HP High Performance Clusters LC Series Design Considerations

March 2004 (First Edition)
Part Number 366713-001

Contents

Introduction

There are three LC Series Solution offerings based on different 1U densely packaged servers called compute nodes.

- The HPC LC 1000 Series solution is based on ProLiant DL140 compute nodes.
- The HPC LC 2000 Series solution is based on ProLiant DL360 compute nodes.
- The HPC LC 3000 Series solution is based on ProLiant DL145 computer nodes.

Each solution offers a unique set of features and performance capabilities.

This document will assist you in determining what size of cluster you need, and which interconnects are best for the applications you intend to operate. It helps you prepare to use the LC Series Design and Configuration Guides to configure an LC Series Cluster.

Migration to Industry Standard Clusters for HPC

The use of industry standard servers in high performance compute clusters for serial, parallel, and message passing applications has been growing rapidly in the last few years. Several areas have contributed to the more extensive use of these servers in high performance clusters.

- The creation of the casual computing market (office automation, home computing, games and entertainment) has provided system designers with new types of cost-effective components. The COTS (Commodity off the Shelf) industry has provided fully assembled subsystems (microprocessors, motherboards, disks and network interface cards). Mass-market competition has driven the prices down and reliability up for these subsystems.
- The latest implementation of Intel Xeon DP and the emergence of the AMD Opteron processor offer incredible price per performance capability that before was present only in expensive RISC processors.
- The development of publicly available software such as the Linux operating system, GNU compilers, programming tools, and the MPI and PVM message passing libraries provide hardware independent software.
- The emergence of the Open Source community as well as the university programs for advanced information technology have spawned huge numbers of libraries and algorithms, which have either been accepted in the industry or extended and productized by hundreds of independent software vendors. Programs like the High Performance Computing Cluster (HPCC) program have produced many years of experience working with parallel algorithms.
- An increased reliance on computational science, and therefore an increased need for high performance computing, has turned more researchers and developers to focus their expertise and experience into making these systems perform and work better.

The combination of these conditions: hardware, software, experience and expectation, has provided the environment needed for the LC Series High Performance Compute cluster which is based on industry standard servers.

The LC Series clusters are based on the Beowulf concept, which is one approach to clustering commodity hardware components to form a parallel virtual computer. It is a system that usually consists of one control node and one or more compute nodes connected via Ethernet or some other network or system area network interconnect such as Myrinet.

LC Series clusters are ideal for tackling very complex problems that can be split up and run in parallel on separate computers. Not every problem can be approached in parallel, however. The LC Series can also be used for consolidated serial applications and complex message passing applications.

For consolidated serial applications, many independent jobs can be allocated and managed within the confines of the cluster. In this form of cluster computing, multiple independent jobs, all running on their own machines with different data inputs and outputs, can execute and save administrative time and resources over a series of independent machines. In this case, a higher latency interconnect is acceptable because there is little to no interaction between jobs operating across the nodes in the cluster. Platform Computing, Altair’s PBS Pro, and other job management systems can provide very efficient and controlled job processing allowing the customer to get the most out of their cluster when the cluster is used in this fashion.

In the message passing case, which is by far the most complex, the cluster interconnect is critical. Programs must interoperate between compute nodes, transferring commands and data to complete their routines. If there is a slow down due to latency or collisions, the application can stall, abort, or in some extreme cases produce an incorrect result.

Parallel industry standard clusters are replacing Massively Parallel Processor (MPP) systems except in the most extreme application cases. An MPP system is typically larger, proprietary, and has a lower latency and higher bandwidth system interconnect network than a parallel industry standard cluster. These MPP, or Vector, computers are needed due to the requirement for highly critical or classified applications where performance is the only concern. As the industry standard processors have been improving, the need for these MPP machines has diminished. Cluster programmers need to consider locality, load balancing, granularity, and communication overhead in order to obtain the best performance. Even on shared-memory machines, many programmers develop their programs in a message-passing style. Programs that do not require fine-grain computation and communication can usually be ported and run effectively on Beowulf clusters.

An industry standard class cluster computer is distinguished from a GRID or NOW (Network of Workstations or servers) by several subtle but significant characteristics. First, the nodes in the cluster are dedicated to the cluster. This helps ease load balancing problems because the performance of individual nodes is not subject to external factors. Also, since the interconnect network is isolated from the external network, the network load is determined only by the application being run on the cluster. This eliminates one of the key flaws of GRID or NOW systems — unpredictable network latency. All the nodes in the cluster are within the administrative jurisdiction of the cluster. For example, the interconnect network for the cluster is not visible from the outside world, so the only authentication needed between processors is for system integrity. On a GRID or NOW, you must also be concerned with network security.

Building your Cluster

To build an industry standard cluster you need the following components:

- Control node
- Compute nodes
- Cluster interconnect

Design considerations for each of these components are addressed in the following sections.

Control Node

In the LC Series clusters, a single control node typically can handle the user interaction and control needs of the cluster. The control node is the basis for all application interface and administration in the cluster. The ProLiant DL380 is the preferred server for the control node because it offers the most operational features in a small 2U space. The DL380 offers up to 12GB of memory, dual processors for 2P performance, and up to 700GB of onboard SCSI storage for data staging and parsing to the compute nodes. The DL380 also has 3 PCI slots for added adapters such as Fibre Channel HBA for SAN extensions, although HP recommends that large storage should be an independent subsystem due to the complexity of cluster compute operations. The DL380 has two onboard 10/100/1000 Ethernet NICs to connect to the In Band (IB) management network and the external LAN for the user interface to submit and retrieve jobs allocated to the cluster. The DL380 also supports the Integrated Lights Out Service co-processor for Out of Band (OOB) management and remote administrative functions, which are critical to cluster health and operations. In addition, the DL380 offers multiple enterprise class features regarding system reliability with optional redundant power, cooling and hot plug disk storage.

The DL145 server can take the place of the DL380 server as the control node in Opteron based clusters if desired. Using this 1U server as the control node keeps the processor type consistent between the control and compute nodes but eliminates some high availability, expandability and management features.

Compute Nodes

Three 1U server offerings are available as compute nodes in LC Series clusters. Each solution offers a unique set of features and performance metrics. The compute node is the most critical component of the cluster configuration. When selecting the compute nodes, you must consider not only the performance of the application, but also the communication between compute nodes and communication to storage. Other important considerations concerning the compute nodes are management, monitoring, and administration of the cluster itself.

In order to determine the number and type of compute nodes, you need to ascertain the performance requirements for the application that is to operate on the cluster. This requires the use of some common metrics and the ability to understand those metrics in terms of peak and sustained performance. Don’t let peak performance — which might never be attainable by a real application — sway a decision to go a certain way when sustained performance is what really counts. Any of the decisions made when designing and configuring a LC Series cluster should rely on a comprehensive understanding of the application to be run. Metrics to consider include but are not limited to

- Floating Point Operations per Second (FLOPS)
- SPECfp performance
- SPECint performance
- Cache size
- Cache bandwidth
- Total memory
- Memory bandwidth
- Storage bandwidth
- System interconnect throughput
- System interconnect latency
- Bisection bandwidth

The following tables show performance summaries for the LC 1000 Series and LC 2000 Series clusters. Updated information will periodically be released to the HP website at www.hp.com.

LC 1000 Series Model Matrix

Cluster			**Processors / Node**	**Memory / Node**		**Disk / Cluster**		**GFLOPS/ Cluster***	
Cluster	**Model**	**Interconnect**		**Base**	**Max**	**Min GB**	**Max GB**	**1P**	**2P**
1 Control Node, 16 Compute Nodes									
LC1016	F	10/100 Fast Ethernet	2	1GB	4GB	131.6	329.6	**35.7**	**65.0**
LC1016	G	Gigabit Ethernet	2	1GB	4GB	131.6	329.6	**44.9**	**82.2**
LC1016	M	Myrinet 2000	2	1GB	4GB	131.6	329.6	**51.1**	**95.2**
1 Control Node, 32 Compute Nodes									
LC1032	F	10/100 Fast Ethernet	2	1GB	4GB	292	585.6	**68.6**	**123.9**
LC1032	G	Gigabit Ethernet	2	1GB	4GB	292	585.6	**87.1**	**159.3**
LC1032	M	Myrinet 2000	2	1GB	4GB	292	585.6	**102.8**	**192.1**
1 Control Node, 64 Compute Nodes									
LC1064	F	10/100 Fast Ethernet	2	1GB	4GB	515.6	1097.6	**131.6**	**237.5**
LC1064	G	Gigabit Ethernet	2	1GB	4GB	515.6	1097.6	**164.6**	**298.4**
LC1064	M	Myrinet 2000	2	1GB	4GB	515.6	1097.6	**205.7**	**384.1**
1 Control Node, 128 Compute Nodes									
LC1128	F	10/100 Fast Ethernet	2	1GB	4GB	1028	3616	**246.8**	**444.6**
LC1128	G	Gigabit Ethernet	2	1GB	4GB	1028	3616	**309.1**	**558.5**
LC1128	M	Myrinet 2000	2	1GB	4GB	1028	3616	**427.8**	**798.3**

LC 2000 Series Model Matrix									
Cluster			Processors / Node	Memory / Node		Disk / Cluster		GFLOPS/ Cluster*	
Cluster	Model	Interconnect		Base	Max	Min GB	Max GB	1P	2P
1 Control Node, 16 Compute Nodes									
LC2016	F	10/100 Fast Ethernet	2	1GB	5GB	36.4	5135.2	**39.7**	**72.3**
LC2016	G	Gigabit Ethernet	2	1GB	5GB	36.4	5135.2	**49.9**	**91.4**
LC2016	M	Myrinet 2000	2	1GB	5GB	36.4	5135.2	**56.8**	**105.7**
1 Control Node, 32 Compute Nodes									
LC2032	F	10/100 Fast Ethernet	2	1GB	5GB	36.4	9813.6	**76.2**	**137.6**
LC2032	G	Gigabit Ethernet	2	1GB	5GB	36.4	9813.6	**96.8**	**177.0**
LC2032	M	Myrinet 2000	2	1GB	5GB	36.4	9813.6	**114.3**	**213.4**
1 Control Node, 64 Compute Nodes									
LC2064	F	10/100 Fast Ethernet	2	1GB	5GB	36.4	19170.4	**146.3**	**263.8**
LC2064	G	Gigabit Ethernet	2	1GB	5GB	36.4	19170.4	**182.8**	**331.3**
LC2064	M	Myrinet 2000	2	1GB	5GB	36.4	19170.4	**228.5**	**426.8**
1 Control Node, 128 Compute Nodes									
LC2128	F	10/100 Fast Ethernet	2	1GB	5GB	36.4	37884	**274.3**	**494.0**
LC2128	G	Gigabit Ethernet	2	1GB	5GB	36.4	37884	**347.2**	**614.4**
LC2128	M	Myrinet 2000	2	1GB	5GB	36.4	37884	**475.4**	**887.0**

The LC 1000 series uses DL140 servers and the LC 2000 Series uses DL360 servers. 2.4 GHz processors were used in obtaining the performance data in these tables. One reason that the performance numbers for the DL140 based systems are different from the DL360 based systems, even though the same processor speed was used, is that the DL360 systems use interleaved memory. This accounts for an 8% to 14% performance difference across these two Xeon DP systems. This is based on multiple execution runs of many types of performance benchmarks since one run will not give a good estimate. This performance difference can be very significant depending on your applications and cluster size. You must assess performance needs, versus expense, versus administrative costs in making your design decisions.

A speed calculation based on processor clock rates and peak GFLOPs may not give you the result you expect. Factor a 30% to 35% degradation factor for estimating sustained GFLOPS, which will be a more accurate measure of performance.

Cluster Interconnect

The Cluster Interconnect is by far the most important factor for applications that rely heavily on message passing.

Cluster control messages and data both move over the Cluster Interconnect. Any network interconnect — local area network (LAN) or system area network (SAN) — can be used to connect the cluster nodes to each other, although some interconnects are faster than others.

The following interconnects are currently supported on High Performance Clusters LC Series.

- Myrinet
- Gigabit Ethernet
- Fast Ethernet

Two very important measures are used to characterize system interconnects. The first is speed. How fast does data pass through the network? Choices of speed run from Fast Ethernet at 10/100 megabits per second to specialized network technologies clocking 1 gigabyte per second or higher.

The other measurement used to characterize system interconnects is latency. How long does it take to prepare a packet of information for transmission and get it onto the network? How long does it take to propagate through the network and all its various potential stages or hierarchies? Latencies can range from milliseconds in some older technologies to just a few microseconds in newer ones.

The two measurements of speed and latency taken together define “throughput” — the total amount of useable aggregate data that can be moved from one system to the other. Clearly, throughput can greatly affect the overall performance of a cluster.

As you review the matrices, you will notice how slow 10/100 Fast Ethernet clusters are compared to the Gigabit and Myrinet clusters. This is due to the protocol speed and latency issue discussed above. If the application target is heavily message passing, such as Fluent or LS-Dyna, then configure the cluster using the Myrinet cluster interconnect. Although expensive, the Myrinet cluster interconnect provides the best bandwidth and latency in this price class of cluster.

You can start with a small cluster. When you need to support more users or need more nodes to operate on a problem, you can simply add them to the existing cluster. Although there may be some small amount of degradation in communications efficiency as more nodes are added to the cluster, scaling is essentially linear. If you double the number of nodes in your cluster, you basically double the performance.

Another aspect of scalability involves interconnecting clusters. System interconnects not only connect the nodes within a cluster, but they can also connect small clusters together to make larger ones. This of course needs to be a consideration when determining the size of the cluster interconnect switch. The LC Series Design and Configuration Guide reference designs specify the maximum number of compute nodes allowed in the design without having to add another interconnect switch.

Using the Design and Configuration Guide reference designs you can easily design a cluster of any size up to 128 nodes with GigE or Myrinet interconnect or 192 nodes with Fast Ethernet interconnect. Scaling over these limits to sizes like 256, 512, or 1024 nodes requires additional engineering and hardware in the form of multiple switches with extensive cabling topologies. Also, additional or different software may be required depending on the problem being solved. HP Consulting and Integration Services have implemented ProLiant clusters in the 1000+ node category and can assist you with designing these complex clusters.

Application and Interconnect Considerations

Knowing the types of applications to be run on an HPC cluster will help you decide what type of cluster interconnect is needed in the cluster.

Clusters can be used to solve one big problem at a time or to solve multiple problems at the same time.

- Solving ONE BIG problem at a time is “capability”
- Solving MULTIPLE problems at the same time is “throughput”

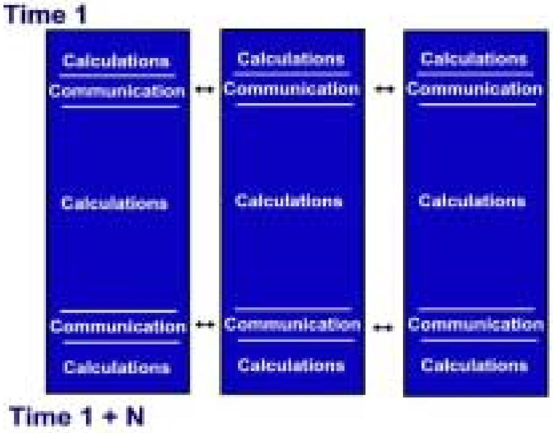
Most clusters are used both ways. Customers with big problems to solve, have to buy a cluster big enough to solve their biggest capability problem. They generally have plenty of smaller problems, however, which get run on the cluster when it is not in use solving the big problem. Throughput usage is more common but capability usage is more challenging for system designers.

Application Granularity

Some applications will work better with different types of cluster interconnect depending on their granularity. A key to efficient distributed computing is data locality. Calculations on local data are often orders of magnitude more efficient than calculations which require data communication. Granularity is the ratio between computation and communication.

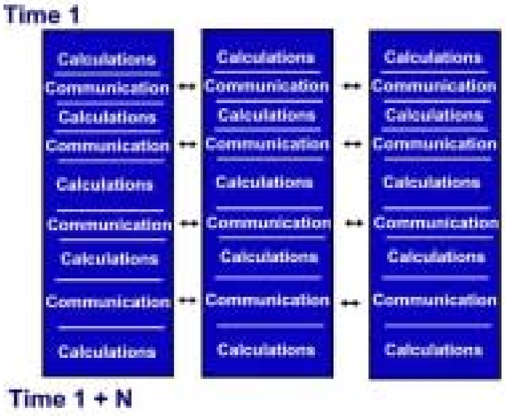
Coarse Grain

Algorithms with a high computation to communication ratio are said to have coarse-grain parallelism. Coarse grain algorithms often offer better scalability even though load balancing may be more difficult. Algorithms with an exceedingly low amount of communication are said to be “embarrassingly parallel”. Examples of highly parallel, coarse grain applications include: image or frame processing, sequence research in life science, and parameter study (stochastic crash tests).



Fine Grain

Algorithms with a low computation to communication ratio are said to have fine-grain parallelism. Examples of fine grain applications include finite element analysis or computational chemistry.



Knowing whether your applications are either fine grained or coarse grained will dictate what type of cluster interconnect is best suited for your cluster. In most groups, there is a mix of application types, but knowing what the majority is will help make the right determination for cluster interconnect, which is an expensive component of the cluster.

In summary, when reviewing your application mix, keep these tips in mind.

- Serial applications (LC Series used as a Compute Farm)
  - Independent processes
  - Performance characterized by CPU performance.
  - System is used for throughput.
- Multithreaded applications (LC Series used as a Parallel Machine)
  - Pthreads, OpenMP
  - One thread per CPU (ideal) working on shared memory.
  - Performance characterized by node performance.
- Distributed Memory applications (LC Series as a Message Passing (MSG) Machine)
  - Cooperating processes.
  - Typically message passing applications (MPI).
  - Application throughput characterized by both node and interconnect performance.
  - Entire system is used by the application

Cluster Size and Interconnect Considerations

The following table will assist you in making the decision on interconnect and cluster size.

Cluster Interconnect and Cluster Sizing Recommendation Matrix						
	Cluster Size					
Job Mix	**16 Nodes**	**32 Nodes**	**48 Nodes**	**64 Nodes**	**96 Nodes**	**128 Nodes**
Serial Only	Fast E	Gig E	Gig E	Gig E	Gig E	Gig E
Serial / Parallel	Fast E	Fast E	Gig E	Gig E	Gig E	Gig E
Serial / MSG	Gig E	Gig E	Myrinet	Myrinet	Myrinet	Myrinet
Parallel Only	Fast E	Fast E	Fast E	Gig E	Gig E	Gig E
Parallel / Serial	Fast E	Gig E	Gig E	Gig E	Gig E	Gig E
Parallel / MSG	Gig E	Gig E	Myrinet	Myrinet	Myrinet	Myrinet
MSG Only	Myrinet	Myrinet	Myrinet	Myrinet	Myrinet	Myrinet
MSG / Serial	Myrinet	Myrinet	Myrinet	Myrinet	Myrinet	Myrinet
MSG / Parallel	Myrinet	Myrinet	Myrinet	Myrinet	Myrinet	Myrinet

Storage Considerations

The LC Series design point was to build a computing machine and allow for local storage to be an option based on the needs of the operating system, cluster manager, and applications. In the LC Series, the DL380 and the DL360 nodes come without storage, which must be ordered as an option. In the DL140 and DL145 nodes, ATA drives are included. The following are some rules of thumb to follow:

- If using Linux, drives are required on the control node but may not be needed on compute nodes, depending upon cluster management and applications to be used. It is recommended that all compute nodes have at least one drive for operating system needs or as a scratch space for applications.
- If using Windows, drives are required on all nodes.

A number of storage subsystems have been defined for use in LC Series solutions to meet customer needs. Divided into a range of subsystem sizes and capacities, these storage options include entry and low-end NFS subsystem options and entry to high-end Global File System solutions.

HP software partners have tested their application codes on the storage subsystems and have given server recommendations for the best data I/O rates. These storage subsystems are designed to link to the cluster through the Gigabit Ethernet In Band Management network. This provides the application full access to the cluster interconnect if needed as in message passing and mixed use clusters. The In Band Management network has been designed in the LC Series for both compute node expansion and storage interconnect.

Next Steps

This document has provided some information to consider when determining the number of nodes and interconnect type to build into an HPC cluster. The next step is to use the Design and Configuration Guide for the compute node of your choice, DL140, DL360 or DL145, to design a cluster to meet your needs. The Design and Configuration Guides allow you to specify the number of compute nodes initially needed in your cluster, the amount of expansion you may need in the future, and the type of interconnect you need. Using the process in the guide you will be able to generate a detailed list of components needed to build the cluster, right down to the numbers of cables and cable lengths needed. Additionally, you will be able to specify storage components and software components that may be needed in the cluster.

The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

HP High Performance Clusters LC Series Design Considerations

March 2004 (First Edition)
Part Number 366713-001