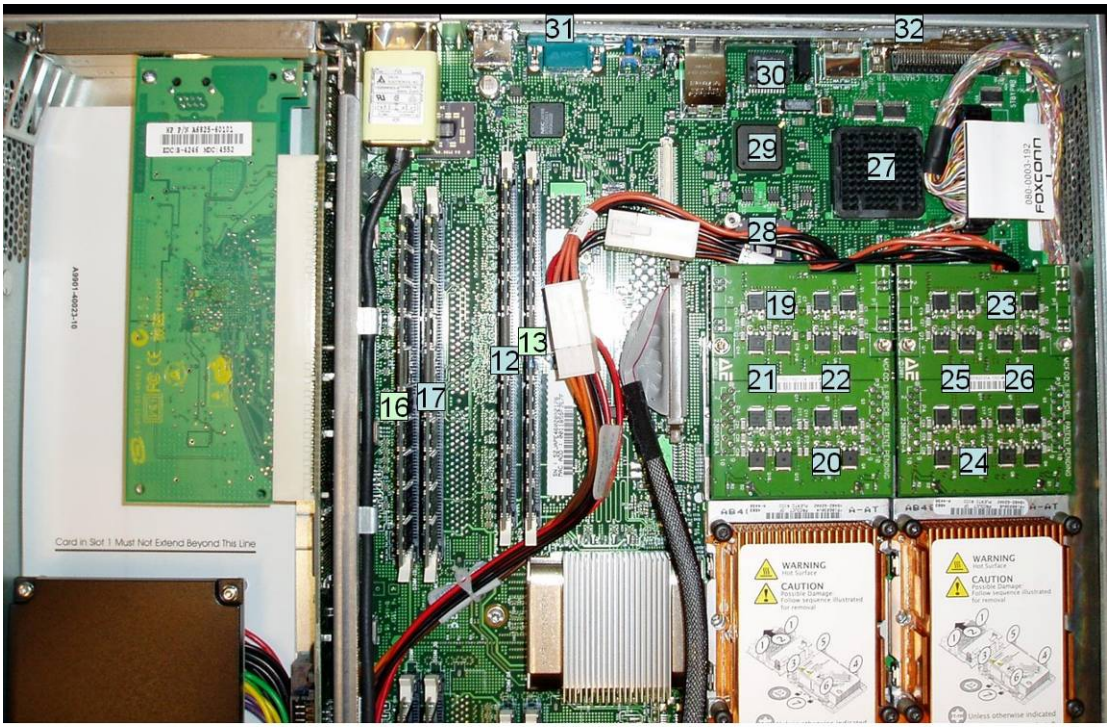
11 DIMM 0A

14 DIMM 3A

15 DIMM 0B

18 DIMM 3B

Figure 12: Fully instrumented server, memory DIMM, Northbridge (MCH), FET, and air exhaust instrumentation



12 DIMM 1A	17 DIMM 2B	27 Southbridge
13 DIMM 2A	19-22 CPU1 power boards, 2 on top, 2 in sandwich	28-30 Assorted ICs on board
16 DIMM 1B	23-26 CPU0 power boards, 2 on top, 2 in sandwich	31-32 Air out

Figure 13: Memory, CPU power board, Southbridge, and Ethernet instrumentation

The three high power servers were distributed throughout the rack, with #1 located at the top, #2 in the middle, and #3 located at the bottom of the stack. Figure 14, shows the delivered configuration along with the node locations. By distributing the nodes throughout the rack, the effects of server location to air inlet temperature and the amount of conductive heat transfer from server surroundings, could be observed. All temperature data was gathered using T-type thermocouples and a Keithley Integra 2750 multiplexing data acquisition system (DAQ). The DAQ was capable of monitoring 200 channels, and outputted data directly to a Microsoft Excel spreadsheet. CPU and server main board temperatures are gathered using the server's built in board management controller (BMC). The BMC interrogates the CPU's on-chip thermal diode and relays that to the user either through an executable which displays the information on screen or via the IPMI driver to the TMU. Thermocouples mounted in the CPU integrated heat sink (IHS, also referred to as the lid) were not used to measure temperatures during the Phase II project. However, extensive testing during the Phase I project revealed that the on board thermal diode was on average 13° C warmer than the lid mounted thermocouple.

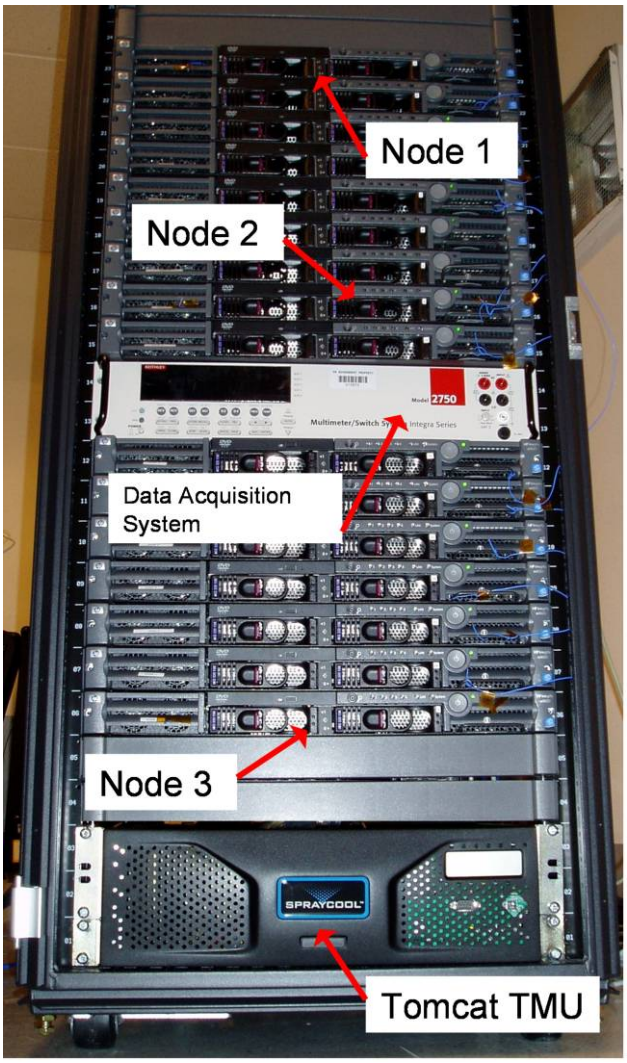


Figure 14: Stack of rx1620s with Phase II TMU in the final rack configuration

Test conditions included operating the servers in an idle state. During the power ranking tests, it was discovered the servers and CPUs dissipated more power and operated at higher temperatures at idle than when being exercised in a fully loaded state. Investigation with the vendor, HP, revealed that these servers in this configuration were always actively looking for jobs via Ethernet. Because of this, they were actually doing more work at idle, than when the CPUs were being exercised by the user. With this knowledge it was decided that the bulk of the air base lining and subsequent SprayCool testing would be done with the servers at idle. By definition the idle conditions consisted of only the Linux OS running with no other executables.

All testing was conducted in the ISR integration lab where air temperatures averaged 23° C. A small floor fan was used to circulate air from the bottom to the top of the rack, minimizing variations in air inlet temperatures.

4.2.2 SprayCool Server Thermal Performance Testing

SprayCool testing employed the same test conditions as the air baseline testing. Ambient air temperatures averaged 23° C in the ISR integration lab. The servers were operated in an idle condition and CPU diodes were interrogated using the BMC bus and IPMI systems. Component temperatures were monitored using the Keithley DAQ system.

The SprayCool system operated using a Fluorinert (dielectric) made by 3M Corporation (PF5060). PF5060 has the following approximate properties given 1 ATM, and room temperature: boiling point of 56° C, specific heat of 946 J/kg-K, viscosity of 0.009 Pa-s, thermal conductivity of 0.055 W/m-K, and a latent heat of vaporization of 94 kJ/kg.

To compare SprayCool operation to air-cooled operation the system was tested in normal operating conditions. The Phase II SprayCool system was designed to operate with the reservoir temperature at 20° C, reservoir pressure at 6.5 psi, and the pump discharge pressure at 20 psid. The system was also tested at reservoir temperatures of 30° C and 40° C, and pump discharge pressures of 15 psid and 25 psid. The reservoir pressure was maintained at 3 psi off of the fluid's saturation pressure based on reservoir temperature. The results of these tests are presented in section.

4.2.3 Burn-In Testing

Burn-in testing was undertaken to reveal weaknesses in the system which could be related to infant mortality as well as help characterize how the system would behave over long periods of time. When not being characterized the system was in burn-in testing. During Burn-in the servers were operated in an idle condition (highest power draw) with the SprayCool system operating normally. For the most part, this system operated 24 hours a day 7 days a week. All system parameters were monitored with emphasis being placed on reservoir pressure, pump discharge pressure, and CPU operating temperatures. Monitoring the reservoir pressure and the status of the system's active venting system provided insight into the system's air in leak rate, as it operated under a vacuum.

4.2.4 Characterization and Operational Testing

Characterization and operational testing included determining the exact capabilities of the SprayCool system as well as how the system would respond to extra-normal operating situations. Tests included testing how the TMU would react to catastrophic failures, the characterization of fluid interconnect leakage, the ability of the TMU to maintain pump discharge pressure, how the system reacts to a facility level water leak, how quickly the TMU could failover pumps, and the amount of power the system uses during normal operation. The test procedures and results are documented in Appendix D, Testing Goals, Scope, and Overview.

4.2.5 Test Results

4.2.5.1 Power Ranking Results

The air-cooled power rank testing identified the three most power servers. The most powerful server, serial # USE4606HJ8, registered 376.3 W (121.8 VAC, 3.10 A) when fully loaded and 383.6 W (121.4 VAC, 3.16 A) while at idle. These test results revealed that the fully loaded servers did not dissipate as much power as they did at idle. On average, the fully loaded servers dissipated 371.1 W and the idle servers dissipated 376.1 W, a difference of 5 W. This trend is also reflected in the CPU power draw and diode temperatures. On average the fully loaded CPUs dissipated 90.6 W and operated at 77.3° C. The idle CPUs dissipated 92.8 W, and operated at 81.1° C. The idle CPUs dissipated 2.2 W more power and were 3.9° C hotter. Table 1 shows the results for all (16) servers fully loaded, the table is organized by server total power draw. Table 2 shows the results for all (16) servers at idle, and the table is organized by server total power draw.

Table 1: HP rx1620 power ranking data with the servers being fully loaded, organized by server total power draw

Server Letter	Node Supply Voltage (VAC)	Node Supply Current (Amp)	Server Calculated Power (W)	CPU0 Diode (°C)	CPU1 Diode (°C)	CPU 0 Calculated Power (W)	CPU 1 Calculated Power (W)
N Loaded	121.8	3.10	377.0	74	82	90.47	95.38
F Loaded	121.8	3.09	376.4	79	78	95.30	90.40
K Loaded	121.4	3.10	376.3	76	80	90.32	95.23
G Loaded	122.0	3.08	375.3	74	83	89.23	95.31
O Loaded	122.0	3.07	373.9	76	84	89.16	94.12
I Loaded	122.0	3.06	373.7	70	78	88.07	90.62
B Loaded	122.3	3.05	373.0	80	82	92.56	90.54
L Loaded	121.8	3.05	371.5	72	76	87.03	89.42
D Loaded	121.7	3.05	371.2	74	83	89.02	92.70
E Loaded	121.9	3.04	370.9	76	77	91.33	90.99
P Loaded	121.9	3.04	370.0	74	82	88.15	91.88
A Loaded	121.2	3.05	369.7	74	79	88.00	92.93
H Loaded	122.3	3.00	366.9	76	79	89.23	89.42
J Loaded	122.0	2.99	364.8	75	79	86.85	90.11
M Loaded	122.0	2.99	364.8	72	78	85.76	91.73
C Loaded	121.6	2.99	363.0	73	78	87.15	90.38

Table 2: HP rx1620 power ranking data with the servers at idle, organized by server total power draw

Server Letter	Node Supply Voltage (VAC)	Node Supply Current (Amp)	Server Calculated Power (W)	CPU0 Diode (°C)	CPU1 Diode (°C)	CPU 0 Calculated Power (W)	CPU 1 Calculated Power (W)
K Idle	121.4	3.16	383.6	78	81	92.56	97.61
N Idle	121.9	3.14	383.1	76	83	92.79	98.96
F Idle	121.9	3.14	382.3	80	80	97.61	93.13
G Idle	121.8	3.11	378.9	75	85	91.55	98.29
I Idle	122.1	3.10	378.9	72	79	90.39	93.00
O Idle	122.0	3.10	378.4	77	85	91.40	96.50
B Idle	122.3	3.09	377.9	81	83	94.87	92.93
E Idle	121.6	3.09	375.1	77	76	93.64	92.18
D Idle	121.6	3.08	375.0	74	83	90.64	95.08
P Idle	121.6	3.08	374.3	74	82	89.23	94.19
H Idle	122.3	3.06	374.2	78	80	92.64	91.81
L Idle	121.9	3.07	373.7	75	82	89.81	93.60
A Idle	121.2	3.08	373.3	74	79	89.86	94.71
M Idle	122.0	3.04	370.3	73	79	86.85	92.85
C Idle	122.5	3.01	369.2	74	80	89.23	93.00
J Idle	121.6	3.04	369.2	76	80	89.09	91.66

The idle test results revealed that CPU powers ranged from a high of 98.96 W to 86.85 W. The CPU temperatures ranged from 85° C to 73° C.

After the ranking test, the final rack configuration was set with the highest power node at the top of the stack of (16) servers. The second highest power server was placed approximately in the middle of the stack, and the third highest power server was at the bottom of the stack. The layout is visible in Figure 14.

4.2.5.2 Air-Cooled Baseline Results

The air-cooled baseline results show that with the server idling CPU 0 averaged 77.4° C and CPU 1 averaged 83.1° C (see Table 3). This data is based on the average temperatures over three tests. There were some temperature variations during the three test runs. Two servers, #1 and #7, have standard deviations which are greater than two degrees. However, the remaining (14) servers have standard deviations ≤ 1.0° C. On average, the data acquisitions system measured ambient air temperatures ranging from 22.9° C to 25.8° C (Table 4), based on the location in the rack.

Table 3: HP rx1620 air cooled baseline CPU temperature data

Cooling	Air	**Reservoir Temperature (°C)**	NA
Ave Amb Temp (°C)	24.1	**Reservoir Pressure (psia)**	NA
Server Load	Burn	**Pump Discharge Pressure (psid)**	NA

CPU0 temperature				**Statistics**			
Node	**Run 1 (°C)**	**Run 2 (°C)**	**Run 3 (°C)**	**Average (°C)**	**Minimum (°C)**	**Maximum (°C)**	**Std Deviation (°C)**
1	83	79	78	80.0	78	83.0	2.6
2	75	76	75	75.3	75	76.0	0.6
3	82	83	83	82.7	82	83.0	0.6
4	76	75	74	75.0	74	76.0	1.0
5	75	76	75	75.3	75	76.0	0.6
6	78	79	77	78.0	77	79.0	1.0
7	83	84	78	81.7	78	84.0	3.2
8	82	82	81	81.7	81	82.0	0.6
9	76	77	76	76.3	76	77.0	0.6
10	74	74	73	73.7	73	74.0	0.6
11	81	82	81	81.3	81	82.0	0.6
12	74	74	74	74.0	74	74.0	0.0
13	76	76	76	76.0	76	76.0	0.0
14	75	76	75	75.3	75	76.0	0.6
15	75	76	75	75.3	75	76.0	0.6
16	76	77	77	76.7	76	77.0	0.6
CPU1 Temperature				77.4			
1	87	83	81	83.7	81	87.0	3.1
2	83	83	83	83.0	83	83.0	0.0
3	81	81	81	81.0	81	81.0	0.0
4	86	87	86	86.3	86	87.0	0.6
5	84	84	82	83.3	82	84.0	1.2
6	86	86	84	85.3	84	86.0	1.2
7	86	86	81	84.3	81	86.0	2.9
8	81	81	80	80.7	80	81.0	0.6
9	86	87	86	86.3	86	87.0	0.6
10	82	82	82	82.0	82	82.0	0.0
11	84	84	84	84.0	84	84.0	0.0
12	82	82	82	82.0	82	82.0	0.0
13	82	82	82	82.0	82	82.0	0.0
14	82	82	82	82.0	82	82.0	0.0
15	82	82	82	82.0	82	82.0	0.0
16	82	82	82	82.0	82	82.0	0.0
Motherboard Ambient				83.1			
1	22	23	22	22.3	22	23.0	0.6
2	21	21	20	20.7	20	21.0	0.6
3	20	20	20	20.0	20	20.0	0.0
4	23	24	23	23.3	23	24.0	0.6

Table 4: Average air-cooled rx1620 component temperatures over 3 test runs

Cooling	Air	**Reservoir Temperature (°C)**	NA
Ave Amb Temp (°C)	24.1	**Reservoir Pressure (psia)**	NA
Server Load	Burn	**Pump Discharge Pressure (psid)**	NA

	Server 1	**Server 2**	**Server 3**	**Statistics**			
Component	Temp (°C)	Temp (°C)	Temp (°C)	Average (°C)	Minimum (°C)	Maximum (°C)	Std Deviation
RH AIR IN	25.8	23.6	22.9	24.1	22.9	25.8	1.5
PS AIR IN	25.4	24.3	23.1	24.2	23.1	25.4	1.2
RAM AIR IN	30.0	27.9	27.5	28.5	27.5	30.0	1.3
HD 0	32.9	30.0	30.1	31.0	30.0	32.9	1.6
PS AIR OUT	44.1	44.2	42.7	43.6	42.7	44.2	0.8
FET 1	31.7	32.4	32.4	32.2	31.7	32.4	0.4
FET 2	38.9	38.8	39.5	39.1	38.8	39.5	0.4
INDUCTOR	32.3	37.1	38.0	35.8	32.3	38.0	3.0
FET 3	30.3	28.4	29.5	29.4	28.4	30.3	0.9
NORTHBRIDGE	46.8	41.7	45.5	44.7	41.7	46.8	2.6
DIMM 0A	45.3	38.0	36.1	39.8	36.1	45.3	4.9
DIMM 1A	46.8	44.6	46.1	45.9	44.6	46.8	1.1
DIMM 2A	51.1	44.7	50.5	48.8	44.7	51.1	3.5
DIMM 3A	42.7	36.7	39.3	39.6	36.7	42.7	3.0
DIMM 0B	45.9	39.7	43.7	43.1	39.7	45.9	3.1
DIMM 1B	51.7	42.9	48.2	47.6	42.9	51.7	4.4
DIMM 2B	51.6	44.1	51.5	49.1	44.1	51.6	4.3
DIMM 3B	45.2	42.6	44.6	44.1	42.6	45.2	1.4
CPU1 PP1	71.9	80.4	70.6	74.3	70.6	80.4	5.3
CPU1 PP2	67.3	69.9	69.5	68.9	67.3	69.9	1.4
CPU1 PP3	67.9	75.4	70.1	71.1	67.9	75.4	3.8
CPU1 PP4	66.8	74.3	74.6	71.9	66.8	74.6	4.4
CPU0 PP1	65.9	62.6	77.6	68.7	62.6	77.6	7.9
CPU0 PP2	62.2	60.2	72.7	65.0	60.2	72.7	6.7
CPU0 PP3	66.7	66.1	77.6	70.1	66.1	77.6	6.5
CPU0 PP4	63.9	63.8	72.4	66.7	63.8	72.4	5.0
SOUTHBRIDGE	66.6	65.6	65.4	65.9	65.4	66.6	0.6
AGILENT CHIP	58.6	58.2	56.4	57.7	56.4	58.6	1.2
ETHERNET CHIP	67.7	60.1	52.4	60.1	52.4	67.7	7.7
PULSE CHIP	53.4	48.8	47.1	49.8	47.1	53.4	3.3
AIR OUT RAM	35.7	33.8	34.2	34.6	33.8	35.7	1.0
AIR OUT CPU	49.6	56.5	51.6	52.6	49.6	56.5	3.5

Air cooled DIMM average temperatures ranged from 36.1° C to 51.7° C depending on the server and DIMM location. The averages were taking over the three test runs. Average Northbridge temperatures ranged from 41.7° C to 46.8° C across the three fully instrumented servers. The warmest components

were the power conversion FETs located on the CPU power pods, where temperatures ranged from 62.6° C to 80.4° C.

4.2.5.3 SprayCool Results

Table 5: SprayCool HP rx1620 CPU diode temperatures

Cooling	SprayCool
Ave Amb Temp (°C)	21.5
Server Load	Burn

Reservoir Temp (°C)	39.3
Reservoir Pressure (psia)	11.8
Pump Dis Press (psid)	14.9

	CPU 0	**CPU 1**	Mboard/Amb
Node	**(°C)**	**(°C)**	**(°C)**
1	78	79	20
2	77	77	18
3	79	78	17
4	78	81	21
5	75	78	23
6	78	80	20
7	80	77	20
8	77	74	18
9	77	77	18
10	76	78	18
11	78	76	18
12	77	80	20
13	76	78	18
14	76	76	18
15	75	75	18
16	76	77	18

Table 6: SprayCool HP rx1620 component temperatures

Cooling	SprayCool	Reservoir Temperature (°C)	40.0
Ave Amb Temp (°C)	24.1	Reservoir Pressure (psia)	NA
Server Load	Burn	Pump Discharge Pressure (psid)	15.0

	Server 1	Server 2	Server 3	Statistics			
Description	Temp (°C)	Temp (°C)	Temp (°C)	Average (°C)	Minimum (°C)	Maximum (°C)	Std Deviation (°C)
RH AIR IN	23.4	20.6	20.4	21.5	20.4	23.4	1.7
PS AIR IN	23.2	21.3	24.5	23.0	21.3	24.5	1.7
RAM AIR IN	27.9	26.0	29.2	27.7	26.0	29.2	1.6
HD 0	29.8	27.6	27.6	28.3	27.6	29.8	1.2
PS AIR OUT	44.9	44.7	45.2	44.9	44.7	45.2	0.2
FET 1	29.6	30.6	30.7	30.3	29.6	30.7	0.6
FET 2	36.5	37.5	37.3	37.1	36.5	37.5	0.5
INDUCTOR	28.2	34.6	36.1	33.0	28.2	36.1	4.2
FET 3	27.9	25.6	25.6	26.4	25.6	27.9	1.4
NORTHBRIDGE	42.2	37.4	41.0	40.2	37.4	42.2	2.5
DIMM 0A	43.3	36.5	37.5	39.1	36.5	43.3	3.7
DIMM 1A	44.9	44.3	NA	44.6	44.3	44.9	0.4
DIMM 2A	49.3	43.9	52.2	48.4	43.9	52.2	4.2
DIMM 3A	40.6	35.0	39.6	38.4	35.0	40.6	3.0
DIMM 0B	43.9	38.4	43.9	42.1	38.4	43.9	3.2
DIMM 1B	49.4	41.9	47.9	46.4	41.9	49.4	4.0
DIMM 2B	49.3	43.3	50.3	47.6	43.3	50.3	3.8
DIMM 3B	43.6	41.5	45.9	43.7	41.5	45.9	2.2
CPU1 PP1	50.5	58.9	51.6	53.7	50.5	58.9	4.6
CPU1 PP2	46.9	45.8	42.8	45.2	42.8	46.9	2.1
CPU1 PP3	55.1	57.0	50.2	54.1	50.2	57.0	3.5
CPU1 PP4	59.4	54.8	56.2	56.8	54.8	59.4	2.3
CPU0 PP1	52.9	44.4	48.0	48.4	44.4	52.9	4.3
CPU0 PP2	45.1	44.3	45.3	44.9	44.3	45.3	0.5
CPU0 PP3	50.7	50.0	52.2	51.0	50.0	52.2	1.1
CPU0 PP4	56.1	51.7	50.5	52.8	50.5	56.1	3.0
SOUTHBRIDGE	50.0	51.1	43.7	48.2	43.7	51.1	4.0
AGILENT CHIP	47.6	44.2	46.3	46.0	44.2	47.6	1.7
ETHERNET CHIP	59.6	52.8	46.9	53.1	46.9	59.6	6.4
PULSE CHIP	39.3	39.0	39.5	39.3	39.0	39.5	0.3
AIR OUT RAM	35.0	33.2	31.3	33.2	31.3	35.0	1.9
AIR OUT CPU	37.3	37.4	35.9	36.9	35.9	37.4	0.8

Table 7: HP rx1620 SprayCool and air cooled CPU diode temperature comparison

CPU0 Server #	Air Cooled CPU0 (°C)	SprayCool #1: 20C, 25 psid (°C)	SprayCool #2: 40C, 15 psid (°C)	Temp Diff Air to SprayCool #1 (°C)	Normalized Diff to amb air (°C)	Temp Diff Air to SprayCool #2 (°C)	Normalized Diff to amb air (°C)
1	78	62	78	16	10.6	0	-4.4
2	76	61	77	15	9.6	-1	-5.4
3	80	62	79	18	12.6	1	-3.4
4	75	62	78	13	7.6	-3	-7.4
5	72	59	75	13	7.6	-3	-7.4
6	77	62	78	15	9.6	-1	-5.4
7	81	64	80	17	11.6	1	-3.4
8	77	61	77	16	10.6	0	-4.4
9	74	59	77	15	9.6	-3	-7.4
10	74	60	76	14	8.6	-2	-6.4
11	78	62	78	16	10.6	0	-4.4
12	75	61	77	14	8.6	-2	-6.4
13	74	59	76	15	9.6	-2	-6.4
14	73	59	76	14	8.6	-3	-7.4
15	74	59	75	15	9.6	-1	-5.4
16	76	60	76	16	10.6	0	-4.4
amb temp	25.9	20.5	21.5	5.4	-	4.4	-
			Average	**15.1**	**9.7**	**-1.2**	**-5.6**
CPU1							
1	81	63	79	18	12.6	2	-2.4
2	83	61	77	22	16.6	6	1.6
3	80	62	78	18	12.6	2	-2.4
4	85	65	81	20	14.6	4	-0.4
5	79	63	78	16	10.6	1	-3.4
6	85	63	80	22	16.6	5	0.6
7	83	61	77	22	16.6	6	1.6
8	76	58	74	18	12.6	2	-2.4
9	83	62	77	21	15.6	6	1.6
10	82	62	78	20	14.6	4	-0.4
11	80	60	76	20	14.6	4	-0.4
12	82	64	80	18	12.6	2	-2.4
13	79	62	78	17	11.6	1	-3.4
14	79	60	76	19	13.6	3	-1.4
15	80	59	75	21	15.6	5	0.6
16	80	61	77	19	13.6	3	-1.4
			Average	**19.4**	**14.0**	**3.5**	**-0.9**

4.3 Global System Upgrade

In Phase I of the ESDC program a development system was deployed at PNNL to show a future path to even greater densities of compute electronics. Known as the Global System, the system was comprised of an 8U chassis that is hermetically sealed and contains a backplane that accepts up to sixteen 6U cPCI based blade servers that are completely spray cooled. So, in contrast to the 1U and 2U servers, all the compute electronics are liquid cooled, which results in higher densities of computing power and greater efficiencies in thermal management. As a reference, the Global System is designed to handle up to 500 W per slot in the chassis, or 8 kW total in an 8U package. The implication here is the demonstration of technology that could soon lead to 30kW+ racks that would enable petascale computing.

4.3.1 Global System Computer Technology Description

In Phase I, four MPB's were delivered, each with dual 2.4 GHz Opteron CPUs and 2 GB of memory per board. Phase II provided an opportunity to upgrade this system to a total of 8 boards, each with 16GB of memory per board for all 8 boards. The result is a theoretical 153.6 GFLOPs per 8U chassis, or 768 GFLOPs per 42U rack.

Achieving 16GB of memory on a 6U cPCI (Compact Peripheral Component Interconnect) board required the development of a custom memory module that utilized stacked memory packages. Furthermore, the density of these packages on the module was such that cooling could only be effectively achieved through the use of evaporative spray cooling. Smart Modular Technologies was contracted to develop these modules with the help of stacked memory from Vertical Circuits Inc.

Figures 15 and 16 show top and bottom views of the 4 GB Smart Modular memory module. This part has completed a pilot run and these modules can now be ordered in production.

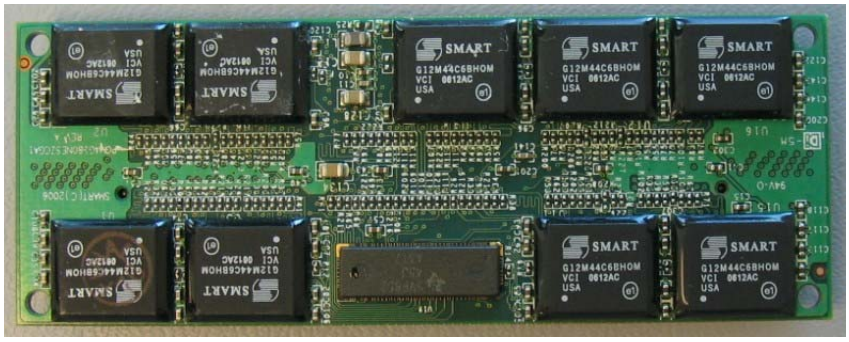


Figure 15: Top Side of 4 GB Memory Module

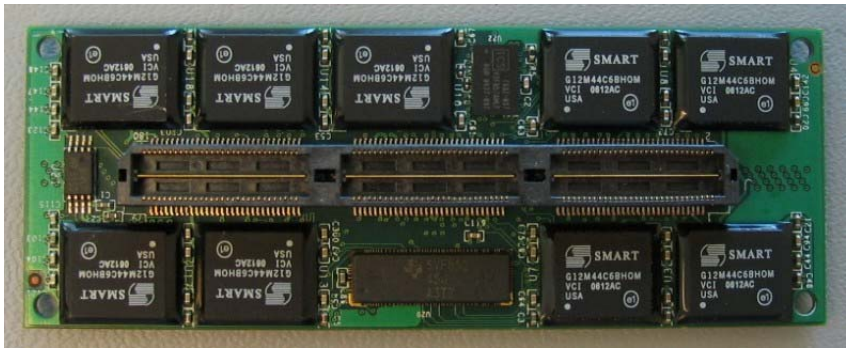


Figure 16: Bottom Side of 4 GB Memory Module

Four of these modules can fit on each blade, with spray directed onto them via the top and bottom card guides in the Global System. A view of the MPB with memory modules loaded is shown in Figure 17.

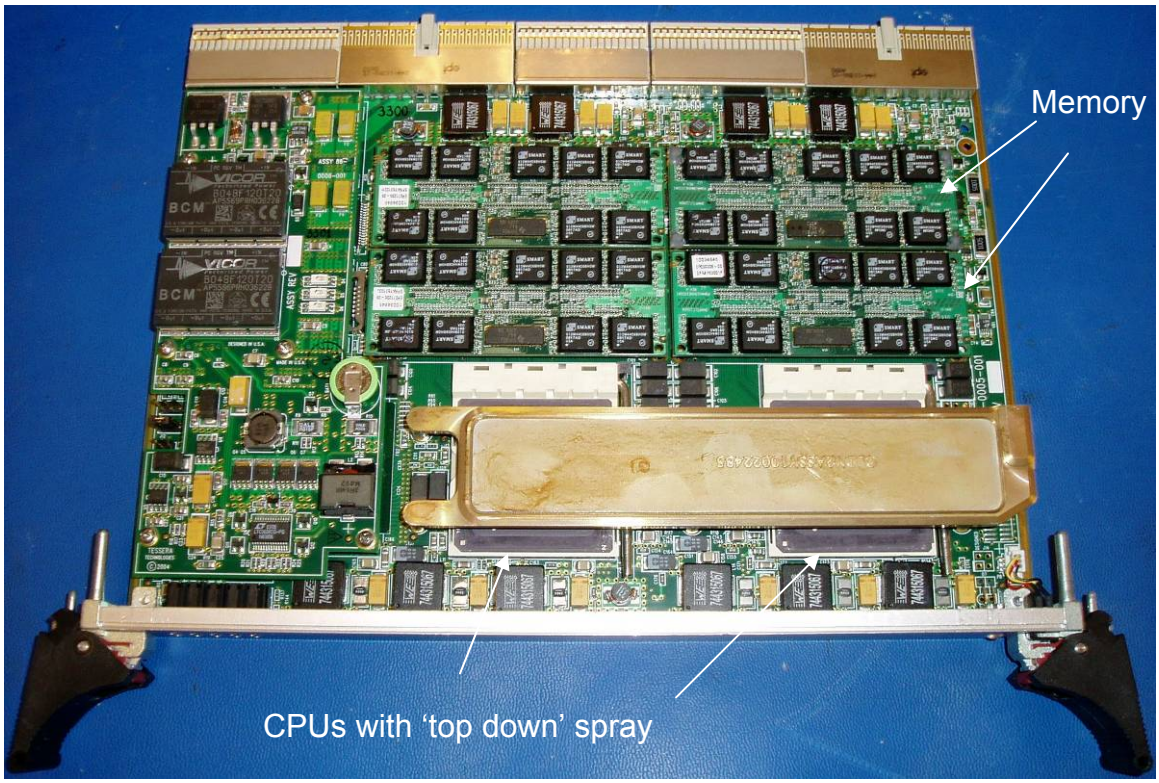


Figure 17: MPB with 16 GB of Memory

4.4 1U SprayCool ATX Server

In support of the goal of easing OEM adoption of SprayCool technology, a 1U SprayCool optimized server (shown below in Figure 18) was developed in Phase II to serve as a reference design for potential manufacturers. The objective was to provide a system that was low cost by leveraging industry standard processors, memory, enclosures and power supplies. Further, this reference

platform can be tested in customer environments for application validation and verification while providing flexibility that will allow for adaptation to a wide variety of enclosures.

Figure 18: SprayCool 1U Server

This was achieved primarily by following the Advanced Technology Extended (ATX) standard, common throughout the industry, together with standard components. What is unique about the design is primarily the board layout, which provides suitable space claims for fluid routing to the spray modules. Furthermore, because the CPU heat is being removed by the fluid, there was less concern for thermal issues on the 'downstream' components. This allowed for selection of high performance application-specific integrated circuits (ASICs) chipsets and ample memory for each processor socket.

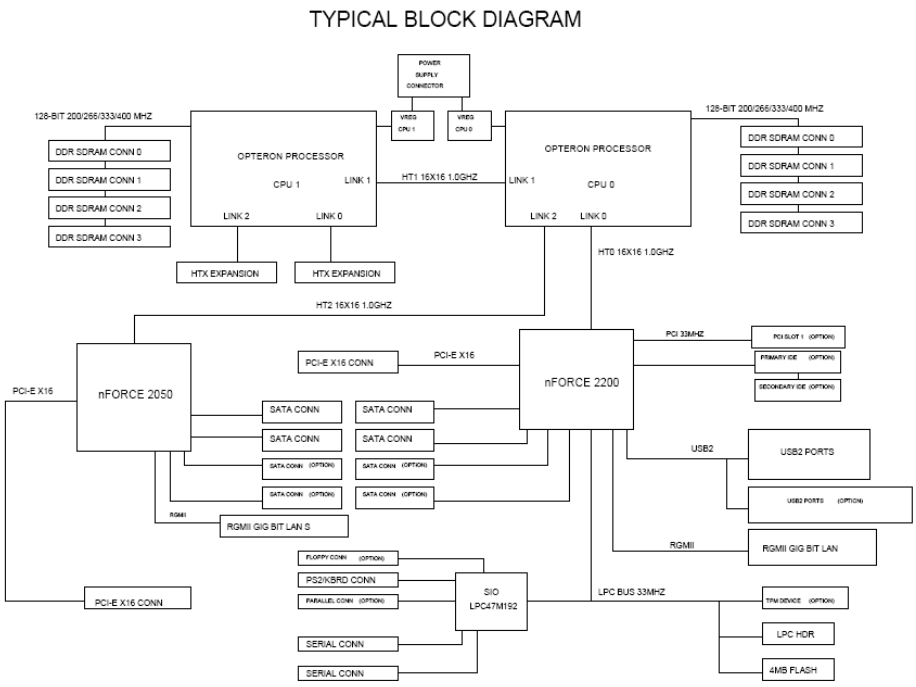


Figure 19: ATX Server Block Diagram

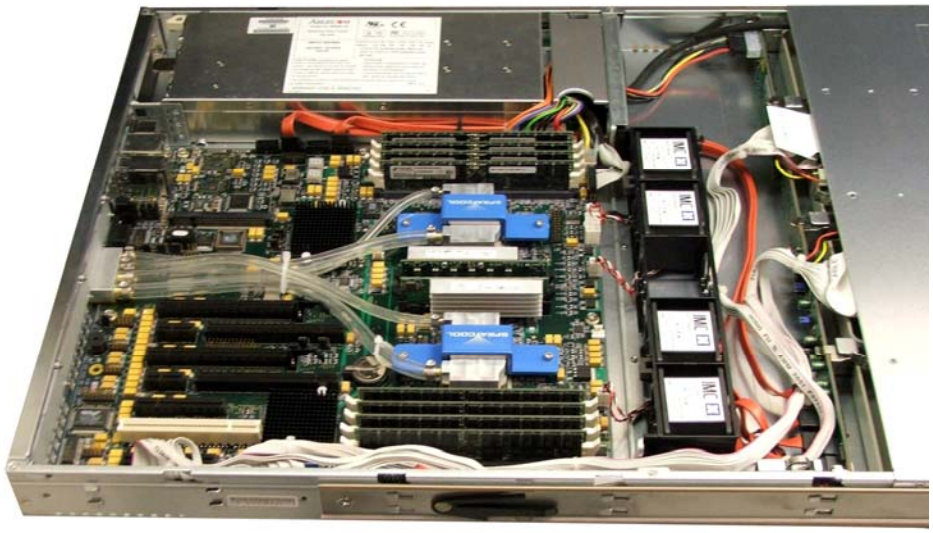


Figure 20: ATX board layout and fluid routing

The initial prototypes delivered under the contract were integrated with a standard chassis offering from Supermicro, Inc. However, due to the industry standard form factor of the board, integration should be feasible with a wide range of vendors with standard chassis.

Appendix A: Thermal Characterization of a Server

Goal:

The goal of this appendix is to explain how to thermally characterize a server for future comparison of different cooling solutions.

Specific Goals:

- Determine power going into server
  - Since the server does no mechanical work, all power is converted to heat.
- Determine the resulting temperatures of the components inside the chassis
  - The specific measurement locations will vary by server; this document will serve as a general guide.
  - The ΔT between ambient air component temperatures is particularly important.

Power Measurement:

The power measurement between the wall and server is the most basic.

- Using two multimeters, connect one to the wall socket below the one used to measure RMS voltage.
- Create a special power cable to measure the current for the server. Cut one of the AC lines in the cord and reroute it through the multimeter and measure the current.
- Optionally, if the current draw is too high for the multimeter, use the clamp on ammeter around one of the AC lines

Some system architectures such as the Itanium allow the power of the individual CPUs to be measured.

- Use a shunt resistor inline with the power going to the CPU.
  - Measure the voltage in the line and voltage drop across the shunt resistor
  - The power calculations are shown below

$$I = \frac{V_{shunt}}{R_{shunt}}$$

$$Power = V_{line} * I$$

Thermocouple Instrumentation:

The air temperatures going into and out of the server should be measured, as well as the temperature of each component. The specific components to be measured will vary by server, but the following are general guidelines. Component thermocouples (TCs) should be attached with a thermal epoxy.

Air Temperature Measurement:

- Air temperatures should be measured at the front and rear of the server. Use two thermocouples on the rear of the chassis and put one in line with the CPU airflow

Components:

- RAM temperatures
  - Measure the temperatures from the center of the top of the stick, one TC per stick
- Northbridge/Southbridge
  - Instrument the heat sink of the Northbridge and Southbridge with a thermocouple. The chipset is one of the hotter components on the motherboard.
- CPU
  - The CPU temperature can either be monitored during runtime through the thermal diode or by grooving the heat sink or processor and installing a thermocouple. The latter gives better results but requires machining.
- Various other components
  - Measure several power FETs on the board, especially if their cooling airflow will be altered by the different cooling solutions.
  - Instrument any other component on the board that appears to have additional heat sinking and could have differing performance depending on the cooling solution.

Appendix B: RX1620 Thermal Characterization

Goal:

The RX1620 chassis from HP is a new SprayCool™ platform that is currently under development. The goals of the first rack are to characterize SprayCool in a new system and compare the performance to HP's air-cooled solution.

Instrumentation:

In order to fully characterize the system, three blades are fully instrumented with thermocouples. The remaining 13 servers are instrumented with air intake and exhaust thermocouples.

Each fully instrumented chassis has two thermocouples at the intake of the server, one in the CPU and motherboard airflow and the other at the inlet of the power supply fans. Additionally, there is a thermocouple behind the hard drives in the air path of the RAM. The exhaust air temperature is measured behind the CPUs, behind the RAM, and behind the power supply.

Many of the components, one being the motherboard, are also measured. Each stick of RAM will have one thermocouple attached to the memory controller chip. The CPUs have a series of five power regulator boards in front of them; four thermocouples are distributed among the FETS and one inductor. The Northbridge and Southbridge are instrumented; the Southbridge is behind the CPU in the air path. The CPUs themselves have power regulator cards attached behind them air-flow-wise. Each card is two boards sandwiched together with an inch of airspace in between. Each card is instrumented with two thermocouples on the top of the top board and two in the middle of the sandwich. The motherboard also has several integrated circuits (ICs) distributed around the inputs/output (I/O) in the back of the system. The Ethernet chip is instrumented, as well as an Agilent chip and power regulator. The main hard drive at the front of the system is instrumented as well.

TC usage:

The thermocouple total for each fully instrumented server is 32. Table 8 shows the total number of thermocouples required, including three fully instrumented servers, and 13 servers with a single air intake and two exhaust temperatures. Table 9 lists each individual thermocouple location.

Table 8: Thermocouple total usage

Where	Number
Fully Instrumented	32 * 3 = 96
Non instrumented	13 * 3 = 39

Total	**96 + 39 = 135**

Table 9: Thermocouple locations for fully instrumented

Location	Number
Rear of hard drives	1
Air intake, right hand side of chassis	1
Air intake, front of RAM	1
Air intake, front of power supply	1
RAM temperatures, 1 measurement per stick	8
Air exhaust, rear of power supply	1
Power boards in front of CPU	4
Northbridge (MCH)	1
Power cards for CPUs, 2 on top, 2 inside sandwich	8
Southbridge	1
Assorted ICs on board	3
Air exhaust measurement, behind CPU	1
Air exhaust measurement, behind RAM	1
Total	**36**

Appendix C: Phase I Data for CPU IHS Mounted Thermocouple

File	Processor	Amb (°C)	Processor Power (W)	Res Press (psia)	Pump Discharge (psid)	Facility Water (°C)	Coolant In (°C)	Cold Plate (°C)	Lid Temp (°C)	Diode (°C)	Spray Module to Coolant (°C)	Diode to Lid (°C)	Resistance Case-to-Water (°C/W)	Resistance Diode-to-Water (°C/W)
SPRAY COOLING														
051905f	CPU 0	24.5	85.6	13.5	20.0	7.0	18.2	37.8	42.2	55.0	19.6	12.8	0.41	0.56
	CPU 1	24.5	82.5	13.5	20.0	7.0	18.2	37.0	41.9	56.0	18.8	14.1	0.42	0.59
051905f	CPU 0	24.5	85.7	13.5	20.0	7.0	18.2	37.9	42.3	55.0	19.7	12.7	0.41	0.56
	CPU 1	24.5	82.5	13.5	20.0	7.0	18.2	36.6	41.5	55.0	18.4	13.5	0.42	0.58
051905G	CPU 0	24.5	85.7	14.1	20.0	7.0	19.8	39.2	43.6	57.0	19.4	13.4	0.43	0.58
	CPU 1	24.5	82.9	14.1	20.0	7.0	19.6	39.0	43.8	57.0	19.4	13.2	0.44	0.60
AIR-COOLING														
012705d	CPU 0	23.9	92.7	NA	NA	7.0	NA	NA	60*	73.0	NA	13.0	0.57	0.71
	CPU 1	23.9	82.8	NA	NA	7.0	NA	NA	58*	71.0	NA	13.0	0.62	0.77

Appendix D: TMU Test Plan

Isothermal Systems Research

			Document No.	10038463	
			Date Release	06/02/2006	
			Revision	T00	Page 1 of 4

Liquid Cooling and Catastrophic Failure Test

APPROVALS

Title	Name	
Program Manager:	Harley McAllister	
Lead Systems Engineer:	Rich Maes	
Originator:	Levi Westra	

DOCUMENT REVISION SUMMARY

Effective Date	Revision Level	Document Change Description
06/02/06	T00	Original Release

File is located in DBWORKS

1.0 GOAL

Test the functionality of the Tomcat TMU/PCS to react to a loss of cooling water or other catastrophic condition

Specific Goals:

- Gather data on the TMU's functionality, specifically on its ability to deal with catastrophic situations (loss of cooling water, system over pressure, pump cavitation, system over temperature, system under temperature, pump failure, low fluid level)
- Improve the PCS action algorithm. What things can be changed to improve the action(s) and their sequence?

2.0 SCOPE

- Engineering documentation of test plan, results, and conclusions.

3.0 OVERVIEW

3.1 Evaluate the reliability, and effectiveness of SC system to protect the computing hardware. Also, to provide a venue for iterating on the intended actions to be taken by the Tomcat/PCS to protect.

3.2 Document results and conclusions electronically in this document and save to DBWORKS for archive.

4.0 TEST SETUP AND INSTRUMENTATION

			Document No.	10038455	
			Date Release	06/01/2006	
			Revision	T00	Page 1 of 3

Leak Rate for Mated Fluid Connector Pairs and Tubing

APPROVALS

Title	Name	
Program Manager:	Harley McAllister	
Lead Systems Engineer:	Rich Maes	
Originator:	Levi Westra	

DOCUMENT REVISION SUMMARY

Effective Date	Revision Level	Document Change Description
06/01/06	T00	Original Release

File is located in DBWORKS

1.0 GOAL

The goal of this test is to gather data on the leak rate of PF5060 from the mated fluid connectors and tubing used in the PNNL PH2 system, and the air ingress into the system through the same artifacts.

Specific Goals:

- Leak rate (mL/yr) of 35C PF5060 out of the system through the Tygothane tubing and faster connectors used in the PNNL PH2 system
  - Note, this may be verified via documentation (OTS) or experimentation.
- Amount of air permeation in through the components

2.0 SCOPE

- Engineering documentation of test plan, results, and conclusions.

3.0 OVERVIEW

3.1 A quantitative understanding of the amount of fluid loss specific components in a SC rack system, specifically the PNNL PH2 modular rack system cooling (16) HP rx1620s, (2) Opus SC 1U servers, and a global system with (8) MPBs. The major components, TMU, SMKs, Global Chassis, are characterized during manufacturing. The OTS components, fluid routing (tubing) and connectors are not.

3.2 A quantitative understanding of the amount of air permeation into the system through the fluid routing (tubing) and mated fluid connectors.

3.3 Document results and conclusions electronically in this document and save to DBWORKS for archive.

			Document No.	10038828	
			Date Release	06/05/2006	
			Revision	T00	Page 1 of 4

Tomcat Pump Pressure Accuracy Test

APPROVALS

Title	Name	
Program Manager:	Harley McAllister	
Lead Systems Engineer:	Rich Maes	
Originator:	Levi Westra	

DOCUMENT REVISION SUMMARY

Effective Date	Revision Level	Document Change Description
06/05/06	T00	Original Release

File is located in DBWORKS

1.0 GOAL

Collect data on the ability of the Tomcat TMU to control pump pressure in steady state operating environment.

Specific Goals:

- Measure the average perturbation of the pump discharge pressure over time.
- Verify that this average perturbation value does not change with or without load being applied the SC system
- Verify the accuracy of the TMU's pressure transducers

2.0 SCOPE

- Engineering documentation of test plan, results, and conclusions.

3.0 OVERVIEW

3.1 Record the normal operating perturbations of the Tomcat TMU while operating in a normal steady state. Compare the results for when the system is cold and when the system is rejecting the full amount of the PNNL Phase 2 setup's heat.

3.2 Verify the accuracy of the TMU's pressure transducers to document the validity of the TMU's status reports.

4.0 TEST SETUP AND INSTRUMENTATION

4.1 SprayCool system configured to PNNL Phase 2 (see ICD document 10038462)

4.1.1 Tomcat TMU

4.1.2 Rack manifold assembly (10034419) plumbed to the SC rack using PF plumbing kit (10035432)

			Document No.	10038456	
			Date Release	06/01/2006	
			Revision	T00	Page 1 of 3

Water Detection Functionality and Reliability Test

APPROVALS

Title	Name	
Program Manager:	Harley McAllister	
Lead Systems Engineer:	Rich Maes	
Originator:	Levi Westra	

DOCUMENT REVISION SUMMARY

Effective Date	Revision Level	Document Change Description
06/01/06	T00	Original Release

File is located in DBWORKS

1.0 GOAL

Test the functionality of the water detection system. This includes the ability of the Tomcat TMU/PCS to detect and react to the notification of a leak, as well as the RLE water detectors ability to detect a leak. Reaction to a leak includes the ability to soft power of the rack of servers, the global spray cool system, notify the facility etc..

Specific Goals:

- Gather data on the TMU's ability to detect and react to a water leak. Does it detect and act accordingly every test?
- Improve the PCS action algorithm. What things can be changed to improve the action(s) and their sequence?

2.0 SCOPE

- Engineering documentation of test plan, results, and conclusions.

3.0 OVERVIEW

3.1 Evaluate the reliability, and effectiveness of SC system to protect the computing hardware. Also, to provide a venue for iterating on the intended actions to be taken by the Tomcat/PCS to protect.

3.2 Document results and conclusions electronically in this document and save to DBWORKS for archive.

4.0 TEST SETUP AND INSTRUMENTATION

4.1 SprayCool system configured to PNNL Phase 2 (see ICD document 10038462)

			Document No.	10038729	
			Date Release	06/02/2006	
			Revision	T00	Page 1 of 3

Tomcat Pump Failover Response Time and Pressure Pertubation Test

APPROVALS

Title	Name	
Program Manager:	Harley McAllister	
Lead Systems Engineer:	Rich Maes	
Originator:	Levi Westra	

DOCUMENT REVISION SUMMARY

Effective Date	Revision Level	Document Change Description
06/02/06	T00	Original Release

File is located in DBWORKS

1.0 GOAL

Measure the amount of time and pressure perturbation generated when the Tomcat TMU/PCS fails over from the primary pump to the secondary pump.

Specific Goals:

- Time record for the Tomcat TMU to recover from a pump fail over
- Maximum pressure perturbation during the fail over process
- Verify functionality of the failed over pump (pressure stability etc.)

2.0 SCOPE

- Engineering documentation of test plan, results, and conclusions.

3.0 OVERVIEW

3.1 Initiate a fail over command on the PNNL Phase 2 SC rack system while it is running at steady state.

4.0 TEST SETUP AND INSTRUMENTATION

4.1 SprayCool system configured to PNNL Phase 2 (see ICD document 10038462)

- 4.1.1 Tomcat TMU
- 4.1.2 Rack manifold assembly (10034419) plumbed to the SC rack using PF plumbing kit (10035432)
- 4.1.3 (16) HP rx1620 servers (10033876) integrated with SC conversion kit (10035428)

4.2 Data acquisition computer capable of communicating with the Tomcat TMU via serial connection.

			Document No.	10038879	
			Date Release	06/05/2006	
			Revision	T00	Page 1 of 4

PNNL PHASE 2 COMPUTING SYTEM AND COOLING SYSTEM POWER DRAW

APPROVALS

Title	Name	
Program Manager:	Harley McAllister	
Lead Systems Engineer:	Rich Maes	
Originator:	Levi Westra	

DOCUMENT REVISION SUMMARY

Effective Date	Revision Level	Document Change Description
06/05/06	T00	Original Release

File is located in DBWORKS

1.0 GOAL

Collect data on the amount of power required to operate the PNNL Phase 2 configured SprayCool computing rack.

Specific Goals:

- Measure the power required to operate the entire compute rack
- Measure the power required by the TMU and the SC cooling system

2.0 SCOPE

- Engineering documentation of test plan, results, and conclusions.

3.0 OVERVIEW

3.1 Measure the amount of power used by the PNNL phase 2 modular rack (16 HP rx1620s, and 2 Opus SC ATX 1U servers, and cooling system)

4.0 TEST SETUP AND INSTRUMENTATION

4.1 SprayCool system configured to PNNL Phase 2 (see ICD document 10038462)
- 4.1.1 Tomcat TMU
- 4.1.2 Rack manifold assembly (10034419) plumbed to the SC rack using PF plumbing kit (10035432)
- 4.1.3 (16) HP rx1620 servers (10033876) integrated with SC conversion kit (10035428)

4.2 Data acquisition computer capable of communicating with the Tomcat TMU via serial connection.

4.3 Fluke Clamp style current meter

Appendix E: References

[1] Roberson & Crowe. 1997. Engineering Fluid Mechanics, 6th edition, John Wiley & Sons, Inc., New York.

[2] Website Wolverine Tube "Engineering Data Book III" (http://www.wlv.com/products/databook/db3/DataBo).