Final Technical Report

Project Title: Recovery Act: Development of a Very Dense Liquid Cooled Compute Platform

Award Number: DE-EE0002896

Project Period: January 31, 2010 to December 31, 2012
Post-Novation Period: May 1, 2012 to December 31, 2012

Technical Contact:

Phil Hughes
3350 Scott Boulevard
Santa Clara, CA 95054
Ph 415 613 9264
Fax 408 327 8101
phil@clusteredsystems.com

Business Contact:

Robert Lipp
3350 Scott Boulevard
Santa Clara, CA 95054
Ph 408 234 6655
Fax 408 327 8101
bob@clusteredsystems.com

Recipient Organization:

Clustered Systems Company, Inc.
3350 Scott Boulevard #30-01
Santa Clara, CA 95054

Partners:

Emerson-Cooligy (Subrecipient and cost-sharing partner)

800 Maude Avenue
Mountain View, CA 94043

California Energy Commission (cost-sharing partner)
Grant Award Number: PIR-10-058

Intel Corporation (cost-sharing source-processors)

SLAC National Accelerator Laboratory (install/testing space)

Date: **March 31, 2013**

Acknowledgment:

This report is based upon work supported by the U. S. Department of Energy under Award No. DE-EE0002896 and Co-sponsored by the California Energy Commission Under Grant Award No. PIR-10-058

Disclaimer:

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. Any findings, opinions, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect those of the United States Government or any agency thereof.

Document Availability: Reports are available free via the U.S. Department of Energy (DOE) Information Bridge Website: http://www.osti.gov/bridge

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange (ETDE) representatives, and Informational Nuclear Information System (INIS) representatives from the following source:

Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831
Tel: (865) 576-8401
FAX: (865) 576-5728

E-mail: reports@osti.gov

Website: http://www.osti.gov/contract.html

Contents

List of Tables

List of Figures

List of Acronyms

CPU – Central Processing Unit
DCIE – Data Center Infrastructure Efficiency
DIMM – Dual In-line Memory Module
Exaflop – 10^{18} floating point calculations/sec
GbE – Gigabit Ethernet
HPC – High Performance Computing
PDU – Power Distribution Unit
PCB – Printed Circuit Board
PCIe Peripheral Control Interface
Petaflop – 10^{15} floating point calculations/sec
PUE – Power Usage Effectiveness
R134a – refrigerant number R134a
SKU – Stock Keeping Unit
Teraflop – 10^{12} floating point calculations/sec
TIM – Thermal Interface Material
VAC – Volts, Alternating Current
VDC – Volts, Direct Current
xU (or xRU) – number of Rack Units (1.75”) where x = the number of units

Executive Summary

The objective of this project was to design and develop a prototype very energy efficient high-density compute platform with 100% pumped refrigerant liquid cooling using commodity components and high volume manufacturing techniques.

We believe that we have surpassed our goal. Testing at SLAC has indicated that we achieved a DCIE of 0.93 against our original goal of 0.85. This number includes both cooling and power supply and was achieved employing some of the highest wattage processors available. This figure of merit is superior even to those of companies such as Google and Facebook who claim a DCIE of around 0.89 using free air cooling with low power processors.

In fact, our relative performance is even better when the non IT load (i.e. server fans and power conversion) is included with the overhead rather than the IT load as it is today. The US average DCIE of 0.55 becomes 0.42 and a free air facility, only 0.80. Clustered's DCIE, after factoring in power conversion to 12VDC, is 0.90.

If the IT industry had our technology, the U.S. could maintain the same compute capacity it has today while releasing only 50% of the CO_2 into the atmosphere.

In another development, a senior fellow at one CPU manufacturer voiced the opinion that our technology could bring forward the delivery of Exaflop computing by a full year.

This project included the development of a fundamentally new cooling architecture based upon pumped refrigerant liquid cooling and encompassed both infrastructure and equipment. The design comprehends mechanisms for heat transfer from the servers' internal components to ultimate dissipation of the heat to the ambient environment. It is powerful enough for HPC (High Performance Computing) applications and cost effective for general enterprise applications. It is relatively insensitive to its environment, thus does not require an expensive conditioned environment.

Energy consumption throughout the entire power and cooling chain has been reduced. Refrigerant touch cooling allows all the power hungry fans to be eliminated, reduces pumping energy costs and can use warm water for cooling, virtually eliminating the energy cost of a chiller. Bulk n+1 redundant 97% efficient rectifiers take 480VAC input and output 380VDC for distribution to on-blade power conversion units, significantly reducing power distribution losses. Virtual I/O reduces the number of high power networking components.

The elimination of air as coolant enables extreme density. This aids inter-processor communication which is very important for HPC, and reduces real estate requirements by up to 90%. Being insensitive to its environment, warehouse style space can be substituted for far more expensive data center accommodation, dropping building costs from around $250 to $80 per square foot.

We can pack 392 CPUs in 6 chassis into one standard 48U data center rack. Our target was 300-400 CPUs. In most applications however, we expect a power supply and network gear to replace one chassis leaving space for 320 CPUs.

The SLAC system has 256 CPUs in 4 chassis. Predicted performance is 47 petaflops. The system is installed in a non-controlled environment in a cage just off the loading dock of Building 950, not in a data center. A Liebert XD pumped refrigerant platform connected to the building's centralized cold water system supplies the coolant.

Besides the results of tests described herein, the compute platform is undergoing long term performance and reliability testing.

The SLAC installation is also at the core of our commercialization efforts as they will evaluate it under real loads and conditions and report on the outcomes. They will allow us to bring visitors to observe the system in use and describe their experience with it.

In order to promote a transition to our technology, Clustered Systems Company will offer customers the option of licensing the technology or purchasing the key elements of the design from us, namely the flexible cold plates and the super compliant TIM.

The data center world is a conservative one with low risk generally the highest priority. This means there is a general reluctance by data center managers to adopt the latest technologies for energy savings. It is not only secondary to their mission, but often hardly on the radar.

Our recommendations mainly reflect this situation and suggest ways to overcome this obstacle without additional investment. These include specifying true DCIEs (i.e. counting server fans and power conversion as infrastructure load) of over 0.9 for future DoE server system purchases, adding a liquid cooling option to your "DC Pro" calculator, and using your "bully pulpit". Finally, improved server proximity will open up new interconnect techniques. These should be studied and as recognized interconnect experts, Clustered Systems wishes to be included in any such RFPs.

Introduction

System Overview

Figure 1 - Rack Illustrationshows the rack as conceived. It is 48U tall, 800mm wide and 1200mm deep.

It has space for six 8U chassis, each containing 16 double height blades. In many cases we expect that 2U of rack space will be used for external switches and 6U for power supplies, replacing one chassis.

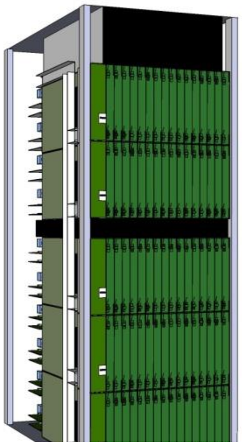


Figure - Rack Illustration

The rack is powered with 3 phase 480VAC. N+1 redundant power converters rated at 115KW rectify the incoming AC to 380VDC that is distributed to the individual chassis in the system. One circuit breaker and eight fuses are provided for each chassis in a PDU blade.

The cooling system is fully interoperable with Liebert, Inc.'s XD pumped refrigerant system.

Cooling Technology

Most electronics systems, including servers, today consist of a baseboard (motherboard) with multiple components mounted upon it. Manufacturing variances cause differences in the height and co-planarity of identical components in the same locations on different boards. Using a single cold plate to touch all of these is problematic as even the smallest gaps create an enormous resistance to the flow of heat.

The workaround has been to use cooling blocks mounted on individual hot components. Coolant is supplied from a manifold in the server rack or chassis but as the servers have to be removed for servicing, liquid connectors must be used. These can leak and cause havoc with the electronics.

Clustered Systems' has developed a contact cooled (touch) system that relies on flexible cold plates covered with a highly compliant and thermally conductive Thermal Interface Material (TIM). Heat is moved from the components to a common plane via heat risers mounted on the hot components. These can be made from simple aluminum blocks. It is thence removed by a refrigerant liquid circulating through a cold plate and TIM combination that is pressed down on the risers and conforms to the uneven surfaces of the plane.

The cold plate and TIM are permanently installed in a chassis, totally eliminating the need for connectors. Servers can be removed without breaking a liquid connection.

The server blade has a slot in the rear that accepts a cold plate. As the blade is inserted, pins in the lid guide the cold plate through the blade above the components but under the lid. After the blade is inserted and locked into position, the cover is cranked down, pressing the cold plate against the heat risers. This force causes the TIM to conform to the various heights and positions of the heat risers, creating a thermally conductive path between the heat risers and the cold plate.

The surface area of the heat risers is designed to be large to minimize thermal impedance. Low pressure is required to avoid a large overall force that might damage the components. Pressure of 0.5psi will compress the Clustered Systems TIM over 75% in thickness. This pressure is considerably lower than commercially available TIMs that typically require over 15psi for about 25% compression.

Clustered Systems had to develop this conformable TIM as no commercially available TIMs met the requirements for this application.

Electronic Sub-Systems

System development required development of six custom PCBs. Four of these boards interface the Intel JP2600JF motherboards used in the compute blade

- User interface and Ethernet connector at blade front (Headlight)
- PCIe switch blade interface and power conversion at rear of blade (Hitch)
- Bridge board, connecting Hitch to the motherboard (Crankshaft)
- PCIe and Gigabit Ethernet switch blade at the rear of the chassis
- Adapter board, providing interface between Hitch and the switch blade.
- 40 port PCIe switch support board (housed in separate chassis)

The PCIe switch blade provides virtual I/O for improved performance, lower cost and power reduction. The last board is used in the 40 port PCIe switches that provide network interface between the compute platform and the outside world.

Background

Electricity consumed in data center and telecom systems is over 2 percent of the U.S. total and growing rapidly. Historically, the energy used to provide cooling for data centers is upwards of 45% of the total facility power. Over the last several years, there have been significant efforts to improve the energy used for IT cooling by making small improvements to the cooling equipment and the control of air movement. However, moving air is energy intensive because air circuits have a low specific heat and density, and high thermal resistance at the gas to solid boundary. Hot and cold air can also mix in unpredictable ways; and air is simply incompatible with high density, high power systems that are required for HPC.

Improvements to increase the cooling power of air cooling systems has meant pushing air through ever decreasing gap sizes, but the required energy is becoming unacceptably large. The net result is that more energy is wasted in the support of a CPU than is used to run it. Unfortunately, the most common standard for measuring data center efficiency, PUE, includes the server fan energy as part of the compute load instead of the cooling overhead. This greatly distorts the final efficiency calculation, hiding a major source of energy consumption.

Server fan load varies from 5% to over 30% of the server compute power depending on a number of factors from power density to fan size to ambient temperature. The chart below illustrates the hidden cost of ignoring fan energy.

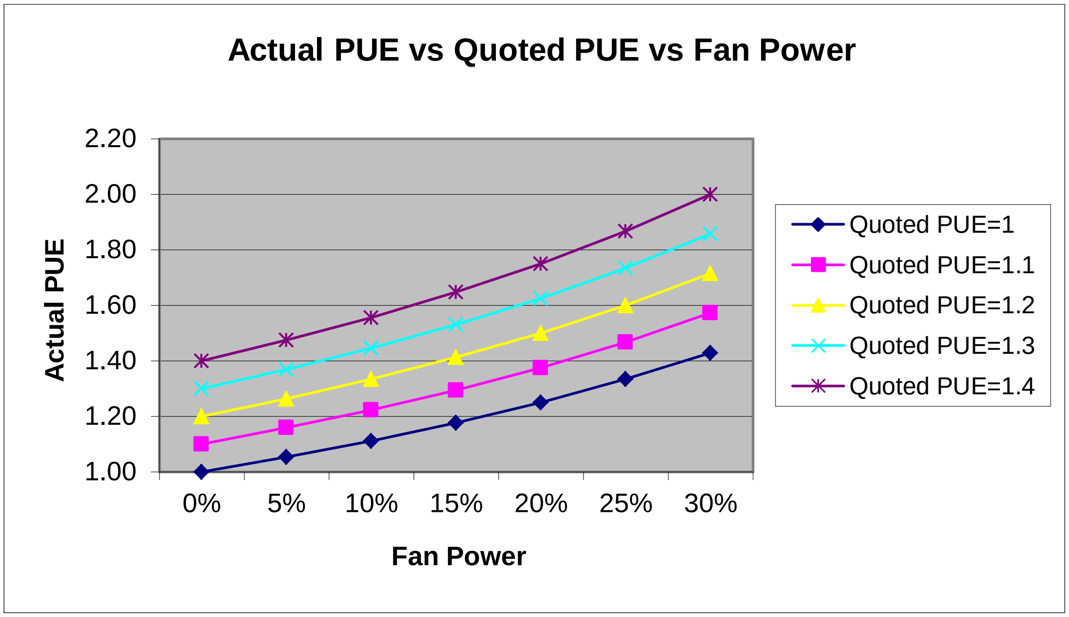
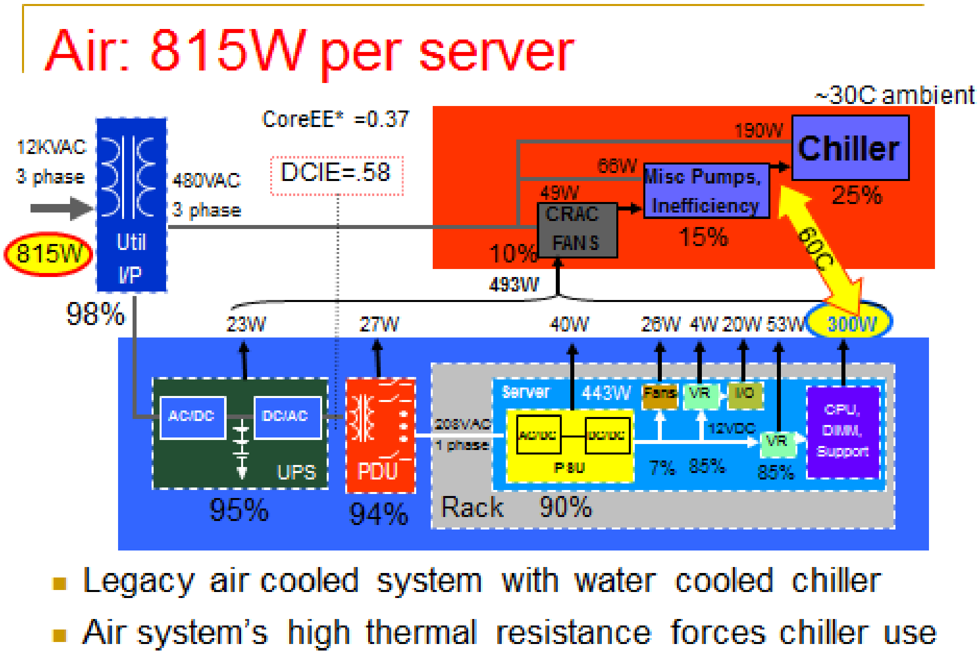


Figure - Actual PUE vs Quoted vs Fan Power

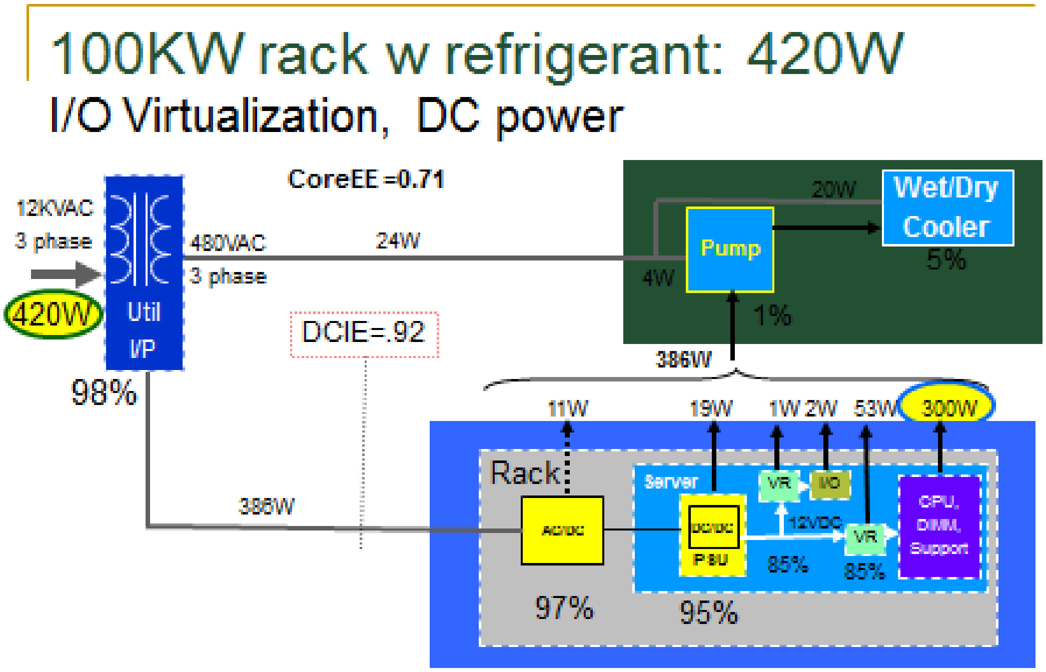
A further examination of the entire energy consumption cycle warrants further examination. Figure - Power Use in an Air Cooled Servershows where the energy is consumed in a typical high performance air cooled server. As can be seen, it requires 815W into a data center facility to run the 300W core consisting of two CPUs and 8 DIMMs.


Air: 815W per server

- Legacy air cooled system with water cooled chiller
- Air system's high thermal resistance forces chiller use

Figure - Power Use in an Air Cooled Server

Figure shows the same server as cooled by the Clustered System's compute platform developed under this program. By making improvements in cooling the entire power train and adopting I/O sharing, total incoming power is reduced by almost 50%.



100KW rack w refrigerant: 420W

I/O Virtualization, DC power

Figure - Power Use in a Refrigerant Cooled Server

Accomplishment, Results and Discussion

The project demonstrates the power of touch based refrigerant cooling as applied to blade servers. Reduced energy consumption, very high density and the ability to support high power processors have been demonstrated. The latest developments in server interconnect, I/O virtualization and power distribution and conversion were incorporated to further reduce energy consumption.

Program Modifications

Learnings and budget limitations during the program led to several changes in the program development deliverables.

We originally were planning on delivering two complete racks to SLAC outfitted with custom motherboards and two 10GbE switches. The system was to be connected either directly to the chilled water system originating from SLAC cooling towers without further cooling by a chiller, or from downstream waste cooling water from the return line of other equipment that had used the water once for cooling.

Mid program we shut down our motherboard development due to technical issues and opted to use a recently released Intel motherboard. Although a difficult switch mid program, it ended up being a wise choice as we now have a much more commercially acceptable solution with the Intel board. Originally we had planned to develop a custom server card that contained the power and networking interface as there was no suitable motherboard available when we started. However, our chosen vendor stumbled in the development of this card and it was significantly behind schedule and in danger of not being functional in a timely manner for this project. Coincidently, Intel completed development of the S2600JF motherboard and in mid project we decided we could more expeditiously complete the project using the Intel board and only develop the much simpler interface boards around it. The basic S2600JF board was designed for a 1U application rather than a blade application and a few of the components are too tall for our application. As Intel was interested in a SKU suitable for blades, they had already designed a modified design that fit our application.

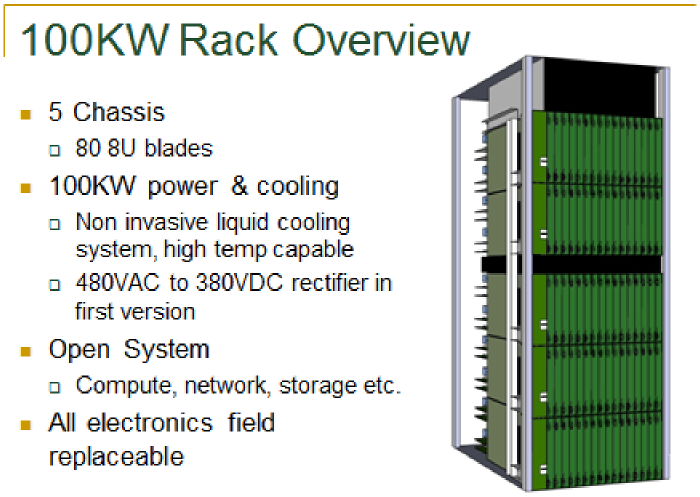
During initial architectural design, our Ethernet switch board vendor advised us they could build a fully PCIe compatible switching system for the same cost as our originally planned hybrid PCI/10GbE system. This had the potential for improved system performance, increased networking flexibility and lower energy consumption, so we proceeded down that path.

Budget constraints at SLAC led to some uncertainty as to where the system would be located. Eventually the LCLS group took ownership. The building was not reasonably amenable to our plans for connecting it to cooling tower water or return water. Physical limitations on the LCLS building in the SLAC campus led to cooling the system using building chilled water. We regretted having to give up these options as operating it without a chiller over the two year test period would have saved energy and provided a demonstration of the power and versatility of our cooling technology. However, in compensation, we now have the ability to control the temperature of the refrigerant going to the system to fully and rapidly demonstrate its cooling operating range and flexibility.

LCLS did not need the full computing power of proposed system so it was scaled down to a single rack. We went ahead with a second rack comprising a single chassis, power supply and switches that we installed in our lab. We were limited by budget overruns due to additional costs incurred primarily labor for fine tuning the design, and did not have sufficient funds to fully populate this second rack. This in-house rack aids our commercialization effort by giving us a unit we can demonstrate to prospective customers in our own facility. It also is a vehicle for further evaluation and system improvements under our control.

Complete System Overview

The final fully configured product as initially conceived is shown in Figure 6

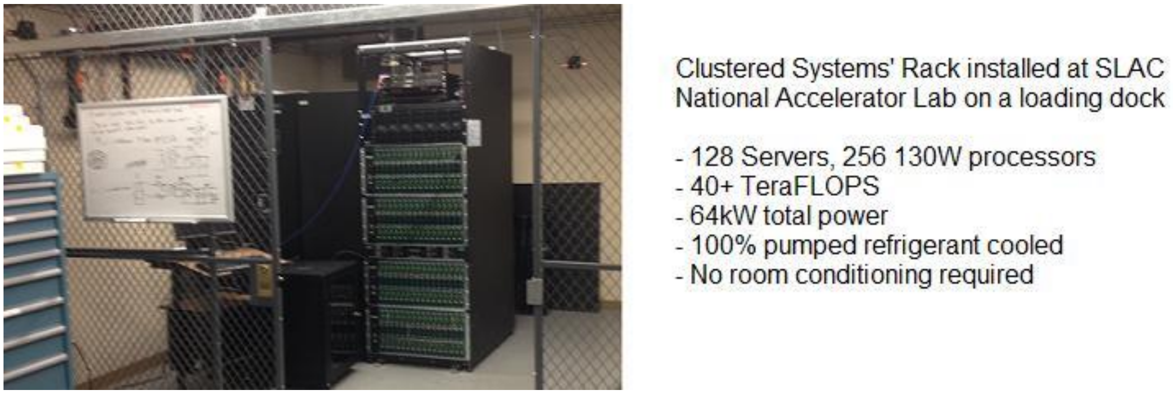


100KW Rack Overview

- 5 Chassis
  - 80 8U blades
- 100KW power & cooling
  - Non invasive liquid cooling system, high temp capable
  - 480VAC to 380VDC rectifier in first version
- Open System
  - Compute, network, storage etc.
- All electronics field replaceable

Figure - 100KW Rack Overview

The system installed at SLAC varies somewhat from this illustration in that only four chassis were installed to make room for the additional switches seen at the top of Figure 5.



Clustered Systems' Rack installed at SLAC National Accelerator Lab on a loading dock

- 128 Servers, 256 130W processors
- 40+ TeraFLOPS
- 64kW total power
- 100% pumped refrigerant cooled
- No room conditioning required

Figure - SLAC Rack Installation

System Components

Compute Blade

There are 16 compute blade enclosures in each chassis.

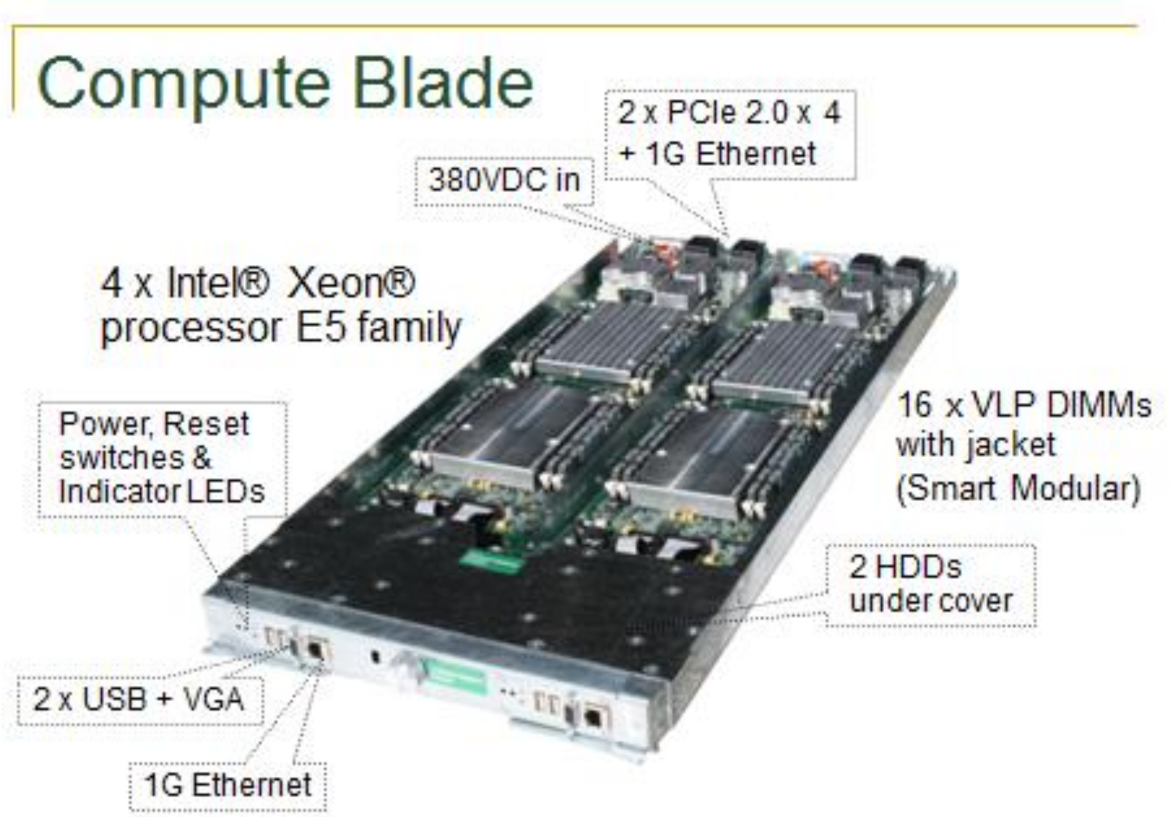


Figure - Server Blade Picture

Configuration and Motherboard

The compute blade contains two identical and independent server sections within one enclosure.

Each server section comprises an Intel S2600JF half width motherboard connected to a rear mounted custom power/networking interface card a by a small connector card, and a further connection in the front to a custom front panel mounted interface card and a hard disk. These enclosures are mounted vertically in the chassis.

The CPUs are Intel E5-2680 2.7GHz 8 core 130W XEON processors. There are 2 CPUs and 8 4 GB DIMMs on each motherboard. The maximum power dissipation for each compute blade is approximately 1KW.

Heat Risers

Note the large silver colored blocks over the CPUs and other hot components. These are simple blocks of aluminum mounted on the components using the existing conventional heat sink mounting mechanisms. They conduct heat from these components to a common plane for heat removal by a cold plate. The DIMMs are jacketed with conventional DIMM heat removal jackets slightly modified such that the top has a flat planar area to thermally interface to the cold plate. Note that the DIMMs, as the tallest components, set the height of the plane that all the risers are designed to.

Safety and Hot Swap

This hot swap enclosure has mechanisms for safely engaging and disengaging the blade from the system, mechanically insuring the enclosure cannot be inserted nor removed while powered on. Additionally, an electrical interlock prevents the card from being powered up until the cooling system is engaged; and assures power is off prior to removal.

Network Interface

The system contains no conventional midplane or backplane. To avoid obsolescence, all the electrical components are removable. The power/networking interface plugs directly into chassis-mounted removable rear power distribution strips and removable active plane networking cards.

The power/networking card plugs converts 380VDC to 12VDC as the primary local power source for the blade. Conversion efficiency is about 96%.

This same card further extends the motherboard PCIe signals to two active plane cards via dual 50 pair high speed differential signal connectors. Management 1GbE signals are also carried across the same connectors.

The power/networking card is connected to the motherboard by a small custom bridge card carrying PCIe and control signals and dual 12V power connectors.

Active Plane

Figure shows the PCIe switch card developed for this application mounted in its blade carrier. The 16 connectors on the bottom provide connectivity to the front CPU blades. The 8 PCIe connectors on the top are for cabling to other switch blades and switches. Two 1GbE sockets provide external management connectivity. High performance inter-processor communication can occur over these active planes without protocol translation, saving energy and improving performance.

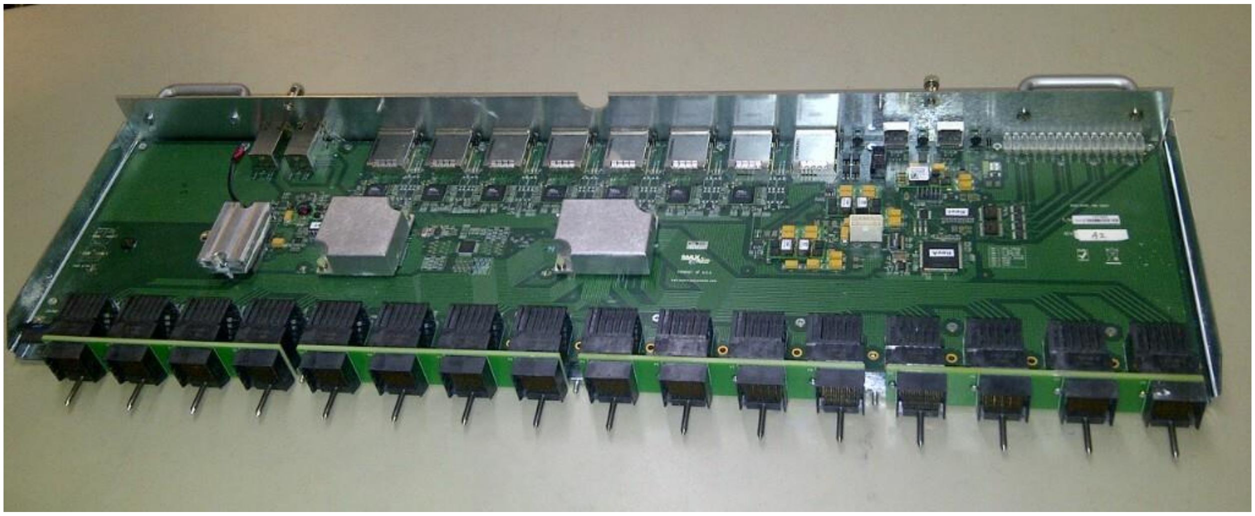


Figure - Active Plane (PCIe Switch) Picture

There are four slots for active planes in the rear of each chassis. A minimum of two are required for full network connectivity from the rear. The cards are inserted horizontally into the chassis, orthogonal to the compute blades. This permits full interconnect between all the motherboards in that horizontal row. An external switch provides connectivity between rows and between chassis.

Power Distribution Blade

The PDU blade was designed and supplied by Panduit, Inc. Although not a formal member of the development team, they agreed to develop and build this blade at no cost to us.

The PDU blade controls power to the chassis. It has a 50A breaker supplying power to a fuse panel contained within. Eight fused circuits distribute power via rear connectors to two power distribution strips and thence to the compute blades and active planes.

It is shown on the left of the picture of the chassis below.

Chassis

Figure 9 shows the front of a single chassis with one empty slot and one CPU blade partially removed.

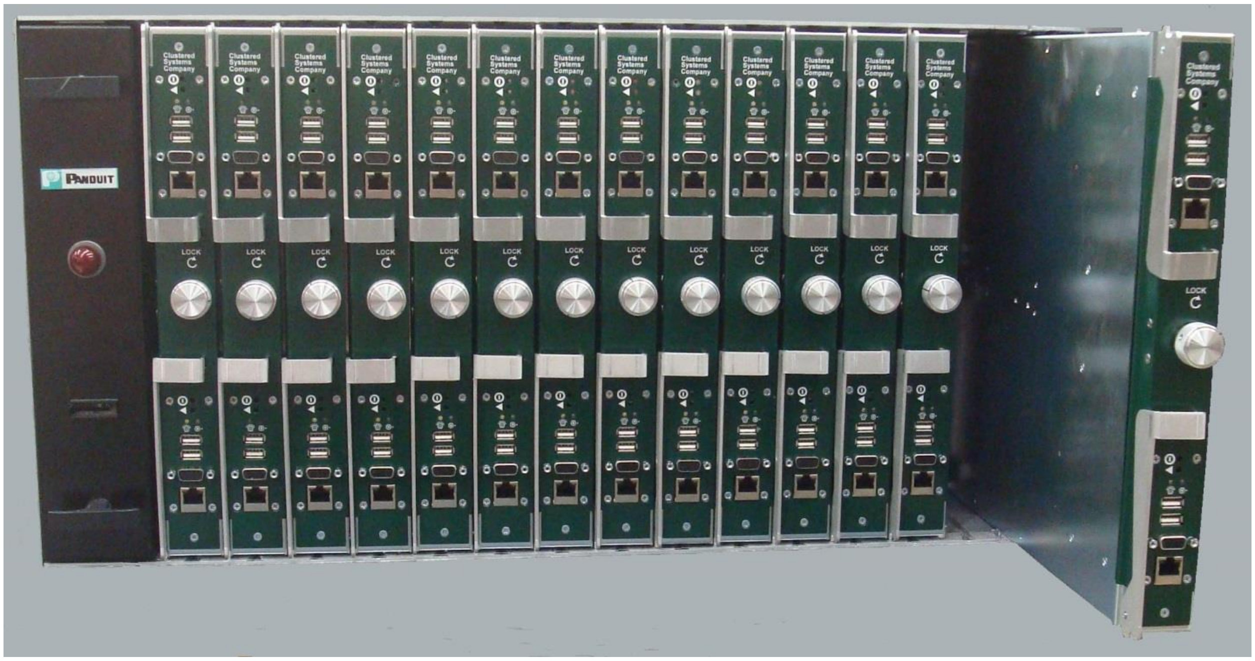


Figure - Chassis Front Picture

Figure 10 shows the chassis rear illustrating how an active plane card is inserted. The plumbing manifolds distribute the refrigerant to the blades.

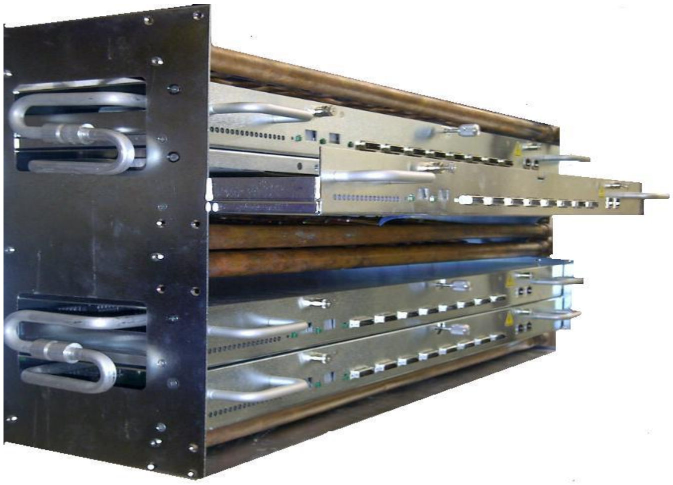


Figure - Chassis Rear Picture

Figure 11 is a picture of a set of rear cold plates. The fluid enters the bottom plate where it flows to the end and returns through the top plate. The return is also shown in the picture above where it protrudes from the left side.

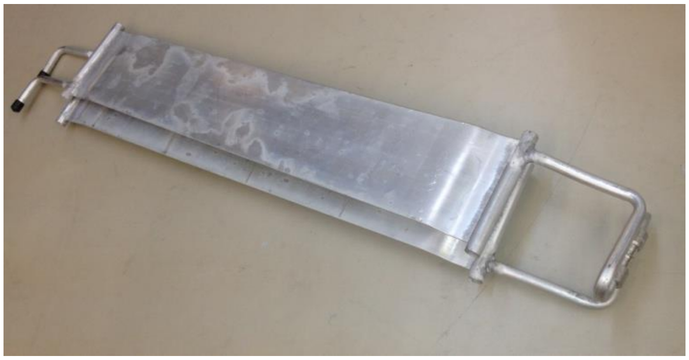
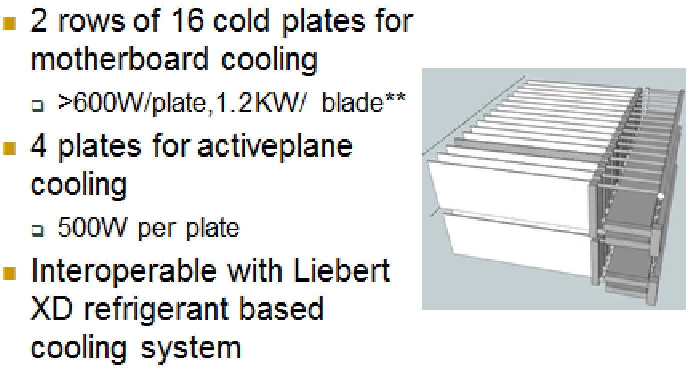


Figure - Active Plane Cold Plates Picture

Figure 12 illustrates the complete chassis plumbing. The two rows of 16 server blade cold plates are on the left. The two pairs of rear cold plates is on the right.



- 2 rows of 16 cold plates for motherboard cooling
  - >600W/plate,1.2KW/ blade**
- 4 plates for activeplane cooling
  - 500W per plate
- Interoperable with Liebert XD refrigerant based cooling system

Figure - Chassis Cold Plate Plumbing Overview

PCIe Switch

A 40 port PCIe switch was custom developed for us by OneStop Systems. Two of these are installed in each rack to provide connectivity between blades and as a virtual I/O uplink to the outside world.

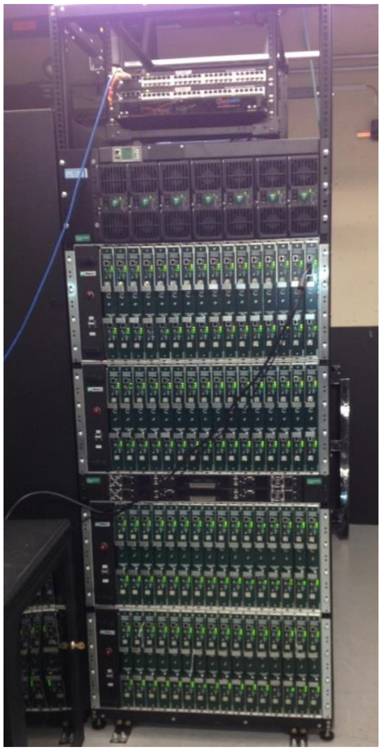
These switches also contain the processors for configuration of the PCIe network. The first level of switching occurs on the active planes that connect all 16 blades on a row and provide 8 uplinks to the switches. All these signals are connected in a Fat Tree configuration by the switches. The switches also have two slots for insertion of a standard Ethernet or Infiniband network card that converts from PCIe to the relevant protocol providing a virtual I/O for uplinks to an external network. This saves energy and is potentially a cheaper and more flexible solution than the current state of the art systems.

The switches currently provide the dynamic virtual I/O function for only 32 blades. Limitations in the PCIe protocol are currently limiting full implementation of this feature. Work is continuing at Clustered Systems and OneStop Systems to fully implement full functionality as we believe it will be an attractive feature for commercialization.

Power Supply

The DC power supply was developed by Emerson Network Power in an informal arrangement. Emerson modified the form factor and size of a product in development to meet our needs. This 480VAC to 380VDC unit is about 97% efficient..

Figure 13 shows the SLAC installation. The rack was built by Liebert, Inc. as a partner in this project.



The components of this rack from the top down are:

- GbE switch with 2ea 10GbE optical uplinks
- Emerson 115KW 380VDC power supply
- 2ea fully configured chasses
- 2ea OneStop Systems 40 port PCIe switches
- 2ea fully configured chasses

Figure - SLAC Rack Configuration

Liebert XDS System

The XDS is installed adjacent to the rack (Figure 14). As a partner in this project, Liebert modified one of their standard XDS systems for compatibility with our system. In particular, they added the sidecar to the right of their standard system and made the system fully redundant.

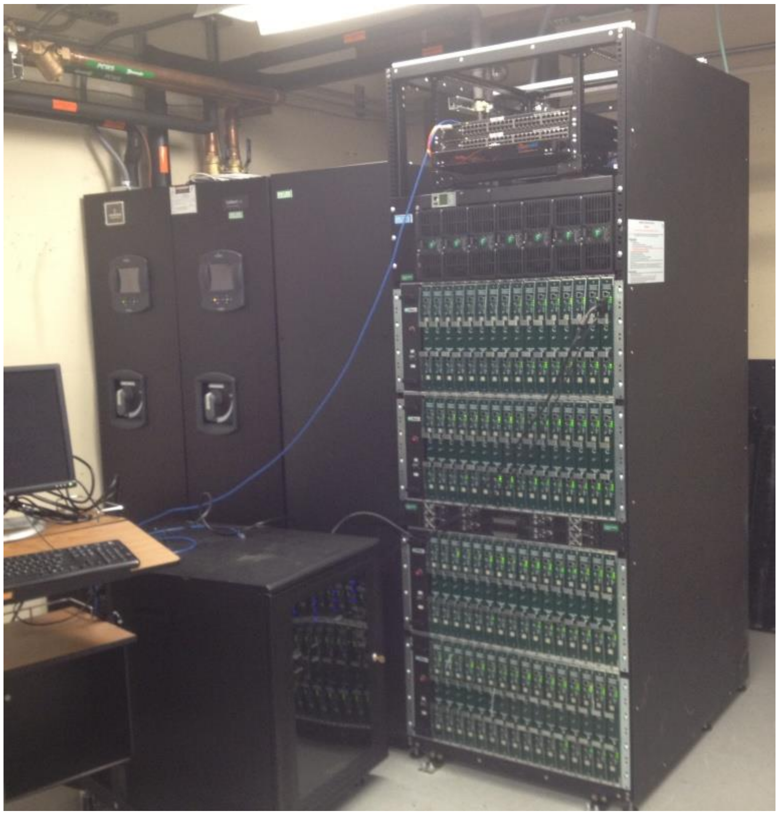


Figure - SLAC Installation Showing the Rack and XDS System

During normal operation, nearly 50% of the refrigerant volume of the system comprises refrigerant gas boiling off the cold plates. However, when the system is off, refrigerant will settle to the lowest and/or coldest spots in the system. A standard Liebert installation requires a minimum of 20KW loading to maintain the proper distribution of gas and liquids in the system. However, in our case the system must be able to be turned off completely. Extra refrigerant storage was added to provide proper functionality over all loads. The sidecar is essentially no more than a large refrigerant storage tank to provide this function.

Technology Development Highlights

Several critical technologies needed to be developed for this project:

- Vertically mounted micro-channel cold plate
- Conformable thermal interface
- Blade installation mechanisms

Micro-Channel Cold Plate

We developed a micro-channel cold plate on an earlier product that is currently in production and commercially available. This earlier product incorporated these cold plates for cooling 1U servers. They were mounted horizontally on the tops of these servers with large manifolds on either side to distribute and collect the refrigerant. They were made from extruded aluminum using mass production facilities and technologies developed for the automotive industry. The aluminum is flexible and will conform to some extent to an uneven surface.

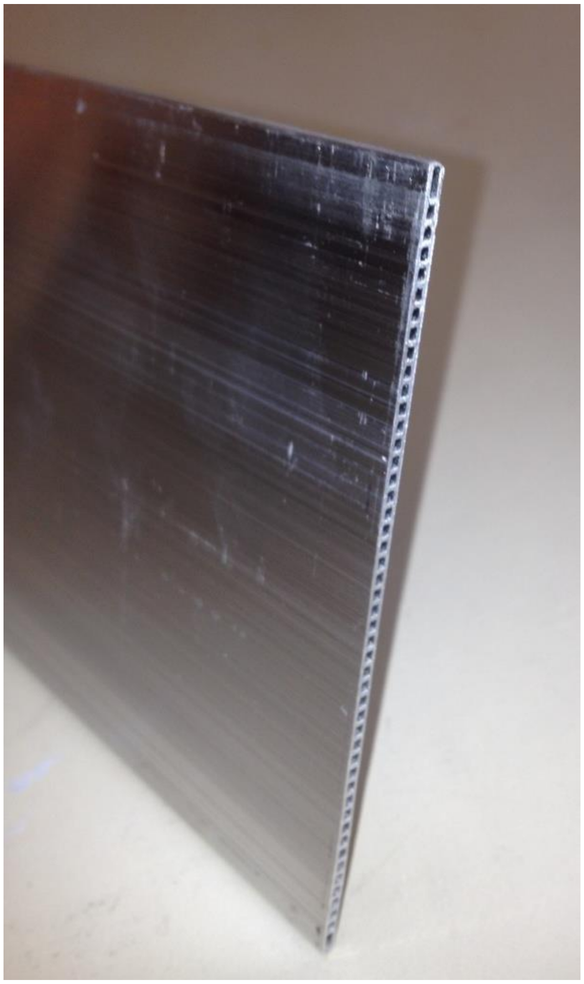


Figure - Micro-channel Close-up View

Figure 15 shows an end-on view of such an extrusion approximately to scale. The extrusion is 100mm wide by 2mm thick and is cut to length as required.

This application required 160mm wide cold plates mounted vertically in the chassis. They could be connected only to a rear manifold to distribute and collect the refrigerant. The front had to remain thin to slide into the narrow slot in the server.

160mm micro-channel extrusions are generally beyond the state of the art for commercial applications. We therefore opted to make the cold plate from two side-by-side 75mm wide pieces with a small space between them. The first plate sources the refrigerant from a manifold and the second side returns the refrigerant to an isolated portion of the same manifold. Rather than retool a new extrusion, for this first limited deployment, we decided to machine our current material down from 100mm to 75mm

The primary challenge was how to connect the front end of the plates such that the assembly remained thin enough to slide into the narrow slot in the rear of the server and maintain continuity of the channels without intermixing between channels. If a simple manifold was used, bubbles from refrigerant heating on the sourcing path would all rise to the top of the manifold, leading to dry out and ineffective cooling on the upper half of the return manifold.

We solved this problem by developing a technology to butt soldering extrusions end-to-end. The cold plate sections are mitered to terminate in 45 degree angles. A third piece is doubled mitered into a wedge shape and placed between them. These three pieces are plated and tinned and soldered together in a temperature controlled press. The other ends are soldered into the refrigerant distribution manifold. Special jigs and techniques were developed to assure an accurate quality solder joint and limit solder flow to unwanted areas such as the micro-channel tubes. As a final quality check, all plates are subject to over-pressure, flow and vacuum tests prior to the next level of assembly.

Figure is a picture of the top of the cold plate prior to TIM installation. The left end (rear) is finished with a manifold that, in turn, connects to chassis supply and return manifolds.

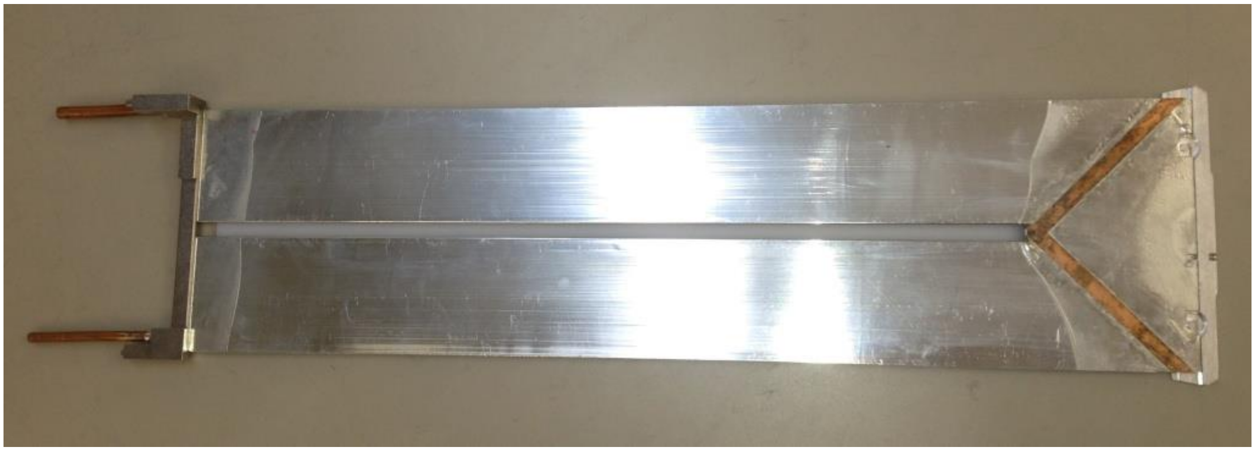


Figure - Server Cold Plate Assembly Bottom

Figure is a picture of the top of the cold plate prior to TIM installation. It is covered with a low friction HDMW plastic. Note the right side (front) is finished with a guide rail that aligns the plate when it is slipped into the server. It also pre-aligns the plate in the proper position in the chassis to receive the server during installation.

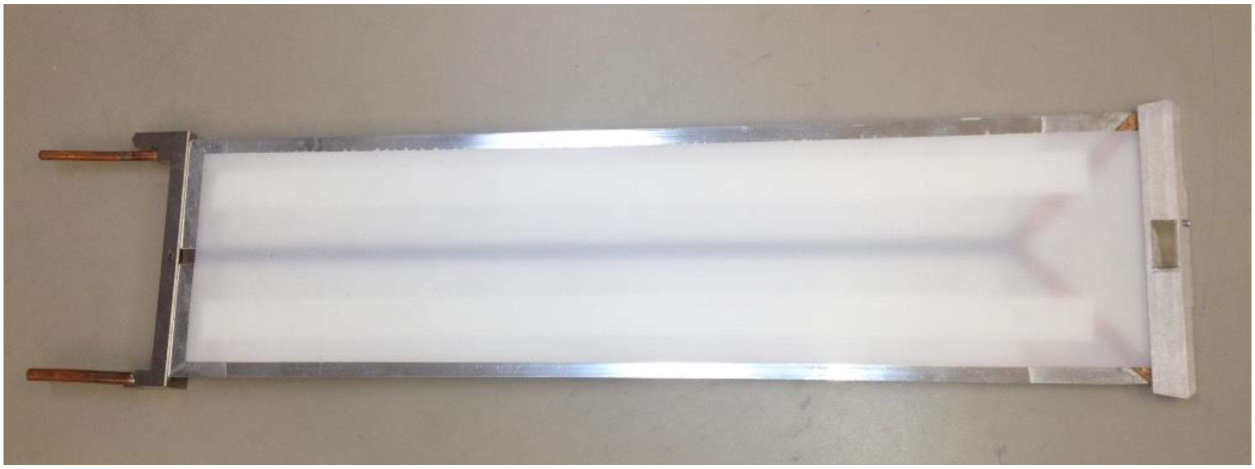


Figure - Server Cold Plate Assembly Top

The 6” wide by 21” long cold plates are mounted vertically two-high in 16 slots in each chassis. The rear of the cold plates are permanently brazed to a common manifold that sources the refrigerant liquid through the bottom pipes and takes away the liquid/gas combination emerging from the top pipe.

Cold Plate Thermal Interface (TIM)

A conformal thermal interface is next installed on the cold plate. It is designed to be readily removable in case of wear or damage.

Figure 18 shows the TIM installed on a cold plate over an aluminum base. Note the brown

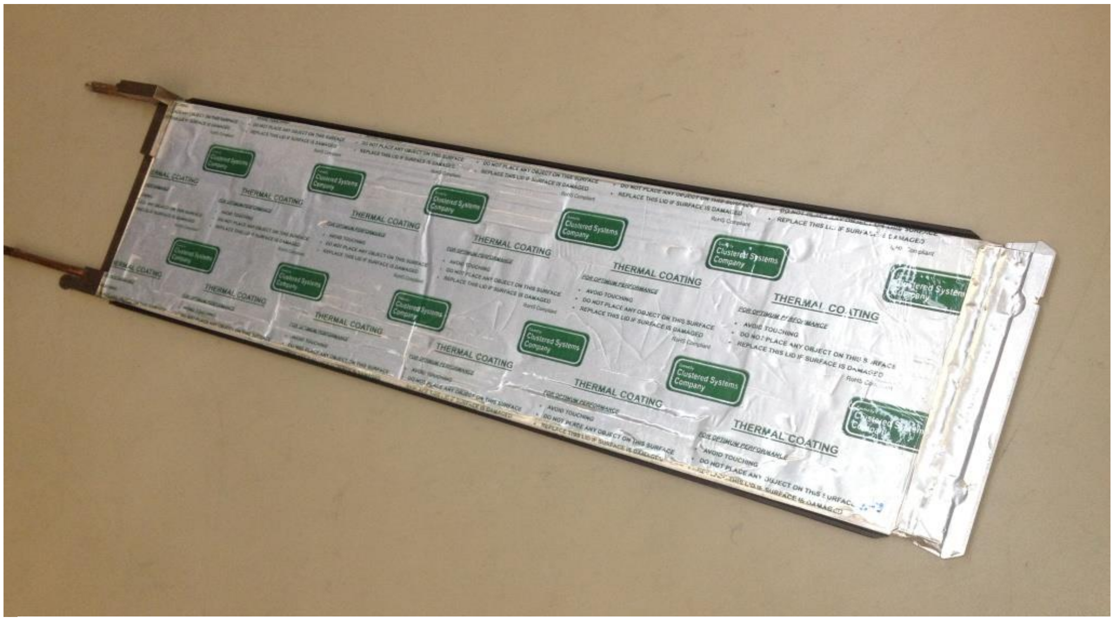


Figure - Server Cold Plate with TIM

plastic edge guards that are attached to the TIM that secure it to the side of the cold plate. Metal aluminum tape over the TIM and guide rail finishes off the unit preventing undue movement and covering all edges that might catch the serve during the installation or removal procedures. The conformability of the TIM is illustrated in the picture by the apparent creases and dents in the surface. These are purely cosmetic in nature and come about in the manufacture and use of these units. They have no measurable effect on operation.

The TIM comprises a proprietary thermal grease inside a containing sleeve. The sleeve is compound material comprising a thin strong thermally conductive plastic coating over a metal layer.

The thermal grease mixture was co-developed with Dow-Corning Corporation with further amendments and processing at Clustered Systems to achieve the appropriate properties. In particular, the material must flow and conform to the irregular surface of the components and heat risers it comes in contact. A large area of thermal contact is required for high thermal conductivity from the components to the TIM and thus the cold plate. It needs a thermal conductance approaching 2 w/m-k. The viscosity must be low enough to quickly conform to pressures as low as 0.1PSI but must be sufficiently thixotropic such that it will never sag or run when placed vertically or possibly damaged by abuse. Lastly, as the material is used in relatively large quantities it must be very inexpensive compared to most commercially available compounds.

Blade Installation Mechanisms

The blade enclosure has a number of features for easy and safe operation. The blade is

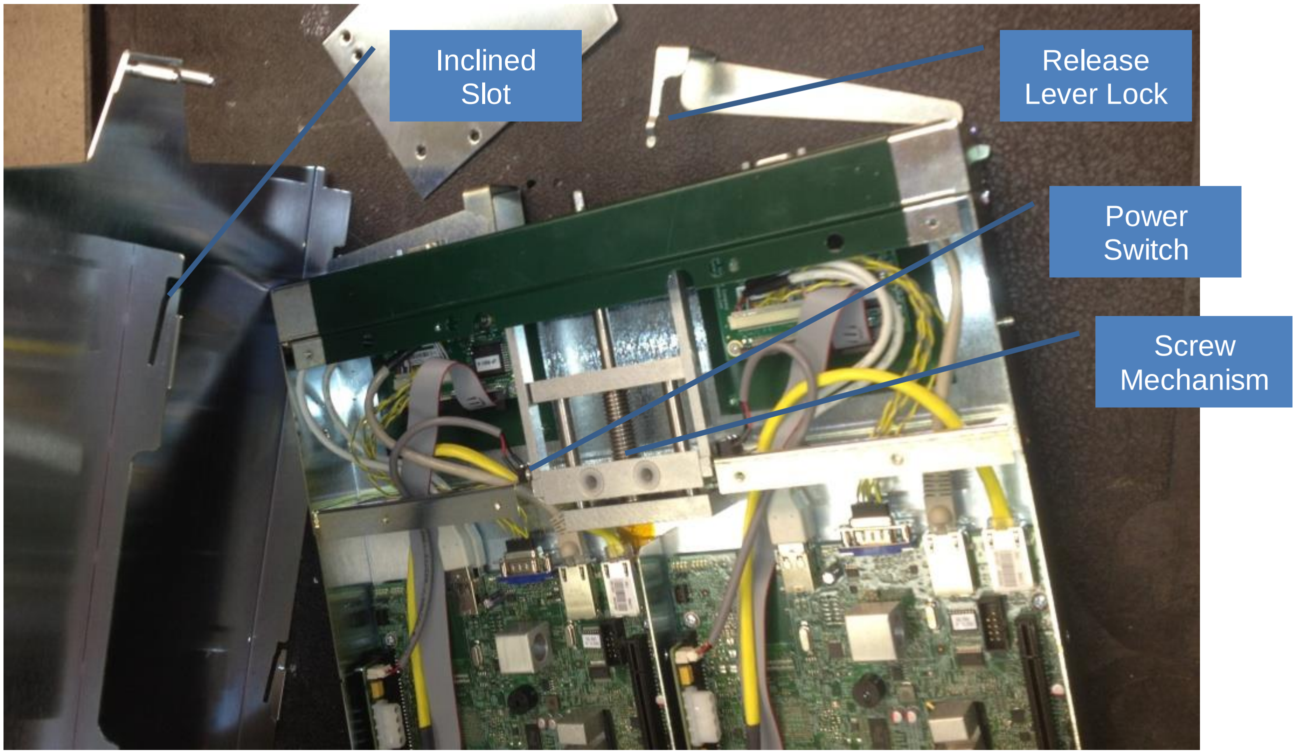


Figure – Close-up of Server Blade Installation Mechanism designed to be hot swappable. Circuitry on the rear power interface board needs to be enabled and disabled to control the 380VDC. It is imperative that the blade cannot be inserted or removed when the power is on or when the engagement mechanisms are not set properly. Additionally, the cold plate must be in contact and properly seated on the heat risers to conduct heat when the server is turned on.

The blade lid is the key operating component for operating the blade cooling system. The cold plate is held up against the lid at all times. When the blade is inserted in the chassis, the lid is held high above the heat risers. After insertion, the lid is lowered, pressing the cold plate against the heat risers. This is accomplished by a screw mechanism in the front of the blade which draws the lid forward and back. Inclined slots in the lid engage pins in the blade base that control the height depending on the relative position of the lid. As the lid moves forward, it descends; it ascends when the direction is reversed.

Multiple interlocks, both mechanically and electrically are incorporated.

- Pins on the rear of the lid prevent the blade from being inserted or removed unless the lid is in the “high” position.
- Bars on the screw mechanism lock the release levers in place and prevent blade extraction unless the lid is in the “high” position.
- Two redundant switches turn off the electrical power to the blade when the screw mechanism is not in the locked position, preventing both removal when electrically “hot” or powering the server when the cold plates are not engaged.

SLAC Installation

Finally installed in late December, the system was charged with refrigerant in mid-January. Due to a quality problem at one of our suppliers, population of the rack was not completed until early March and test software loaded mid March. This revealed a secondary problem with the network adapter cards. Twenty seven of them would not communicate via the backplane. We resolved this by adding cables to connect the front Ethernet connector to an external switch. This allowed all the servers to communicate with the management system and upload the test software.

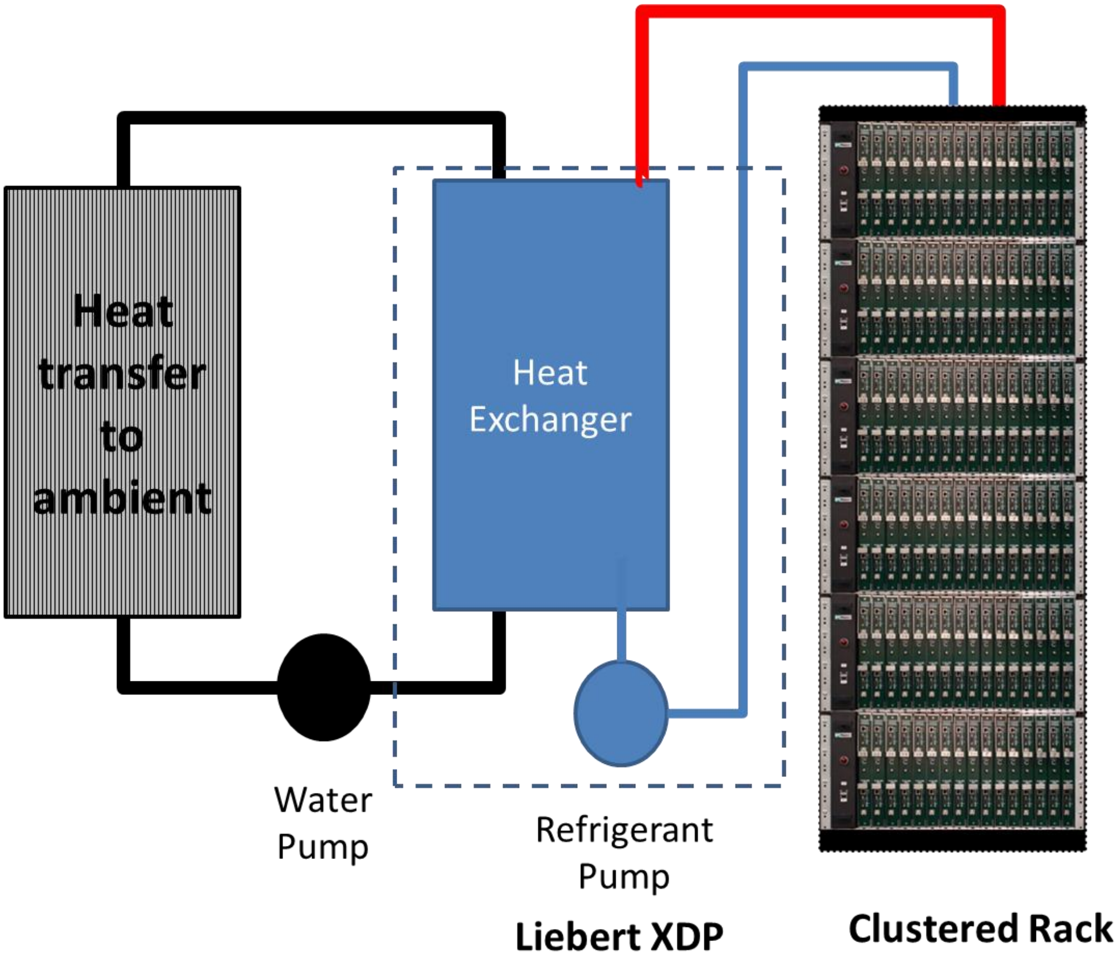


Figure - Generic Rack Cooling Circuit

Figure illustrates how Clustered's rack connects to the rest of the cooling system. In the case of the SLAC installation, the "Heat Transfer to Ambient" is another heat exchanger that interfaces with the building chilled water supply. The controls available to us were a) water pump speed and b) a flow valve on the building chilled water circuit. Both had to be manually operated. We ran tests over two days. Compute load was supplied by using "Prime95" in torture test mode.

Data was collected using power, temperature and flow loggers loaned by PG&E as well as the internal temperatures of the various components as reported by the servers' BMCs (Baseboard Management Controller).

SLAC Installation Measured Results

On the first day, we did an operating range check by allowing the refrigerant temperature to climb to about 60^{o}C and observing at what temperature the CPU would begin to reduce its clock speed.

Exploration of maximum coolant temperature performance

Table Temperature Cycle

Refrigerant Return Temp	Motherboard Inlet Temp	BMC Temp	P1 VR Temp	P1 Thermal Margin	P2 Thermal Margin	LAN NIC Temp	DIMM 1 Thermal Margin
17.0	36	31	44	-53	-53	34	-41
17.3	37	32	46	-51	-51	34	-40
18.5	38	32	46	-50	-50	35	-40
20.0	39	33	49	-44	-44	36	-39
22.5	41	35	52	-41	-41	39	-36
26.2	44	38	55	-38	-38	42	-34
29.0	47	41	56	-38	-38	45	-31
31.0	50	43	57	-39	-38	47	-29
35.8	53	46	60	-34	-34	51	-26
41.3	58	50	66	-26	-26	56	-22
53.4	68	58	76	-12	-12	66	-14
57.5	74	66	82	-12	-12	72	-6
55.2	75	67	82	-15	-15	71	-4
47.0	71	65	80	-19	-19	65	-7
48.9	70	64	78	-21	-21	67	-8
44.5	71	64	79	-20	-20	63	-8
30.7	63	56	72	-32	-32	53	-16
24.6	54	49	64	-39	-39	47	-23
21.5	48	44	58	-44	-44	43	-28
19.8	45	40	54	-47	-46	41	-32
18.9	43	38	52	-48	-48	39	-35
18.3	41	36	50	-49	-49	38	-36
17.9	40	35	49	-50	-50	37	-37

This table shows a typical server (#5). Measurements were made over a two hour time period at 5 minute intervals. The highest coolant temperature is highlighted in yellow. All temperatures are in ^{o}C. All but the refrigerant temperature were logged using Intel management software.

Table column headings are as follows:

Refrigerant Return Temp – Temperature of the cooling refrigerant as it exits the rack
Motherboard Inlet Temp – Temperature sensor on the motherboard
BMC Temp – Baseboard Management Controller chip temperature
P1 VR Temp – Processor 1 Voltage Regulator temperature
P1 Thermal Margin – Processor 1 thermal margin
P2 Thermal Margin – Processor 2 thermal margin
LAN NIC Temp – NIC chip temperature
DIMM 1 Thermal Margin – DIMM #1 thermal margin

The temperature of the CPU is reported as a margin as Intel does not release the maximum allowable junction temperature. The more negative, the better. When a CPU heats up enough to force the margin to zero, the BIOS reduces its power consumption (throttles performance) to avoid overheating. As can be seen, the margin disappears at approximately 60°C. The key here is that full performance is maintained to 60°C.

In practice it is not recommended to run a server continuously at such elevated temperatures. While the semiconductor components may survive, the ancillary components, such as electrolytic capacitors, may dry out and fail.

Design Repeatability

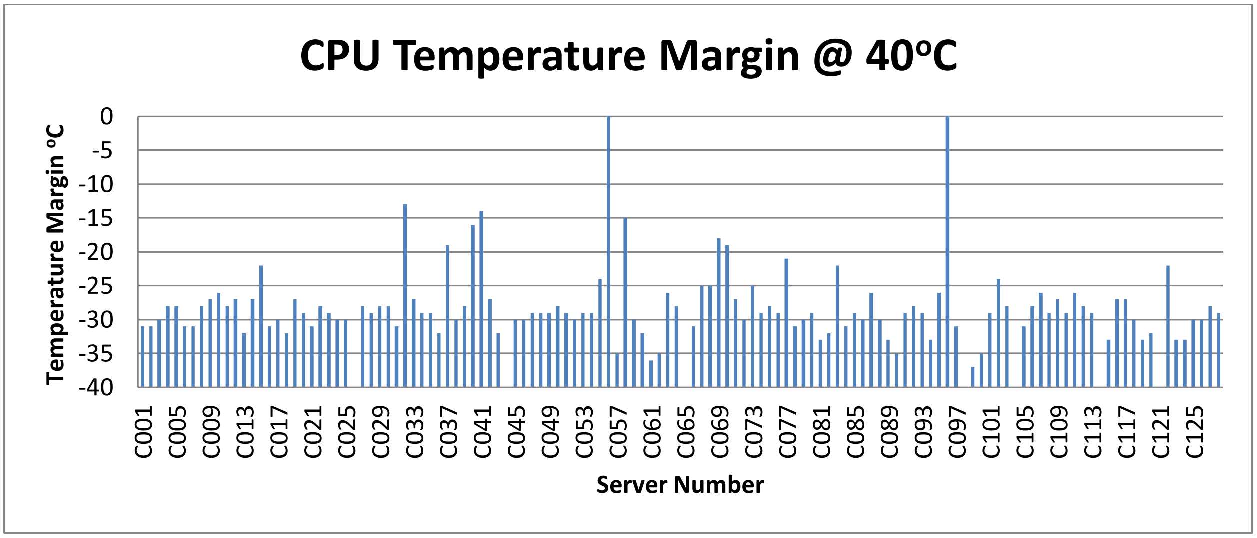


Figure - CPU Temperature Margin

Here we took a simultaneous sample of all 122 active servers. Of the 128 total, two had power connection problems that could not be repaired in time and four were not brought up due to an oversight by the systems administrator. As can be seen about 10% of the servers had significant excursion from the -30°C margin.

The servers with zero margin have blocked inlet metering devices to the cold plates. This was due to debris from the SLAC installation process blocking refrigerant flow. This will be eliminated in future by installing a filter in the inlet manifold to each chassis.

Investigating the cause of the less significant excursions needs further work. Possible causes include blade and TIM manufacturing variances, CPU power differences and possibly debris restricting refrigerant flow through the metering devices.

As all the cold plates, TIMs and servers were prototypes assembled by hand, such variations in performance would be expected.

Power Conversion Loss

We attached a flow meter and temperature sensors to the chilled water supply and return, and a power meter to the 480VAC to 380VDC converter (Courtesy PG&E). We then varied the refrigerant temperature slowly over four hours in order to detect any possible temperature related anomalies. There were none.

Table Power Supply Efficiency

Average				
Delta T	GPM	BTU	Calculated kW	Actual kW
37.84	7.02	131,152	38.44	39.22
Powr Supply Efficiency				98%

As can be seen, the power supply was even more efficient than specified in Emerson's literature (97%) even at 40% of full load (note however that the stated accuracy of the power meter was +/-1% and the flow meter about the same).

Overall Power and Cooling Efficiency

As it was too time consuming to set up measurements for the refrigerant and water pumps, we relied on earlier measurements reported by Lawrence Berkeley Laboratories. We also used their curve for chiller and economizer power requirements.

In this computation we assume that the system will be run with 40^{o}C refrigerant temperature. This implies that the cooling water temperature must be about 85^{o}F as shown in Table .

Table Cooling Water Temperature Computation

Refrigerant feed	40	^{o}C	Recommendation
Refrigerant feed	104	^{o}F	
XDP HX drop	10	^{o}F	Data Sheet
Water to ambient	9	^{o}F	
Ambient	85	^{o}F	Target

The total power required by the Liebert XDP is 0.81kW. This is capable of cooling 160kW, so we pro rate this to 0.33kW.

The cooling water pump is assumed to be compliant with ASHRAE's 90.1-2008 Chapter 11 rules which specify 0.022kW per GPM.

Finally, the heat rejection to ambient is calculated using a model chiller as defined in Appendix B of LBNL document "Demonstration of Rack-Mounted Computer Equipment Cooling Solutions", by H. Cole 2010.

The 3 term polynomial equation used derives the kW/ton has the coefficients shown in Table

Table Coefficients for deriving kilowatts required per ton of cooling

	Coefficient
T^3	0.00000628
T^2	-0.00105494
T^1	0.04389858
T^0	0.10162531

The efficiency is thus:

Table Efficiency Calculation

	kW	
System power	38.44	Measured
Power conversion	0.79	Measured
Refrigerant Pump	0.33	Reported (LBNL)
Water Pump	0.64	Computed
Cooling Plant	0.77	Computed (LBNL)
Total power	40.97	
Efficiency	**94%**	
PUE	**1.07**	

Power Consumption Variation with Temperature

There was a fairly good correlation between the refrigerant return temperature and the power consumed by the unit even though the power difference was less than 10%. The power consumption leading the refrigerant temperature can be explained by its low flow rate.

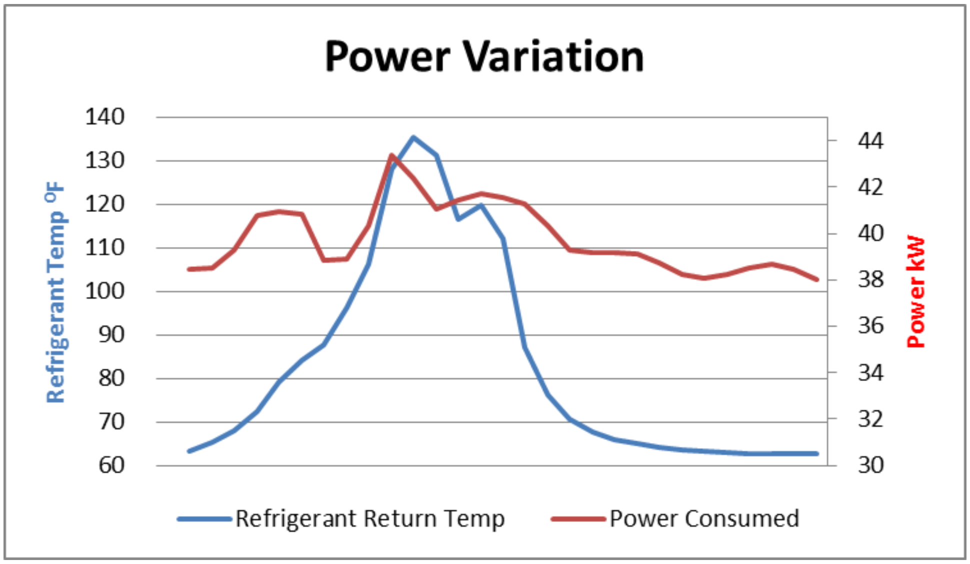


Figure - Power Variation with Temperature Over a Two Hour Test Cycle

Publications:

These are being prepared in anticipation of announcing the placing in service of the rack at SLAC. They will be accompanied by additional collateral

- “Survey of liquid cooling systems” paper written, seeking publisher
- “Data Center in a Box” Paper written, seeking publisher
- “High Density Rack Cooling – choices” seeking publisher

Patents:

USPTO Application No. 13448353 “COLD PLATE WITH REDUCED BUBBLE EFFECTS” was filed for work done under this contract.

Benefits Assessment

The results showed a true PUE of 1.07 for our system.

According to the Uptime Institute's 2012 Data Center Survey, the global average of respondents' largest data centers is between 1.8 and 1.89.

As an example of state of the art, Google quotes an average PUE for its data centers of 1.12. http://www.google.com/about/datacenters/efficiency/internal/index.html However, even though it claims to include all sources of overhead in its calculations, it does not appear to include AC-DC power conversions in the servers and the energy consumed by moving air through the servers. Energy Star labeling requires a minimum internal power supply efficiency of 85% at 50% of rated load. This is from the AC input to the 12VDC output. Fans consume an additional 10-20% of the total server load.

If we generously assume a 90% power supply efficiency and 10% fan load as a state of the art example, only about 79% of the input power is actually delivered to the compute elements (motherboard) of the server. The true PUE is then not 1.12, but:

$PUE_{Google} = (1.12 / 0.9) / (1.0 – 0.1) = 1.38$

A similar argument could be used on the global average with less efficient subsystems to produce a true:

$PUE_{global\ average} = (1.85 /.85) / (1.0 -0.2) = 2.72$

The measured PUE of 1.07 for our system includes the first stage of AC-DC conversion from 480VAC to 380VDC. We have a second stage on the board of 380VDC-12VDC. To make a comparable calculation we need to add that efficiency into the calculation:

$PUE_{Clustered} = 1.07 / 0.96 = 1.11$

According to a report by Jonathan Koomey, data centers in the U.S consumed about 2% of the total electricity in the country.

According to the EPA, 3,856 billion kWh were consumed in the U.S. in 2011. If we assume similar numbers for the two years, approximately 77 billion Kwh were consumed by U.S. data centers in 2011. Each billion Kwh generates an average of 1.22 million pounds of CO_2 per EPA data (eGRID2012). If 50% of data centers adopted technology such as herein described, about

50 million pounds less of CO_2 and other greenhouse gases would not be released into the atmosphere every year – just in the U.S.

Table U.S. Data Center Energy Usage

Total US Energy Generation 2011	3,856	MMWh
DC Energy Consumed (Koomey 2010) %	2%	
Data Center Portion	77	MMWh
CO_2 per MMWH (EGRID 2012)	1.22	lbs/MWh
Data Center Portion	94	Mlbs

Other secondary but nonetheless important benefits of our system include:

- The design is very high density. Greater density offers:
  - shorter inter-processor wire lengths which require less driver power while speeding communication, improving compute throughput, a very important factor in HPC.
  - a smaller footprint; data center buildings can be made much smaller resulting in less material input (especially cement, the manufacture of which produces large quantities of CO_2) and cost. In addition, they are also much simpler.
- Complex air handling systems replaced by simple plumbing.
- Considerably reduced cooling energy as liquid requires much less energy to circulate and is a much more efficient heat remover; these energy savings combined with the elimination of complex, ongoing air management processes result in greatly decreased operational expenses.
- Energy savings through equipment upgrade could be repurposed to run more computers, postponing the need to build new data centers.

In summary, our system uses only 34% of the energy consumed by the global data center average and 73% of the energy of the best of breed.

Commercialization

The technology is positioned as an open system cooling architecture that will accept any mechanically compliant designs. An OEM might decide to make a totally proprietary design for both front and rear blades or follow an as yet to be defined standard and make units compatible with others' products.

We have only recently focused heavily on commercialization. Even so, we have attracted a large data center Infrastructure OEM to evaluate out technology. So far, they have bought a rack and two chassis plus a couple of blades and are proceeding with evaluation.

Intel has also introduced us to a few OEMs. Feedback from them is our technology will be useful for generations emerging in 2014.

A Europe based company has designed an adapter for our blades so we can populate a single blade with 8 dense form factor (DFF) GPU boards.

These however are just part of the picture. Our strategy, as a small company with limited resources is based upon three tactics:

a) Contact end customers to drive demand

b) Increase our visibility through our web presence and guerrilla marketing

c) Approach second tier OEMs and VARs to offer unique advantage

d) Work with Intel, AMD, nVidia, etc. to gain access to larger OEMs

e) Attempt to influence legislation favoring increased energy savings

In the short term, we may also make copies of the SLAC installation for early adopters.

We offer our customers the option of either a royalty bearing license to our technology or to purchase the core technology items from us. These include the cold plates and TIMs.

The SLAC installation is at the core of our commercialization efforts. They are using it not only for real ongoing number crunching for their large R&D compute needs, but are providing independent testing and validation of the system and will be writing a series of reports on the system. They are allowing us to invite prospective customers to view the system. This will be an invaluable sales tool.

We have installed a chassis with one rack in our showroom/laboratory to showcase the product and customer evaluation. As opposed to the SLAC installation that is being actively used by SLAC as part of their data processing facilities, this system is available for customers to evaluate in great detail. We can run their programs on it and demonstrate in real time its performance.

A website has already been established (www.clusteredsystems.com) and profiles established on Linked in and Facebook.

A comprehensive set of data sheets and white papers have been prepared.

Conclusions

The project has been a total success in its cooling and power supply aspects. The stated goals were exceeded and additional benefits of the technology were discovered. These include very high density that can cut capital expenditure by 50%, and the flexibility of the cold plate system. We believe that we can double the cooling capacity to 200kW per rack, making it suitable for GPU intensive computing.

We are disappointed that the work on the I/O virtualization and processor-to-processor communication was not finished prior to the end of the contract. However, we are working with our supplier to complete the work and install it at SLAC.

We have noted considerable resistance to change in the Data Center facilities community. This is due to multiple considerations including fear of job loss, lack of understanding and resistance to change what is good enough (even though very expensive and unreliable).

The many expert Data Center designers whose whole experience has been with air cooling will find that their cooling skill set would no longer be needed. Most facilities personnel have not had a strong grounding in basic physics and properties of matter, hence are skeptical of the claims made by the liquid cooling community. Finally, change is resisted because nobody wants to be first, resulting in a “Catch 22” situation.

Recommendations

1) The DoE should specify that PUEs for any new HPC project should not exceed 1.15. This would help develop the market, not just for Clustered Systems but for any company with a viable liquid cooling solution.
2) Energy calculations for liquid cooled systems should be added to the DoE “DC Pro” program.
3) The orthogonal switch architecture, while it has been used for several years in network switches and routers is new to the compute space. This can introduce a new degree of architectural flexibility, enabling higher speeds due to increased server density, different interconnect topologies and even improve software defined networking. DoE should solicit proposals along these lines.
4) DoE must use its “bully pulpit” for liquid cooling promotion.
5) Fund additional development in the cooling of GPUs in the Clustered architecture.

Acknowledgements

We would like to thank our partners and their employees whose participation in this project helped make it a success:

California Energy Commission – additional funding
Intel Corporation – CPUs, motherboards and technical assistance
Emerson Network Power – custom rack, modified XDP and bulk 380VDC power supplies
One Stop Systems – PCIe switches and boards
Panduit Corporation – PDU development and manufacture
PG&E – for the loan of thermal, power and flow data loggers
Dow Corning Company – custom thermal grease formulations and technical support
SLAC National Accelerator Laboratory – system installation and evaluation

The many small machine and metal worker shops that gave us invaluable advice and support including California Machining, Spacesonics, Eclipse Metal Fabrication, Sematech, etc.

References

http://uptimeinstitute.com/2012-survey-results

http://www.google.com/about/datacenters/efficiency/internal/index.html

Jonathan Koomey http://www.analyticspress.com/datacenters.html

http://www.energystar.gov/ia/partners/downloads/V5_0_Computer_Clarification.pdf?d5d9-e03b

ASHRAE’s 90.1-2008 Chapter 11

Appendix B of LBNL document “Demonstration of Rack-Mounted Computer Equipment Cooling Solutions”, by H. Cole 2010

Appendices

Data Sheets

CLUSTERED SYSTEMS COMPANY, INC

Extreme Green Cooling

ExaBlade XCS-3000 100kW Blade Rack

XCS-3000 Overview

The ExaBlade XCS-3000 is a very dense, very energy efficient, blade product family. This blade power and cooling platform uses Clustered Systems' phase-change Touch Cooling™ technology. The power and cooling overhead is so low that it is an excellent choice for any multi-server application.

It supports up to 192 dual CPU server boards in 96 dual board blades, 24 network switch blades, plus power distribution and cooling. Alternately, power supplies and switches may be substituted for one or more chassis to provide a fully integrated Super Node in a single rack.

All electrical and electronic components and boards are readily replaceable, ensuring a multi-generational platform life.

The figure on the right illustrates a XCP-R3006 rack fully populated with six XCP-C3000 chassis, each outfitted with 16 XCP-CB3014 blades comprising two Intel S2600JP server boards.

Standard configurations

Standard configurations support 1.2KW/blades (>100KW total) power and cooling.

XCP-3080 Configuration

- 80ea 8U cooled front blade slots
- 20ea 1U cooled rear network blade slots
- 2ea 1U slots for network switches
- 1ea 6U 105KW 480VAC-to-380VDC bulk power supply

XCP-3096 Configuration

- 96ea 8U cooled front blade slots
- 24ea 1U cooled rear network blade slots

Target Applications

The XCS-3000 is ideal for hosting High Performance Computing, Cloud Computing, Routing and Switching, Storage Systems, or combinations thereof.

Clustered Systems has partnered with leading hardware and software vendors to offer complete turn-key systems.

Cooling System

The platform is cooled by Clustered Systems' Touch Cooling™ technology. The XCS-3000 can be configured to cool up to 2.4KW per blade and up to 200KW per rack.

Two cold plates cool each front blade position. Pumped R134a refrigerant carries away the heat via highly efficient refrigerant phase change. One cubic centimeter of refrigerant can carry off the same amount of heat as 5.5 liters of air using only a small fraction of the pumping energy.

The thermal resistance between the hot chips and cold plate is designed to be very low. Often the returning refrigerant can be re-cooled by ambient air without the use of a chiller. Typically, cooling overhead is 5%-10% of IT load. Chips dissipating up to 300W each can be accommodated and cooled.

XGS-3000 100kW Blade Rack

The system is plug compatible with Liebert Inc.'s XDP and XDC pump/heat-exchanger /chiller units.

Front Blade Slots

These 16 slots per chassis are designed to house compute servers, line cards or disk drives. Each has two built-in cold plates capable of providing effective cooling in excess of 1.2KW. Special thermal pads are mounted to the cold plates. The cold plates slip into slots in the rear of the blades upon blade installation, whereupon mechanisms then clamp the cold plates down upon the hot active components

Rear Blade Slots

These are orthogonally mated to the front blades without an intermediate backplane or mid-plane. They provide interconnect between the front blades and external network(s). As they are field removable, they may also contain active components, unlike conventional blade products which require the chassis to be dismantled to extract the mid-plane. Each slot has a cold plate above it.

Bulk Power Supply (3080 only)

Provision is made in the rack for a 6U high unit. A 105KW 480VAC-to-380VDC rectifier unit is available as an add-in to the XCS-3000. Eliminating intermediate conversion and distribution steps can cut energy consumption by 10%.

Cable raceways are installed on each chassis. Up to 1.2KW (2.4KW) 380VDC is available to each blade. An interface module in each blade converts from 380VDC to 12VDC blade power.

Serviceabilty

The XCS-3000 is designed to be highly service friendly. All electronics, including backplanes are field removable making it very easy to execute a total system upgrade in situ.

Other benefits

- Quiet! (no fans in blade chassis)
- Room neutral, deployable anywhere
- Eliminates power waste associated with air movement; 20% to 50% energy savings in the data center
- All connections are permanent, hermetic and virtually leak proof.

Availability

Available now.

Companion Products

Clustered Systems offers the following introductory products for the XCS-3000:

- XCP-C3000-s – standard blade chassis mechanical enclosure that may be loaded with subsystems of choice.
- XCP-C3000-g – 2.4KW/blade chassis version for GPU subsystems
- XCP-PS3105-e 105KW 480VAC-380VDC Emerson power supply
- XCP-PB3001 – Chassis level PDU
- XCP-CB3014 – Dual Intel S2600JF blade
- XCP-CD3042 – Intel S2600JF and dual Intel PHI 7100 blade
- XCP-CB3082P – Intel S2600CP GPU controller blade
- XCP-CB3082G – Octal Intel Xeon PHI 5100 blade
- XCP-NBR3024 – a 24 port PCIe rear mounted switch blade
- XCP-SP3040 – a 40 port external PCIe switch and I/O virtualization engine
- XCP-CB3000X – standard front blade mechanical enclosure that may be loaded with subsystems of choice.
- XCP-NBR3000X – standard rear blade mechanical enclosure that may be loaded with subsystems of choice.

XGS-3000 100kW Blade Rack

Feature Summary

Rack Standard	EIA-310-D, EIA-310-E Compliant
Height	48U, 84 in (2134 mm)
Width	31.5 in (800 mm)
Depth	47.2 in (1200 mm)
Enclosure Style	Rectangular
Rack Style	Free Standing
Number of Chassis Per Rack	6
Chassis Height	8U (13.95 in)
Number of Front Slots Per Chassis	16
Inside Height of front blade slots	13.58 in (340mm)
Inside Width of front blade slots	1.6 in (40.6mm)
Inside Depth of front blade slots	29.9 in (759mm)
Number of rear slots per Chassis	4
Inside Height of rear blade slots	1.5 in (38.1mm)
Inside Width of rear blade slots	27.1 in (688 mm)
Inside Depth of front blade slots	7 in (177.8 mm)
Doors	Optional lockable front & rear doors
Side panels	Included
Construction Material	Powder coated steel
ROHS Compliance	Optional
Leveling Feet	Supplied
Rack Hole Type	N/A
Gross Weight (when fully outfitted with blades)	4000lbs
Coolant	
Type	R134a
Refrigerant ingress connection	2ea 7/8" Quick Connect male
Refrigerant egress connection	2ea 1¼" Quick Connect male

Bldg 30-01, 3350 Scott Boulevard, Santa Clara, CA 95054

www.clusteredsystems.com
info@clusteredsystems.com
Tel +1-408-327-8100 Fax +1-408-327-8101

Extreme Green Cooling

ExaBlade XCP-C3000 Chassis

XCP-C3000-x Overview

The XCP-C3000 is a member of Clustered Systems' XCS-3000 ExaBlade product family. The XCS-3000 is an open platform that accepts any mechanically conforming device. This very dense, energy efficient blade power and cooling platform uses Clustered Systems' phase-change Touch Cooling™ technology.

The XCP-C3000 is designed to be mounted in a XCP-R3006 rack along with other chassis, power supplies or switches.

All electrical and electronic components and boards are readily replaceable, ensuring a multi-vendor compatibility and multi-generational platform life.

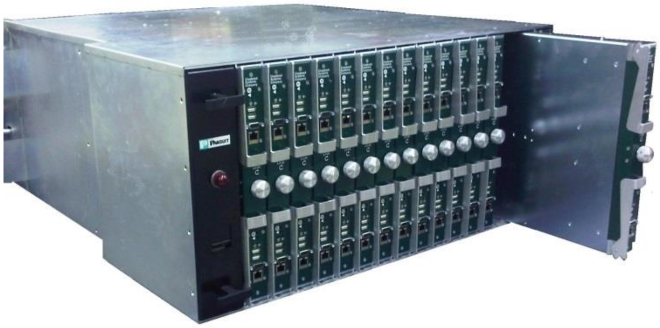
The XCP-C3000 Blade Chassis comprises the mechanical support, power distribution and cooling for 16 XCP-CB30xx blades in the front and 4 XCP-NBR30xx orthogonal active plane cards in the rear. The active planes are typically networking cards providing the first layer of switching in a larger system.

The chassis has built-in power distribution with the leftmost slot reserved for a PDU blade.

Two versions are available, the –s version that powers and cools up to 1200W per blade slot, and the –g version that powers and cools up to 2400W per blade slot.

Pumped R134a refrigerant carries away the heat via highly efficient refrigerant phase change. One cubic centimeter of refrigerant can carry off the same amount of heat as 5.5 liters of air using only a small fraction of the pumping energy. Operation is completely silent.

The thermal resistance between the hot chips and cold plate is designed to be very low. Often the returning refrigerant can be re-cooled by ambient air without the use of a chiller. Typically, cooling overhead is 5%-10% of IT load. Chips dissipating up to 300W each can be accommodated and cooled.



Chassis Components

- 16ea 8U cooled vertical front blade slots
- 1ea 8U uncooled vertical front PDU blade slot
- 4ea 1U cooled horizontal rear blade slots

Cooling

The system is plug compatible with Liebert Inc.'s XDP and XDC pump/heat-exchanger/chiller units.

Chasses are factory installed in the XCP-R3006 rack. Refrigerant is distributed to each chassis in the rack via fixed flow regulators sized to the chassis application. Refrigerant connections to each chassis are permanently and hermetically sealed by brazing to minimize the possibility of leaks and consequential refrigerant loss. All connections within each chassis are likewise permanently and hermetically sealed.

Front Blade Slots

These are designed to house compute servers, line cards or disk drives. Each has two built-in cold plates capable of providing effective cooling in excess of 1.2KW. Special thermal pads are mounted to the cold plates. The cold plates slip into slots in the rear of the blades upon blade installation, whereupon mechanisms then clamp the cold plates down upon the hot active components

The figure above illustrates a XCP-3000-s fully populated with sixteen XCP-3014 blades and a PDU blade.

Rear Blade Slots

Rear blades directly mate to the front blades without an intermediate backplane or mid-plane. They provide interconnect between the front blades and external network(s). As they are field removable, they may also contain active components, unlike conventional blade products which require the chassis to be dismantled to extract the mid-plane. Each slot has a cold plate for cooling.

Chassis Power Distribution:

Cable raceways are provided with each chassis. Up to 1.2KW (2.4KW) is available to each blade. A PDU blade distributes the power to the individual blades.

.

Feature Summary

Chassis Height	8U (13.95 in)
Chassis Depth	38.6 in (980.3mm)
Chassis Width (mounting clearance)	28.7 in (728.8mm)
Number of Front Server Slots	16
Number of Front PDU Slots	1
Inside Height of front blade slots	13.58 in (340mm)
Inside Width of front blade slots	1.58 in (40.1mm)
Inside Depth of front blade slots	29.9 in (759mm)
Number of rear blade slots	4
Inside Height of rear blade slots	1.5 in (38.1mm)
Inside Width of rear blade slots	27.1 in (688mm)
Inside Depth of rear blade slots	7 in (177.8mm)
Construction Material	Zinc coated steel
ROHS Compliance	Optional
Coolant	R134a
Power/Cooling Option	20KW / 40KW
Power Distribution	380VDC
Rear Blade power / cooling	400W
Front Blade power / cooling options	1200W / 2400W

Bldg 30-01, 3350 Scott Boulevard, Santa Clara, CA 95054

www.clusteredsystems.com

info@clusteredsystems.com

Tel +1-408-327-8100 Fax +1-408-327-8101

CLUSTERED SYSTEMS COMPANY, INC

Extreme Green Cooling

ExaBlade XCP-3014 Dual Intel S2600JF Blade

Overview

The XCP-CB3014 is a member of Clustered Systems' XCS-3000 blade product family. This very dense, energy efficient blade power and cooling platform uses Clustered Systems' phase-change Touch™ Cooling technology.

The XCP-CB3014 Quad Socket Compute Blade has the highest power density of any available blade server. It is designed to plug into the XCP-C3000 Blade Chassis. It contains two independent dual socket Intel S2600JF server boards and two optional hard drives. All communication to and from each board is carried over two rear PCIe ports and two 1GbE ports, one on the front panel and one in the rear.

The external PCIe network is configured to provide shared memory operation, very low latency inter-processor communication and I/O virtualization. This approach also cuts energy usage by about 8%.

A 380VDC power connector provides power for each board.

CPU and Memory

Each server board supports two of Intel's Xeon EP-26xx processors and eight DIMMs.

System Interconnect

The rear of each server has two high speed connectors for redundant system interconnect. Each carries a 4 lane PCIe 2.0 link plus a single 1GbE connection.

These connectors plug into a chassis mounted Network Blade, such as the XCP-NBR3024, a 24 Port PCIe switch. This in turn connects to an external XCP-SP3040, a 40 Port PCIe network switch. This latter switch has a controller to configure the PCIe network and provide I/O virtualization facilities.

The rear Ethernet connection may be passed on to an external management network.

Front Panel

USB, VGA and 1GbE ports are brought to the front panel from each server and multiple LED indicator lights provide system and server status.

Hard Disk Drive

Each server board supports a 2.5" form factor hard disk drive.

Cooling

Cooling is supplied by the blade chassis using Clustered Systems' Touch Cooling™ technology. Power and cooling will support full CPU turbo mode of 3.1GHz and 135W power dissipation.

Power

Each server board connects to a 380VDC power supply via a connector on the rear of the blade. This voltage is converted to lower voltages in the blade. This approach eliminates the need for heavy cabling or bus bars and lowers resistance losses.

As a bulk 380VDC power supply may be directly connect to native 480VAC, several intermediate conversion stages are eliminated, resulting in about 10% energy savings.

Safety

At high DC voltages, measures must be taken to ensure personnel safety and to prevent arcing which may cause connector erosion or overheating.

The XCP-CB3014 contains one electrical and two mechanical interlocks that prevent removal of a powered board, and prevent power being applied to the boards unless the blade is properly seated and the cold plate is in position to cool the boards.

Each Blade	
Form Factor	13.55” high x 1.57” wide x 29.9” deep
Weight	28 pounds
Number of Server Boards	2
Each Server Board Slot	
Server Board	Intel Server Board S2600JF (“Jefferson Pass”)
Front Panel	Connectors:1GbE, VGA, 2ea. USB Switches: Reset, Power, Selected LEDs: Power, Status, ID, High Voltage, Network Activity
Rear I/O	2ea PCIe x4 Gen 2 2ea 1GbE management network 1ea 380VDC power connector
Hard Disk	2.5” Form Factor, 7200 RPM, 160GB, SATA 3Gb
Power input	380VDC, 1.3A
Cooling System	Clustered Systems Touch Cooling™ technology
Remote Management	IPMI 2.0 interface

Bldg 30-01, 3350 Scott Boulevard, Santa Clara, CA 95054

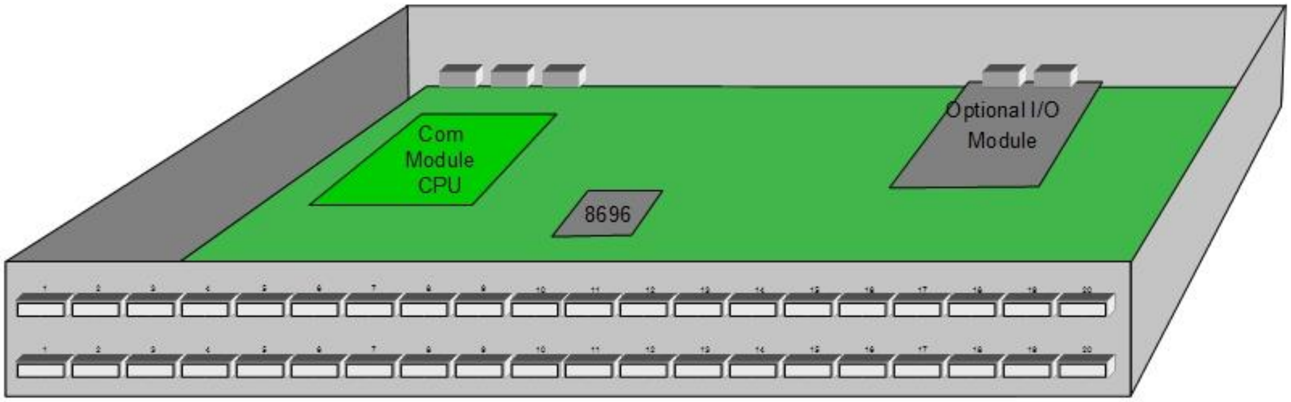
www.clusteredsystems.com

info@clusteredsystems.com

Tel +1-408-327-8100 Fax +1-408-327-8101

Extreme Green Cooling

ExaBlade XCP-SP3040 External PCI Switch



Overview

The XCP-SP3040 Switch is a PCIe 2.0 x4 switch containing two independent 24 port switches in a 1U high box. 20 ports from each switch terminate to external PCIe connectors in the rear while the other four connect to the COM module and optional I/O module.

COM Module

Each switch has a COM module for control and configuration.

The COM module performs network discovery and configuration functions for the switch network. This information is shared with other interconnected switches to provide a unified platform for blade-to-blade communication.

PCIe Switch Network

The PCIe switch network comprises all the interconnected switch blades and external switches. Every server connected to the network has access to every other server through PCIe packet switching.

Intel® I/O Expansion Module

An Intel® I/O Expansion Modules slot in the front enables external connectivity via dual10G Ethernet, or any other I/O type compatible with the Intel® I/O Expansion Module format, including InfiniBand, SATA, etc.

Applications

The XCP-SP3040 interconnects with, and controls the XCP-NBR-3024 Switch Blade. It provides switching and access to an external network.

Power and Cooling

The box is powered by 120 VAC and is air cooled.

References

Intel® I/O Expansion Modules
http://www.intel.com/products/server/io/index.htm

Features

Each Blade	
Form Factor	1U high x 26” wide x 30” deep
PCIe 2.0 connections	40
Intel® I/O Expansion Modules slots	1
Power input	120 VAC, 200W