

# Horizontal Scaling Fabrics for Sun Fire™ V60x and V65x Servers: InfiniBand

## Solution Brief



### Key Feature Highlights

#### Clustering

With horizontal scaling, computing can be scaled to thousands of CPUs.

#### Sun Fire™ V60x and V65x Servers

- Dual 3.06-GHz or 2.8-GHz processors
- 533-MHz Front Side Bus
- Up to 12 GB of ECC memory (V65x only)
- Built-in dual Gigabit Ethernet
- Dual PCI-X buses
- Separate on-board service processor for remote management
- Supports standard Linux distributions and Solaris™ 9 Operating System (x86 Platform Edition)

#### Cluster Fabric Application Factors

- Scalability
- CPU utilization
- Latency
- Bandwidth
- Price/performance

#### Application Industries

- Education
- Energy
- Entertainment
- Financial
- Government
- Healthcare
- Life sciences
- Manufacturing

### Scaling CPUs to Harness Massive Compute Power

Solving scientific, engineering, and business problems requires ever-increasing compute power. More and more of today's emerging applications use computers to model real-world phenomena such as weather simulations and to analyze and search for solutions hidden in the enormous amounts of data gathered through data collecting sensors, computer transactions, and various human endeavors such as locating a new gene in the human genome database. Similar computing techniques are used to design new products, such as the latest wireless chip, or to create blockbuster movies in the entertainment industry.

The need for large compute power can be addressed with high-performance compute clusters using a large number of independent compute servers connected over a high-speed network fabric. Thousands of CPUs can be configured to provide gigaflops of compute power using appropriate interconnect technology. For many companies and institutions, scaling the computing infrastructure by adding independent servers, or horizontal scaling, has become an effective means to address the increased demand for compute power.

A leader in helping companies scale vertical and horizontal computing environments, Sun™ is committed to providing an end-to-end architecture based on open technology and standards for building a computing infrastructure. With vertical scaling, services are scaled within the system—resources such as CPUs, memory, and I/O can be incrementally added to the server over time to increase performance. And with CPUs sharing the same memory, applications can leverage the extremely fast data transfer speeds between processors and memory. To support vertical scaling, Sun offers servers that provide scaling of up to 106 processors, support for 64-bit computing, and the Sun Fireplane Interconnect, which provides maximum throughput among the processors, memory, and I/O subsystem.

With horizontal scaling, independent servers are interconnected in a high-speed network fabric—managed by a software layer to coordinate the computation in the cluster—to provide even more computing power. Computing can be scaled to thousands of CPUs to meet the demand for compute resources. Horizontal scaling enables low-cost, entry-level servers to be added as incremental compute components to a cluster fabric.

**INTERCONNECT  
OPTIONS:  
ETHERNET  
INFINIBAND**



**Figure 1:** Sun Fire V60x and V65x servers are designed to work with Ethernet and InfiniBand interconnect technologies. InfiniBand technologies are supported by their respective third-party vendors.

With high levels of computer power and memory density, Sun Fire™ V60x and V65x servers provide an ideal platform for building clusters. Sun Fire V60x and V65x servers feature dual 3.06- or 2.8-GHz Intel Xeon processors, a 533-MHz Front Side Bus, up to 12 GB of memory, and support for the Linux operating system. In addition to built-in support for 100-BaseT and Gigabit Ethernet, Sun Fire V60x and V65x servers also feature two PCI-X buses to support high-speed, high-performance interconnect technology such as InfiniBand. A separate on-board service processor is also included for remote management. In addition, the small form factor of the rack-optimized Sun Fire V60x and V65x servers helps companies to assemble large numbers of CPUs within a limited space.

Yet one of the challenges of implementing an effective horizontal scaling infrastructure is choosing the right fabric interconnect technology. The type of interconnect can be one of the most important factors in performance when considering such factors as CPU utilization, latency, and bandwidth. This brief provides insight into technical applications and their cluster requirements and the criteria for selecting a cluster fabric interconnect.

### Cluster Components

A cluster is composed of a managed collection of computer systems and typically consists of the following components (see Figure 2):

- Multiple compute nodes
- Cluster interconnect infrastructure, which can include Ethernet and InfiniBand or other interconnect technology such as from Myricom or Quadrics
- Management infrastructure, which can include terminal servers, for server health management and software provisioning
- Optional storage infrastructure, which can include storage area networks (SAN) or network attached storage (NAS) or emerging InfiniBand interconnect switches that provide gateways to consolidate paths to external storage

### Cluster Software

Cluster systems also require resource and management software such as Sun ONE Grid Engine and MPI (Message Passing Interface) to manage workloads.

Sun ONE Grid Engine software is a distributed management tool designed to optimize utilization of software and hardware resources in a network. MPI is an open-source, industry standard that allows a cluster of servers to use a distributed memory model for message passing.

### Cluster Optimization

Cluster systems can be optimized for throughput, performance, or high availability.

“Embarrassingly parallel” or throughput clusters are ideal for applications that start multiple, independent processes running on multiple servers but require little communication between the server nodes.

MPI-parallel or performance clusters are ideal for applications that divide tasks into many parallel tasks that require interdependent communication. MPI-parallel clusters require high-performance interconnect and interprocess communications. A Beowulf compute cluster, which runs on the Linux operating system and other open source soft-

ware such as the MPI standard, is an example of an MPI-parallel cluster.

In the event of a failed server or failed component, clusters configured for high availability continue to provide service. Although this brief focuses on clusters based on embarrassingly parallel and MPI-parallel computing, the concepts discussed can also apply to high availability cluster systems.

### Application Requirements

Applications running from different cluster systems, whether embarrassingly parallel or MPI-parallel, have different requirements.

For example, BLAST (Basic Local Alignment Search Tool), a bioscience application used to search DNA patterns in gene databases, requires very large data sets. BLAST jobs typically run on embarrassingly parallel cluster systems, which can perform multiple processes at the same time.

Table 2 at the end of this brief lists applications in different industries, ideal cluster type, and typical requirements. The table shows example applications requiring Ethernet or InfiniBand interconnects. Some applications require larger memory capacity and 64-bit computing, which can be provided

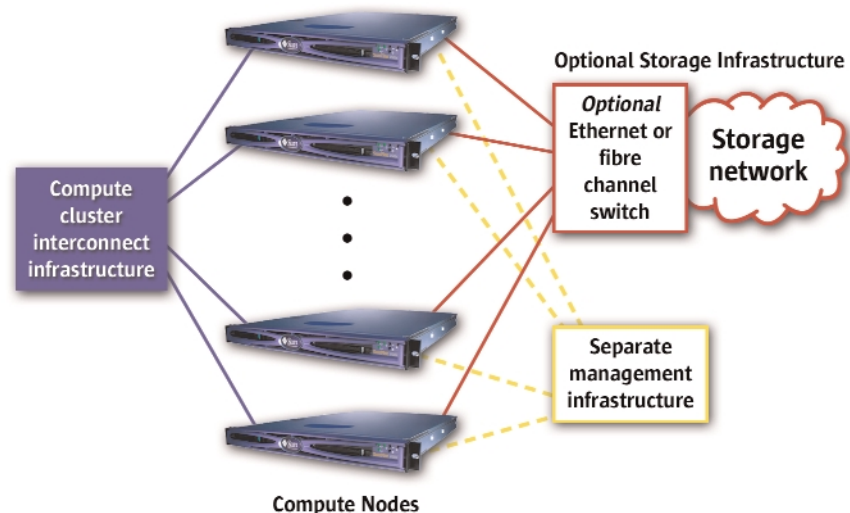


Figure 2: Cluster components

by SPARC®-based Sun Fire servers. In cases where applications require large data sets, a separate storage network may be needed. And for applications with heavy graphics requirements, a separate graphics system may also be appropriate.

### Cluster Interconnect Considerations

The type of interconnect used in a cluster system is a critical consideration as applications scale out to larger numbers of nodes. Organizations that are implementing horizontal scaling must address several important factors when considering the type of interconnect to use, including:

- Number of nodes to be scaled
- CPU utilization of the interconnect, which is the overhead of the interconnect (for example, Ethernet uses TCP/IP which has high CPU utilization)
- Latency that can be tolerated as messages pass between nodes
- Bandwidth of the interconnect (how quickly data can move between nodes)
- Price/performance

### Type of Interconnect

When choosing an interconnect, organizations must consider the kind of applications that will run on the cluster. For example, standard Ethernet can typically handle the load generated by embarrassingly parallel applications in which very little data traffic is needed and CPU utilization, latency, and bandwidth are not significant factors.

When more data traffic, less CPU utilization, lower latency, and higher bandwidth are needed, organizations should consider higher-performance interconnects, which can support a broader range of applications and accelerate application performance.

For MPI-parallel applications—in which less CPU utilization, higher bandwidth and lower latency are required—specialized interconnects, such as InfiniBand-based switches, are more appropriate. These interconnects generally

have a PCI card installed into the server that connects to a high-speed switch.

The InfiniBand architecture is a new industry standard for high-speed, point-to-point switching architecture similar to mainframe network channels. InfiniBand overcomes traditional server networking CPU bottlenecks while providing considerable economies of scale. The result is less CPU utilization, higher bandwidth, and lower latency.

Table 1 compares some of the different parameters of Ethernet and InfiniBand.

### Applying the Appropriate Interconnect

Table 2 lists the cluster interconnect type(s) recommended for example applications in different industries.

Whatever the type of interconnect used, Sun and its partners can help companies build robust and flexible cluster systems. Sun Fire V60x and V65x servers, featuring dual, 10/100/1000-BaseT Ethernet ports, provide on-board support for Gigabit Ethernet. Sun Fire V60x and V65x servers are also compatible with InfiniBand-based cards and switches from Topspin Communications that use silicon (chips) from Mellanox Technologies. These partners provide full

support for their interconnect technology. Sun provides support for Sun Fire V60x and V65x servers.

For applications requiring 64-bit computing, Sun also offers SPARC-based servers.

### Conclusion

Increasingly, organizations are turning towards horizontally scaled systems to address the demands that high-performance applications place on their computing infrastructures. By choosing the cluster fabric interconnect appropriate for its needs, companies can build compute clusters that enable them to be more efficient, deliver products and services faster to market, and remain competitive.

Sun Fire V60x and V65x servers, based on an open-standards architecture, the Linux operating system, and fast computing, are ideal for placement as compute nodes within cluster systems.

For contact information on Mellanox, go to [www.mellanox.com](http://www.mellanox.com). For more information on Topspin, go to [www.topspin.com](http://www.topspin.com). For more information about Sun Fire V60x and V65x servers, go to [www.sun.com](http://www.sun.com) or contact your local Sun representative.

Parameter	Ethernet	InfiniBand (IB)
Scalability (number of nodes)	1000's demonstrated	1000's projected
CPU utilization	High (TCP/IP)	Low (DMA)
End-to-end latency (zero byte)	60 µs	5.4 µs
Unidirectional bandwidth (large packets): Bidirectional bandwidth (large packets):	~100 MB/sec. ~200 MB/sec.	864 MB/sec. 780 MB/sec.
Scalability of CPU/memory limited applications	High	High
Scalability of latency sensitive applications	Low	High
Scalability of bandwidth sensitive applications	Low	High
Scalability of latency and bandwidth sensitive applications	Low	High
Price	Low	Medium
Grid and MPI support	Yes	Yes
Standards organization	IEEE 802.3	IB Trade Assoc.
Acceptance/Maturity	High (since 1970s)	Emerging (since 2001)

**Table 1:** Parameters of Ethernet and InfiniBand. Note that the content of this table is subject to change as InfiniBand interconnect technology continues to evolve. Data provided by Mellanox and Topspin, based on industry-standard Intel servers.

Application Area	Type of Application	Type of Interconnect	Other Requirements	NAS or Fibre Channel
<i>Bioscience</i> • BLAST • FAST, FAST-A	• Embarrassingly parallel	• Ethernet	• Large data sets	Yes
<i>Scientific Research and R&amp;D</i> • Weather simulation • SETI	• MPI-parallel • Embarrassingly parallel	• InfiniBand • Ethernet	• Large data sets • Large data sets	Yes
<i>Government and Defense</i> • Data analysis	• MPI-parallel	• InfiniBand	• Large memory (64-bit computing) • Large data sets	Yes
<i>Electrical Design and Engineering Analysis</i> • Verifications (VCS) • Timing	• Embarrassingly parallel • MPI-parallel	• Ethernet • InfiniBand	• Shared data	Yes
<i>Geoscience and Geoengineering</i> • Seismic processing	• Embarrassingly parallel • MPI-parallel	• Ethernet • InfiniBand	• Large data sets	Yes
<i>Mechanical Design and Engineering Analysis</i> • Computational fluid dynamics (CFD) • Crash simulation • Structure analysis	• MPI-parallel	• InfiniBand	• Large memory (64-bit computing) • Large data sets	Yes
<i>Chemical Engineering</i> • Molecular dynamics	• MPI-parallel	• InfiniBand		No
<i>Simulation</i>	• MPI-parallel	• InfiniBand	• Large data sets	Yes
<i>Mechanical Design and Drafting</i>	• Stand alone	• None	• Large memory (64-bit computing) • Large data sets • Graphics	Yes
<i>Imaging (Visualization Clusters)</i>	• MPI-parallel	• InfiniBand	• Large memory (64-bit computing) • Large data sets • Graphics	Yes
<i>Software Engineering</i>	• Embarrassingly parallel	• Ethernet	• Large data sets	Yes
<i>Digital Content Creation (DCC) and Distribution</i>	• MPI-parallel	• InfiniBand	• Large data sets	Yes
<i>Economic and Financial Modeling</i> • Economic simulation • Stock analysis	• MPI-parallel • Embarrassingly parallel	• InfiniBand • Ethernet	• Large data sets	Yes
<i>Industrial Process Analysis</i> • Product lifecycle management (PLM) • Data mining	• Stand alone	• None	• Large memory (64-bit computing) • Large data sets	Yes

**Table 2:** Sample application areas and types, interconnect types, requirements, and storage support for different applications

Company	URL	Product
Sun	www.sun.com	Sun Fire V60x and V65x servers
Mellanox*	www.mellanox.com	InfiniBand silicon (chips), cards, switches (from OEM or major value-add resellers)
Topspin*	www.topspin.com	InfiniBand cards, switches, software

\* Provides full support for respective interconnect technology

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA Phone 800 786-7638 or +1 512 434-1577 Web sun.com



**Sun Worldwide Sales Offices:** Africa (North, West and Central) +33-13-067-4680, Argentina +5411-4317-5600, Australia +61-2-9844-5000, Austria +43-1-60563-0, Belgium +32-2-704-8000, Brazil +55-11-5187-2100, Canada +905-477-6745, Chile +56-2-3724500, Colombia +571-629-2323, Commonwealth of Independent States +7-502-935-8411, Czech Republic +420-2-3300-9311, Denmark +45 4556 5000, Egypt +202-570-9442, Estonia +372-6-368-900, Finland +358-9-525-561, France +33-134-03-00-00, Germany +49-89-46008-0, Greece +30-1-618-8111, Hungary +36-1-489-8900, Iceland +354-563-3010, India-Bangalore +91-80-2298989/2295454; New Delhi +91-11-6106000; Mumbai +91-22-697-8111, Ireland +353-1-9055-666, Israel +972-9-9710500, Italy +39-02-641511, Japan +81-3-5717-5000, Kazakhstan +7-3272-466774, Korea +82-2-193-5114, Latvia +371-750-3700, Lithuania +370-729-8468, Luxembourg +352-49 11 33 1, Malaysia +603-21161888, Mexico +52-5-258-6100, The Netherlands +00-31-33-45-15-000, New Zealand-Auckland +64-9-976-6800; Wellington +64-4-462-0780, Norway +47 23 36 96 00, People's Republic of China-Beijing +86-10-6803-5588; Chengdu +421-2-4342-94-85, South Africa +27 11 256-6300, Spain +34-91-596-9900, Sweden +46-8-631-10-00, Switzerland-German 41-1-908-90-00; French 41-22-999-0444, Taiwan +886-2-8732-9933, Thailand +662-344-6888, Turkey +90-212-335-22-00, United Arab Emirates +9714-3366333, United Kingdom +44-1-276-20444, United States +1-800-555-9SUN or +1-650-960-1300, Venezuela +58-2-905-3800

**SUN** THE NETWORK IS THE COMPUTER © 2003 Sun Microsystems, Inc. All rights reserved. Sun, Sun Microsystems, the Sun logo, Solaris, and Sun Fire are trademarks, registered trademarks or service marks of Sun Microsystems, Inc. in the United States and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the US and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc. in the United States and other countries. Printed in USA 6/03, Solution Brief, SunWIN 379990