HPC Software Requirements to Support an HPC Cluster Supercomputer

Susan Kraus, Cray Cluster Solutions Software Product Manager
Maria McLaughlin, Cray Cluster Solutions Product Marketing

Cray Inc.

Table of Contents

Overview

The dramatic growth of the high performance computing (HPC) market has been powered almost exclusively by the adoption of Linux clusters. The low cost and high performance of these systems has been the major driver for this growth. This advantage is always in danger of diminishing if system support software does not keep up with the rapid growth in hardware performance and complexity. A major cause of cluster complexity is proliferating hardware parallelism with systems now averaging thousands of processors. Each of the processors is a multicore chip with its own memory and input-output resources. The memory size, number of processing cores and overall performance of these nodes has been increasing for more than 20 years at a rate that doubles the performance approximately every 18 months. Add to this trend the additional issues of controlling power consumption and the difficulty of scaling many applications and cluster management tools quickly becomes a major problem.

The global HPC server market in 2012 reached a record level of $11.1 billion, en route to more than $14 billion in 2016 (7.3%CAGR). Most market research firms emphasize the importance of alleviating cluster complexity through a coordinated strategy of investment and innovation to produce totally integrated HPC systems. They stress the software advances will be more important for HPC leadership than hardware progress. Countries and companies that underinvest in cluster software will lose ground. This makes it both important and urgent for companies building supercomputer-class HPC clusters to provide the right cluster software, management tools, system integration support and professional services to provide datacenter administrators and end users with the tools they need for problem solving. [1]

This paper is intended for end users who are interested in implementing an HPC system and need to learn about the essential cluster software and management tools required to build and support cluster architecture. You will quickly understand why the software provided for an HPC cluster has become much more important than the differentiated cluster hardware. You will explore how Cray (www.cray.com), a leading provider of innovative supercomputing solutions, combines HPC open source software with key compatibility features of the Advanced Cluster Engine™ (ACE) management software to provide a complete software stack for its Cray CS300™ cluster supercomputer product line. This paper will clarify why an HPC cluster software stack is required to build a powerful, flexible, reliable and highly available Linux supercomputing cluster architecture. This paper will also provide you with a close examination of Cray's HPC cluster software stack and management software features that have been addressing many customer pain points for medium to large HPC deployments.

Cray HPC Cluster Software Stack

The Cray HPC cluster software stack is a validated and compatible set of software components below the end user application layer and essential to support an entire supercomputer operation. This software stack includes programs that are unique to the architecture and are required to support the applications that will run on the system. Each node of the Cray CS300 cluster supercomputer is a shared memory multiprocessor with its own memory and independent operating system software. The nodes are interconnected by a high performance communications network that supports the rapid exchange of data between the individual computers. This network is normally based on InfiniBand technology and allows the individual computers to work together on problems that can be divided across multiple computers. The parts of these parallel programs execute independently and synchronize using the interconnect network fabric only when it is necessary for data sharing.

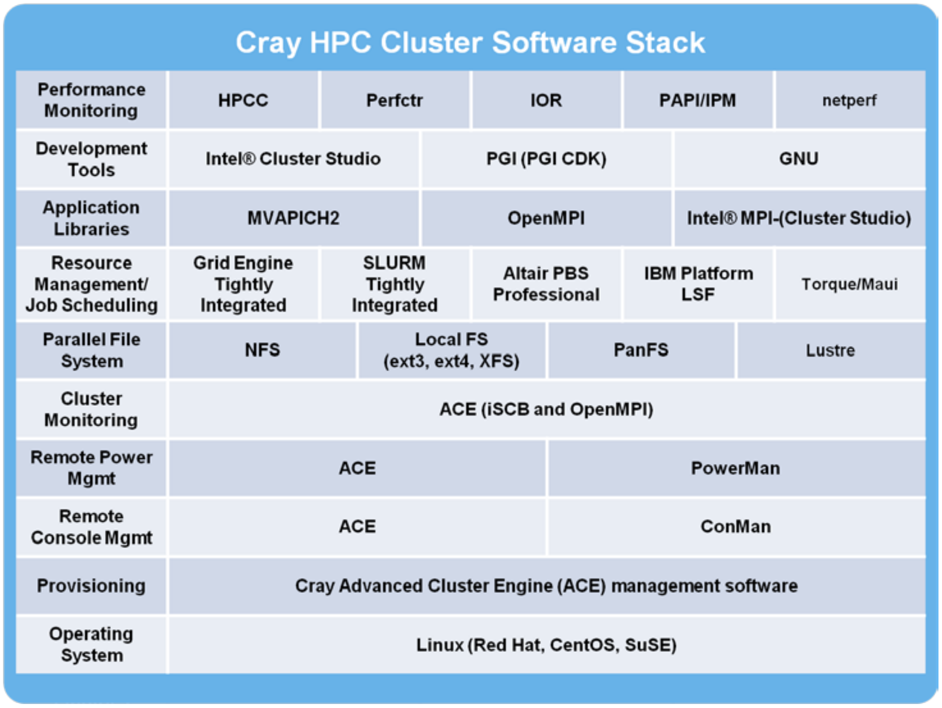
The software stack for one of these machines can be divided into two groups of programs. The first is the software that runs on the individual nodes. The second is the software that aggregates the large number of individual computers into a single system. The software that makes up this second group very often runs on nodes that are specifically dedicated as management nodes.

First let's discuss the software that runs on the individual nodes. Each node runs an independent operating system. The operating system used for most HPC systems is Linux. When Linux is installed on the nodes it is configured with a local file system and networking software that supports internode communications. The operating system, file system and network software are all part of the HPC software stack. In addition, this first software group includes libraries of routines optimized for the specific processors being used in the system. These libraries include math functions and other utilities that are repeatedly used by multiple applications. They also include Message Passing Interface (MPI) programs that support efficient message passing and synchronization between the nodes in a cluster. Software development is supported by codes that are also part of the software stack that runs on the individual nodes. These programs support the development of new codes and the maintenance of existing software. They include "C Language" and "Fortran" compilers plus debugging, test programs and performance verification programs that can be used to test system performance

The second group of software includes three major codes. The first is the management system that allows all of the system hardware to be managed from a single point. The second is the system resource management and scheduling software that provides a single user interface for scheduling and executing parallel problems across all of the processors in the system. The third is the software that supports the routing of messages over high-speed networks such as the network fabric management.

The following software components are part of the HPC cluster software stack for its Cray CS300 cluster supercomputer product line. Refer to Figure 1 for the Cray HPC Software Stack and Figure 2 for the entire Cray system architecture:

- Operating Systems: Linux – Red Hat Enterprise Linux, CentOS and SUSE
- Application Libraries: MPI and Math Libraries
- Development Tools: Compilers, Debugging Tools
- Performance Verification Tools and Monitoring
- Cluster Management: System Management, Monitoring, Provisioning
- Network Fabric Management
- Resource Management and Scheduling
- Parallel File System



Cray HPC Cluster Software Stack

Performance Monitoring	HPCC	Perfctr	IOR	PAPI/IPM	netperf
Development Tools	Intel® Cluster Studio	PGI (PGI CDK)	GNU		
Application Libraries	MVAPICH2	OpenMPI	Intel® MPI-(Cluster Studio)		
Resource Management/ Job Scheduling	Grid Engine Tightly Integrated	SLURM Tightly Integrated	Altair PBS Professional	IBM Platform LSF	Torque/Maui
Parallel File System	NFS	Local FS (ext3, ext4, XFS)	PanFS	Lustre	
Cluster Monitoring	ACE (iSCB and OpenMPI)				
Remote Power Mgmt	ACE	PowerMan			
Remote Console Mgmt	ACE	ConMan			
Provisioning	Cray Advanced Cluster Engine (ACE) management software				
Operating System	Linux (Red Hat, CentOS, SuSE)				

Figure 1: Cray HPC Software Stack

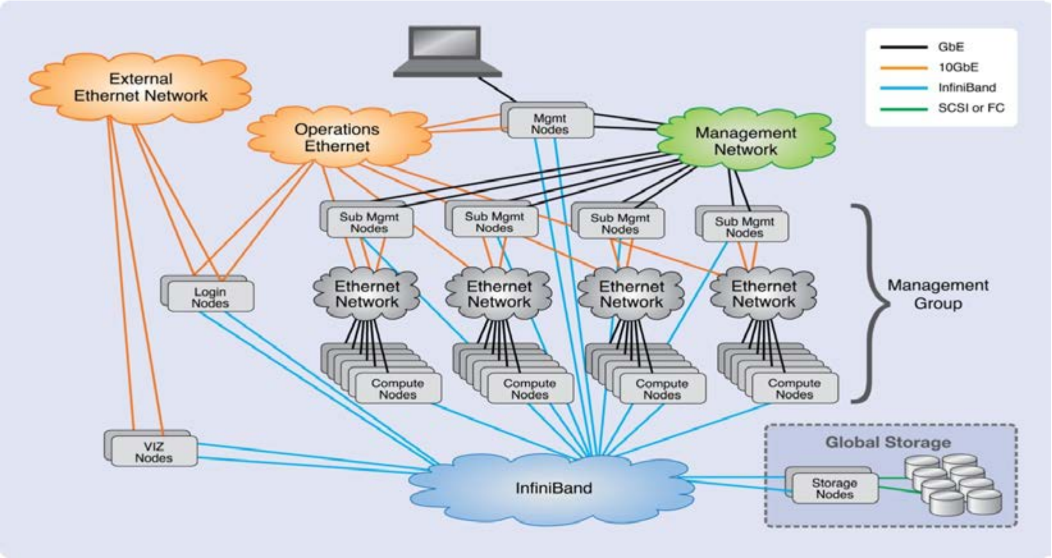


Figure 2: Cray CS300 Cluster Supercomputer Architecture

Operating System

Almost all HPC clusters use the Linux operating system. More than 10 versions of the Linux operating system are supported by specific vendors and by the open source community. The versions used for most commodity clusters are Red Hat Enterprise Linux, SUSE, CentOS and occasionally Scientific Linux. These versions are very similar and differ mostly in the installation programs supplied with the operating system software. A larger difference may be the level of support provided by the vendor. The same versions of these operating systems run on workstations, servers, and on the compute nodes of supercomputer clusters.

The Cray CS300 cluster supercomputer supports Red Hat Enterprise Linux, SUSE and CentOS operating systems. The Cray Cluster Management System is delivered with Red Hat Enterprise Linux installed on the management nodes. It was selected because of the higher level of support available. Cray delivers preconfigured Red Hat, SUSE or CentOS compute node images. Red Hat and SUSE are priced on a per node basis. CentOS operating system is available as open source software that can be used at no cost.

Application Libraries

Computer applications make use of libraries to perform commonly used functions. These common functions such as math functions have been designed and optimized to make the most efficient use of a particular processor's capability. Libraries are available for message passing, mathematical functions, and the OpenFabrics Enterprise Distribution (OFED™).

The Message Passing Interface (MPI) libraries are especially important to cluster operation since they support efficient messaging and synchronization between codes running on different processors in a cluster. These libraries simplify user programs since they can be included where required with little programming effort. Cray includes the open source MPI libraries available for the Linux operating system and offers the MPI libraries available with the Intel® and PGI® development suites.

Cray also offers the Intel® Math Kernel Library which provides Fortran routines and functions that perform a wide variety of operations on vectors and matrices including sparse matrices and interval matrices. The library also includes discrete Fourier transform routines, as well as vector mathematical and vector statistical functions with Fortran and C interfaces.

Development Tools

Development tools include C Language, Fortran and other compilers that are used to create software for HPC applications. These compilers include features that support creating and debugging code for parallel computing. Cray's HPC cluster software stack provides open source development software that is available with the Linux operating system and offers compilers and development tools from Intel or PGI.

Intel® Cluster Studio provides the tools needed to develop, test and run programs on clusters based on Intel processors. It bundles Intel's C, C++ and Fortran compilers and includes Intel's version of the MPI message passing protocol that allows server nodes to share work. It also includes various math and multithreading libraries to improve applications performance. The Intel® MPI stack can scale to more than 90,000 MPI cores. It also includes support for redundant networks and supports both load sharing and failover. The debugger tools not only provide a single node view, they also provide a cluster-level view of multiple data gathering and reporting mechanisms. This software can support finding threading errors and memory leaks across an entire cluster. These tools are essential for supporting parallel computing. The Intel math libraries are essential to getting maximum performance from Intel processors.

The PGI development suite provides compilers, MPI libraries, and math libraries for non-Intel based clusters and also for Intel-based clusters that use NVIDIA® GPUs. These systems are supported by the PGI CDK® Cluster Development Kit®, which includes support for C++ and Fortran compilers plus parallel libraries and debugging tools. The PGI CDK includes the tools needed to support NVIDIA GPU devices and the NVIDIA CUDA language. [2], [3]

Performance Verification Software and Monitoring

Cray's HPC cluster software stack includes the standard benchmarking and performance tools used to measure and verify cluster performance. Programs included are HPCC, Perfctr, IOR, Bonnie, PAPI/IPM and netperf. The stack also includes the OFED library that measures the latency and bandwidth for InfiniBand networks.

Cluster Management: System Management, Monitoring and Provisioning

Cluster management software started as simple remote server management programs based on the Intelligent Platform Management Interface Protocol (IMPI) and Simple Network Management Protocol (SNMP). The programs allowed servers that were remotely located to be turned on, turned off, reset and rebooted. They collected status from the servers and provided a serial I/O connection to the servers for interacting with the operating systems or with programs running on the servers. When clusters began to appear, these remote management programs were adapted to allow all of the nodes in a cluster to be managed from a single point. As clusters grew larger and more complex so did the cluster management programs. [4]

Early clusters booted the operating system for a node from a local disk connected directly to the node. This meant that a copy of the operating system had to be stored at each node. Functions were added to the cluster management software to support updating and verifying the software on the nodes from a master copy. As clusters grew to thousands of nodes it created configuration control problems. It was a very slow process making it impractical to quickly load a different OS or library to the nodes. With the continued growth of clusters and I/O performance improvements, diskless compute nodes were introduced. The cluster management software for a diskless system needed to be able to quickly load a new operating system image to the nodes when required. This evolution first led to operating systems using so-called lightweight kernels to simplify the loading process, but these lightweight kernels loaded only the basic operating system kernel capabilities in memory.

Newer cluster management systems allow full OS support with the ability to load transient routines quickly from high performance external storage devices managed by the cluster management system. Modern cluster management programs like Cray's Advanced Cluster Engine™ are able to reboot any or all of the nodes in a cluster within a few minutes. Refer to Figure 3 for the feature diagram of Cray's Advanced Cluster Engine (ACE) management software.

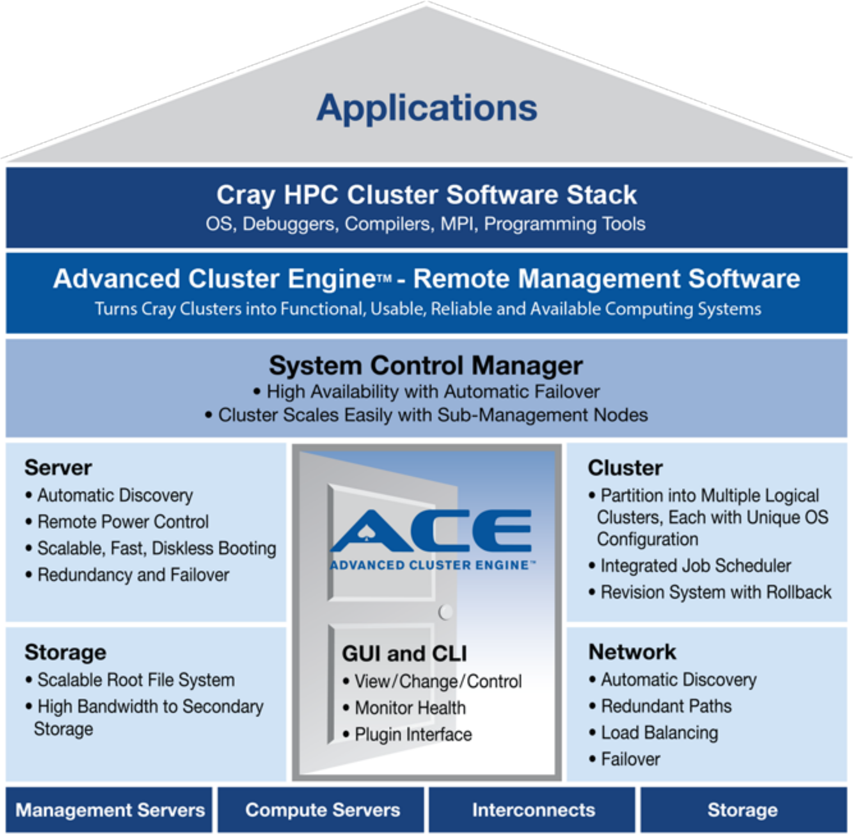

Applications
Cray HPC Cluster Software Stack
OS, Debuggers, Compilers, MPI, Programming Tools
Advanced Cluster Engine™ - Remote Management Software
Turns Cray Clusters into Functional, Usable, Reliable and Available Computing Systems
System Control Manager
• High Availability with Automatic Failover
• Cluster Scales Easily with Sub-Management Nodes
Server
• Automatic Discovery
• Remote Power Control
• Scalable, Fast, Diskless Booting
• Redundancy and Failover
ACE
ADVANCED CLUSTER ENGINE™
Cluster
• Partition into Multiple Logical Clusters, Each with Unique OS Configuration
• Integrated Job Scheduler
• Revision System with Rollback
Storage
• Scalable Root File System
• High Bandwidth to Secondary Storage
GUI and CLI
• View/Change/Control
• Monitor Health
• Plugin Interface
Network
• Automatic Discovery
• Redundant Paths
• Load Balancing
• Failover
Management Servers
Compute Servers
Interconnects
Storage

Figure 3: Advanced Cluster Engine Management Software

This feature makes it very easy to install and maintain the software on a cluster. The administrator needs to install or update only one library copy of the OS and applications. All of the servers get the new version as soon as they boot an updated image. This allows a single configuration controlled copy of the software to be loaded to the nodes in a cluster. The node's unique data such as /etc files are created by the cluster management software as a node is booted. This minimizes the storage requirements for the operating system image. This capability also allows logical clusters to be supported. Logical cluster operation allows one or more of the nodes in a cluster to be loaded with a specific operating system configuration while other nodes run different operating system images enabling multiple OS versions to be supported simultaneously on the same cluster.

This architecture makes the use of a large cluster much more flexible since nodes can be reconfigured in minutes to best support different users. For example, a large cluster with thousands of nodes can be divided into several logical clusters running the same or different versions of Linux to support different working groups. A set of nodes can be configured to support a specific application and then loaded only when required. Ultimately, the entire machine could be reconfigured in minutes into a single cluster to run large overnight jobs. [5]

ACE can simultaneously support hundreds of logical cluster configurations. Some of these may be active while others may be loaded onto nodes only infrequently as they are required to support specific user requirements. The integrated configuration management functions in the ACE software support up to 10 revisions of a compute node image for each logical cluster along with a rollback capability, making it ideal for testing new software packages and providing customized environments.

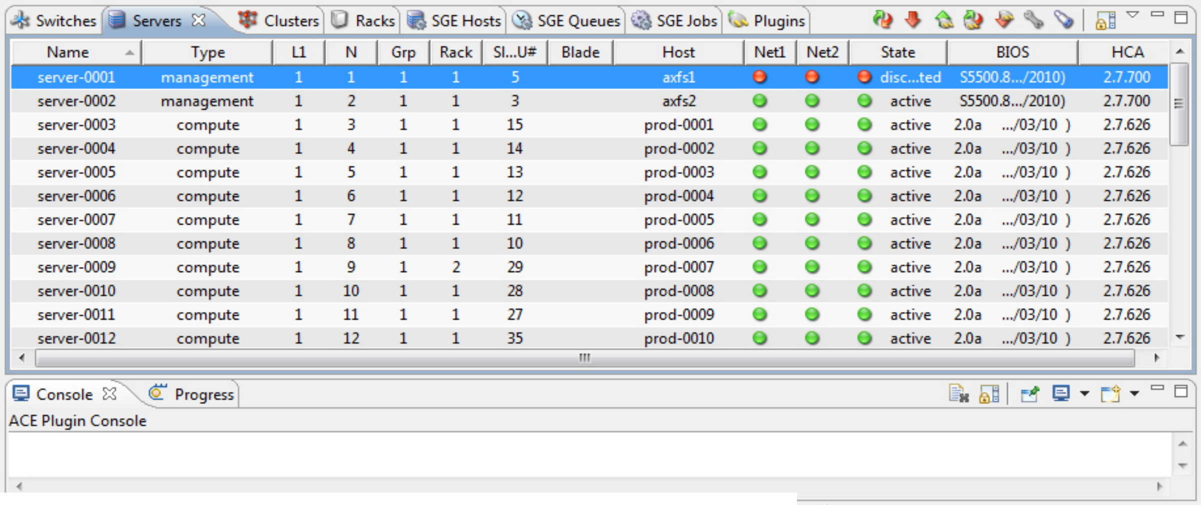
The cluster configuration and topology are predefined in a database during the ACE installation. ACE verifies that the servers are properly connected and configured. It supports the capability to turn each individual server on and off, reset the server, or reboot the server. ACE has three interfaces, a command line interface, a graphical user interface, and a file system interface into the database. These interfaces provide the system administrator with a great degree of flexibility to customize the operation and monitoring of the cluster. It also supports remote BIOS and firmware updates.

ACE includes the capability to discover and check the interconnect configuration and to recognize switches or links that are down, not operating properly or incorrectly connected. It can support fat tree and torus configurations in single-rail or multi-rail configurations. ACE completely supports diskless operation thereby simplifying the administration of the cluster. It provides a multilevel cached root file system, which contributes to the scalability of the system. Local disks are supported for scratch or temporary space.

As clusters have grown to thousands of nodes, availability has become a major consideration. The failure of a single server, data link or switch cannot be allowed to stop the operation of the other nodes in a cluster. ACE supports dual management servers with automatic failover in the event that one of the servers fails. The servers are configured as a replicated pair with the standby server always keeping its database in sync with the active management server. This allows the standby server to take over with no interruption. Since the compute nodes in a cluster are inherently redundant, jobs can always be restarted by the resource manager and scheduler from a previous checkpoint or from the beginning.

ACE allows servers that are part of an active logical cluster that have failed to be replaced with no changes required to the configuration. Networks can be configured with redundancy to support continuous operation. It is also designed to work with hierarchical active/active sub-management server pairs to allow systems to scale without diminishing performance.

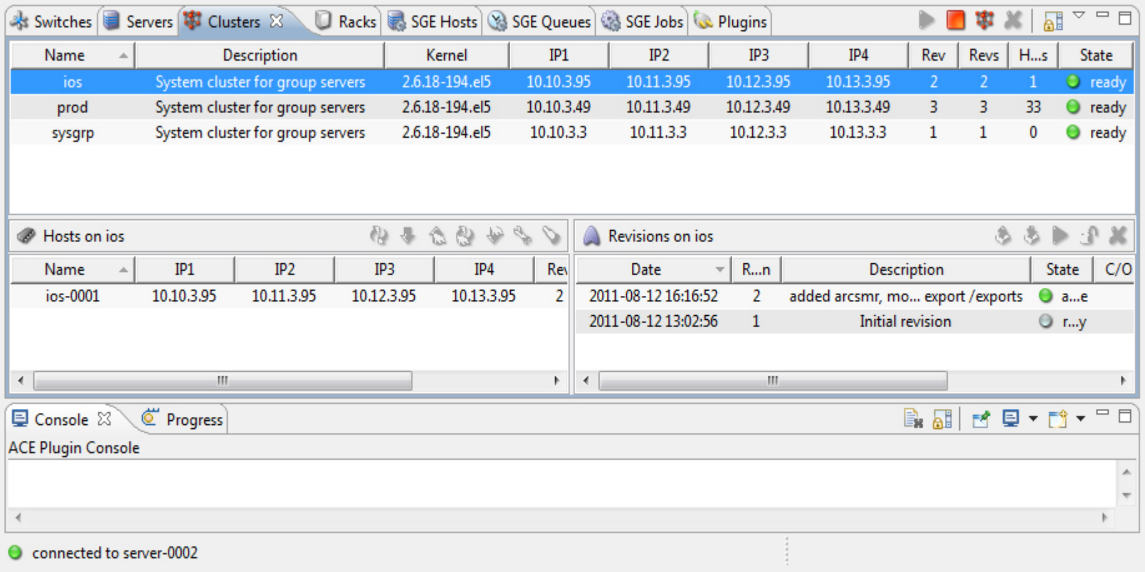
ACE is designed to display the status of servers, networks and logical clusters in an intuitive format. This allows the state of the entire cluster to be determined in a glance and problems to be recognized and dealt with quickly and easily.



Name	Type	L1	N	Grp	Rack	Sl...U#	Blade	Host	Net1	Net2	State	BIOS	HCA
server-0001	management	1	1	1	1	5		axfs1	🔴	🔴	🔴 disc...ted	S5500.8.../2010)	2.7.700
server-0002	management	1	2	1	1	3		axfs2	🟢	🟢	🟢 active	S5500.8.../2010)	2.7.700
server-0003	compute	1	3	1	1	15		prod-0001	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0004	compute	1	4	1	1	14		prod-0002	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0005	compute	1	5	1	1	13		prod-0003	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0006	compute	1	6	1	1	12		prod-0004	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0007	compute	1	7	1	1	11		prod-0005	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0008	compute	1	8	1	1	10		prod-0006	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0009	compute	1	9	1	2	29		prod-0007	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0010	compute	1	10	1	1	28		prod-0008	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0011	compute	1	11	1	1	27		prod-0009	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626
server-0012	compute	1	12	1	1	35		prod-0010	🟢	🟢	🟢 active	2.0a .../03/10)	2.7.626

ACE Plugin Console

Example 1: ACE Servers Display



Name	Description	Kernel	IP1	IP2	IP3	IP4	Rev	Revs	H...s	State
ios	System cluster for group servers	2.6.18-194.el5	10.10.3.95	10.11.3.95	10.12.3.95	10.13.3.95	2	2	1	● ready
prod	System cluster for group servers	2.6.18-194.el5	10.10.3.49	10.11.3.49	10.12.3.49	10.13.3.49	3	3	33	● ready
sysgrp	System cluster for group servers	2.6.18-194.el5	10.10.3.3	10.11.3.3	10.12.3.3	10.13.3.3	1	1	0	● ready

Hosts on ios

Name	IP1	IP2	IP3	IP4	Re
ios-0001	10.10.3.95	10.11.3.95	10.12.3.95	10.13.3.95	2

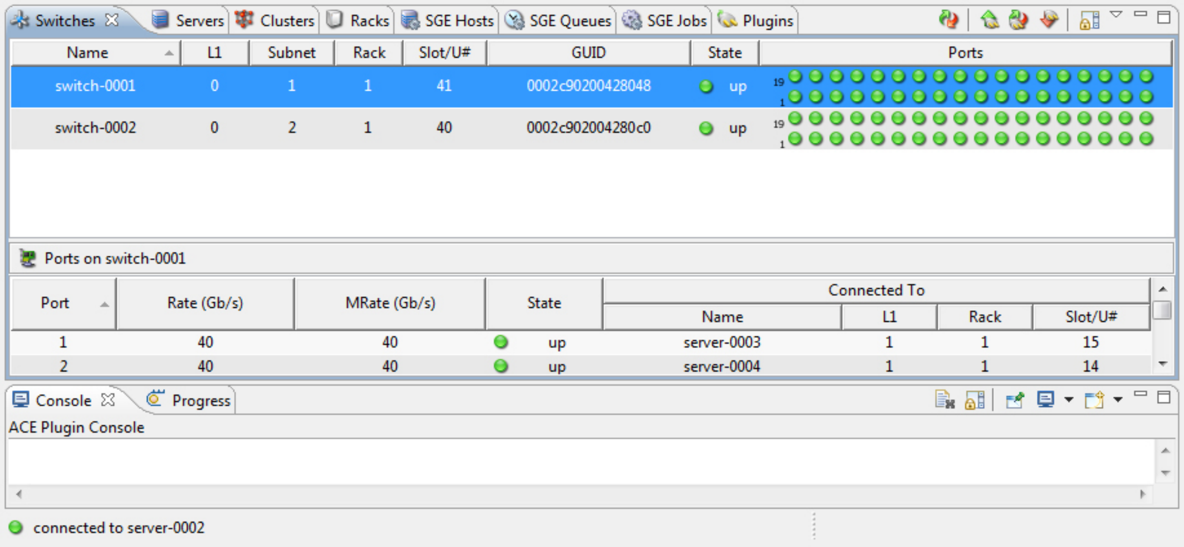
Revisions on ios

Date	R...n	Description	State	C/O
2011-08-12 16:16:52	2	added arcsmr, mo... export /exports	● a...e	
2011-08-12 13:02:56	1	Initial revision	○ r...y	

ACE Plugin Console

● connected to server-0002

Example 2: ACE Clusters Display



Name	L1	Subnet	Rack	Slot/U#	GUID	State	Ports
switch-0001	0	1	1	41	0002c90200428048	● up	
switch-0002	0	2	1	40	0002c9020042 80c0	● up	

Ports on switch-0001

Port	Rate (Gb/s)	MRate (Gb/s)	State	Connected To			
				Name	L1	Rack	Slot/U#
1	40	40	● up	server-0003	1	1	15
2	40	40	● up	server-0004	1	1	14

ACE Plugin Console

● connected to server-0002

Example 3: ACE Network Status Display

Network Fabric Management

The OpenFabrics Enterprise Distribution (OFED) Software is also included in the Cray HPC cluster software stack. Cray works closely with the OpenFabrics Alliance, non-profit organization to provide architectures, software repositories, interoperability tests, bug databases, workshops, and BSD- and GPL-licensed code to facilitate development. The OFED stack includes software drivers, core kernel-code, middleware, and user-level interfaces. It offers a range of standard protocols, including IPoIB (IP over InfiniBand) and DAPL (the Direct Access Programming Library). It also supports many other protocols, including various MPI implementations, and many file systems, including Lustre, and NFS over

RDMA. This collection of codes is designed to support parallel processing on HPC clusters that use InfiniBand networking technology for interconnecting the nodes. Codes are included for managing both fat tree and torus network architectures. This includes managing the routing tables on the InfiniBand switches such as initializing the routing tables and updating the tables to route around connection failures. The library also includes the device driver software for the InfiniBand Host Channel Adapters used to connect the servers to an InfiniBand network. Routines are included to collect status and performance data from InfiniBand networks and programs to test specific network functions such as latency and bandwidth.

Resource Management and Scheduling

The operating system on a single computer manages the processor, memory and I/O resources. It schedules programs that are submitted for execution only when the resources needed to run the program are available and then monitors the execution of the programs.

An HPC cluster needs a similar resource management and scheduling capability for the entire cluster. This has led to the development of applications such as LSF, PBS and Grid Engine. These programs keep track of the processor and memory resources for all of the nodes in a cluster. They submit jobs to the operating systems on the individual nodes when all of the resources are available to run a parallel application. They provide users with a single interface for submitting and monitoring jobs.

ACE is designed to provide the system information needed by the cluster resource manager to properly schedule jobs. It is built with a specific interface for the Grid Engine Enterprise Edition resource manager and scheduler, and SLURM (Simple Linux Utility for Resource Management) that tightly integrates the scheduler into the management system. It is also compatible with other resource management and scheduling software, such as Altair's Portable Batch System (PBS Pro), IBM Platform's LSF, Torque/Maui, etc.

Parallel File System

ACE supports scalable, cached root file systems for diskless nodes, multiple global storage configurations, and high bandwidth to secondary storage. The Cray HPC cluster software stack offers support for common file system options such as the Network File System (NFS), Panasas PanFS™ and Lustre. Lustre is the most widely used file system on the TOP500® fastest computers in the world with over 60 percent of the top 100, and over 50 percent of the top 500. Lustre enables high performance by allowing system architects to use any common storage technologies along with high-speed interconnects and can scale well as an organization's storage needs grow. By providing multiple paths to the physical storage, the Lustre file system can provide high availability for HPC clusters. ACE supports all of the standard file systems that are used with the Linux operating system including ext3, ext4 and XFS. [6]

Conclusion

In this paper you had a complete overview of the essential cluster software and management tools that are required to build a powerful, flexible, reliable and highly available Linux supercomputer. You learned that Cray combines all the required HPC software tools including operating systems, provisioning, remote console/power management, cluster monitoring, parallel file system, scheduling, development tools and performance monitoring tools with key compatibility and additional powerful features of the Advanced Cluster Engine™ (ACE) management software to deliver a best-of-breed HPC cluster software stack to support its Cray CS300 supercomputer product line. Cray has also taken a customer-centric, technology-agnostic approach that offers the customer a wide range of hardware and software configurations based on the latest open standards technologies, innovative cluster tools and management software packaged with HPC professional services and support expertise.

References

[1] IDC Maintains HPC Server Revenue Forecast of 7% Growth Despite Flat Second Quarter http://www.idc.com/getdoc.jsp?containerId=prUS23671312#.US1c4DDvsk0

[2] Intel® Xeon® Processor E5 Family http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-5000-sequence.html

[3] HPC Accelerating Science with Tesla GPUs http://www.nvidia.com/object/tesla-supercomputing-solutions.html

[4] Mesos, A Platform for Fine-Grained Resource Sharing in the Data Center UC Berkeley Tech Report, May, 2010

[5] Adaptive Control of Extreme-scale Stream Processing Systems Proceedings of the 26th IEEE International Conference on Distributed Computing Systems.

[6] "Rock-Hard Lustre: Trends in Scalability and Quality". Nathan Rutman, Xyratex..

Improving HPC Cluster Efficiency with 480V Power Supplies

Giri Chukkapalli, Cray R&D Principal Engineer
Maria McLaughlin, Cray Cluster Product Marketing

Cray Inc.

Table of Contents

Introduction

Traditional AC power distribution systems in North America bring 480V AC power into a universal power supply (UPS), where it is converted to DC to charge batteries, and then inverted back to AC. The power is then stepped down to 208V within the Power Distribution Unit (PDU) for delivery to the IT equipment. The power supplies in the IT equipment convert the power back to DC and step it down to lower voltages that are consumed by processors, memory and storage.

Server power supplies that operate directly from 480/277V power distribution circuits can reduce the total cost of ownership (TCO) for a high performance cluster by reducing both infrastructure and operating cost. Most servers used in high performance clusters were originally designed for desktop applications and for operation in office environments. They normally operate from 120/240V single-phase AC power and are cooled by room air conditioning. The reason for using low voltage 120/240V distribution circuits is safety. This equipment is connected and disconnected by untrained personnel creating the potential for electrical shock. When large clusters are constructed using these same commodity servers, the legacy power and cooling designs create inefficiencies. The 120/240V circuits were never designed to deliver large quantities of power. These inefficiencies are acceptable for a single server but the aggregate losses in a large system with thousands of servers are too large to be acceptable. The higher voltages do not present a safety issue for these larger systems since they have better facilities and trained maintenance personnel. These servers need an alternate power supply design that better fits a computer room environment.

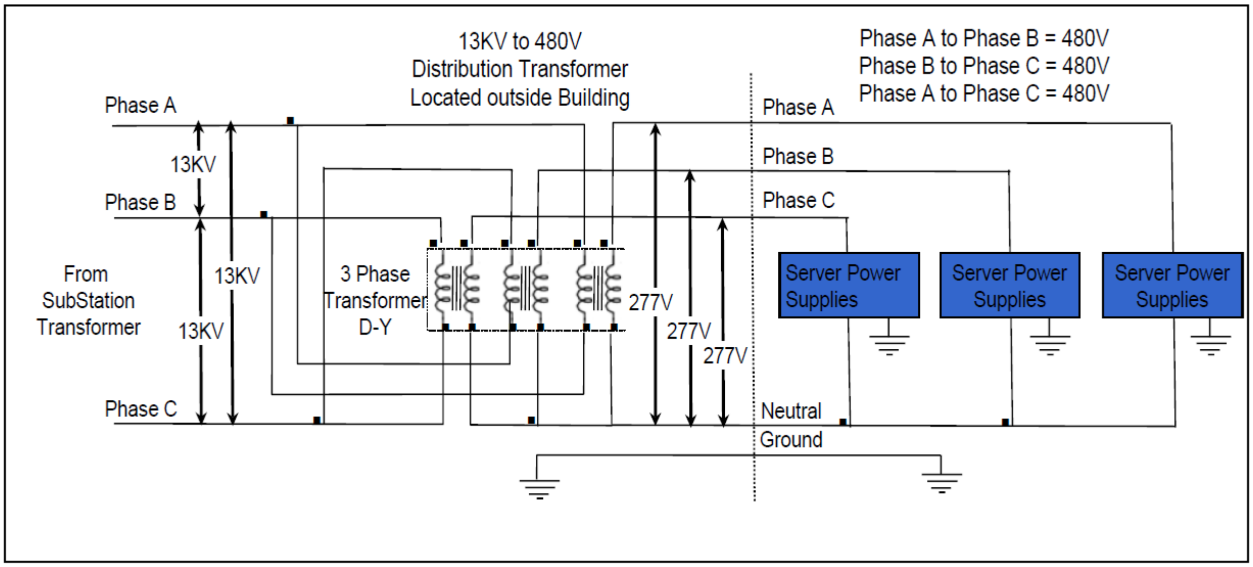


Figure 1: 480V three-phase power distribution directly to server power supplies

Power Distribution

Commercial and industrial users have large power demands. Figure 1 illustrates a typical 480V power distribution system used in commercial facilities. The distribution transformers are generally located outside the buildings. They receive high voltage power directly from a power company substation. The 480V secondary circuits may be capable of delivering up to 2,000 amps, and fuses that protect the transformers are located at the transformers while circuit breakers that protect individual power distribution circuits are located inside the building. The power is distributed from the circuit breakers to equipment inside the building using 480V three-phase circuits. The use of the higher voltage 480V circuits reduces resistive I2R losses by a factor of four to five times compared to using 240V or 208V circuits. This is because 480V circuits require less current to deliver the same amount of power. A load requiring 10KW of power would require 28 amps at 208V, 24 amps at 220V, but only 12 amps at 480V. Most large industrial equipment and mainframe computers are designed to operate directly from 480V three-phase circuits.

Step-down transformers are necessary to convert the 480V three-phase circuits to 208V three-phase and 120/240V single-phase circuits. These circuits are used to power office equipment, desktop computers and coffeemakers. Resistive losses are minimized by locating the 480V to 208V step-down transformers near the loads. Using these lower voltage circuits to power a high performance computer results in unnecessary cost compared to using the 480V circuits directly. It also results in decreased efficiency when 208V circuits are used to operate 120/240V power supplies because they are operated at the low end of their operating range. [1]

The HPC Power Issue

A large high performance computing (HPC) cluster may have thousands of individual servers. Sixty-four or more servers can be located in a single equipment rack requiring up to 30KW. A single equipment rack can require more power than is required by a small office. A typical HPC cluster with 1,024 compute nodes may require 16 or more equipment racks and need more than 500KW of electrical power. Figure 2 illustrates how 208V three-phase power is distributed in current computer rooms. These applications typically use multiple step-down transformers with ratings between 15KW and 150KW to reduce the voltage from 408V three-phase circuits to 208V three-phase circuits. The only reason for the step-down transformers is to accommodate the 120/240V power supplies used in the servers. In fact, current power supplies would be more efficient if the transformers output was 415/240V. Figure 3 illustrates the reduced efficiency of operating 120/240V power supplies on 208V or on 120V. The efficiency quoted for 120/240V power supplies is always measured when running on 240V. They are never as efficient when running on 208V or 120V.

A 1,024-node HPC cluster consisting of 16 equipment racks with 64 compute nodes in each rack would typically require up to approximately 28KW per equipment rack. The servers in each equipment rack would be powered by between 24 and 32 separate power supplies, each with a 1,600- to 1,800-watt capability. The power would typically be distributed to each rack by two 208V 50-amp three-phase PDUs. In many clusters built with commodity servers, up to four distribution circuits per rack are required because of the limited size of commercial rack PDUs. Our example would require approximately 32 separate 208V three-phase power distribution circuits with circuit breakers rated at 50 amps. The power would typically be supplied by approximately eight 75KVA step-down transformers with each transformer supporting four 208V three-phase outputs.

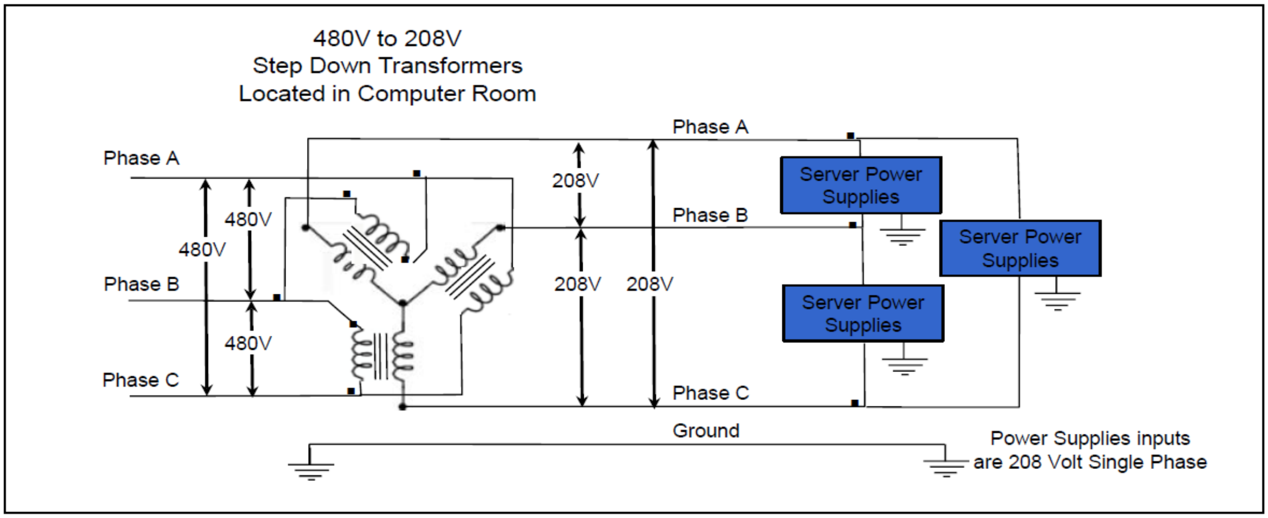


Figure 2: Power from 480V three-phase to 208V three-phase step-down transformer

The 208V three-phase circuits from the step-down transformers are four-wire circuits, which have conductors for each of the three phases plus a safety ground. The server power supplies are connected between two of the phases. This feature provides them with 208V single-phase power. Since the 208V circuits are three-phase circuits, it is important that the server power supplies be distributed equally across the three phases to balance the current in the phases. The PDUs located in the racks would be 208V three-phase units delivering an aggregate of approximately 40 amps to 12 to 16 outlets. Under the North American electrical codes, the branch-level circuit breakers are 80 percent rated so that a 50A circuit can deliver only 40A to a load.

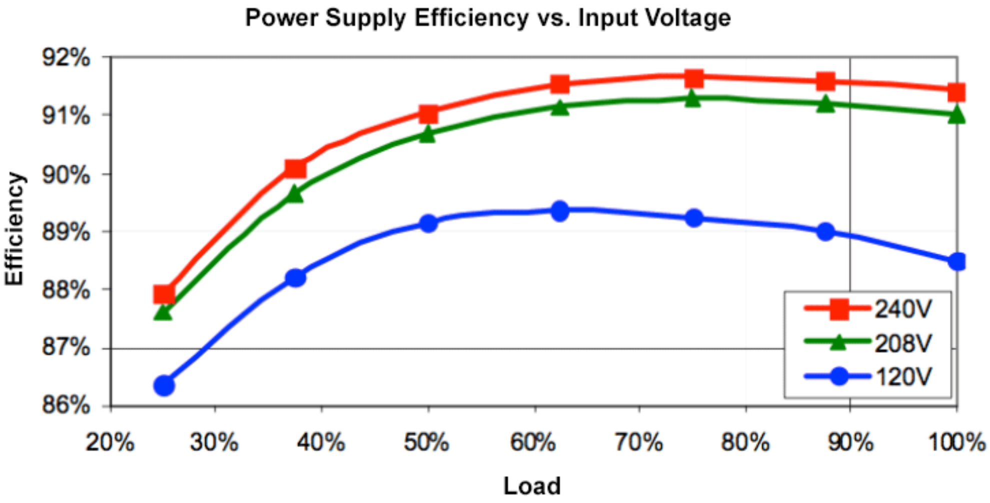


Figure 3: Power supply efficiency versus input voltage

The 480V to 208V transformers used in these applications are designed and rated on their ability to deal with non-linear loads, such as those found in the switching power supplies used for servers. Transformers are rated by the K-factor system, which ranges from K-1 units, which are designed to work only with resistive loads, to K-40 units, which are best suited for use with non-linear loads. Non-linear loads cause harmonics in the transformers that result in increased losses in both the wiring and the transformer cores. In datacenters the commonly available transformer types are K-1, K-13 and K-20. K-1 transformers, which are not designed for non-linear loads, are the least expensive and are normally used for resistive heating and lighting loads. K-13 transformers are somewhat better but more expensive, while K-20 or K40 transformers would be the preferred choice for datacenters using switching power supplies.

Most low-cost transformers used for these applications have aluminum windings and no better than K-13-grade core material. This is driven by installation cost considerations. Users have not been so concerned about the design of these transformers in the past and electrical contractors tend to choose the least expensive units. Using more expensive, higher efficiency transformers could cut transformer losses in half. Transformers that use copper windings instead of aluminum have lower I2R losses. Transformers that use higher quality core material have lower magnetic circuit losses. These transformers are also more expensive, and may offset any savings in electrical power, resulting in approximately the same overall cost. Contractors will typically install K-1 or K-13 transformers if possible because they are more concerned about the installation cost than the operating cost. The best solution, however, is to reduce the number of transformers required.

In a typical 1,024-node cluster, requiring 500KW, the power loss in the 480V to 208V step-down transformer is approximately 3 percent. The losses are due to a combination of resistive and magnetic core losses. A 500KW system would typically have transformer losses of approximately 15KW. This power loss is converted to heat inside the transformers. If the transformer is located inside the computer room, the heat must then be removed by the computer room air conditioning equipment. The losses in the transformers and in the wiring both add to the total load on the 480V three-phase circuits and to the air conditioning load. [2]

The typical life of a high performance cluster is three years. Approximately $47,000 is wasted over the life of a 500KW computer based on a power cost of 12 cents per kilowatt hour. If step-down transformers are located inside the computer room, the increased air conditioning load could cost an additional $20,000 to remove the heat generated by the transformers. For a 500KW system the installed cost of the step-down transformers would add a minimum of $40,000. This adds up to a total life cycle cost of $107,000. This is wasted money. It is wasted as heat and the purchase of equipment that would not be required if the server power supplies operated directly from the 480V three-phase circuits.

The configuration in a system using 277V power supplies is different from the 208V system. The power supplies connect from the phases to neutral instead of between the phases. This provides 277V single-phase power to the power supplies. Only a small increase in the input voltage rating of the power supplies is required to accommodate the change from 240V to 277V operation. In fact, when operating from 208V most server power supplies operate at the low end of the 240V input range and are not as efficient as when operating on 120 or 240 volts. Even though the phase-to-phase voltage is 480V, the voltage at the receptacles for the server power supplies is only 277V.

A 208V system has higher I2R wiring losses than the 480V system. This is because the lower voltage power supplies require more current to deliver the same power. The current required for the 208V power supplies is 277V/208V=1.33 times higher than for the 277V power supplies. Each of the 208V three-phase circuits carries 1.33 x 1.73 = 2.3 times the current because of the current sharing in a three-phase circuit with the power supplies connected phase to phase instead of phase to neutral. Since the losses are proportional to the current squared times the circuit resistance, the losses for a 208V system is up to 5 times greater than for a 480V system. The loss will actually be a bit less since the 208V system will need more lower-resistance wiring, which also adds to the installation cost.

To determine the savings, one first needs to compute the losses for the 208V system. Then, subtract the losses for an equivalent 480V system. This example still uses the same 1,024-node 500KW system used in the previous example. Thirty-two 208V three-phase 50 amp PDUs would be required for each equipment rack in the system. Each 208V three-phase circuit would have a maximum rating of approximately 18KW. The wiring on each phase would be required to handle a total of approximately 40 amps per phase. This would require No. 4 gauge wiring that has a resistance of 0.25 ohms per 1,000 feet. Assuming the step-down transformers are located within 50 feet of the power supplies, the wiring resistance would be 0.0125 ohms per phase. The resistive losses in the 32 circuits would be 1.9 KW.

Loss = Phases x Current2 x Resistance x Number of Circuits
Loss = 3 x 40 x 40 x 0.0125 x 32 = 1.9KW or approximately 0.4 percent

When using 480V power supplies, illustrated in Figure 4, the system still requires 32 power distribution circuits. The number of circuits is determined by the number of servers being supported by the circuit. These circuits are each required to supply a total of approximately 19 amps. If No. 8 wires were used for these connections, the same length of wire would have 0.03 ohms of resistance, resulting in a loss of 1KW or approximately 0.2 percent.

Loss = 3 x 19 x 19 x 0.03 x 32 = 1KW

Since the 480V wiring may need to be routed further to supply the 480V to 208V step-down transformers, the 480V wiring is necessary in both 480V and 208V systems. The additional losses in a 208V system are approximately 1,900 watts. The savings over three years at 12 cents per kilowatt hour are approximately $6,000.

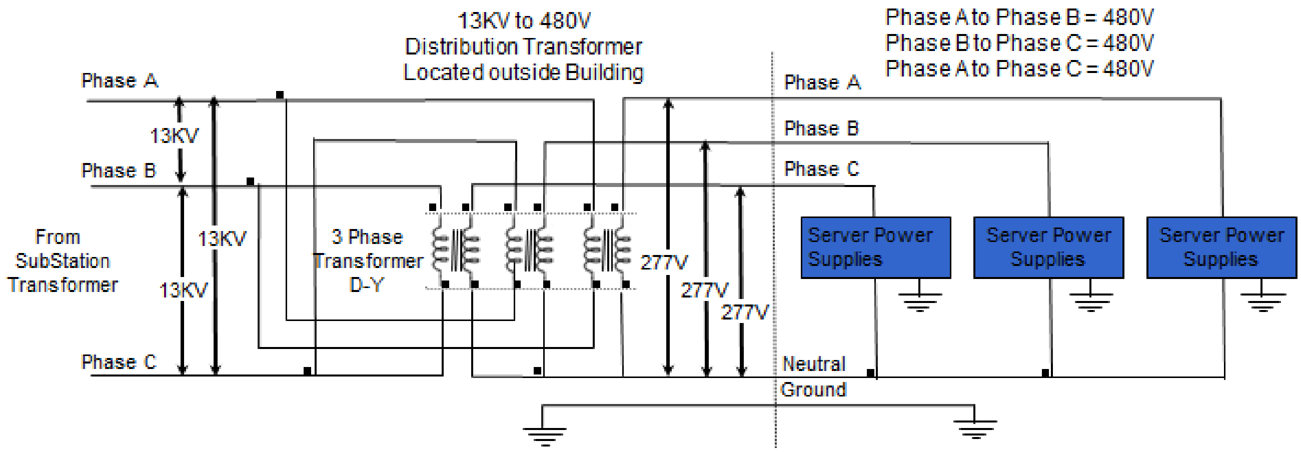


Figure 4: 480V distributed directly to the server power supplies

This makes the total savings from lower resistive losses and eliminating the step-down transformers more than $110,000. Additional savings for the 480V system would also result from the use of less expensive wiring and power distribution units. The savings would be even higher if the wiring runs were longer than the estimated 50 feet.

Power Distribution Examples

Figure 5 illustrates a one-line diagram for a 208V power distribution system with 480V to 208V step-down transformers that provide power to the PDUs and servers. Figure 6 illustrates the same diagram for a system with 277V power supplies that use the 480V circuits directly. The visible difference is the requirement for step-down transformers in the 208V system. The less obvious difference is the increased resistive loss in the 208V system.

Another system that should be mentioned is one with an uninterruptible power supply (UPS), as illustrated in Figure 7. Many UPS systems have 480V inputs and supply 240V to the load. These systems require transformers inside the UPS system to transform 480V to 240V. This is particularly true for bypass mode, where the UPS is bypassed during maintenance. The transformers in a UPS may be of higher quality but still experience the same losses. The best UPS solution is a 480V UPS with 480/277V outputs and power supplies that use the 480/277V outputs directly. These 480/277V power supplies are an integral part of any high-efficiency power distribution system for a datacenter. [3]

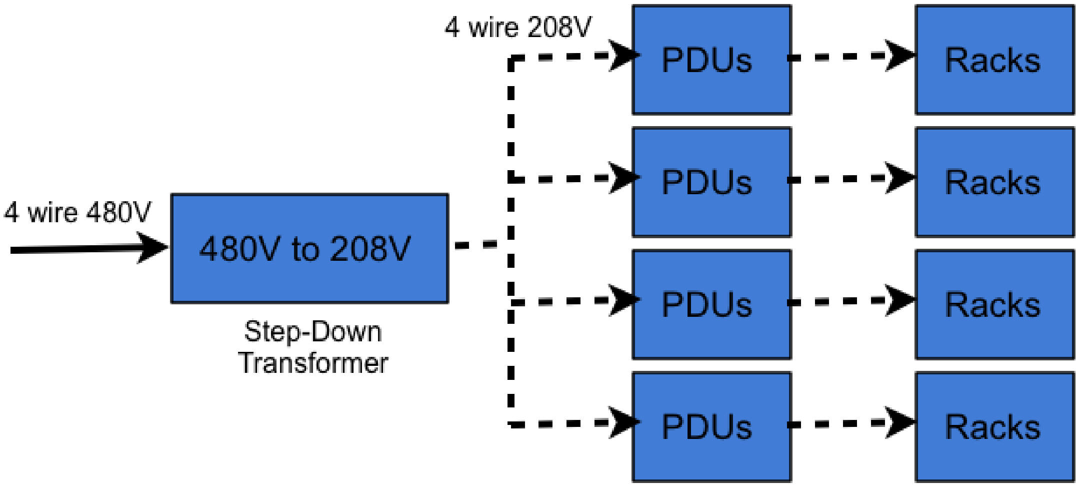


Figure 5: One-line diagram for 208V power distribution

Total cost of ownership is improved by:

1) Elimination of the 480V to 208V three-phase step-down transformers

2) Lower resistive losses because of lower current in the power distribution circuits

3) Lower cost power distribution units

4) Lower copper wiring cost

5) Higher power supply efficiency by operating the power supplies at the appropriate input voltage; power supplies designed for 240V operation are less efficient when operated at 208V.

Negative impacts of using 480V power supplies:

1) Miscellaneous equipment such as data routers, switches and storage still require 120/240V inputs. This will require separate 120/240V PDUs. This has an upside because the power supplies in this equipment will be more efficient operating from 120/120 than on 208V circuits.

2) The 480V systems require better trained personnel and more concern with the higher voltages because of the consequences of shorts or arcing.

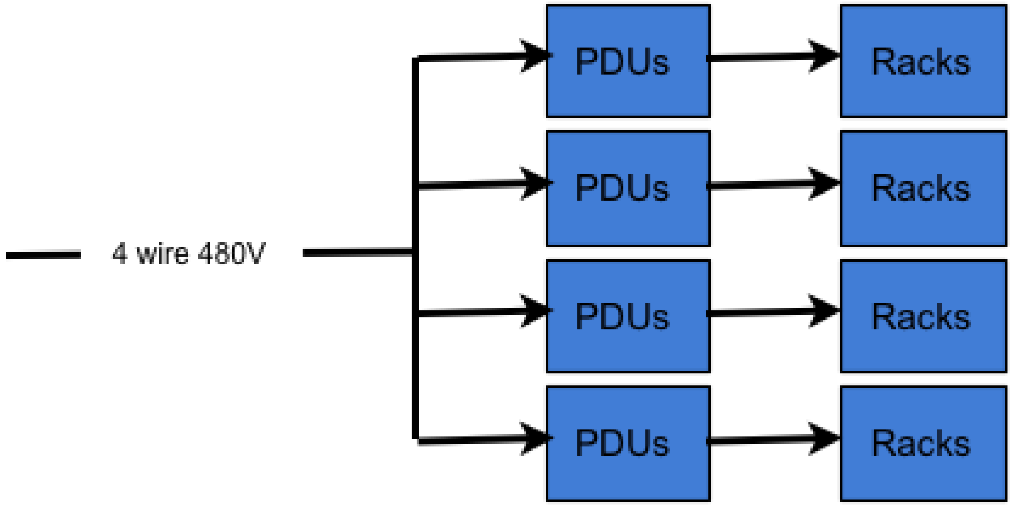


Figure 6: One-line diagram for 480V power distribution

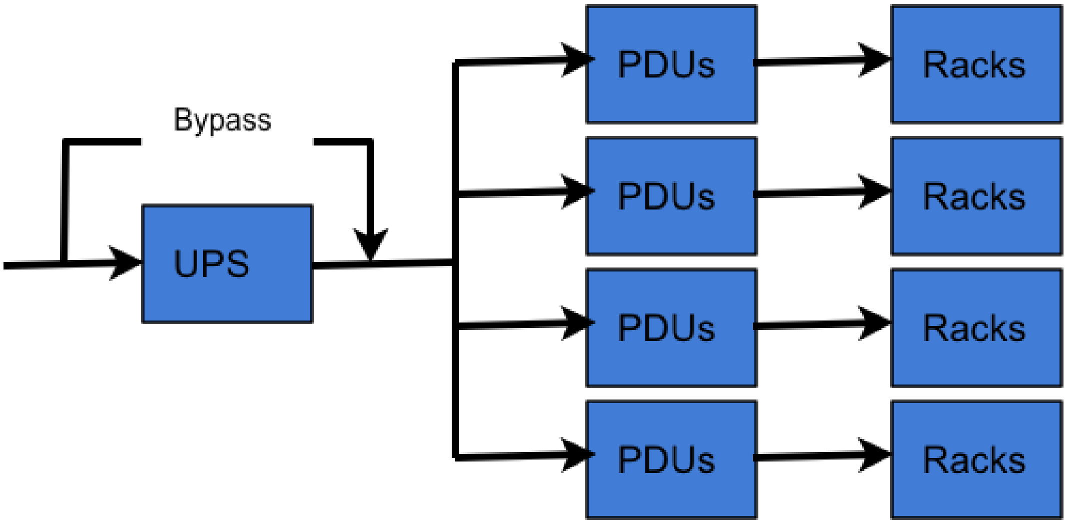


Figure 7: System with uninterruptible power supply

Cray introduces 480V power supplies to be used in the Cray CS300™ cluster supercomputer series. This has been driven by requirements from customers such as U.S. national labs and the U.S. Department of Defense for more efficient power supplies and power distribution systems.

Summary

Many of the servers used in computers on the HPC TOP500® list use 208V power supplies and very inefficient air cooling solutions. Some of these systems have tens of thousands of servers. This represents millions of dollars in wasted electric power and added equipment infrastructure costs. In today's energy-sensitive environment, these practices need to change. The switch to 480/277V power supplies for the next generation of high performance clusters is a small change with large rewards.

References

[1] Qualitative Analysis of Power Distribution Configurations for Data Centers – The Green Grid. Retrieved 2007
http://www.thegreengrid.org/~/media/WhitePapers/TGG_Qualitative_Analysis.pdf?lang=en

[2] Data Center Transformer Types and the K Factor – Search Datacenter
http://searchdatacenter.techtarget.com/tip/Data-center-transformer-types-and-the-K-Factor

[3] The Role of Isolation Transformers in Data Center UPS Systems – APC by Schneider Electric
https://www.insight.com/content/dam/insight/en_US/pdfs/apc/apc-role-of-isolation-transformaers-in-data-center-ups-systems.pdf

Acknowledgements

This material is based upon work supported by Lawrence Livermore National Laboratory.