



Detailed Buyers Reference

Aspen Systems Inc
© Aspen Systems 2009

Table of Contents

<u>Detailed Buyers Reference</u>	1
<u>Applications</u>	3
<u>Budget</u>	4
<u>Processors</u>	5
<u>What are your Processor Options?</u>	5
<u>Interconnects</u>	6
<u>Gigabit Ethernet</u>	6
<u>InfiniBand</u>	8
<u>Myrinet</u>	8
<u>Other 10 Gigabit Ethernet</u>	9
<u>Ethernet Bonding</u>	9
<u>Memory</u>	10
<u>Distributions</u>	11
<u>Aspen Supported Distributions</u>	11
<u>Do you want to use diskless or single image compute nodes?</u>	12
<u>Do you want to use OpenMosix, Rocks, OSCAR, or OpenSSI?</u>	14
<u>Cluster Management</u>	16
<u>Cluster Management Hardware & Software</u>	16
<u>Warranty and Support</u>	21
<u>Warranty and Support</u>	22
<u>Hardware Support</u>	22
<u>Technical Support</u>	23
<u>Cluster Racking</u>	26
<u>Density</u>	30
<u>Cluster Logical Layout</u>	31
<u>Small Cluster</u>	31
<u>Medium Cluster</u>	32
<u>Large Cluster</u>	32
<u>Storage and Database Clusters</u>	33
<u>Visualization Cluster</u>	34
<u>Master Node</u>	36
<u>Storage</u>	39
<u>Internal RAID Systems</u>	39
<u>Hot Spare Disks</u>	40
<u>External RAID Systems</u>	41

Table of Contents

<u>Storage</u>	
<u>Network File System</u>	41
<u>Parallel File Systems</u>	42
<u>GFS</u>	42
<u>GPFS</u>	42
<u>Lustre</u>	43
<u>OCFS</u>	43
<u>PVFS</u>	43
<u>Parallel File System Hardware Requirements</u>	44
<u>Other Options</u>	44
<u>Backup</u>	45
<u>Backup Hardware</u>	45
<u>Backup Software</u>	46
<u>Nodes</u>	47
<u>Power</u>	50
<u>UPS Systems</u>	50
<u>A.C. Power - 120v Circuits</u>	51
<u>A.C. Power - 208v Circuits</u>	54
<u>240v Circuits, 3-phase Supply, or Direct Current (D.C.) Feed</u>	54
<u>International Power Standards</u>	55
<u>What power options to use</u>	55
<u>Plug Types</u>	56
<u>Cooling</u>	59
<u>Small Room Cooling</u>	61
<u>Compute Facility Cooling</u>	64
<u>Commercial Software</u>	66
<u>Distributions</u>	66
<u>Compilers, Utilities, and Debuggers</u>	67
<u>Commercial MPIs</u>	68
<u>Parallel File Systems</u>	68
<u>Commercial Backup Software</u>	68
<u>Schedulers</u>	69
<u>Torque/Maui</u>	70
<u>Moab</u>	70
<u>Sun Grid Engine (SGE)</u>	70
<u>SLURM (Simple Linux Utility for Resource Management)</u>	71
<u>PBS Pro</u>	71
<u>Platform LSF</u>	71
<u>Account Management</u>	72

Table of Contents

<u>Security</u>	73
<u>Windows Integration</u>	75
<u>Shipping and Delivery</u>	76
<u>On-Site and Training</u>	78
<u>Next Steps</u>	79

Detailed Buyers Reference

Specifying and purchasing a High Performance Computing (HPC) cluster that meets your needs can be time consuming and labor intensive. The Aspen Systems Detailed Buyers Reference helps to explain the options you have for your new cluster and some of the choices you will need to make.

Like most HPC vendors Aspen offers a standardized build and package selection that follows HPC best practices. However, unlike some other HPC vendors, *we also offer you the opportunity to customize your cluster hardware and software, with options and capabilities tuned to your specific needs and your environment.*

This is a more complex process than simply providing you a “canned” cluster, which might or might not best fit your needs. Many customers value us for our flexibility and engineering expertise, coming back again and again for upgrades to existing clusters, or new clusters which mirror their current optimized solutions. Other customers value our standard cluster configuration to serve their HPC computing needs, and purchase that option from us again and again.

To help with this process, Aspen provides you with four tools that can help you at different periods in your configuration and procurement.

- **Buyers Guide (PDF) (One Online Page)** – The Buyers Guide is a tool that can help you summarize your requirements, functioning as a quick reference or memory aid for the choices you need to think about for your new HPC system. Even advanced users find the Buyers Guide helpful as a quick reminder of their configuration choices, and your sales engineer may have you look at the Buyers Guide on-line while he or she discusses your options and requirements. Every item in the Buyers Guide is linked to the appropriate location in the Detailed Buyers Reference so that you may easily access more detailed information about your choices as well as our specific recommendations for certain configurations.
- **Detailed Buyers Reference** (*this document*) – Use this document to investigate different technologies and configuration options. Our web site links to specific sections of this document to explain specific configuration choices, or you can read it on-line or download it to peruse at your convenience. Reading the entire document may provide useful information, or perhaps combine some of your current knowledge in ways that will help you better understand your needs and how Aspen HPC solutions can be used to meet them.
- **Configuration Guide** – The Configuration Guide is a detailed explanation of specific information we will need to know to actually build your system when you decide to purchase. Each section of the

Configuration Guide corresponds to a section of the Statement of Work. Your Aspen sales engineer will provide you the Configuration Guide and a Statement of Work with their first quotes to you.

- **Statement of Work** – The Statement of Work is used to record your specific requirements and configuration choices for your purchase, and Aspen will use your completed Statement of Work and your quotes to perform a final engineering review of your system before it is built. We ask that you provide us a completed Statement of Work before or with your Purchase Order so that we may complete your final engineering review as expeditiously as possible.

Applications

This may be one of the most important questions you can answer when configuring a new cluster. Your application needs can drive many of your configuration choices while obviating others, so it is critical to research your application requirements.

User groups and software vendors often publish performance results for different architectures, compilers, and interconnects running their codes. This can be quite valuable information, as different combinations of chip architecture, memory and disk configuration, interconnect, and software environments can yield quite different results. Sometimes community or in-house developed code(s) are developed primarily on a single architecture, compiler, Message Passing Interface (MPI) implementation, or interconnect, and exhibit erratic behavior when ran under a different configuration.

Your Aspen Systems Sales Engineer can help you to determine what configurations we know to function properly with your applications. We also highly recommend that you take advantage of our benchmark clusters in order to determine real-world performance differences. Benchmarking your code(s) on an Aspen Systems benchmarking cluster can quickly show what combination of architecture, compiler, MPI implementation (if needed), and interconnect works best for your problem.

Budget

What is your budget?

There may be budget trade-offs we, *and you*, need to make to your systems design based on the budget you have available for your new purchase. Your sales engineer will work with you to prioritize your needs and examine these trade-offs, because your budget will directly affect what options you are able to afford.

The best technical solution for your particular application may not fit within your budget. Your Aspen sales engineer will work with you to design the best possible solution for your requirements that fits within the budget you give us. Many design decisions, such as your choice of specific file systems, storage solutions, memory per core, processor type and speed, and even chassis form factors can dramatically affect the final solution price.

There are certain features you should *never* contemplate doing without. One big area to watch carefully is storage reliability. On smaller cluster designs, customers often consider removing hot spare disks from the design to save several hundred dollars. Perhaps that money can be used to good effect for more processing power or another node, but how valuable is your data? Is it better to have more processing power and have the cluster down or your data lost because you didn't spend that money for extra data reliability? No, it isn't, because your data output is the entire reason you purchased your solution.

As a general rule, lower your processing power goals, *not* your reliability goals when faced with budget crises.

How about that whatchamacallit that does the newfangled other thing that you just read about on HPC Wire that you really really want? Hey, we're gear heads too! We *love* new stuff. But please keep in mind that technology is almost always more expensive during its initial adoption phases, and tends to become much cheaper later in its market life. Please concentrate on getting a solid cluster design that has the reliability and basic features you need before adding additional "nice to have" requirements that can adversely impact your budget. Your sales engineer can discuss these decisions and outline some of the options you need to think about as they help you specify your system design.

Processors

What are your Processor Options?

Your applications again play a critical role in your processor selection. Many times, your code(s) will execute faster on one processor architecture vs. another, and your user community may have specific opinions about which processor is better for your particular code(s). Use your Aspen Systems sales engineer as a resource for this decision. We build many clusters every year, and often will have worked with other customers to define the optimum configuration for particular codes or applications.

Current offerings from AMD and Intel are both extremely competent processor architectures, while new models in the pipeline will offer even more capabilities. Historically, better integer performance has been achieved by higher clock speeds, which Intel offers, and AMD has had better memory bandwidth leading to better floating point performance. These generalizations are almost invalid, however, as new processors with new and better capabilities are introduced regularly. Trade-offs such as power usage and acquisition cost can also play a big factor in your selection.

Processor architectures change rapidly, with new models and capabilities appearing almost quarterly. Your Aspen Systems Sales Engineer can be one of your best resources to determine what the latest technology is, and how best to configure the technology for best price/performance in your particular circumstances. Use them, and our benchmarking clusters, to help make your decision easier.

Interconnects

Many HPC applications and codes are programmed to take advantage of parallel computation, and utilize the Message Passing Interface (MPI) Applications Programming Interface (API) as their primary method of communications between processes running on the same or different nodes in your cluster. Other applications are serial in nature, with an entire computational job running as a single process. These serial applications can have specific data or processing requirements that mandate a high speed Interconnect as well.

The most basic HPC cluster will utilize a single Gigabit Ethernet network for administrative traffic, data sharing (NFS or other protocols) and MPI or applications processing traffic. If your applications are bandwidth or latency sensitive, using only a single Gigabit Ethernet for your cluster network is perhaps the least desirable of your choices.

Often, an HPC cluster will be configured with two (2) internal networks. The first, a Gigabit Ethernet network connected to an on-board Network Interface Card (NIC) on each node and identical to the single Ethernet network used by the basic HPC cluster, is used for scheduling, node maintenance, basic logins, and perhaps data sharing, while the 2nd internal network is dedicated to computational traffic. This configuration ensures that critical computational traffic is not hampered by other traffic (which is normally much less bandwidth or latency sensitive).

The 2nd internal cluster dedicated network could also be Gigabit Ethernet as well. This option has one advantage, cost. Most systems used for HPC processing have two on-board Gigabit Ethernet interfaces, so the cost of an additional, dedicated Gigabit Ethernet network is only additional switch ports and cabling. For MPI based code(s), Aspen recommends a second internal Gigabit Ethernet network dedicated to your MPI or applications processing traffic as a minimum.

Gigabit Ethernet

Gigabit Ethernet provides full duplex communications at 1Gb/s (or 1000Mb/s) and latencies ranging from ~40us to ~300us. However, many codes require higher bandwidth or lower latency than standard Gigabit

Ethernet interfaces and switches can provide to operate efficiently. Both InfiniBand and Myrinet technologies are commonly used in these cases.

InfiniBand

InfiniBand is a switched fabric link topology network which utilizes Host Channel Adapters (HCAs) installed in each node of the cluster to communicate. InfiniBand is made up of 2.5 Gb/s Lanes which are used in parallel to communicate between nodes.

A Single Data Rate (SDR) 4x (4 Lane) InfiniBand connection provides 10 Gb/s (or 10,000Mb/s) raw full duplex bandwidth, with 8 Gb/s usable to processes. An InfiniBand Dual Data Rate (DDR) 4x connection provides 20Gb/s (20,000Mb/s) raw full duplex bandwidth, with 16 Gb/s usable to processes. InfiniBand latencies range from ~1us to ~5us, depending on the HCA and switch topology used, and relatively large non-blocking fabrics can be constructed. InfiniBand also supports other protocols to facilitate capabilities such as access to remote memory, sockets, and storage.

An InfiniBand HPC network is commonly implemented using CX-4 copper cables, which are thicker than a standard IEEE Cat 5 cable, and there are length limitations. Maximum cable length using CX-4 cables is 15 meters for 4x SDR, and 10 meters for 4x DDR. Fiber optic cable options for InfiniBand networks do exist, but are quite expensive. Aspen recommends using the Open Fabrics Enterprise Distribution (OFED) InfiniBand stack on InfiniBand clusters unless your code(s) or application(s) are not supported.

The InfiniBand specification is supported by multiple vendor implementations, and current information on vendor implementation and MPI selection is necessary to determine support for and performance of your code(s) on any given implementation. For instance VASP, a molecular dynamics package from the University of Vienna, currently seems to run best on InfiniBand with Intel compilers (9.1, specifically), using Open MPI version 1.2.6 and the Intel Math Kernel Library on OFED version 1.3.1 using fftw version 3.1.2. Over a hundred combinations of different compilers, MPIs, and utilities were tested to arrive at this selection. As shown in the examples above, some codes can be quite complex to build for best performance.

Many gateway solutions also exist that can connect your clusters InfiniBand directly to other Enterprise networks if needed.

Myrinet

Myrinet is a high speed interconnect supplied by Myricom, an HPC interconnect company. Myricom originally manufactured a 2 Gb/s technology (Myri-2G), which was arguably the most widely deployed low latency clustering technology of its time. Myricom now provides Myri-10G solutions, which combine Myricoms Myrinet (MX) capabilities with near wire-speed 10 Gigabit Ethernet. Myri-10G NICs include

processors and firmware to offload network protocol processing, lower node CPU utilization, and provide communications paths that bypass the host kernel. Myri-10G also supports fiber optic cables with a maximum cable length of 85 to 200 meters (depending on protocol used), and can provide ~2.3us MPI latency at 9.6Gb/s(9,600 Mb/s) raw full duplex bandwidth. Myrinet switches are used inside the Myrinet network and software encapsulation is used at the node to utilize 10 Gigabit Ethernet protocols if the node is configured to utilize MX. The ability to mix and match 10 Gigabit Ethernet and Myrinet protocol on the same network is a major advantage of Myrinet technology.

Other 10 Gigabit Ethernet

10 Gigabit Ethernet networking (other than from Myricom) can also be used to interconnect your HPC cluster, however the latency incurred by the protocols and switches does not currently lend itself well to the requirements of most MPI codes, and the price/performance ratio can be high. Low Latency switches are available, and we have had some success with low latency drivers such as Open-MX on commodity 1 and 10 Gigabit Ethernet. Contact Aspen for more information if you are interested in this type of solution.

Ethernet Bonding

Some clusters utilize Ethernet bonding, which bonds two Gigabit Ethernet interfaces together to provide more bandwidth than a single Gigabit Ethernet interface can provide. Your switch must provide this capability, and your distribution must support the capability, which most do. This is an efficient method to provide additional bandwidth for the higher data transfer requirements some applications exhibit. Bonding, however, does not help, and can interfere with, your MPI implementations. If your applications utilize MPI and you do not intend to configure a high speed interconnect, Aspen recommends that you do not utilize channel bonding on your Gigabit Ethernets, but instead utilize two dedicated networks, one for storage and administration, and one dedicated to your application MPI traffic.

It is difficult to say with any certainty which Interconnect will give your cluster the best price/performance ratio without knowing your specific code(s), situation, and requirements. Low latency interconnects can add significant per node cost to your cluster. Your Aspen Systems Sales Engineer can work with you to determine your requirements, and customize a solution that will serve you well. We also provide benchmarking across the different interconnects so that you can see the differences and implementation specifics of these Interconnects with your code(s). Ask your Aspen Sales Engineer about accessing our benchmarking clusters. We highly recommend that you benchmark your code to determine your best configuration if you are at all unsure of your selection choices or which Interconnect will serve you best.

Memory

The amount of memory needed per node is also based on your application requirements. Some applications are very memory intensive and require quite large amounts of memory per processing core, which other applications can run comfortably in 512 MB memory per core. Aspen recommends that a general use HPC cluster be configured with a minimum of 1 GB memory per core, which comfortably accommodates most HPC applications. Consult your vendor, user group, or your Aspen Systems Sales engineer for specific information about the applications you wish to run on your new cluster.

To control costs, you may also configure one or several nodes with a larger amount of memory per core to be used for those applications that require it, and configure the remaining nodes with a lesser amount of memory. Using your scheduling system or other methods, you can direct codes which require larger memory footprints to those few nodes that are equipped appropriately, and route applications with smaller memory requirements to other nodes.

Front-end nodes on your cluster are nodes which users log into to compile codes (if needed), run their applications, or perform other memory intensive operations. Larger numbers of users on the cluster can expand the memory requirements on the interactive node(s). Aspen recommends a minimum of 2 GB per core on multi-user front-end and interactive nodes.

Distributions

Unlike many cluster vendors who offer only one particular distribution for their cluster offerings, Aspen Systems supports several different distributions. You may have a site support contract for a particular commercial distribution, or perhaps your systems administrators are more familiar with one specific distribution. Those are valid reasons to select a particular distribution. However, each distribution may introduce limitations in the capabilities of your cluster. Perhaps your unmodified distribution of choice doesn't support the hardware you have selected for your cluster, or a particular application has not been ported to your distribution. As always, consult your software vendor or user group for supported distributions and versions and talk to your Aspen Sales Engineer about our offerings. Aspen Systems recommends the CentOS distribution for most cluster uses due to its wide use in clustering, long support life cycle, wide user base, and the large number of HPC utilities it provides.

Aspen Supported Distributions

- [RedHat Enterprise Server](#)
- [SUSE Linux Enterprise Server](#)
- [Centos](#)
- Centos Stateless (NFS root, diskless or hybrid compute nodes)
- [Fedora](#)
- [OpenSUSE](#)
- [Warewulf / Perceus](#)
- [Scientific Linux](#)
- Other

Some distributions, such as those targeted more toward general user desktops, are not well supported in the HPC community. If you can't easily find or read about another cluster running your code(s) on a given distribution, that may indicate that the distribution is unsuitable for your purposes. Aspen can integrate other distributions than those shown above, but another distribution may not have all of the functionality we normally provide. Should you select an unsupported distribution, Aspen will not guarantee that all the capabilities outlined in our standard build will work.

The update period, or patch life cycle, is extremely important for Linux distributions deployed in an Enterprise environment, but is less of an issue with clusters. A cluster does not normally need to be managed as you would the same number of individual enterprise nodes. Only the master node and any specialty nodes are treated as unique nodes for update purposes. The compute nodes are normally treated as a single upgrade target, and Aspen provides tools on our default distributions to help you do that. Many HPC users stabilize their system(s) on a distribution and optimized code-base which changes very little throughout the life cycle of the system.

Do you want to use diskless or single image compute nodes?

You may also choose a single image cluster, where only a single image is kept on the master node of the cluster, and all compute nodes network boot that image. Aspen Systems supports the Perceus / Warewulf clustering tool kit installed on CentOS or RedHat Enterprise, as well as stateless boot (NFS root) on RedHat derived distributions. Some training is necessary to utilize a single image system, and there can be limitations to your configuration flexibility in return for scalability.

With Perceus / Warewulf, no operating system is installed on the disks in your compute nodes. In this configuration, all user data space is contained on the master node or a storage node, and network mounted to the compute nodes. Perceus / Warewulf can be configured with no hard drives at all in the compute nodes (diskless) , or with local disks in the compute nodes (hybrid). When no node disks are installed, you must ensure that all your codes are memory disciplined, and can execute in the physical memory you have installed in the compute nodes with no swap space needed. Some physical memory is taken by the running operating system in a diskless compute node as well.

If local disks are used in a hybrid configuration, they are normally configured with swap and scratch space on the node hard drive. Some applications, Gaussian for instance, need and use local scratch space on each node to speed up calculations and reduce network traffic. Some of the advantages of a single-image system are configuration consistency and ease of expansion. Any change you make in the node image environment on the master can be deployed to your entire cluster by a simple reboot of all your nodes. You have to learn a few more commands to work in the node image environment, but the commands are relatively easy.

Adding additional nodes of the exact same hardware configuration is a simple matter of installing the new hardware, connecting the new nodes to your network, and booting them in the order you want them to be identified as additional nodes. Adding additional nodes with different motherboards and/or different processors, not an uncommon occurrence in cluster upgrades, will almost always require additional and possibly extensive modifications to the node image.

One of the advantages of a diskless compute node cluster is higher reliability due to the lack of disks in the compute nodes. Approximately 50% of all node failures are caused by failing hard drives. Configuring your cluster as a hybrid cluster, which has local disks used for local scratch, required by many applications, eliminates this advantage.

Some of the disadvantages of a single-image system is a certain lack of flexibility, some configuration complexity, and a slightly less user friendly and more advanced user experience. Perceus / Warewulf clusters are always configured in certain ways, naming is preset, and it is critical that the master node “own” the internal network used by the compute nodes, meaning that only cluster internal nodes should be connected to that network. It might be difficult to add specific applications or utilities in the node images, and some knowledge of kernel modules and boot sequences becomes necessary in more customized environments. Documentation is geared toward the more advanced cluster user, sometimes making it difficult to troubleshoot problems. Aspen clusters are equipped with command line utilities, or a GUI if the Aspen Beowulf Cluster Management System (ABC) is purchased, that allow you to copy one disked node then quickly re-image all other disked nodes with that copied image. These utilities are installed on all our default distribution selections, and work exactly the same across all environments. This option gives you configuration simplicity, ease of customization, and ease of expansion while using more traditional Linux administration skills you may already have. For many users, this option is more cost efficient and productive than deploying a single image cluster.

Only procure a diskless single image cluster if:

- you intend to scale to a large number of nodes on this cluster
- you know that your code(s) will easily reside in physical memory w/o accessing swap space
- you have few or no site specific or unusual requirements for the configuration of the cluster
- you have no current or planned applications that need to use node local scratch space
- you have or intend to have slightly more advanced Linux HPC administration skills in your organization

Only use a hybrid single image cluster if:

- you intend to scale to a large number of nodes on this cluster
- you have few or no site specific or unusual requirements for the configuration of the cluster
- you have or intend to have slightly more advanced Linux HPC administration skills in your organization

Use a disked cluster if:

- you have site specific or unusual requirements for the configuration of your cluster which might require special host naming, network configurations, or unique node configurations.
- you do not have advanced Linux HPC skills in your organization, or you intend to contract Aspen to administer your cluster

Do you want to use OpenMosix, Rocks, OSCAR, or OpenSSI?

The OpenMosix project ended on March 8th, 2008. OpenSSI supports older versions of Fedora, Debian, and RedHat 9, which may not have device drivers for newer hardware in today's clusters. OpenSSI plans to add 64 bit support soon, and plans to support CentOS and Red Hat Enterprise Linux 5 in the future, check the OpenSSI web site for up to date information.

OSCAR supports Fedora Core 4 and 5, RedHat Enterprise Linux 4, CentOS 4. The latest release was on November 12, 2006, although Version 5.2 is now in alpha and ported to the debian distribution. Check the OSCAR website for up to date information.

The Rocks Cluster Distribution (originally called NPACI Rocks) is a popular open-source Linux cluster distribution based on CentOS, and sponsored by an National Science Foundation award. Rocks is a diskless cluster deployment and management solution, and utilizes the concept of "rolls", which are pre-configured sets of RedHat Package Manager (RPM) packages with specific changes made to integrate into a Rocks cluster. The Rocks goal is to simplify building a cluster, and it succeeds. However, Rocks, much like Perceus / Warewulf, makes specific assumptions about how your cluster will be configured, and your cluster will be configured in that manner if it is to operate properly. Additionally rolls released by vendors or user groups, may be valid for only certain Rocks versions, and some rolls can conflict with other rolls, so some knowledge is necessary to successfully build and deploy a Rocks solution that fits your needs. As with all in-progress development efforts, bugs exist. Newer hardware and driver requirements can also require customization of the Rocks images, and make deployment of older Rocks versions, perhaps necessitated by roll compatibility, more difficult on the latest hardware configurations. There are two very good reasons to select Rocks.

1. First, you may belong to a specific user community which has standardized on Rocks. For instance, the Rocks "bio" roll contains a suite of bio-informatics applications most commonly in use by the bio-informatics community, such as MpiBLAST, Emboss, Glimmer, HMMER, and NCBI BLAST. If your community routinely uses Rocks to satisfy its HPC needs, then you will most likely have already heard of Rocks clusters being used with your applications when you researched your applications requirements.
2. Secondly, your organization may have standardized on Rocks, and already have specific administration experience with it.

Rocks clusters, as any deployed cluster management solution, rely on standardization, and some customizations may be very difficult or failure oriented. For instance, the standard solution to renaming your cluster master node host name is to re-install the cluster from scratch. Rocks requires very specific directory

structures, and almost all rolls are configured in standard ways that may or may not meet your approval.

Specific information, such as the permanent IP address and fully qualified domain name of the master node, is necessary to know before we start building your Rocks cluster. If you require a configuration that is unusual to Rocks, or have unique organization or site-specific requirements, we may charge you extra to implement those customizations based on the complexity of your request. Some vendor Interconnects and specific HPC utilities may not be supported, and we, and you, will need to carefully research your roll selection, paying special attention to the version of Rocks your roll selections support. Speak to your Aspen Sales engineer for more information.

Some of the advantages to Rocks is its ability to quickly add additional nodes of the same hardware configuration to your cluster or re-image existing nodes, and the availability of packaged HPC applications (rolls) for specific user communities.

Rocks images nodes via RPM packages and kick start scripts, so any node customizations must be scripted in order to be present in any new image. The Aspen utilities utilize an actual node image, which contains all the customization that had been done to that node previously, changing only the IP address and host name. This means that a standard disked image cluster using Aspen utilities is more easily customizable than Rocks and just as scalable.

Cluster Management

The type of cluster you operate can drive different cluster management and support requirements.

A “lights out” cluster is one that is installed in a remote or co-location environment, with limited physical access by you or your administrators. Lights out clusters need remote power management, remote Keyboard Video Mouse (KVM) or serial port access, and automated tools to monitor and alert when problems are encountered.

“In house” clusters are located in areas that are easily accessible by you or your administrators. While full remote capabilities may not be as critical in an in house cluster, they may be desirable.

Cluster management and support is perhaps one of the most overlooked facets of operating a cluster. Two questions must be answered for your successful cluster deployment. What hardware and software capabilities will be installed on your cluster to facilitate successful management and support, and what are your cluster management, warranty, and support options?

Cluster Management Hardware & Software

Intelligent Platform Management Interface

Aspen highly recommends that you configure Intelligent Platform Management Interface (IPMI) on your cluster nodes. IPMI is a specification for a set of common interfaces for system administrators to monitor and manage any system. IPMI operates independently of the operating system, using a Baseboard Management Controller (BMC) on each node. A BMC is a small solid state card with a network interface that plugs into the

node motherboard and is powered by the node power supply. The node does not have to be operational or booted for this access to work.

The IPMI BMC provides remote access to serial ports (Serial over LAN, or SOL) or even the actual video display on the node (KVM over LAN), allowing you to use a web browser or IPMI client to troubleshoot boot problems, diagnose hardware faults, or even modify BIOS settings.

IPMI is also used to remotely power on or off the node, and to retrieve sensor values from the motherboard such as voltages, temperatures, and fan speeds.

IPMI interfaces utilize an Ethernet interface for communications with the node. On some nodes, the IPMI BMC can be “vampire” to the primary Ethernet interface, so that no additional cabling is needed. The Ethernet interface can still be used by the node for communications as well. Almost all nodes can be configured with IPMI “3rd LAN” interfaces, which are a separate Ethernet interface dedicated to IPMI communications. Utilizing vampire IPMI interfaces can cut cost, while utilizing a 3rd LAN interface allows the IPMI network to be separated from the operational network for security or traffic purposes.

IPMI KVM over LAN interfaces can be used instead of or in conjunction with local KVM console capability on the cluster. The cost of a local KVM connection to each node is roughly equivalent to the cost of an IPMI BMC with KVM over LAN capabilities, and IPMI provides remote power and sensor capabilities to the node in addition to the console capability that a KVM solution provides.

Many customers think that the most cost effective solution for remote (via network) and local (at the cluster) management is to configure IPMI for their cluster, then install a single 1U integrated video console (TFT) unit in the cluster which is attached to a master or administrative node with a long cable. In normal operations, the console unit remains attached to the master or administrative node, and a web browser on that node can be used to access every nodes console via node IPMI interfaces. If desired for convenience or on-site work, the console cable can be reattached directly to any node in the cluster and used on that specific node for a time. The master(s) and administrative nodes are normally configured with IPMI “3rd LAN” interfaces, and an additional Ethernet connection from your organization is ran to those interfaces to allow remote console and power control from your organizational network should a fault situation occur.

Aspen Beowulf Cluster Management System

The Aspen Beowulf Cluster (ABC) Management System is a commercial Aspen web based application suite that you can purchase and use to monitor and manage your Aspen cluster. ABC requires IPMI interfaces on all nodes in the cluster, and uses them to present all your cluster management tools in a single secure web browser connection that you can access from anywhere you allow.

Using ABC, you can remotely clone and install nodes, attach to any nodes video screen (if KVM over LAN IPMI is purchased) or serial port, attach to any nodes via ssh, monitor all cluster hardware, upgrade software packages, define alarm thresholds for every monitored item, submit, review and manage scheduled jobs, and set up alerts for different fault conditions.

ABC is especially valuable to the beginning cluster user. It transforms a sometimes complex set of software tools into a converged, homogeneous environment that does not require the in-depth knowledge normally needed to operate, manage, upgrade, or maintain a cluster.

Every user on your cluster automatically has an ABC account. ABC can also provide a web portal for submission of scheduler jobs for your cluster users if torque or Sun Grid Engine schedulers are used. Only certain users are administrators by default (root), but any user account can be configured as an administrator in order to configure ABC and maintain devices and nodes in the cluster.

Using the ABC “Tools” proxy web service, RAID systems, Ethernet, Myrinet and InfiniBand switches and other peripherals which run web servers for their management interface can be displayed through the ABC UI even though they do not have an externally reachable IP. Access to these sensitive configuration interfaces is normally limited to administrators.

One of the many strengths of ABC is the ability to quickly copy any node and use that copy to add a new node or re-image an existing node or all nodes in the cluster. This makes major node upgrades, maintenance, or node recovery extremely quick and easy. Aspen also provides command line tools on our clusters for imaging, remote power, and sensor programs. These are often used by more advanced cluster users to quickly check status on nodes, remotely power them on or off, or to re-image large groups of nodes.

Ganglia

Aspen normally also installs and configures Ganglia on your cluster, and can make Ganglia available as a Tools menu option inside ABC, or externally available as a default web page for organizations who are used to seeing Ganglia as the front end web page for their clusters. Ganglia is a quite popular scalable distributed monitoring system for clusters and grids, and many HPC customers do not consider a cluster complete

without it. Aspen will turn Ganglia on or off on your cluster based on your Statement of Work selection.

Switched and Metered Power Distribution Unit (PDU) Options

Your cluster can also be configured with remotely switchable PDUs, which can turn power off or cycle power to any system connected to it. Prior to the wide adoption of IPMI, any cluster which needed the capability to remotely power on and off nodes required these PDUs, and they are still often used on peripherals which do not support IPMI. If you wish to monitor power consumption on circuits, you may configure either switched or metered PDUs for power connections. Metered PDUs can be polled to determine total power consumption on a circuit, but cannot be used to power off an attached system as switched PDUs can. Any Un-Interruptible Power Supply (UPS) system(s) installed in your cluster can also be remotely polled for circuit power consumption.

Serial Console Servers

Serial console switches or servers can be configured and installed in your cluster, which provide serial console access, and logging of any console events in a central location. In this case nodes are configured identically to IPMI Serial Over LAN (SOL) equipped nodes, with BIOS, the boot loader, and the operating system redirected to a serial port on the node. ABC supports these serial consoles, or Aspen can install the “ConMan” console manager for console logging and command line connections if your organization prefers.

KVM (Keyboard, Video, Mouse) Switches

Aspen can configure your cluster with a KVM system which can be connected to some or all nodes on your cluster. This will allow a user located at the cluster to utilize a local console unit to “hot key” switch between all connected nodes video consoles for maintenance purposes. If you desire remote console connectivity to your cluster KVM, an additional remote unit can be installed which can be remotely accessed by a web browser just as IPMI interfaces are, allowing you to remotely access your KVM, then hot key between displays to different hosts in your cluster. Normally, the remote KVM unit is connected to your organization Ethernet and IP addressed within that space, not your cluster internal IP space, to allow remote access from administrators on the organizational network.

[<< Previous](#) | [Next >>](#)

Warranty and Support

Warranty and Support

Cluster Warranty, Management, and Support Options

The hardware and software options you select for your cluster should support the cluster management model you intend to use. If you have administrators on-site with the cluster, some of the hardware or software outlined in the last section may be redundant. However, if your administrators sometimes work from home or other locations and have remote access to your cluster, the remote control and access capabilities may be very useful.

Aspen Systems provides warranty and technical hardware and software support for any software or systems you purchase from Aspen. Support can be obtained by calling Aspen at (800)992-9242 Monday through Thursday, 8:30 A.M. to 5:30 P.M. Mountain Standard Time, or 8:30 A.M. To 5:00 P.M. Mountain Standard Time Friday. You can also get support by e-mailing support@aspsys.com. When you e-mail, a support ticket is automatically opened, and a return email is sent back to you with an automatically assigned ticket number from the Aspen trouble ticketing system. All engineers at Aspen see all support tickets, so there is no need to mail to any particular engineer. In fact, Aspen recommends that you always use the support@aspsys.com email address even if you have any engineers specific e-mail address. Your particular engineer may not be available or may be working another case, and we want to help you as quickly as possible.

Hardware Support

Aspen provides Bronze (1 year), Silver (2 years), Gold (3 year), Platinum (4 year), and Diamond (5 year) hardware warranties on systems purchased from Aspen.

We also offer advance replacement parts for customers with good credit standing. You must obtain a Return Material Authorization (RMA) number from Aspen to get an advance replacement. Aspen pays for standard ground shipping both ways when an RMA is processed, shipping the new part to your organization with an RMA and return shipping label already included.

Aspen can also process an RMA for you to return an entire node or peripheral to Aspen for repair or replacement should that be needed. Aspen pays for standard ground shipping both ways in this case as well, and this service is available for the term of your hardware warranty.

Aspen also offers on-site hardware support through select hardware maintenance partners. We can tailor your on-site hardware maintenance coverage days, hours, and response times to fit your budget and needs. Aspen recommends the use of on-site spares kits for all organizations who opt for on-site support. Talk to your sales engineer for more information.

Technical Support

When you purchase a cluster or system from Aspen, you get software and hardware technical support as well. Aspen supports the hardware, operating system (O.S.), and applications that were installed on your system(s) or cluster(s) at time of delivery. Aspen software support covers fault correction and minor revision O.S. and application upgrades to the delivered configuration. Support for software installation and O.S. upgrades that were not originally contracted for is available, but will result in additional charges.

As an example, let's say you purchased a cluster from Aspen that has RHEL 5.0, InfiniBand, and runs the NAMD and Charm applications. A year after the cluster is installed, you wish to upgrade to RHEL 5.2. That will involve upgrading the operating systems on all nodes, rebuilding or upgrading the InfiniBand drivers, rebuilding any associated MPI implementations and perhaps some other basic utilities, rebuilding NAMD and Charm, and performing regression testing to ensure functionality. This upgrade would be covered in your Aspen software support warranty, and we would help you perform this upgrade in one of several ways.

If Aspen has remote access to the cluster, we can log in and perform the upgrades on specific nodes, test the upgrade, then re-image the rest of the cluster based on those upgraded nodes. If no remote access is allowed, it becomes much more difficult. We would have you take an image of the current cluster or use the image we took of your system prior to shipping, emulate your hardware environment at an Aspen facility, perform the upgrades, then transfer the modified images so that you could re-image the cluster with the new software. This approach is more time consuming, as we might encounter delays allocating hardware resources, or have to assemble matching systems to perform your upgrades.

Now, let's say that 2 years have passed, and you wish to upgrade your cluster from your installed RHEL 5.0 to the now current RHEL release, say, RHEL 6.3. Many things such as hardware driver implementations, base system development libraries, or even basic utility functionality can change between distribution major revisions. This upgrade (from RHEL 5.0 to RHEL 6.3) would not be covered by the base Aspen's software support warranty, and would result in additional charges. Aspen supports upgrades within the major revision level of the distribution you have selected free of charge, but will charge you for distribution upgrades that

cross a major revision level.

Many customers use Aspen technical support as their technical backup, contacting Aspen when problems occur that they do not know how to solve. Routine administration, such as adding users or performing system backups, are not included in the standard Aspen software support offering. However, Aspen can provide additional support above and beyond the standard support options you get with the cluster purchase. Some of the options Aspen offers are;

- **Blocks of Support Time**

These are 1, 5, 10, 15, and 25 hour blocks of time that you may pre-purchase, then use only as you need them, to have an Aspen engineer complete any cluster administration or upgrade tasks you wish

- **On-Site Visit**

On-site visits are normally used to accomplish a specific task, or perhaps to perform additional informal on-site training for users or administrators after or at the same time your cluster is delivered.

- **Full Support Contract**

Aspen can support your entire cluster administration needs as well. If you do not have an on-site administrator, and do not wish to support your cluster yourself, Aspen offers full cluster administration and support contracts. These will normally include a on-site hardware maintenance contract as well, and require Aspen remote access to your cluster.

In all cases, Aspen recommends that you allow remote access to your cluster from Aspen support engineers if possible or allowed. While this is not possible in all situations due to security or organizational requirements, remote access will make your life easier in the event of a problem occurring, and allow your Aspen support engineers to work directly with you on problems to quickly solve them.

Aspen can provide a VPN (Virtual Private Network) client on your cluster that connects back to dedicated Aspen support systems after the cluster is installed and operational. This VPN client can be turned on or off at will by you to permit Aspen access only when you wish, or can be left on and connected to Aspen at all times. Many customers have found this “call home” solution more flexible and easier to implement than making the local security configuration changes necessary to allow incoming connections. Consult your organizational networking personnel to determine what configuration is acceptable and most effective at your site.

The Aspen VPN client connects your cluster master or administrative node back to a secure network at Aspen, and allows 4 dedicated support systems there to access your cluster. Aspen delivers your cluster with an “aspensys” account, which is used by Aspen support engineers to log in to your cluster. All logins are done via secure shell and authenticated via keys, resulting in a doubly encrypted path between your cluster and your Aspen support engineer.

If you select the Aspen VPN option and your system is equipped with ABC, Aspen will add your system to our central monitoring system so that we may monitor your cluster in real time. This will allow your Aspen support engineers to help you even more quickly should a problem occur.

If an outbound connection is not allowed, Aspen can work with your networking personnel to coordinate an inbound connection from Aspen through your organizations firewall(s). Aspen will connect to your cluster from a single dedicated IP address which your network personnel may use to control access to the port, and requires only port 22 access (secure shell) to your internal master or administrative node. This port can be different on your external firewall, and simply mapped to port 22 on your cluster end point. Alternatively, Aspen maintains multiple customer specific VPN and remote access solutions that enable us to meet specific customer access requirements and allows us to provide you with the best possible service.

Speak to your sales engineer about what hardware and software support options will best fit your needs.

Cluster Racking

Aspen racks all clusters in a standardized way, but can customize your clusters rack layout to meet your needs. Aspen is an APC™ partner, and utilizes APC™ AR3100 series 42U (rack unit) racks for standard installations. Rack units are 1.75” in height, and are used as a standard in-rack height measurement for rack mounted equipment. A 42U rack contains 73 1/2” of inside vertical space available for installing equipment. These racks are quite strong, with a static load of 3000 pounds and a dynamic load of 2250 pounds, and are also one of the most compact full height racks available, which facilitates rack delivery and installation.

We also can provide racks in less than full height sizes, such as 25U, but do not recommend utilizing less than full height racks unless you have a requirement to specifically do so. Floor space is becoming more and more valuable in facilities, and a shorter rack occupies the same floor space while offering less expandability for future needs. If your organization utilizes specific or specialized racks, we can also accommodate your needs. Speak to your sales engineer about your requirements.

Your cluster may come in one or more racks. Each standard rack is designed, and includes hardware for, baying to other racks on either a 24” (U.S. standard) or a 600 mm (International standard) grid, allowing an exact one to one rack to floor tile placement in any computer room. Each standard rack has two side panels, a lower and an upper, on each side. These panels can be removed when baying racks together to allow for more dense wiring bundles between each rack or to allow rack to rack cooling flow.

Aspen places heavier hardware, such as master nodes, UPS systems, external RAID units, or storage nodes low in the rack to remove any danger of tipping when units are extended in the rack, and nodes are normally installed above these systems. Racks are numbered left to right, and lower numbered nodes are placed lower in the rack. Node numbering sequences are always ordered from the front, left to right, bottom to top.

If an integrated console (keyboard, mouse, and display) unit is configured for your rack, it is installed so that it's bottom is at rack unit 22 or 23 (40 1/2” or 42 1/4”) whenever possible. This height allows most people to comfortably stand at the console and type, or to use an extended height office chair to sit at the keyboard. This

location can be customized to your specific needs.

High speed interconnect switches such as InfiniBand or Myrinet are installed in such a way as to minimize cable length and simplify cable routing. In larger clusters, these switches will be located toward the center of each rack row and toward the bottom of the rack if underfloor wiring is utilized, or toward the top of the rack if rack top cable routing accessories or ceiling mount cable trays are used.

Gigabit Ethernet switches used for the administrative, computing network, or Intelligent Platform Management Interface (IPMI) networks are normally installed in the top rear of racks, and are located to reduce cable lengths as well. Many of these switches are not full depth installations, so allow the same rack space in the front of the rack to be used for additional units such as other network or Keyboard Video Mouse (KVM) switches.

All nodes, and every system that allows it, is mounted on slide-out rails to allow for maintenance. The units slide from the front of the rack, and have two or more attachment screws securing the unit for shipping and normal operation located at the front of each unit. Specific hardware such as UPS systems and other very heavy units are mounted in the racks using fixed rails for safety. In almost all cases, power, network, and other cables must be detached from the unit before it is extended from the rack, although cable management solutions can be configured that will allow your nodes to be extended in the rack while all cables remain connected at additional expense.

Your cluster cabling is bundled between racks and formed for easy connection to the appropriate switch or KVM, and Aspen utilizes maintenance loops on most cables which are routed down the inside of the frame rails on each side of the rack. Cluster Ethernet cables are color-coded as follows;

- Yellow: inside network #1 (1st Ethernet network, administrative and NFS network)
- Blue: inside network #2 (2nd Ethernet network, Data access or MPI)
- Green: inside network #3 (3rd Ethernet if needed)
- White: IPMI network (3rd LAN Ethernet network dedicated to IPMI)
- Red: Outside world connection (if this cable is supplied by Aspen)

In addition to color codes, each Ethernet cable inside your cluster is labeled at both ends with a letter-number combination which uniquely identifies the cable within its color group. For instance, an Ethernet cable connected to your single master node might be identified with an “M”, and the equivalent node1 Ethernet cable would be identified with a “1”.

Power Distribution Units (PDUs) are mounted vertically in the rear of each rack (up to 6 in some cases), and can be reversed to allow connection to overhead or underfloor power. If switched or metered PDUs are used,

their Ethernet connections will be color coded to match the network to which they are cabled.

Aspen's default rack front and rear doors are perforated to allow full cooling flow to the air intakes of all units, and free egress of rear exhaust air. Aspen can provide additional cooling options, such as raised floor helper fans, rear fan doors with rout-able plenum's to control exhaust air direction and destination, front or rear self-contained rack cooling doors which are connected to building chilled water, in-row Computer Room Air Conditions (CRAC) units, or even fully enclosed "cold aisle, hot aisle" solutions to meet your facility needs.

You may also request that nodes be racked with 1/3 Rack unit spacing between each node in the rack. This is done to eliminate metal to metal contact between node cases so that heat transfer is minimized as well as to provide greater front to rear air flow within the rack. This may be desirable in facilities with air flow issues or less adequate cooling.

You or your organization may have special racking or unit location preferences or requirements that are not reflected in our standard rack layout. In that case, your Aspen sales engineer will arrange a conference with you, the sales engineer, and our hardware engineers to determine your specific needs. Aspen will generate a custom rack layout diagram that reflects your specific requirements, then rack your cluster according to that layout diagram after your approval and acceptance.

Lets pretend that you have configured a 46 node InfiniBand cluster from Aspen, with 2 UPS systems, a 4U storage node and master, external RAID system, and KVM capabilities. The following diagram illustrates a standard physical rack configuration for this example cluster.

In this example cluster, the InfiniBand switch selected can support a maximum of 48 ports, so connecting the master and storage nodes to the high speed interconnect only left 46 ports available. What advantage did we gain by connecting the master to the low latency interconnect in this case? Your sales engineer can answer that question. A larger 7U switch chassis placed in R2 might be a more cost effective option for future expansion, as it can be populated with additional InfiniBand port cards when additional nodes are purchased, and allow the new nodes to be added to this cluster without any rewiring of current node connections. The 7U switch supports 96 ports, so a third rack could be procured at a later time with up to 44 additional identical nodes and bayed to your existing cluster as R3, expanding this cluster to 90 nodes.

Your Ethernet infrastructure needs to be examined as well. If non-blocking Gigabit connectivity were not needed, a multi-port trunk between the existing switch and a new switch needed to connect the additional nodes might be sufficient. This would imply that your data access on the compute nodes is to be over the InfiniBand interface, not the Ethernet. If your data mounts are to be done over the Gigabit Ethernet, a stackable switch with a high speed stacking interface might be a cost-effective solution depending on the number of nodes to be added, while a larger chassis based Gigabit Ethernet switch could guarantee nonblocking connectivity. As your storage or master node data serving speed on Gigabit Ethernet would be limited to 2 Gb/s (if 2 Ethernet interfaces were bonded) in any case, it would be more cost and performance effective to utilize the InfiniBand interface for data access given this number of nodes.

Density

Racking density is a major concern of many HPC users due to limited space availability and rising facility expenses. Aspen has several solutions for this environment, including Blade server offerings. While more expensive, a blade server solution might be for you if space savings are the primary consideration for your cluster. However, blades often will not support the highest performance processors, and there will most likely be storage, access and other compromises you will need to make in order to utilize blade servers for your cluster. A more cost effective alternative for your cluster might be “twin” systems.

Aspen offers twin systems, which combine 2 nodes into a 1U chassis sharing a common power supply. In the example cluster above, the nodes inhabit 46U of rack space. Using the twin systems, the same number of nodes would inhabit only 23U of rack space, and unlike many blade solutions, no processor or storage compromises need be made. In more dense clusters, it is critical that your facility offer adequate cooling or that we provide you with one of our additional cooling solutions. Additional rack space could be saved by utilizing a 3U or 2U master and storage node, or changing out the UPS systems for a larger single system. Your sales engineer can help you decide what options best fit your needs and budget.

If your facility has cooling issues or a history of overheating, Aspen can also rack all units with a 1/3U spacing between every node. In this case, 3 1U nodes would require 4U of actual rack space, and no heat transfer between nodes due to metal to metal contact will occur.

Cluster Logical Layout

Aspen builds, sells, and supports HPC, storage, database, visualization, and special use clusters. With certain exceptions, they all are configured much like HPC clusters, with some additional capabilities or different hardware. There are three general sizes of clusters, small, medium and large. Each type and size has specific uses and is often configured in similar ways.

Small Cluster

The “small” cluster has 32 or less nodes, and is usually used for either a small work-group or perhaps even a single user. One, or only a few codes are used on the cluster, and the problem size(s) the cluster is needed to solve are generally not that large.

The small cluster can have a low latency high-speed Interconnect (depending on application) and is serviced by a single front-end, or “master” node, which performs multiple functions for the cluster. The master node services interactive cluster user logins and perhaps user compilation needs, performs job scheduling for the entire cluster, can have the long term data storage and backup systems for cluster data, and performs systems monitoring and fault correction. The master node also firewalls the compute nodes from your organizational network.

Normally only the master node is visible to the network at your organization. The networks inside the cluster are normally set to private IP space, and used only for internal communications between the nodes. The master can also operate a Network Address Translation (NAT) gateway for the nodes, allowing them to access the outside world while not being visible themselves. Figure 1 illustrates a generic representation of a small cluster.

Medium Cluster

The “medium” cluster has between 32 and 256 nodes, and is usually used as a organizational resource, and many codes may be used on the cluster. At this size, a single master node may not be able to scale properly, so some functions may be removed from the master node and encapsulated into separate nodes, such as a dedicated storage node, login, compilation, display, or an administration node. A storage node is a dedicated data serving resource and may or may not be connected to the organizational network for file sharing purposes. An administrative node is used for cluster monitoring and fault correction. Login nodes remove the interactive login load generated by cluster users to a separate node, and dedicated compilation nodes might be used just for compilation of various codes or development purposes.

Any, or all of these disparate node types might be present in any cluster, depending on your specific needs, and variations of all of these configurations are possible. Some customers run codes that access data kept on external storage, so each node might need an external connection. Or perhaps a router, layer 3 switch, or gateway is used on a cluster internal network so that each node is directly reachable from your organizations computers. In some cases, a dedicated compilation node or nodes might be needed to serve development needs, or perhaps login nodes are used for compilation and direct cluster user logins.

There are no hard and fast rules, a storage node might be configured in a small cluster due to data sizing or transfer requirements, and an administrative node might not be needed in a medium cluster. Perhaps the storage node is connected to the high speed Interconnect for data sharing across that network, and perhaps the administrative node is also configured as a fail-over master. There are myriad options, and your Aspen Sales Engineer can help you decide what options best fit your needs.

Large Cluster

The “large” cluster has more than 256 nodes, sometimes thousands. At this size, scaling is critical, so the cluster could contain multiple storage nodes, dedicated fail-over masters, and perhaps more than one administration node.

The challenges of large clusters are many, but Aspen engineers can help you design and deploy your large cluster successfully. The scope of the problem sets and level of detail necessary to successfully deploy a large cluster require much closer design interaction than smaller clusters, as improper design decisions that might only cause annoyance in a smaller cluster can cause performance, operations, or execution issues in a large cluster. Your Aspen sales engineer will schedule several requirements meetings with you as well as have our senior cluster designers work directly with you and your organization to define your large cluster solution.

Storage and Database Clusters

A storage cluster is used to service larger data space or faster access requirements, and can be configured multiple ways. Sometimes multiple nodes are connected to a Storage Area Network, which is implemented by Fiber Channel or InfiniBand switches to connect multiple RAID systems to multiple hosts. A parallel file

system might be deployed, which allows multiple hosts to access the same data space simultaneously. Aspen supports GFS, GPFS, Lustre, OCFS, and PVFS parallel file systems as well as other commercial offerings. If you have a specific parallel file system requirement that are not on this list, ask your Aspen Sales engineer. We probably have experience with that file system and have deployed it.

A single large RAID system might be deployed which has host ports for storage nodes to connect to. This option, while less scalable, removes the cost of a SAN switch, associated software, and support costs. This can be a significant savings. Storage clusters themselves can be integrated into the HPC cluster to serve data processing needs for the cluster and other organization computers. Figure 4 shows both the SAN storage cluster option as well as the smaller single RAID with host ports. Database clusters hardware configuration can resemble storage clusters in many ways.

Visualization Cluster

Visualization clusters often have one or more Graphics Processing Units (GPU) installed in each compute node. Normally a visualization cluster that is installed with more than one GPU in each node must be

implemented using 3U or 4U compute nodes. The GPUs may be used for general computation on certain codes, or as graphics processors which perform rendering and push images to front end display nodes which are connected to the cluster. Aspen Systems has deployed visualization clusters in multiple environments, and can help you with your visualization cluster needs as well. If you have visualization questions, use your Aspen Systems sales engineers expertise to design and optimize your visualization solution.

Master Node

Your master node is one of the most critical systems in your cluster. Aspen provides tools on all our default distributions to rebuild all other nodes in the cluster from the master node, but a failed master could require manual configuration and involve some downtime if a fail-over master is not used. Aspen can configure your cluster with a fail-over master at additional cost. While the failover master option requires additional hardware as well as specialized software configurations to operate, many customers may find it to be cheap insurance to ensure that their cluster continues to function in any hardware failure situation.

The master operating system file systems should always be mirrored on two disks. RAID 1, or mirroring, can be done via a hardware RAID card, or by software raid. The master O.S. RAID will contain not only the distribution, configurations for all your cluster utilities, and source for your particular utilities or codes if applicable, but also the single image for network booting compute nodes (if a single image system is used), or multiple snapshots of node images that are used to restore or upgrade your compute nodes. This, combined with your data, *is your cluster*.

Cluster images can be extremely customized, containing site, facility, code, or performance specific modifications that Aspen, and you, have spent many hours completing. Figure 5 shows an average functional master node software stack.

Aspen automatically keeps an image of your cluster as it was shipped on secure storage at our facility, and can retrieve additional images at later dates, perhaps after upgrades or site customization, to facilitate disaster recovery should that be needed. Cluster images do not normally contain any application or model output data, so data needs to be backed up using some other mechanism.

If current images have been taken of your master node(s), a compute node, and any additional specialty nodes, and those images have been copied to secure storage at your site or at Aspen, we can always re-install your entire cluster should that be needed. Of course we don't wish to do that, and a properly configured master node will help ensure that that eventuality never occurs. Using the Aspen image is also an efficient way to implement additional clusters should you need them, as any customization we have done for you or any site localization or customization you have performed resides in these images.

If you have only a single master node, your master node should have redundant power supplies which are connected to different breakers, and if possible, those breakers should feed from different panels in your facility. If your single master has only one power supply, your entire cluster can be rendered inoperative by the failure of that power supply. If you have redundant power supplies but they all are connected to a single circuit breaker, your cluster can fail if that single circuit breaker becomes faulty. A single panel normally is powered by a single master breaker, so if at all possible, ensure your redundant supplies are fed from different electrical power panels in your facility.

Your master node in a small cluster will function as the interactive login node for all your cluster users, so additional memory is needed, especially if you expect to deploy Virtual Network Compute (VNC) servers for individual users, or large code compilations are performed. In a small cluster configuration where the master

serves all front end functions, any performance degradation on the master can affect job execution on the entire cluster, so it is better to over specify the amount of memory in your single master node rather than the reverse.

The master node may be used to burn data sets onto Blu-ray, DVD, or CDROM, so a burning unit might be needed as well, and a single master can contain the data storage for the cluster, which we will discuss next.

Storage

Almost all clusters require shared data space that is exported to all nodes in the cluster. This space can be the users home directories, dedicated data space mounts, or combinations of both. Modern HPC clusters require larger and larger amounts of data space, and also require as much performance as is possible within the allowed budget.

The use of 1 Terabyte (TB) and larger Serial ATA (SATA) II and smaller Serial Attached SCSI (SAS) drives in combination with high performance internal RAID cards or external RAID systems helps make these goals possible. In a small cluster single master configuration, your master node can be configured with either internal RAID storage, or external RAID storage.

Internal RAID Systems

In internal RAID storage configurations, the master node can have up to 24 drive slots, two of which may be used for the mirrored O.S. drives. Aspen highly recommends that any RAID 5 or 6 sets always be configured with at least one hot spare drive, more if you can afford it. Given a 24 slot drive bay master node, subtracting 2 drive slots for the RAID 1 O.S. drives, using one drive as a hot spare, and using RAID 5, between 3 and 19 TB can currently be configured on a single master node. RAID 6, which provides fault tolerance from two drive failures (RAID 5 protects you from only one disk failure), can be configured on most systems as well, and would yield approximately 18 TB with one hot spare disk. These numbers represent raw space, and your choice of file system will affect how much is actually usable for user data. Most current 15, 16, and 24 slot (3.5" disk) chassis have limited front panel space, so the addition of a burning unit may not be possible. A slimline DVD or CD unit is normally configured into these chassis.

Hot Spare Disks

Configure as many spare disks as you can afford. All disks eventually fail, and the use of hot spares for critical data partitions is a small investment compared to the importance of your data. Aspen will always configure the RAID systems to notify you, and Aspen if possible, in the event of any disk failure, and we will ship you a new disk to replace the failed unit (unless you have contracted for on-site maintenance). So why use hot spares? You use hot spares for several reasons;

- all disks are mechanical devices, and all mechanical devices eventually fail
- the same model disks have a tendency to fail within a short time of each other
- other bad things happen, sometimes at the same time your disks are failing

Let's say that you do not have a hot spare disk configured in your RAID system, and leave town for a 3 day weekend on Friday afternoon, a well earned vacation. Friday night, a single drive in your data RAID 5 set fails, notifications are duly sent to you via e-mail, and the RAID recovers. This particular drive is model "XXXX" from manufacturer "YYYY", and it has a previously unknown adverse reaction to the slightly higher temperatures that it must endure in your facility.

You perhaps haven't been as diligent as you could have been about backups, because you've been busy. We all know how that goes. Perhaps your organization does not allow the cluster to e-mail an external source, and Aspen does not receive a notice of the disk failure either. Perhaps your organization mail server is down for maintenance that weekend, so no mail can be sent externally. Or perhaps Internet routing between your organization and Aspen is down that weekend.

On Tuesday morning you return, read your e-mail, and contact Aspen for a replacement drive. At this point, your best case scenario is that the replacement drive will arrive mid-morning on Wednesday, leaving your data unprotected, and open to loss from failure by a single disk drive, for over 4 days!

But perhaps something else bad happens. Lets say that the work on the mail server trips a breaker, which also happens to power the temperature monitoring for your building cooling system, which in turn caused a spike in your temperature, and another model "XXXX" drive to fail. The result is total loss of your current data set, perhaps including the applications that were running over the weekend while you were on that well deserved vacation, and a period of downtime for your cluster while you retrieve backups, receive additional drives, and rebuild the RAID system. A single hot spare disk is cheap insurance, and would have saved this users data.

If your cluster data is critical,

- use at least one hot spare disk per RAID set unless the set is RAID 10 or RAID 1
- use RAID 6 vs. RAID 5 if performance requirements allow
- allow Aspen to configure the RAID to notify the appropriate users or administrators in case of disk failure
- if you can, allow the RAID to be configured to notify Aspen Systems when a drive fails (via email to support@aspsys.com, which automatically opens a support ticket)

External RAID Systems

External RAID systems can be configured on your master node or a dedicated storage node. While often more expensive than an internal RAID solution, these systems can offer more expandability (additional expansion chassis can be added), better performance in some cases, and more streamlined manageability with an embedded web server, telnet and ssh access, and the ability to be managed independently of the master node. Some internal RAID cards can also utilize external expansion chassis as well, which can eliminate the expandability advantage. External RAID systems can also be configured to attach to multiple nodes, normally via Fiber Channel, allowing for node fail-over or parallel file system configuration.

Which RAID to select is determined by your storage requirements and budget. If you intend to greatly expand your data storage or upgrade the master node in the future, an external RAID might be better for your purposes. If you intend to utilize storage fail-over or connect the same RAID system to multiple hosts, an external RAID system is especially needed. If your data requirements are known, and within the capacity of an internal RAID solution, an internal RAID system might be more cost effective. In all cases, the minimum rule of one hot spare drive per RAID set should be followed.

Network File System

Many HPC clusters, especially small and medium clusters with data access requirements that can be served by a single host, utilize the Network File System (NFS) to mount data directories on the compute nodes. NFS is the de facto standard for data sharing in HPC clusters because of its ease of configuration and ubiquitous support. An NFS server runs on the node that serves the data space, and an NFS client runs on the compute nodes. This allows the users home directory, or a shared data directory, to be mounted to all compute nodes so that applications running on the compute nodes have access to the same data. The NFS directories can be shared over the clusters administrative network, which is often Gigabit Ethernet, or over the clusters high speed interconnect, which can significantly increase data access speed.

While there are projects to make NFS servers function in parallel so that more than one NFS server may be used, this capability is not main stream at this time. NFS relies on a single server to serve any particular data space, which means that the protocol overhead, local RAID speed, bus architecture, network interfaces, memory, and Central Processing Unit (CPU) speed of that server limits the speed as which NFS data can be accessed. If your application(s) data requirements or your overall data requirements for the cluster are higher than this single server can accommodate, the NFS server becomes a performance bottle neck for the cluster. Parallel file systems can be used to solve this issue.

Parallel File Systems

Aspen offers GFS, GPFS, Lustre, OCFS, and PVFS parallel file system options as well as other commercial products. Each parallel file system solution has distinct characteristics, and is used for specific types of data serving needs, so close consultation with your Aspen sales engineer is necessary to select the proper parallel file system to fit your needs. Parallel file systems are complex, and can require specialized knowledge to configure and maintain, so some additional organization training may be necessary. In many cases, a specific parallel file system can be obtained both in an open source, unsupported version, and as a commercial product with support. Direct commercial support for your parallel file system may be necessary to achieve optimum performance and reliability in your configuration.

GFS

GFS is the Red Hat Global File System, and supports shared disk access from multiple nodes to a single RAID. GFS is available on RHEL servers along with their Red Hat Cluster Suite as a supported commercial application, or can be installed in a more limited fashion as an open source application on Red Hat or Red Hat derivatives such as CentOS. A client on each compute node can be used, or each individual GFS server can also be an NFS server. GFS is normally deployed on a maximum of 8 servers, although larger deployments are possible and do exist.

GPFS

GPFS is the IBM General Parallel File System, a commercial product from IBM. Aspen is an IBM partner, and can build your cluster with IBM components and a customized software stack as outlined here and in our

Configuration Guide and SOW, and include the GPFS file system. GPFS is a licensed commercial product, and GPFS servers can also serve as NFS servers to compute nodes.

Lustre

Lustre is a parallel file system originally developed by Cluster File Systems, Inc., and now owned by Sun Microsystems, with both commercial licensing and open source versions. Lustre is used in some of the largest HPC clusters in the world, and while considered by some to be difficult to configure, tune, and maintain, it is used in many very high performance environments. Aspen Systems also partners with Terascale, and can integrate the Terascale high throughput, scalable, Lustre parallel storage appliances into your cluster to serve your performance needs. The Terascale appliance removes many of the pitfalls of managing a Lustre implementation yourself while providing the superior speed and scalability of a Lustre parallel file system implementation.

OCFS

OCFS is the Oracle Cluster File System, an open source project from Oracle. OCFS is meant for use in an Oracle database environment, not as a general use file system.

PVFS

Parallel Virtual File System (PVFS) version 2 is an open source project design to provide high performance for parallel applications, where concurrent, large IO and many file access are common. PVFS is designed as a set of clients and servers, so normally a subnet of dedicated nodes provide the storage space and act as PVFS servers, while all other nodes function as clients to access data. PVFS can be configured in multiple ways, but it is recommended not to use PVFS servers themselves as compute nodes, as any crash of a running application on that node could cause the entire cluster to become inoperable. There are high availability configurations for PVFS which can be configured, but PVFS is not designed as long term storage, but rather as very fast scratch space for parallel applications.

Parallel File System Hardware Requirements

Almost all parallel file system implementations other than PVFS will require the use of an external RAID unit that can be connected to more than one host, either by SCSI (not recommended due to speed issues) or by Fiber Channel or InfiniBand.

Other Options

Other commercial software or hybrid software and hardware products exist, such as [Panasas](#), which can provide extreme reliability and access speed for your cluster. Speak to your Aspen Sales engineer about other solutions we offer to meet your requirements. Almost all parallel file systems require additional hardware as well as custom software configurations, and Aspen can help you design your storage to meet your needs and wishes.

Backup

Data backup is a needed part of any HPC cluster deployment, however the size of today's storage solutions can make that problematic. Aspen Systems makes disaster recovery images of your cluster when it is shipped. These images do not contain your user data in most cases, such as /home or /data, but instead are meant to rebuild the cluster should a catastrophic hardware or software failure occur. You may also contract with Aspen to store additional disaster recovery images if the cluster environment is changed in any significant way. These images are normally smaller than 20 Gigabytes, and often less than a gigabyte for a compute node. So, where does that leave your data in event of a total RAID failure or other catastrophic failure? You need a backup solution for at least some of your most critical data, and there are several options you might consider.

Backup Hardware

Often, larger organizations will already have an enterprise-wide backup solution, and your new cluster can be licensed or configured as a client to the existing resource. If that resource does not exist, Aspen can provide your cluster with Digital Linear Tape (DLT), Advanced Intelligent Tape (AIT), and Linear Tape-Open (LTO) tape drives and tape drive libraries, depending on your needs. Tape capacities range from 160 to 800 GB per tape, with write speeds from 12 MB/s to 120 MB/s. Optical worm drives are available as well.

Almost all of the single tape drive solutions can be mounted as internal devices in your master or storage node(s) if they are 3U or 4U chassis, while tape libraries are normally configured as rack mount units or on rack shelves. Internal drives are normally Small Computer Systems Interconnect (SCSI) interfaces, while external drives can be SCSI or Fiber Channel.

You may need to decide what data on your cluster is critical and back up only that data. Often, saving only results or application output data can significantly reduce your backup requirements, allowing even a single tape drive of the right type to be used for your backup needs.

Aspen can provide your system with larger tape drive libraries, which can back up your entire system, including your data storage. Some of these systems can archive petabytes of data, speak to your Aspen sales engineer about your specific backup needs.

Backup Software

You have many choices for your backup software solution as well. If you wish an open source solution, Aspen recommends [Amanda](#) or [Bacula](#), and Aspen offers [Storix](#), [IBM Tivoli](#), and [Veritas Backup](#) as commercial products. Amanda is command line based, while Bacula has X windows clients and a command line text console user interface. The commercial products all offer Graphical User Interfaces (GUIs) as well as command line based options. Veritas, now owned by [Symantec](#), is arguably the most widely deployed commercial backup solution, while IBM Tivoli Storage Manager is ideal for an IBM hardware based cluster. Storix is a relatively full featured and price competitive commercial offering, and has a web user interface.

As data sets become larger and larger, full backups of all data spaces becomes problematic, forcing some organizations to rely only on their RAID systems, with no additional backup capabilities. In these cases, it is critical that hot spares be configured on all RAID sets, notification to your administrators and to Aspen be configured and routinely tested, and that the RAID system(s) be located on conditioned power sources.

Nodes

Compute nodes are almost always considered as replaceable, with little or no customization for any particular node. Using the Aspen utilities, Rocks, or Perceus / Warewulf, it is possible to define a particular group of nodes (a group could be one node) which have some unique properties, such as an extra external network connection, a different processor architecture, or a different software build, which can then be used for different functions you might need for your cluster.

These nodes are considered unique recovery targets. If you are using Aspen utilities, these nodes are backed up as a unique image which is tied to that particular node or nodes, just as a storage or fail-over master node is itself a unique image. In a Perceus / Warewulf system, another image of these nodes are kept on the master node, and in Rocks a separate class of node is created, which has its own unique properties.

This does not preclude another node of similar hardware configuration being used as a replacement for that node should it fail, although some physical intervention might be required. For instance, lets say that an external network as well as a SAN connection is attached to node1, and node1 fails. These connections could be serviced by node2 if;

- the external Ethernet connection on node1 is moved to node2
- node2 contains a fiber channel card identical to node1 or,
- the node1 fiber channel card is moved to node2
- node2 is re-imaged with a current node1 image

If you are going to utilize any of your compute nodes in unique hardware configurations for mission critical tasks, Aspen recommends that you configure more than one node with that exact configuration in order to facilitate fail-over in the case of node failure. In the above example, adding a fiber channel card to node2 and attaching another external Ethernet connection to it, even though neither is actually configured up during normal operation, would allow node1 functionality to be returned to the cluster in as little as 5 minutes. An even better option is to place this functionality on the master or another front end node which itself is set up for high availability fail-over.

It is possible to utilize the same software build, and image, on nodes of different clock speeds and memory configurations with no modification. Nodes of different chip architectures will require different images to accommodate different hardware drivers and to take advantage of specific processor performance attributes.

Your nodes memory configurations should be based on your applications memory requirements. Many customers have some additional applications that require more memory per core than might be economical, but do not run many instances of this application compared to other code(s) they run. In this case, configuring a single or several nodes with more memory to accommodate these requirements is more economical, and you may utilize your scheduler or other utilities to route that particular application to that node or set of nodes.

While it is uncommon, nodes can fail in your cluster. There are two separate types of compute environments that affect your node hardware configuration choices.

1. **Non-critical codes:** In many environments, applications are easily re-ran in the case of a node failure. Applications are usually of limited duration and can be re-submitted or reran if a node involved fails, or possibly the code has checkpoint and restart capabilities.
2. **Critical codes:** In this environment, applications are not easily re-ran, perhaps because of the number of applications ran, duration of application execution, time sensitivity, input data storage requirements, or model preparation complexity.

In a non-critical job environment using disked clusters with our standard distributions, there is no need to have a hardware or software RAID 1 environment on compute nodes. Aspen provides command line utilities, or a GUI when ABC is purchased, to restore any node very quickly, and the added expense of hardware RAID cards and/or additional disks can be utilized to procure an entire spare compute node. Nodes do not need to have redundant power supplies, and this cost savings can also be utilized to purchase additional compute nodes which can function as complete spares in the case of node failure.

In a critical code environment, node failures are taboo. Utilize software or hardware RAID 1 configurations on your disks, and ask for your nodes to be configured with redundant power supplies. In cases where 1U nodes with a high speed interconnect card are used, your expansion slots may be limited, making software RAID 1 more attractive. If your nodes are configured with redundant power supplies, they should be powered by different rack Power Distribution Units (PDUs) within the cluster, which are in turn connected to circuits connected to separate circuit breakers and if possible different electrical power panels within your facility.

Occasionally, extremely fast local scratch space is needed by your application(s). A single SATA II disk on a compute node can provide ~50 to ~70 MB/s sustained write throughput, depending on the disk model used; that might not be fast enough for some applications. In these cases, compute nodes can be configured with multiple disks and RAID 5 or RAID 0 sets to achieve the desired level of reliability and performance.

Many single 1U servers can be configured with four 3.5" disks, which allow for an entire RAID 5 set that includes 3 drives with one hot spare disk and includes the operating system build and scratch space (more reliable, faster than a single drive), or a less reliable but quite fast single O.S. drive and a 3 disk RAID 0 (striping) set.

Hardware or software RAID can be used in either of these scenarios, although when configuring with software RAID 5, the /boot partition is configured as RAID 1 to allow correct booting. If you opt for a RAID 5 compute node solution, Aspen recommends that you configure your nodes with a hardware RAID controller if possible. Software RAID 5 will almost always incur higher overhead on your node, and can slow down code(s) execution, so Aspen recommends that you utilize a hardware RAID solution if RAID 5 is needed. Your application requirements and node expansion slot availability will drive this choice.

Power

Power Distribution Units (PDUs)

Modern clusters can require significant power. Each rack in your Aspen cluster is normally equipped with rack mounted PDUs which provide power to one or more nodes. Normally, one or more PDUs are installed in the rear of the rack behind the node mounting infrastructure, and do not impact the rack space available for mounting your other hardware. Aspen can provide your cluster with switched PDUs. [ABC](#), or cluster administrators can use these switched PDUs to remotely power off any system in the cluster. Metered PDUs are available as well, which ABC can poll for circuit status and load.

These PDUs are usually connected directly to outlets on the wall, under your raised floor, or in your overhead rack infrastructure. They can also be connected to Un-interruptible Power Supply (UPS) units which are located in your Aspen rack(s) or elsewhere.

UPS Systems

You may have a facility that already has a UPS or even a generator. A UPS unit is necessary to insure that no power interruption occurs, even if you have a generator. A UPS unit alone will only keep the units powered for a limited amount of time, usually less than 30 minutes, but can be used in conjunction with a facility generator to insure that your infrastructure continues to run even in extended power outages.

It is always a good idea to protect your critical systems, normally the master, any fail-over masters, administrative, and storage nodes, with UPS systems if your facility does not have them. Operating these nodes on UPS ensures that a sudden power blackout (power is lost completely) or brownout (low voltage levels), or dropout (momentary total loss of power) does not cause these nodes to crash, which risks file system or hardware damage. Brownout or dropout situations are transparent to a node protected by UPS, and if the blackout lasts long enough to drain the UPS battery, monitoring software can effect an orderly shutdown of the node(s) to minimize possible file system damage and facilitate a clean reboot process later.

You may purchase UPS systems from Aspen that are integrated into the cluster rack(s) and monitored by ABC. If your nodes are not protected by UPS, then current jobs running when the power outage occurs will fail. It is possible to equip every node in your cluster with UPS protection so that jobs are not interrupted by power outages, however this will adversely affect your rack space utilization (UPS systems can be up to 5U in height) and can be costly, depending on the number of nodes in your cluster. Unless ABC is used, some complexity is introduced into your system in monitoring and controlling multiple UPS systems with the same master (an SNMP management card is required in each UPS in most cases), and we will need to carefully balance your run times for each UPS under normal load conditions by moving nodes from one UPS to another.

Additional UPS runtime can be configured for your cluster by adding additional battery packs to your UPS system(s) as well. Speak to your Aspen sales engineer about your specific UPS needs.

In North America, your facility may provide A.C. power at 120 volts or 208 volts. There are reasons to prefer one voltage over another.

A.C. Power - 120v Circuits

Convenience and circuit availability are the most usual reasons to power your cluster with 120v circuits. 120 volt (v) 15 amp (a) office power outlets are ubiquitous in almost all environments, and if you are installing your cluster in a converted room that does not have raised floor or was not designed specifically for computers, these circuits will most likely already exist in your facility. However, there are significant limitations to the amount of power available on these circuits, and you may require more of these circuits than you have available.

The majority of existing office 120v circuits are rated at 15a. The Underwriters Laboratory specifies that a circuit can not draw more than 80% of a receptacles rating for safety, leaving 12 amps available.

Total watts available is derived by multiplying the voltage (120) by the amperage (12), so we have the equation;

$$120(v) \times 12(a) = 1440(w)$$

This means that a standard 15a receptacle can provide 1440 Watts. Now, two terms are used to describe electrical equipment power ratings. “Watts” are the true, or maximum, power of the circuit, and “volt-amps” is the apparent power, or what the circuit can really produce given the effects of capacitance and inductance by the components in the load. Watts are the maximum the circuit can produce, while Volt-amps are the actual power that the circuit can produce under real world circumstances, and depends on what type of equipment is attached to the circuit, and what that equipments “power factor” is.

Most Aspen servers have power factor corrected power supplies with a .85 or higher correlation between Volt Amps and Watts, however some inevitable internal power supply loss occurs. We'll use a power factor of .95 to allow for lower efficiency in some of the electrical devices in the cluster.

Total volt-amps is derived by multiplying the voltage (120) by the amperage (12) by the power factor (.95), so we have the equation;

$$120(v) \times 12(a) \times .95(pf) = 1368(va)$$

This 120v 15a circuit can provide approximately 1368 volt-amps of actual power. 120v 20a (1824 volt-amps) and 120v 30a (2736 volt-amps) rated circuits do exist in some facilities, although they are a bit less common.

Most 1U nodes that Aspen sells are equipped with a 650 watt power supply, This is the maximum power the node power supply can draw, but a nodes true average running load is between ~300 and ~322 watts (equipped with a single hard drive and 8 GB memory) under load. Additional memory or hardware options, or processor intensive code execution, can increase this load significantly. As a general rule of thumb, no more than 4 1U compute nodes should be connected to any single 120v 15a circuit (1368 / 322 = 4.2).

Your master or storage nodes, or “twin” systems (two systems in 1U sharing a power supply) have 800 to 1200 watt power supplies. A master or storage node, and its associated external RAID if so equipped, can draw up to ~600 watts under nominal load due to being equipped with additional memory and disks as well as RAID cards and other accessories. If your master or storage node has redundant power supplies, you should budget the total power load on every circuit that powers one of the power supplies.

Your Aspen sales engineer can provide much more exact numbers, and will do so when he or she computes the power load for your configured cluster. Aspen also provides a [Facilities Estimator](#) that will estimate your power needs based on your selections.

Please note that if you intend to over-subscribe your cluster or routinely perform processor intensive

calculations, your power demands will be higher.

A.C. Power - 208v Circuits

Power capacity is one major reason to power your cluster with 208v circuits. Almost all node power supplies are capable of auto-switching between 120v and 208v power sources, and common 208v circuit sizes are 20a (3161va) and 30a (4742va), but even larger circuits are available. Any node will draw less current ($\sim\frac{1}{2}$) at 208v than it will at 120v, reducing waste heat and increasing power supply component life, and allowing more nodes to be fed from a single circuit and its associated UPS or PDU.

It is also common for each 208v receptacle to have its own circuit breaker, whereas 120v 15a circuits are often wired with multiple outlets per circuit. Many people have probably had the experience of turning on one appliance in an older house which blew the circuit breaker, only to find that electrical devices in a totally different room became non-functional. This is not a good thing to have happen to a circuit that is powering part of your cluster.

208v circuits are superior to most 120v circuits in other ways as well. 208v circuits are commonly equipped with twist lock plugs, which are turned, or twisted, to lock into the receptacle. Although 120v twist lock plugs do exist, many 120v circuits employ straight plugs, which do not lock. If your plugs are not twist lock, an inadvertent bump or cable pull can remove power from parts of your cluster, especially in crowded underfloor conditions. 120v electrical service components are often “consumer” grade, while 208v receptacles are most likely industrial grade, so the quality of 208v electrical components are higher, reducing the chance of intermittent power connection issues.

240v Circuits, 3-phase Supply, or Direct Current (D.C.) Feed

In some unusual cases, your facility may provide 240v circuits to supply your compute power needs. Virtually all single phase 208v equipment will operate from a 240v circuit, and your cluster can almost always be configured identically to one powered from 208v circuits.

Your facility may provide 3-phase power (a power feed characterized by 3 “legs” which provide their peak power at different times in the cycle). In this case, the power must be converted to single phase (a 3-phase 208v 50a circuit will convert to 3 separate 120v 50a single-phase to neutral circuits or a combination of double pole 208v and 120v circuits) either by your facility infrastructure, or via a converter Aspen can configure for your cluster.

While much more common in telecommunications facilities, your facility might provide only D.C. power for your computing needs. Normally, D.C. power is used in facilities with battery backup and/or a generator, as its use can simplify power generation and storage. Specialized node power supplies, along with inverters for peripherals and devices which cannot be configured with D.C. power supplies, are necessary to configure a fully D.C. powered cluster, and not all node configurations are supported. Aspen can help you with your D.C. powered computing needs.

International Power Standards

Many countries have different power options, plug configurations, and connection requirements. Aspen builds systems that are installed in countries all over the world, and can configure your cluster so that it matches the power supplied in your facility. Discuss any specific power connection requirements you may have for your cluster with your sales engineer.

What power options to use

Use 208v circuits:

- whenever possible and economically feasible
- if you have to install additional power circuits to provide power to your new cluster, and your facility and electrical panels will support the addition of 208v circuits

Use 120v circuits only if:

- a sufficient number of existing 120v circuits already exist in your facility to power your cluster, or
- you will have to install additional power circuits to provide power to your new cluster, and your facility can not provide a sufficient number of 208v circuits in your facility electrical panels to support your cluster

Use 240v circuits only if your facility is equipped only with 240v infrastructure. Work with your Aspen sales engineer to determine interoperability of all selected cluster components with the 240v power source. Use 3-phase to single phase conversion only if your facility provides only 3- phase power. The additional equipment needed for 3-phase to single phase conversion can be expensive. Use D.C. as your cluster power source only if that is the only power source available in your facility, as that option will limit your hardware configuration choices.

In all cases, your Aspen sales engineer will speak with you to understand your facility power options, and configure the cluster to that specification. They will also provide you with the number and type of circuits required. Aspen recommends using 208v single phase power, and 30 amp or larger circuits for your cluster(s) power requirements.

Plug Types

One of the most common customer errors is to specify incorrect plug types for their cluster installation. International customers do not have this problem as often, as they are accustomed to stating specifically what standard plug type they require.

National Electrical Manufacturers Association (NEMA) plugs and receptacles are commonly used in North America, and use designators such as “NEMA L6-30R” to identify receptacle and plug types. The “R” stands for receptacle, which is the receptacle you provide at your facility to plug your cluster into, while “P” stands for plug.

The most common plug/receptacle types you will encounter in North America are;
120v – 15a to 30a

- NEMA 5-15r (120v 15a standard office wall plug)
- NEMA 5-20r (120v 20a circuit installed in some more modern facilities)
- NEMA L5-20r (120v 20a twist lock receptacle)
- NEMA L5-30R (120v 30a twist lock receptacle)

208v – 20a to 50a

- NEMA L6-20R (208v 20a twist lock)
- NEMA L6-30R (208v 30a twist lock)
- NEMA L14-30R (208v 30a twist lock 4-Wire Grounding)
- CS6364 (208v 50a twist lock “California Style” 4-wire Grounding)

The CS6364 receptacle and plug combination is not a NEMA designation, but is becoming more common as power needs grow. There are many other types of electrical receptacles and plugs. For instance, the back of most nodes are equipped with International Electrotechnical Commission (IEC) C14 chassis plugs, while the power cords connecting the node to its PDU are normally C13 line socket to NEMA 5-15P.

You may need to consult with your electrical personnel at your facility to determine exactly what receptacles you have or can support. In all cases, speak with your sales engineer if you are confused about your options or what type of receptacles your current facility has.

Cooling

Modern clusters are more powerful than ever before. The amount of processing power, memory, and the interconnect options available in a single node today far exceed anything available even 2 or 3 years ago. However, that additional processing power comes with a price, more power usage. In electronics, power is transformed into heat. More power usage means more heat to dissipate. As we pack more and more performance into smaller packages, which we then rack as tightly as possible to take advantage of all available space, cooling becomes critical. **Inadequate cooling is the primary cause of cluster hardware failures.** Aspen Systems partners with APC™ and Liebert™ to solve your cooling needs.

The optimum ambient temperature for your cluster is 68° to 77° F (Fahrenheit) (20° to 25° Celsius). Maximum ambient temperature should not exceed 80.6° F (27° Celsius) for any length of time. While your cluster can be operated in environments with above maximum ambient temperatures if necessary, doing so will adversely affect your code(s) performance and your clusters long term hardware reliability. UPS batteries are especially sensitive to higher temperature environments. A UPS battery deployed in a 90° F ambient temperature environment might only last 1 to 2 years, versus the 3 to 5 year normal battery lifetime. System memory and disk drives also will fail significantly sooner than normal in higher temperature environments.

Many existing raised floor computing facilities were designed for an estimated heat load of 3 to 4 kilowatts (kw) per rack. Using our standard 1U server loaded power usage (322 watts), a rack of 40 of these systems would produce a 12.88 kw heat load, over 3 times the maximum amount of heat the facility was designed to handle in that one rack. Using our twin systems, the heat load would be more than 6 times the amount of heat the facility can reliably dissipate from a single rack. If your systems are heavily utilized and as densely racked as possible, goals most HPC users aspire to, a rack could produce over 27 kw of waste heat. High density racks often overwhelm the cooling capacity in a single area of a traditional computer room, causing hot spots. Aspen can upgrade or supplement your facilities current cooling solutions so that this does not occur.

In non-standard compute facilities such as converted closets, or rooms that were originally designed to house personnel, the situation can be even worse, and can lead to errant code behavior and equipment failure. We call this the “small room” cooling problem, and Aspen has specific recommendations and solutions for this situation.

Small Room Cooling

Let's pretend that we're going to install the example cluster outlined in the physical layout section into a small 10 by 10 foot room that we've re-purposed for our cluster. This is a 46 node InfiniBand cluster, with 2 UPS systems, a 4U storage node and master, external RAID system, and KVM capabilities. Under full load, this cluster would generate approximately 16.363 kw of waste heat. We have a building air conditioning system, and the room is located in the center of your building, so no solar gain or additional heat is being produced. We would like to have the cluster maintained at an average temperature of no more than 77° F.

Conductive cooling, where heat flows through the walls of the space, can remove approximately 400 watts in a room this size. Passive ventilation, where the heat generated by the cluster flows into cooler air via a door, wall, or ceiling vent without a helper fan, could accommodate approximately 800 watts. Adding a fan to this vent could make this approach accommodate about 2 kw of waste heat. So combined passive (400w) and fan-assisted (2 kw) cooling could provide only 1/8th of the heat removal we need. In addition, the following factors must be taken into account.

- Room size – temperature increases as the room gets smaller
- walls, ceiling, floor – temperature increases as thermal resistance increases
- AC setback – if your building turns down, or off, the building air conditioning on nights and weekends, your room temperature will increase proportionally
- exposure – if a wall or walls is subject to sun exposure or heat transfer, temperatures will increase.

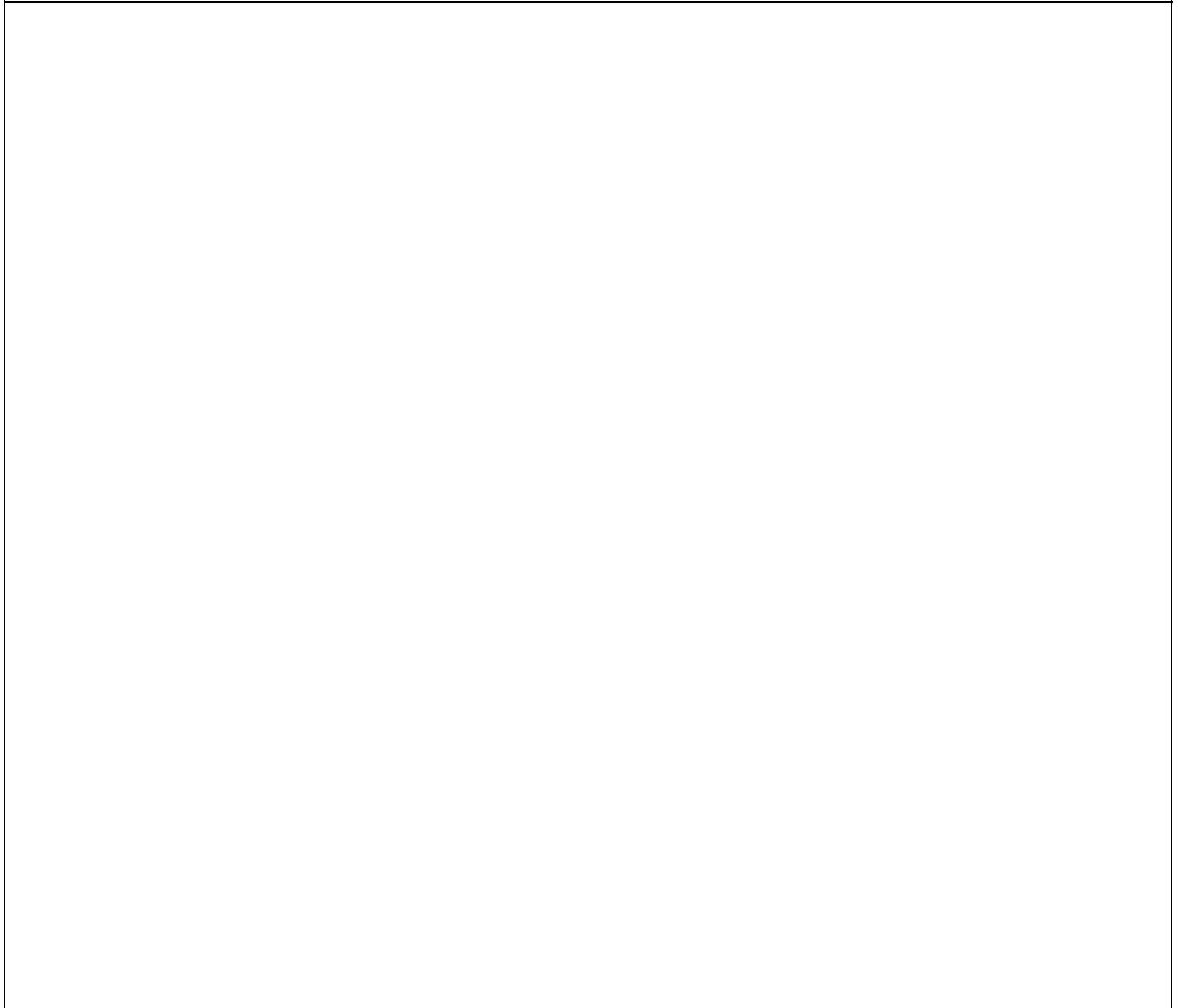
It is pretty obvious that we will need a dedicated air conditioner of some type for this cluster. How many tons of cooling would be necessary to remove the heat? Multiply your wattage by 3.413 to convert to British Thermal Units per hour (BTU/H).

$$16363(w) \times 3.413 = 55847 \text{ (BTU/H)}$$

$$55847 \text{ (BTU/H)} / 12000 = 4.65 \text{ tons of cooling}$$

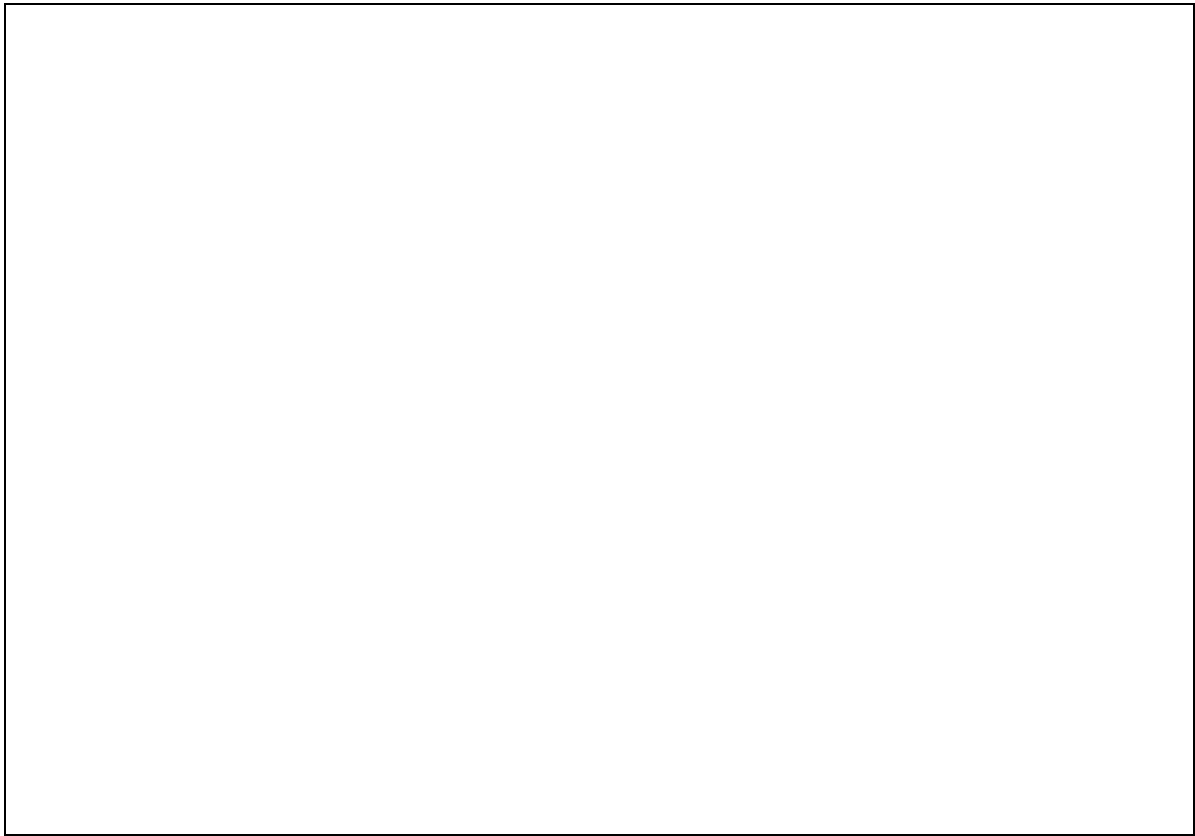
Modifying your building air conditioning system to provide this 5 tons might be expensive or problematic. A supplementary 5 ton air conditioning unit, perhaps attached to your buildings chilled water supply, would adequately cool this cluster in this room, although altitude may play a part, as most air conditioner units are de-rated to lower tonnage numbers at higher altitudes.

If there is a return air plenum available to building air with sufficient capacity, Aspen can provide you with in-row, portable, rack, or ceiling mounted air conditioning or fan assist units. One popular option is a fan assisted rear door which attaches to every rack, allowing the exhausted heat to be plumbed into the building return air system. The APC™ Rack Air Removal Unit can remove up to 16.5 kw of exhaust heat per rack, and can be vented into building return air space.



If there is access to building chilled water, condenser water, or a glycol loop with enough capacity, Aspen can use chilled water, condenser water, or glycol units mounted in-row (in between racks), as rack doors, or mounted on the floor, wall, or ceiling to cool your cluster. These options circulate air within the room and the rack(s) through condenser units which are plumbed to the building supply, and the return fluid carries away the heat to a roof or outside mounted cooling tower or cooling system.

If neither of these options are available, Aspen can design an entire system for you, plumbing racks or in-room cooling units to an outside wall or roof where condensers or fan units are installed.



For a single rack or a few racks, Aspen can integrate your cluster with totally enclosed self-contained rack enclosures such as the Liebert MCR™ or XDF™ systems. These rack enclosures have integrated cooling and optional UPS systems, and in some cases do not need access to building water or utilize remote heat exchange units, so they can provide a totally stand-alone cooling environment for your cluster, and requiring only electrical power for the air conditioner. Totally self contained units are normally limited to 3 tons (10548 kw) per rack, so our example cluster could be installed in a small room with no external cooling integration by using two of these racking systems.

In most cases, Aspen recommends directed cooling units be installed for your cluster if possible, not room cooling units, to alleviate any possible hot spots inside the room. Directed cooling is installed with your cluster, as rear rack doors or in-row cooling units, or as the rack itself, and allow the hot air to be chilled very close to the heat source, your cluster. This increases effectiveness and results in less overall cost.

Directed cooling solutions for this small room environment can also be applicable to much larger computer facilities.

Compute Facility Cooling

Larger compute facilities, such as existing raised floor computer rooms, often have cooling issues as well. They often were designed for much lower density, power, and cooling requirements than today's high performance clusters have, so supplemental cooling may be needed.

Fan assist doors, in-row cooling units, rear air conditioner doors (plumbed to existing glycol, chilled water, or condenser water sources), totally enclosed racks, and other cooling assist technologies can help ensure that your cluster runs cool and efficiently. Totally self-enclosed cooling racks, such as the Liebert MCR™, can allow you to place a high density Aspen cluster into an over-stressed compute facility without additional demands on an already over-stressed facility.

For even larger installations, Aspen recommends that you consider “cold aisle / hot aisle” solutions, such as the APC InfrStruXure™ or Liebert XD™ Systems. These solutions concentrate the exhaust heat into “hot aisles”, increasing cooling efficiency, and can be retrofitted into almost any computer room. APC InfrStruXure™ Hot Aisle Containment Systems are integrated racks which are installed back to back, and have row end doors and a roof over the hot aisle between the rack rows. In-row cooling units are used to cool the exhaust air from the hot aisle and circulate it back to the front of the rack for re-use by the equipment. Cabling can be accommodated within the racks, on dedicated cable management trays installed on top of the racks, or ran under the floor in a raised floor facility.



No matter what your cooling needs, Aspen sales engineers can help you design the most cost effective solution for your problem. Talk to them about your cooling needs, they can help you.

Commercial Software

Your application(s) may either be open source or commercially licensed. Mathematica, Gaussian, and Ansys CFX, for instance, are all commercial packages that require a system or site license to install and run. Other packages, such as VASP, require site licenses or agreements to operate. In some cases, your organization may already have a site license or purchasing agreement with the vendor who supplies your application(s). Some vendors require direct purchase, while Aspen can resell other applications to you.

Some open source packages are written and supported by companies who themselves can be contracted for support or technical help. In certain unusual cases where you are attempting to compute non-standard problem sets with the application, this type of support may be necessary to achieve optimal performance. Your Aspen sales engineer can help you determine what licenses you need for your application or if you require additional support resources.

Distributions

Some distributions, such as RedHat Enterprise Linux Server, or Novell SUSE Linux Enterprise, are commercially licensed distributions, and require licenses to be installed on your cluster. One or more of your applications may support only specific distributions, and only specific versions of that distribution. For instance, Ansys Release 11.0 on Opteron processors is not certified on RedHat Advance Server 5.x, which is the current enterprise Linux version RedHat sells. While the application may run on RHEL 5.2, Ansys has not yet certified the application on this version, and so will not support it if problems should occur. Carefully check your application requirements before selecting your distribution.

RedHat Enterprise Linux Server is installed and licensed on the master node(s) only, and your compute nodes are licensed as RedHat Enterprise Linux for HPC Compute Nodes. You must purchase a minimum of 4 nodes, but additional compute node licenses can be purchased incrementally. RedHat Enterprise Linux for HPC Compute Nodes requires identical nodes, and is licensed by the number of CPU sockets per node, either 2 per system, or 4 per system (for quad socket systems).

Novell SUSE Enterprise Linux requires that SUSE Enterprise Linux server be installed and licensed on your head nodes, then each compute node is licensed as a Novell HPC licensed node.

Compilers, Utilities, and Debuggers

Some applications have been built, perform best, or only compile, with specific commercial compilers, such as PGI, Intel, Pathscale, NAG, or Absoft. Sometimes applications can achieve considerable performance gains by using a commercial compiler versus GNU C, C++, and Fortran compilers. Applications may work well on one version of a compiler while not compiling with a newer or older version.

Your licensing model for any commercial compiler should match your planned usage pattern. Most compiler vendors offer single seat node-locked licenses, where only a single compilation task at any one time can be ran on one specific node, and multiple seat floating licenses, where many compilations can take place concurrently on one or multiple hosts. Multiple seat floating licenses can be considerably more expensive than single seat, but may be needed for your cluster due to your planned usage.

Compiler vendors can offer discounted pricing for specific customer types, such as academic users, or free for personal use licenses. Free for personal use licenses are almost never appropriate for a cluster installation. Your organization may have a site licensing agreement with a compiler vendor, or may have a current licensing server with floating licenses that your new cluster can access. Other vendors, such as PGI, offer a compiler license that is valid for a free trial period, 2 weeks in the case of PGI. After that trial period, any applications built with that license will stop working.

In almost all cases MPI implementations that will be used to build your application are built with your selected commercial compiler for optimum performance. Aspen can include any of these compilers with your cluster purchase, or utilize existing licenses you transfer to your cluster.

Intel offers the Math Kernel Library (MKL) as a commercial product. The MKL is a suite of highly tuned libraries for performance on scientific, engineering and financial applications that is especially effective on Intel platforms. Intel also offers the VTune™ Performance Analyzer, which can be used to help you speed up some applications run time on Intel processors.

The TotalView Technologies Multi-process debugger can be licensed for your cluster and used to debug both MPI and OpenMP processes. TotalView only supports specific MPI implementations, so be aware that this

might limit your MPI selection.

Research your application requirements and what compilers, utilities, and debuggers you might need to purchase for your cluster by consulting user groups, talking to your software vendor, or speaking to your Aspen sales engineer.

Commercial MPIs

Commercial MPI implementations, such as Platform MPI (formerly Scali), HP MPI, or Intel MPI may provide additional performance for your code(s) in some cases, or be mandated by your application. Some commercial software packages are only ported to specific commercial MPI implementations, and do not work well, or at all, with open source alternatives such as Open MPI, MVAPICH2, MPICH-GM, or other freely available MPI implementations. The research you undertake for your application requirements will generally point out these limitations, or speak to your sales engineer about your applications MPI implementation requirements. Aspen can include any commercial MPI implementations needed with your cluster purchase, or transfer existing licenses.

Parallel File Systems

Commercial parallel file systems, such as GPFS (IBM), or a commercially supported release of GFS (RedHat) and its associated RedHat Cluster Suite require purchase and licensing. In the case of Lustre, the file system itself is open source, but you can purchase direct software service plans if your deployment is non-traditional or highly complex.

Commercial Backup Software

Aspen provides several commercial backup solutions by default, and can implement others based on your requirements. The Storix, Veritas Backup, and IBM Tivoli backup solutions require licensing, and normally employ a client-server model. One or more nodes function as backup servers, with other nodes as clients. As little or no data is kept on the compute nodes, it is not necessary to license each compute server as a backup client. Normally only the node supporting the data storage area for the cluster needs to have a client license.

Schedulers

Aspen highly recommends that you have us install and configure a resource manager / scheduler on your cluster, even if you do not currently use one. Schedulers and resource managers are used within HPC clusters to automate your application execution and allocate cluster resources between different users and groups. Resource managers normally handle knowing what the current state of the cluster is, what resources are currently being used, and asks the scheduler portion to decide how those resources will be used based on rules or policies defined by the cluster administrators. Some utilities combine both functions into one software suite, while others separate the resource manager and scheduler into different applications. To add to this confusion, the term “batch queuing system” is also sometimes used to describe these utilities.

Users on your cluster can submit applications to a queuing system. The scheduler will determine whether resources (nodes, memory, CPUs, interconnect, or any other property) are free on your cluster that can successfully execute the application now. If so, the application is ran; if not, the application is queued for later execution.

All schedulers and resource managers provide command line tools to submit, control, and review job status. Some provide web pages, X graphical user interfaces (GUIs), or even remote clients which can be used to perform those same tasks.

The application output is directed to where the user specifies, perhaps into a user specified file in a specific directory, or into standard output and captured by the resource manager. If e-mail is configured and operational on your cluster, an e-mail can be sent to the user notifying them of the status of their job.

Use of a scheduler and resource manager on your cluster can greatly increase your application productivity by queuing jobs that will be ran as soon as resources allow, not when the users get around to running a job. The scheduler system allows you to utilize your cluster much more efficiently while removing the need for users to be available to start their applications interactively.

Specific open source projects and applications have been written that utilize or must have specific scheduler installations. FMRIB Software Library (FSL), for example, is a brain imaging analysis suite distributed by Oxford. FSL has been written to utilize the Sun Grid Engine (SGE) scheduler, and will not function in a cluster environment without that scheduler installed and operational.

Check to see if your particular application requires or interfaces with any particular scheduler, as that may drive your selection.

Aspen can install and configure several different resource manager and scheduler combinations on your cluster. Some are open source and no charge to you, while some are commercial products which you must purchase. Aspen can procure and install these utilities on your cluster for you, or transfer licenses from existing licenses you might have.

Torque/Maui

The Torque Resource Manager and Maui scheduler are open source projects maintained by Cluster Resources. This is a very popular scheduling solution which scales well to large clusters and provides all the basic scheduling requirements normally needed on a cluster. The Aspen ABC suite connects to Torque to allow your cluster users to submit jobs and review job status via a web GUI if that is desirable, and an X Windows GUI program called “xpbsmon” can be used to gain a graphical view of your current cluster usage. Specific node allocation and other policies can be set by the Maui scheduler.

Moab

Aspen recommends the Moab Cluster Suite© for more complex scheduling needs. Moab is a commercial product supported by Cluster Resources, and runs in conjunction with Torque, replacing the Maui scheduler. Moab provides simple web based job management, graphical cluster administration, and management reporting tools as well as remote clients which can be used on Windows, Linux, or other Unix systems. Some clusters must support multiple users who each have differing resource requirements. Moab can be used to simplify the administrative overhead of larger clusters that are oversubscribed or have many different departments within an organization who might compete for resources on your cluster.

Sun Grid Engine (SGE)

Sun Grid Engine is a scheduler project supported by Sun Microsystems that is used by many cluster communities for scheduling and resource management. SGE supports all the basic scheduling and resource management needs just as Torque/Maui do, but also supports usage accounting and reporting and advanced scheduling algorithms much as Moab does. SGE provides an X Windows GUI for scheduler configuration, job submission, and job status, and is available both as an open source version and as a supported

commercially licensed product.

SLURM (Simple Linux Utility for Resource Management)

SLURM is an open-source resource manager used on many Linux clusters. SLURM is not a sophisticated batching system, but does provide an interfaces to the Maui scheduler and Moab Cluster Suite© . Some user communities rely on SLURM for their batching needs, and SLURM is used on some larger clusters.

PBS Pro

PBS Pro is a commercial grid and cluster resource manager offered by Altair Engineering. PBS Pro excels at connecting different clusters and workstations across your organization into a cohesive managed application execution environment.

Platform LSF

Platform LSF is a resource management and scheduling suite offered by Platform Computing, Inc.. Platform LSF is arguably the most widely deployed commercial batch processing implementation on some of the larger clusters. Platform LSF is also available across your entire infrastructure, including workstations and clusters, and has been integrated with many commercial HPC applications.

Account Management

By default, Aspen delivers your cluster with a standard simplified user schema based on password, shadow, and group files on the master node(s). After adding a user on the master node using standard distribution tools such as “useradd”, executing an Aspen supplied script called “authcopy” propagates the user information to all other nodes in the cluster.

Aspen clusters are configured for host-based authentication between all cluster nodes by default, so any user account that exists on the master node is automatically allowed to log into any node that has that same account. The user password, ssh keys, authorized_keys file, and .rhosts files are not checked, so the user can change their password on the master node(s) at any time without affecting any node connectivity within the cluster.

In normal cases, “authcopy” does not need to be ran after every user password change. If other externally accessed nodes exist in the cluster, password changes can be aliased to automatically perform an “authcopy” after each password change is successfully completed so that any other nodes a user might log in to from outside the cluster immediately receives the new password.

Removing a cluster user is just as simple. Utilize the distributions command, such as “userdel” to remove the user, then run “authcopy” again. The user is removed from every node in the cluster.

Some organizations may operate single sign-on or centralized user management mechanisms such as LDAP, Kerberos, or Network Information System (NIS) which are used to authenticate all users in your organization. Your cluster nodes may be configured to access an external server for user authorization in this case, but slow performance or reliability issues on your organization servers may affect code execution speed or reliability on your cluster.

Secondary or slave servers can be configured on your master node(s), however, Aspen cannot perform this configuration for you without interaction with your organizational user administrators. Extensive coordination between your organization user administrators and Aspen engineering will be necessary to successfully deploy your site specific user authorization schema, and final integration may only be possible after your cluster is installed in its final location.

Security

Aspen normally configures your externally connected hosts with packet filtering firewalls implemented with IP Filters and a Network Address Translation (NAT) gateway that allows internal nodes to access external network resources. IP Filters are used to allow only specific ports to be accessed on your cluster external access points. Your firewalls are configured to allow;

- any communications from internal nodes to external destinations
- ICMP (for pings)
- multicast DNS (optional, needed for some sites)
- Internet Printing Protocol (optional, needed for some sites)
- Secure Shell
- SMTP (e-mail)
- http (web server)
- https (secure web server)
- ABC (if ABC is installed, it utilizes ports 10140 and 10150 for specific ABC access, and ports 40000 through 40500 to proxy other internal cluster web pages)

Additional port rules may be added based on your cluster customization requirements in order to allow communications between your cluster and your organizational network for any additional applications you specify.

In some cases, the Aspen cluster firewall may be customized or disabled based on your custom requirements. Perhaps your administrators wish to perform all security filtering on a centralized firewall system, or your requirements mandate a specific firewall solution be placed in front of your cluster external connections. If the firewall is disabled, your cluster master can be configured as a NAT gateway only, and allow all communications. Your external node firewalls may be customized by filling in your requirements in your Statement of Work.

Additional tools are included based on your distribution, such as tripwire or auditd. These utilities can be configured to your specification, but Aspen engineers will require specific coordination with your site security administrators and clearly stated rule sets to work from. Due to the complexity of some site specific rule sets, this integration is not normally included with your cluster purchase, and may result in additional charges based on your site requirements and the engineering hours needed to perform the integration.

Aspen can configure your cluster to meet National Industrial Security Program Operating Manual (NISPOM) Chapter 8 requirements, but this may limit your distribution choices. This integration may result in additional engineering charges based on your requirements. Other utilities, such as SNORT, or Port Sentry can be configured based on your organizational requirements for additional charges as well.

Speak to your Aspen sales engineer about your specific security requirements. Aspen can help you meet your site-specific security requirements for your cluster deployment.

Windows Integration

Aspen master or storage nodes can be configured with SAMBA to allow your organizations Windows hosts to access and mount data shares on your cluster, or to allow your cluster to mount organization data sources that are located on Windows Servers. This integration is outside our normal cluster build scope, and in complex cases may result in additional charges based on engineering time necessary to perform the integration. Coordination with your organizations Windows administration staff will be necessary, and full integration verification will only be possible once the cluster is installed at your location.

Shipping and Delivery

Unlike many other vendors, Aspen builds your cluster at our facility, not yours. We believe that this results in a better and more standardized product.

After we have finished, you are encouraged to perform remote systems testing of your running cluster while it is still at Aspen. This allows us to fine tune or change any configurations that do not meet your needs before the system is shipped to you, or help you with applications porting or any other issues that may come up.

After build, unit testing, cluster testing, and remote customer testing, your cluster is broken into single racks. Our systems are normally shipped as single racks which are re-assembled at the customer site. Each rack is then palletized, wrapped, and banded for transport, then shipped to you. Aspen normally utilizes FedEx Freight for our system shipping, but we may use other carriers as needed.

Our rack pallets are designed to allow the use of a fork lift or pallet jack to move them. One or more of your palletized racks will have ramps attached to it, which are used to roll the rack off of the pallet. Accessory boxes will be banded and shrink wrapped to one or more of your racks. ***The accessory boxes are also node shipping boxes. Keep them in case you need to ship a node back to Aspen for maintenance.*** International shipments will look different than the figure shown below, as we utilize fully enclosed wooden shipping crates for all International shipments.

Aspen must contract for a delivery truck with a lift gate if your receiving facility does not have a loading dock. Utilize a pallet jack to move the racks while they are still on their shipping pallets. Aspen can arrange for your truck to have a pallet jack at time of delivery if needed. You will need a minimum of a six foot by eighteen foot area to successfully unload a rack from a pallet.

Once your racks have been delivered and removed from the pallets, the racks can be rolled into the final installation location using the integrated rack casters. The Aspen standard 42U rack is 6', 7" tall. Your facility must have adequate clearance through all doorways and hallways between the unloading area and final destination. Please check for obstacles such as door jambs and automatic door mechanisms which might cause clearance problems. A minimum of two personnel is normally needed to move a fully loaded 42U rack. If your facility does not have adequate overhead clearance to roll the rack into its final installation location, systems can be removed from the rack so that it can be tilted, or special dollies can be used to tilt the rack while it is being transported.

Aspen can receive and install your cluster as part of an on-site visit should that be desired. Discuss your shipping and delivery options and any special needs you have with your Aspen sales engineer.

On-Site and Training

Integration of your Aspen cluster into your site infrastructure can be labor intensive. Aspen can assist you in these tasks if you need help. Aspen can coordinate with you to have one or more cluster engineers meet the truck delivering your cluster, and take your cluster from one or several disconnected racks sitting on your computer room floor to a fully functioning HPC resource integrated into your site network and performing useful work.

You may customize the task list for the Aspen on-site visit in any way you wish. If you have personnel available for such physical tasks as helping our engineer move the racks into position, it might be more cost effective to contract for only one engineer. If the scope of your project is larger, or you wish all installation tasks be handled by Aspen, multiple engineers can be dispatched to your facility.

Tasks such as SAMBA integration for Windows host communications, local backup system integration, site network implementation, modifying your cluster to interface with your site user authentication schema, or specific site firewall and security customization can best be accomplished on-site by Aspen engineers in conjunction with your site personnel. Such site specific tasks are in many cases more difficult or impossible to verify if they are done prior to shipping, because they can really only be tested when the cluster is on-site and attached directly to your organizational network. By knowing your requirements in advance, Aspen can prepare your cluster as much as possible while it is at Aspen to allow on-site engineers to quickly and efficiently finalize the localization.

An installation on-site visit can also be used to familiarize your administrators with your new Aspen cluster and teach them how best to interface with and manage the system. Many customers also find this visit an excellent time to have our engineers work directly with your users, covering such topics as code execution, proper environment initialization, and job submission if a queuing system is involved. If this is your first HPC cluster, or your users are not familiar with cluster application execution, Aspen engineers can conduct informal training sessions in addition to normal installation tasks to get your users up to speed and running code as quickly as possible.

If a more formalized training environment is desired, Aspen can tailor a specific formal training class for your users or administrators, with a curriculum that you and Aspen have worked out before hand. Please be aware that some preparation is necessary in order to allow Aspen to assemble your custom curriculum, so this option will result in some additional expense and incur some delay in the scheduling of the site visit.

Your Aspen sales engineer can include 2, 3, 5 day or longer on-site visits as part of your cluster purchase for one or more Aspen engineers. You may use this time to have your on-site Aspen engineer(s) accomplish any tasks you wish. Work with your sales engineer to define the goals you wish to accomplish during the on-site visit so that Aspen can better prepare to meet those goals.

Next Steps

We've covered a lot of different areas in this [Detailed Buyers Reference](#). Hopefully some of the information presented here has helped you to better understand your cluster requirements as well as some of the options available to make your Aspen cluster purchase best fit your needs.

Your Aspen sales engineer is there to help you navigate the choices necessary to define the best possible solution for your needs, not some hypothetical standard cluster user. Your sales engineer will work with you through the quoting process, helping you every step of the way, and adjusting your configuration to meet your budget while achieving the best possible technical solution within the constraints you specify. This process often entails multiple quotes and requirements review sessions with your sales engineer to help you define your system more and more tightly.

Your sales engineer will furnish you with two more documents with your first quote(s) or shortly after. They are the [Aspen Configuration Guide](#), and the Statement of Work for your specific order. Both documents can be accessed as web links or downloaded from our web site for your convenience.

We ask you to fill out and submit a Statement of Work for your system by the time you submit a purchase order. Aspen uses your completed Statement of Work to perform a final detailed engineering review of your hardware and software configuration, matching them to the requirements you outline in your Statement of Work. This process helps to ensure that we fully understand your requirements and can meet them to your satisfaction.

The Aspen Configuration Guide explains the Statement of Work questions in more detail, and provides additional and specific technical advice and recommendations for your cluster configuration. Don't worry too much if you can't answer some of the questions in your Statement of Work. Your Aspen sales engineer can help you, or coordinate teleconferences or meetings with Aspen cluster engineers to address these questions. Aspen can even assign a specific cluster engineer to work with you and explain every option in more detail, or perhaps work with other site or organizational administrators you identify to determine your technical requirements and configuration.

Aspen deals with many customers, some with complex requirements requiring specific and involved technical configurations, and others who are well served by our standard cluster implementations. Our process works well for both types of customers, and we will do whatever we can to make your specific Aspen purchase and ownership experience as trouble free and productive as possible. Call Aspen to discuss your clustering needs today, we stand ready to help you.

We congratulate you for reading this far! Many people will not have the patience to do so, but we hope that this document has provided some useful information and will help you in your cluster configuration and acquisition.

