Abstract

While the 90s were about information, today is about answers. Answers drive our society. Whether you are a search engine, a Fortune 500 company, or a researcher at a national laboratory, answers are your business and the ability to provide them quickly and cost effectively is a distinct competitive advantage.

Today’s high performance data center needs to generate answers rapidly and requires that the right people can seamlessly access those answers. Answers are dynamic and must be accurate. Additionally, the more accurate answers a business can produce in a given time, the more revenue it can generate. And as the world and workforces disperse, there must be fast and reliable connections from anywhere at any time for the data center to help increase efficiency and competitiveness.

Enabling the rise of this high performance data center are the following technological changes:

• High performance, cost-effective computing
• High-speed storage access and backup
• Embedded security and dynamic denial of service protection
• High-speed, congestion-free networks
• Virtualizing a centralized network and assets

To ensure accuracy and speed of answers, a high performance cluster computing architecture that leverages the total processing power of a business is required. With a cluster computing architecture in place, businesses can then feed it with information.

The ability to rapidly access stored information and answers is crucial to interpreting available information. This requires high-speed access and backup to a storage network. If answers are revenue, it is important to ensure that the right people are accessing the right content. Advanced dynamic denial of service protection and embedded security at the system level are required to limit and allow access where, when and by whom it is needed.

Next, making the disparate pieces of the data center infrastructure work together requires a high performance network. No longer simply providing transportation, but linking tens of thousands of servers, databases and other computational resources, the network becomes involved in the creation of the answers. As such, it must be both fast and congestion-free to ensure accurate answers are delivered rapidly. And it is resilient, high capacity systems that provide the scalability and availability to seamlessly interconnect the pieces of potentially geographically distributed infrastructure.

Finally, the distinction between centralized storage and computing versus distributed storage and computing must be transparent to the user community. By enabling anytime access to the network from any external locations, businesses can more effectively leverage their infrastructure to create answers.

In today’s high performance data center, where information transformed into answers is the currency, information technology is regaining its role as a strategic asset that enables competitive advantages. This white paper explores the technologies needed to ensure businesses are prepared to meet the new challenges and opportunities of the emerging “answer economy”.
Introduction

Many IT organizations are currently working to solve the problems created by the explosive expansion of IT infrastructure in the late 1990's. In many instances this rapid expansion to meet business unit demand for IT has resulted in a highly complex infrastructure that offers sub-optimal reliability and security while being overly difficult to manage and expensive to operate. Some of the symptoms of excessive IT complexity are the following:

- Numerous data centers distributed throughout the organization, including ad hoc patchworks of servers located in wiring closets and computer equipment rooms.
- Large numbers of servers dedicated to single applications and provisioned for "worst-case" workload levels. Average utilization of servers in this category run as low as 10%.
- Storage subsystems that are dedicated to single servers as embedded storage or directly attached storage (DAS) systems. Single application servers with dedicated storage often utilize only 20% of available storage capacity. Low utilization has contributed to the meteoric growth in storage capacity (e.g., from 7 TeraBytes (TB) in 1996 to 48 TB in 2002 for the average Fortune 500 company). Dedicated storage also requires dedicated back-up facilities and makes consistent data protection policies difficult to implement.
- Numerous older generation fixed configuration and modular Ethernet switches providing connectivity within the data center and to the remainder of the enterprise network. Older generation Ethernet switches may support only 10/100 server connections and often have relatively low port counts. A patchwork of switches from different technology generations and vendors add to complexity by having disparate functionality and management interfaces.
- Highly heterogeneous server environments with multiple versions of Windows, Unix, Linux, plus legacy operating systems running on multiple brands and models of server hardware. Heterogeneous systems complicate system management, including patch roll-outs.

IT departments with symptoms similar to these are performing internal assessments to determine whether Total Cost of Ownership (TCO) savings and operational efficiencies can be gained by implementing a range of data center and server consolidation options. A TCO approach is needed in order to make the right tradeoffs between the additional capital expenditure required and the savings to be gained by improving resource utilization and reducing operational costs via simplification of the infrastructure. In particular the TCO assessment has to address the hidden costs of configuring, managing and supporting distributed and dispersed environments as new technologies are considered for deployment.

Data center consolidation and architectural re-assessment is often planned in the context of an overall IT program to improve the alignment of IT with business goals, reduce operating costs, and improve business continuity, security, and manageability.

Technologies for the Next Generation Data Center

The best general approach to planning a consolidation program for the data center starts with developing a vision of what the future data center should be like. Based on that vision, an architectural framework can be developed that guides near term consolidation decisions while providing a flexible path to future generations of the data center.

Most of the current rhetoric on the evolution of the data center is focused on three technology areas: **infrastructure virtualization**, **cluster computing**, and **Grid computing**. These concepts all share the notion that enterprise applications should be able to tap a common pool of computing resources (servers, storage, and networking) as if it were a single large virtual system. Depending on the application, the resources might be located in a single data center or distributed across a number of data centers, or even the entire enterprise network. For all of these technologies, the single system view is accomplished by interposing a middleware layer between the application and the resource pool as shown in Figure 1.

![Figure 1. Single system vision for distributed computing](image-url)
While the middleware can take various forms depending on the type of application and the specifics of the resource pool, its primary function is to provide an abstraction layer or virtualization interface between the application and the resource pool. Virtualization supports the allocation of physical devices to logical resource entities, thereby simplifying the alignment of physical resources to application needs. The middleware may in some cases be embedded within the application, layered on the operating systems of the servers and other intelligent devices in the resource pool, or installed on specialized servers/appliances. Virtualization middleware has a number of potential benefits, including:

- Shielding applications from the details of the physical resources in the pool
- Providing flexibility for static, or even dynamic, changes to the infrastructure with minimal disruption of applications
- Simplifying centralized management and providing the basis for automating management functions

There is a rapidly proliferating range of sources for virtualization middleware. All the major system vendors are building middleware suites that include a number of software components, including vendor-developed proprietary software, proprietary software from partner ISVs, and some industry standard or open source software.

In spite of the common single system philosophy, there are considerable differences among the middleware suites being developed for infrastructure virtualization, cluster computing and Grid computing, as outlined below.

**Infrastructure Virtualization** software is intended to de-couple existing enterprise applications and software services from underlying infrastructure by virtualizing server processing, storage and networking. This de-coupling allows hardware management to be completely separated from software management, with the hardware allocated and re-allocated to various software services on the fly via a management console that provides the required resource allocation, load balancing, and monitoring of resource utilization.

Server virtualization is based on "virtual machine" software which provides a standardized software representation of a reference server hardware configuration to the server operating systems and applications, as shown in Figure 2.

![Figure 2. Virtual machine architecture](image)

With virtual machine software, a single server can support multiple instances of either a single operating system or multiple instances of multiple operating systems as indicated in the figure. In addition, virtual machine software makes it possible to extend the uniform virtual hardware platform model across the entire data center, allowing software to be installed on or moved from one physical system to another without requiring reconfiguration of the operating system or applications. Application availability is enhanced by the capability to move running applications by transferring the entire system and memory state of a running virtual machine from one physical server to another. The system's disk space, including all of its data, software and boot partitions, must be stored on a shared storage infrastructure, such as a SAN, to allow a high speed, transparent transfer.

Storage virtualization software is another important aspect of infrastructure virtualization. Storage virtualization software includes volume controller functionality that allows files and disk storage capacity to be pooled into a single reservoir under the control of centralized management. From the central management console, administrators can migrate data from one disk system to another, add/remove disk systems, upgrade disk system software, and backup data without disrupting application availability. Therefore, with SAN volume controller software the group of heterogeneous servers at the right side of Figure 3 could each be using a virtual disk distributed to access the SAN-attached storage. A cluster file system can also allow homogeneous servers in the cluster at the left side of the diagram to share files via a NAS server...
that uses a virtual disk that spans both DAS and the SAN. As a further enhancement, a SAN file system can also enable SAN-wide sharing of files so that the heterogeneous servers shown can access shared files, making the SAN look like "server-less" NAS.

Although today storage virtualization software is proprietary, the Storage Network Industry Association (SNIA) is a vendor-neutral industry forum that is working on standards to improve multi-vendor interoperability for storage management, including storage virtualization.

Network virtualization can take a variety of forms. Both server and storage virtualization are highly dependent on a high performance switched network to provide flexible interconnection among the physical devices that may belong to a variety of dynamic logical groupings. These data interconnects provide shared access to data and files and the high speed transport to allow migration of image files and processes among systems. For Ethernet switches, VLAN functionality can be used to partition and isolate resources into multiple groupings thereby allowing network administrators to allocate data center and enterprise network bandwidth to ensure predictable performance for specific critical applications. Another form of network virtualization may be based on switches or other physical devices that can perform a range of networking and security functions, such as web load balancing, firewall filtering, and intrusion detection. A virtual network device therefore can be configured to perform the functions that are required at different points within the business cycle, in a manner that is analogous to server virtualization.

Infrastructure virtualization is not necessarily an automated process. The re-allocation of resources may involve periodic intervention by a system administrator based on predictable changes in data processing requirements that are synchronized with quarterly, monthly, weekly, or daily business cycles. In addition, resource monitoring tools are needed to help administrators identify longer term trends and respond rapidly to unexpected changes in demand for computing resources. As the technology continues to evolve, the re-allocation of resources and the provisioning of additional resources may become fully automated under the control of a real-time, policy-based management system that would seek to protect the service levels contracted for by the various business units. At this point in time, infrastructure virtualization could begin to be described as "utility computing".

**Cluster Computing.** There are essentially two types of computer cluster: the load-balancing cluster and the computational cluster. Both types of cluster are based on software that makes the array of homogeneous servers appear as a single virtual system.

Load balancing clusters provide high availability and scalable performance by distributing client connections among the application servers in the cluster. Various load balancing algorithms may be used to distribute the client sessions to achieve relatively equal distribution of the workload. Because each client session is independent of the others, the servers can operate in parallel, but there is no requirement for interprocess communications (IPC) among the servers or modification of existing applications to run on the cluster.

Computational clusters offer an entirely different dimension of virtualization in which a tightly-coupled cluster of computers provides supercomputer-like power in executing applications specifically programmed to run
on parallel computers. Computational clusters are generally comprised of highly homogeneous systems built with commodity hardware and software (typically, but not exclusively, based on the Linux operating system and Intel architecture processors). Cluster middleware provides the single system image (SSI) virtualization needed to make a multiplicity of networked computer systems appear to both the user and application as a single parallel computing system. A key aspect of the computational cluster middleware is a message passing system for IPC between the servers in the cluster. The speed of IPC is critical for a class of parallel applications known as tightly-coupled synchronous applications. These applications require communication among all of the cluster processors (global synchronization) before the computation can be completed.

Synchronous applications include clustered databases such as Oracle’s 9i RAC and 10g and IBM’s DB2, which are typically run on small clusters with less than 40 processors. Other synchronous applications are scientific/engineering applications involving complex simulations of discrete event systems, particle systems, and lumped/continuous variable systems. The technical applications generally run on much larger clusters with as many as hundreds or even thousands of cluster nodes.

Due to its homogeneity, the computational cluster can readily present a highly transparent single system image (SSI) to the user/application. For example, a cluster-wide file system can provide each node of the cluster with ready access to shared data and a single namespace for the cluster resources including file systems, devices, processes, networking, and IPC objects. SSI middleware can also provide resource allocation/monitoring and fault tolerant services such as checkpointing, automatic fail-over, and recovery from failure among all nodes of the cluster.

Grid Computing is taking shape as a general purpose distributed computing model where heterogeneous systems on an Intranet or the Internet can publish services (computational services, data services, or other types of service) that may be accessed and utilized by other systems participating in the Grid. With Grid computing, computer systems and other resources therefore aren’t necessarily constrained to be dedicated to individual users or applications, but can be made available for dynamic pooling or sharing in highly granular fashion to meet the changing needs of the organization. The vision of the Global Grid Forum (GGF) is that standards-based Grid middleware will allow Internet-wide resource sharing and collaborative problem solving by multi-institutional “virtual organizations”.

With the approach being taken by the GGF, the Grid concept is really modeled on the idea that ubiquitous shared computing services can turn the entire Internet into a single large virtual computer.

However, most system vendors and ISVs approach the Grid from the perspective of marshalling the resources within the enterprise network. This has led to the creation of a second industry forum, the Enterprise Grid Alliance (EGA). The EGA is more focused on adoption of Grid technology by the enterprise, which involves more of a focus on interoperability between the proprietary products of the system vendors and ISVs. While some of the tools and concepts may be derived from the standards work of the GGF, it is very likely that proprietary Grid middleware will continue to dominate the enterprise market for some time to come.

One class of enterprise application of the Grid is essentially a generalization of the computational cluster where the servers are not necessarily homogeneous or directly attached to the same data center switched interconnect system. In this case, the Grid would typically be used for executing parallel applications that are more loosely coupled than those assigned to a computational cluster. An example would be an application using Monte Carlo simulation for calculating the risk/reward potential of an investment portfolio. For such applications the Grid offers the advantage that the number of compute servers involved can be much larger than for the cluster because the Grid can extend well beyond the data center to include multiple clusters and desktop systems or other systems dispersed throughout the enterprise. For a Grid of clusters, each cluster could be specialized for a specific compute-intensive task (e.g., simulation vs. visualization).

A second enterprise application of the Grid provides a federated approach to data integration where data from different sources (relational data bases, files, or application data) can be consolidated in a single data service that hides the complexities of data location, local ownership, and infrastructure from the consuming application. With data federation, the data remains in place at its source, with no disruption of local users, applications, or data management policies. Integration of data from multiple sources and locations facilitates a wide range of integrated applications, including corporate performance dashboards, marketing analysis tools, customer service applications, and data mining applications. Because the data resource is accessed as a service on the network, minimal modifications of existing data-publishing or data-consuming applications is required.
The Virtual Data Center

While all of these virtualization technologies are still in the process of being refined and extended, it is fairly clear that the data center of the future will be largely based on a virtualized structural model similar to that shown in Figure 4. As the Virtual Data Center model is adopted, these technologies will continue to mature, and system vendors and ISVs will collaborate to offer integrated middleware that will unify the three modes of virtualization. Integration will greatly facilitate the task of deploying a Virtual Data Center that leverages all the available technologies.

For example, a combination of infrastructure virtualization software and clustering software would allow servers to be reallocated among clusters running different versions of cluster middleware and different applications, as shown in Figure 5. In this scenario, a group of servers currently participating in load-balancing Cluster A is re-allocated to a computational Cluster B running database software. Virtual machine software is used to de-install the software images from the servers being transferred and then install the required computational cluster image including application, middleware, and operating system. As part of the process, the VLAN membership of the transferred servers is changed from VLAN A to VLAN B. Although not shown in the diagram, each of the servers shown would be connected to a SAN for shared storage or storage virtualization.
Data Center Consolidation: Toward the Virtual Data Center

In the context of a strategy to implement a Virtual Data Center model, data center consolidation can be viewed as a program to simplify and improve the flexibility of the pool of computing resources that will subsequently be virtualized. Therefore, consolidation can not only reduce TCO in the near term but also prepare the physical infrastructure for deployment of virtualization technologies, which can be expected to yield additional rounds of consolidation and further reductions in TCO.

There are two general approaches to consolidation: location consolidation and infrastructure consolidation or rationalization. Location consolidation involves reducing the number of data centers in combination with a program to centralize the location of as many servers as possible within the data center. Location consolidation generally results in improved security and control, better adherence to standards and best practices, and significant economies of scale in system management. This form of consolidation lowers TCO through reducing aggregate data center facilities costs, including floor space and telecommunications costs. Co-location of larger pools of computing resources also makes it more practical to provision the high speed interconnect required by virtualization.

Infrastructure consolidation focuses on removing as much of the complexity of the data center as possible. There are a number of ways that this can be accomplished:

- **Consolidation of Server Operating Systems:** Industry standard operating systems (such as Linux and Windows) running on commodity Intel-architecture hardware platforms tend to offer the best available price/performance as well as the broadest support for enterprise applications. Migrating legacy applications from proprietary platforms to standard platforms can therefore reduce server costs as well as result in significant simplification of the infrastructure. Data centers that are planning to take full advantage of load balancing and computational clustering technology will tend to gravitate toward some version of Linux because of its highly mature clustering technology. Database applications, ERP applications, file services, and web services are all prime candidates for migration to clusters. Operating systems running on Intel-architecture processors also have the advantage of being supported by EMC’s VMware virtual machine software, with the future option of using Microsoft’s Windows Virtual Server software (currently scheduled for shipment in 2005). Operating system standardization also implies minimizing the number of operating system versions that are deployed and supported.

- **Server Consolidation:** Replacing several older servers dedicated to a single application with a single more powerful server running multiple applications reduces complexity as well as space and power requirements. More powerful servers are also advantageous for server virtualization where servers will tend to run multiple instances of one or more operating systems.

- **Consolidation of Server Hardware:** Standardization of server hardware configurations is another way to simplify the data center. One of the platform types that should receive strong consideration is the blade server. Blade servers conserve rack space, simplify power and system management, and are a very suitable platform for high density cluster computing.

- **Storage Consolidation:** Storage Area Networks (SAN) offer an excellent solution for consolidating storage resources to improve storage management, flexibility, and cost-effectiveness. SANs allow data from disparate storage subsystems to be migrated to large enterprise-class shared disk arrays that offer advanced functions including RAID, remote mirroring, and instantaneous data replication. As noted earlier, deployment of SANs also greatly simplifies storage virtualization that further enhances consolidation.

- **Consolidation of Data Center Interconnect:** Most data center networks have already been consolidated based on Ethernet and TCP/IP as the standard network transport. However, further consolidation is possible by replacing older models of Ethernet switches with current generation families of Ethernet switches that support high densities of Gigabit Ethernet (GbE) and 10 Gigabit Ethernet (10 GbE) interfaces. High density switches can simplify the infrastructure by replacing numerous older generation switches. High performance Ethernet switches in conjunction with TCP/IP can also provide the high speed interconnect required for virtualization, including IPC for computational clusters and Grids, and SANs with iSCSI. A more complete discussion of the tradeoffs between Ethernet and alternative IPC and SAN switching technologies is included in the next sections of this document dealing with Cluster Interconnect and Storage Networking.
High Speed Cluster Interconnect

As data centers begin to deploy server virtualization and computational cluster computing, the existing Ethernet switching fabric will be required to support two new types of traffic:

- Transfers of context (state) to/from virtual servers
- Cluster IPC for synchronous parallel applications

A high performance, non-blocking, data center class Ethernet switch can easily provide the bandwidth for these additional traffic types. However, for tightly coupled parallel applications that require very low end-to-end latency (equal to switch latency plus the delays required for host message processing) more specialized IPC fabrics, such as InfiniBand may need to be considered. However, before increasing the complexity of the data center with an additional switching fabric, a careful analysis should be performed to understand the performance benefits that can be expected both today and with future generations of the various technology options. The remainder of this section of the document provides an overview of the primary options for computational cluster interconnects.

IPC over Ethernet

Switched Gigabit Ethernet is an excellent choice for computational cluster configurations such as the one shown in Figure 4. GbE switching has the advantages of being a highly scalable, very low cost alternative for a switched IPC fabric. The cost for switch ports and high performance server adapters is expected to continue to drop rapidly as more enterprise campus LANs and desktops migrate to GbE. 1000Base-T supports cable lengths of up to 90m for Cat5e and 1000Base-SX supports cable lengths of over 500m over multi-mode fiber facilitating the configuration of large clusters without the added expense of single mode fiber optic cabling.

Gigabit Ethernet supports system-to-system data transfer approaching wire speed at ~1000 Mbps and send/receive message latencies on the order of 60 microseconds for short messages. The send/receive latency (or "ping pong" latency) is measured as one half the round trip delay between a message being sent and a reply received (i.e., a LOFI measurement). The round trip latency includes the switch latency (on the order of 10 microseconds in each direction) and the receive/send latency of the remote host due to TCP/IP processing and buffer-to-buffer transfers, both of which typically involve software processes. Bandwidth and message latency in the cited range are quite compatible with a wide range of synchronous parallel applications that don’t involve very high degrees of coupling between individual processes. Of the November 2003 Top500 list of the most powerful computers in the world, over 100 employ Gigabit Ethernet as the primary cluster interconnect technology, including #25 and #43 on the list. Maximizing Ethernet throughput performance requires very careful tuning of the TCP protocol parameters (such as send/receive buffer space) for the applications in question. According to the Web100 organization, expert tuning can result in more than an order of magnitude throughput improvement vs. accepting default parameters. TCP tuning is becoming more important as Ethernet speeds increase and for long haul applications of TCP/IP over Ethernet or other high speed WAN technologies. Web100 is a software suite for Linux and is available at www.web100.org. Web100 provides the instrumentation and diagnostics required to allow non-TCP gurus to optimize the configuration of TCP for maximum throughput. Support for other operating systems is in the planning stages.

As noted above, most of the end-to-end latency incurred with Ethernet data transfer and IPC is due to software processing on the hosts. Another drawback of network I/O processing in software is that it results in fairly high CPU utilization during intense network activity. The usual rule of thumb is that each bit per second of TCP/IP bandwidth consumes one Hz of CPU capacity. Thus a sustained Ethernet transfer at 800 Mbps would involve approximately 80% CPU utilization of a 1 GHz CPU. High CPU utilization can have an impact on applications that require high levels of concurrent computation and IPC, but is less of a concern when intense computation and communication do not overlap. The obvious implication of the 1 Gbps per GHz rule is that 10 GbE NICs will have to be designed to eliminate host processing in software in order to achieve line rate. NICs that offload the host CPU, will also be of benefit for GbE end-to-end performance.

Intelligent hardware-assisted GbE and 10 GbE NICs are expected to become widely available over the next year or two. These NICs will leverage the iWARP specification issued by the RDMA Consortium, including hardware-assisted Transport Offload Engines (TOE), Remote Direct Memory Access (RDMA), and kernel bypass. iWARP includes support for multiple application interfaces: MPI for IPC, Socket Direct for general IP applications, iSER (iSCSI over iWARP), and NFS/RDMA. With these enhancements a iWARP NIC will feature hardware-assisted throughput at or near wire speed along with low CPU utilization and much lower end-to-end latency (see Table 1 below for some performance comparisons for iWARP NICs). The first iWARP Ethernet host adapters were introduced in the latter half of 2004 with support
for 1000Base-T and 10GBase-CX. When the 10GBase-T standard for 10 GbE over twisted pair is ratified, it is expected that the first implementations will be iWARP NICs. As more iWARP NICs are introduced and 10 GbE becomes more widely adopted in enterprise LANs, Ethernet product volumes are expected to result in rapidly decreasing costs for both NICs and 10 GbE switch ports.

**IPC over InfiniBand**

InfiniBand (IB) is a standards-based switching technology intended as a general-purpose "server area" I/O fabric envisioned to function as an IPC fabric, a SAN fabric, and even as a possible replacement for the PCI bus for embedded system interconnects. InfiniBand was evidently modeled to some degree on the proprietary cluster/system IPC interconnects from Myricom (Myrinet) and Quadrics (QsNet) used extensively in High Performance Computing (HPC) and high end computational clusters. All of these technologies employ switch fabrics based on cut-through switching across “fat tree” configurations of relatively small crossbar switching chips assembled to build a modular chassis switch with as many as 128 to 288 ports. Larger switching fabrics with thousands of ports can be assembled as fat tree federations of multiple stand-alone switches. InfiniBand HCAs also feature hardware-supported network transports and RDMA support pioneered by Myrinet and QsNet. As a result, InfiniBand offers very low end-to-end latency, very low CPU utilization, and good throughput, while supporting three link speeds: 2.5, 10, and 30 Gbps.

InfiniBand has had some early successes as an IPC fabric for large scale computational clusters, such as the Virginia Tech Macintosh-based cluster that placed third on the November 2003 Top500 list. For enterprise applications of computational clusters, InfiniBand has also achieved some promising benchmark results with cluster-aware database software such as Oracle 9i RAC or 10g running on small clusters of Linux servers. According to these test results published by Oracle, the performance of clustered database applications is quite dependent on the latency of IPC. IPC is required among cluster nodes to establish exclusive access (or lock) to a block of data. These tests have shown that a small Oracle 9i RAC cluster with 10 Gigabit InfiniBand IPC can process 100% more block transfers per second (at 16KB per block) than a similar cluster with GbE IPC. These peak transfer rates were achieved with CPU utilizations of 60% and 90% for InfiniBand and GbE respectively. It is worth noting that the application performance advantage for InfiniBand is relatively modest in view of its superior benchmarks in terms of throughput, latency and CPU utilization vs. GbE as shown in Table 1. This provides a good example of why

<table>
<thead>
<tr>
<th>Crossbar Chip</th>
<th>Mellanox InfiniBand</th>
<th>GbE Ethernet</th>
<th>10 GbE with RiWARP DMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switch Topology</td>
<td>Fat tree</td>
<td>Multiple ASICs</td>
<td>Multiple ASICs</td>
</tr>
<tr>
<td>Max Cable Length (Switch to Host)</td>
<td>10m copper</td>
<td>90m cat 5e</td>
<td>10-25 m 10GBase-CX</td>
</tr>
<tr>
<td>Max Size Standalone Switch</td>
<td>288 ports</td>
<td>1,260 ports</td>
<td>224 ports</td>
</tr>
<tr>
<td>Host Adapters</td>
<td>PCI-X</td>
<td>PCI-X</td>
<td>PCI-X</td>
</tr>
<tr>
<td>Port Speeds</td>
<td>2.5, 10, 30 Gbps</td>
<td>1 Gbps</td>
<td>10 Gbps</td>
</tr>
<tr>
<td>MPI Throughput of Large Messages &gt;64KB</td>
<td>6.6 Gbps (10 Gbps ports)</td>
<td>1 Gbps</td>
<td>10 Gbps</td>
</tr>
<tr>
<td>CPU Utilization at Max Throughput</td>
<td>3%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Send/Receive Latency 16B Message</td>
<td>5 µsecs</td>
<td>~60 µsecs</td>
<td>20 µsecs store and forward 10 µsecs cut through</td>
</tr>
<tr>
<td>Send/Receive Latency 4KB Message</td>
<td>15 µsecs</td>
<td>na</td>
<td>na</td>
</tr>
</tbody>
</table>

10 GbE data from Chelsio

Table 1. Comparison of cluster/system interconnects
application-level performance testing is far more reliable as a basis for comparison of IPC switching technologies and cluster and Grid software suites.

Table 1 also includes some performance results for 10 GbE IPC with a newly introduced iWARP NIC from Chelsio Communications. The results show higher throughput with 10 GbE vs. 10 Gigabit InfiniBand and comparable latency over a cut-through 10 GbE switch. Using a store-and-forward 10 GbE switch adds 10 microseconds to end-to-end latency.

**Storage Networking Fabrics**

While direct-attached storage (DAS) works well in smaller environments with a limited number of servers, the solution does scale well for data centers with large numbers of servers that require easy access to a common pool of data. Storage for each server must be managed separately and cannot be shared without the intervention of the host server. As enterprise data centers move away from the "one server-one application" paradigm, they are typically leveraging the superior data/storage management offered by a networked storage approach, such as NAS or SAN.

Network Attached Storage (NAS) was originally designed to allow Ethernet-attached desktop systems to share data residing on a networked file server with its own dedicated DAS. The client and server incorporate support for network file systems such as NFS and CIFS. NAS is a very efficient means of sharing read-only access to files such as CAD/CAM files and image files for websites. For these and other similar applications, NAS has been adapted to the data center environment with servers accessing files stored on other servers or a file server appliance. Because NAS is based on familiar technologies, it is relatively easy to deploy and manage. The primary difficulty has been that the file server CPU often proves to be a bottleneck that limits scaling of the system. In order to address this problem, vendors of file server appliances have replaced the general purpose operating system in the NAS device with a "storage operating system" kernel and are starting to offer NAS storage devices in which the single file server CPU is replaced by a cluster of embedded processors running a cluster file system. Clients can access the NAS cluster via NFS or via the same distributed file system used within the NAS cluster.

A Storage Area Network (SAN) can be used to replace DAS (or augment DAS) with a switched network of shared storage devices. Using the SAN, the server accesses files by making block level requests directly to the remote storage device. The overhead processing on the requesting server is low and no processing is required on the storage node to satisfy the request. The original SANs were based on Fibre Channel (FC) switches with Fibre Channel Host Channel Adapters (HCAs) on the servers. Therefore the FC SAN is essentially a second high speed network fabric that off-loads storage access traffic from the LAN. Off-loading the LAN was an important benefit when FC at 1 Gbps was first introduced because at that time 100Mbps was the maximum speed for Ethernet LANs. The current versions of Fibre Channel support link speeds of 1 and 2 Gbps, with subsequent generations scheduled to support 4 and 10 Gbps. In spite of the good performance, SAN deployments have tended to be restricted to higher-end data centers because of the relatively high cost of the FC switches and HCAs, plus the added management burdens inherent with an additional, relatively complex switching fabric technology.

A second SAN technology is iSCSI which uses TCP/IP over Ethernet as the local switched transport for SCSI block level requests. The advantages of using TCP/IP/Ethernet as the storage transport are numerous:

- The need for an additional storage-specific fabric (including switches, cabling, and HCAs) is eliminated
- Ethernet is a familiar technology with a mature set of management tools
- Ethernet economics have made 1 Gbps switch ports and adapters very low cost, thereby lowering TCO
- 10 Gigabit Ethernet is maturing rapidly, with 100 Gbps a likely next generation
- Storage requests can be routed over any existing IP MAN or WAN to reach remote storage facilities without requiring additional gateway devices
- iSCSI can be leveraged to provide access to stranded DAS capacity via the IP network; iSCSI also facilitates storage virtualization strategies that incorporate DAS

In spite of these advantages, iSCSI SANs are usually relegated to the edges of the enterprise network (often as an alternative to NAS) rather than being used as primary SAN for a large data center. The reasons for this are that the FC SAN is often the incumbent data center technology, which also offers better performance than an iSCSI over 1 Gbps switched Ethernet. As in the case of IPC messaging, iSCSI over a TCP/IP software stack running on the host yields relatively high end-to-end latency and throughput generally below GbE wire speed.
As discussed in the section of this document dealing with IPC switching, iSCSI over IWARP NICs will result in iSCSI performance becoming very competitive with that of FC. Performance results from Chelsio Communications and Neterion show iSCSI throughput of 10 Gbps over 10 GbE with end-to-end latency of approximately 20 microseconds via a store and forward switch.

There are two other TCP/IP storage protocols, FCIP and iFCP. FCIP is a gateway protocol that tunnels FC through a TCP/IP network to connect FC SANs over a TCP/IP backbone. FCIP uses TCP/IP mechanisms for congestion control and management and a combination of TCP/IP and Fibre Channel mechanisms for error recovery. iFCP, is a second gateway protocol that addresses FC SAN connectivity over IP by replacing the Fibre Channel transport layers with TCP/IP.

InfiniBand is another technology that could potentially be used as a SAN switching fabric. However, in the SAN arena, InfiniBand has not yet had much of an impact, partially because storage vendors haven’t been motivated to offer InfiniBand-attached storage in view of the nascent installed base of InfiniBand switches. Another impediment that InfiniBand will face in the enterprise SAN market is the resistance to another new technology whose advantages may not prove to be sustainable without a sizable share of the market to justify on-going development.

As noted earlier, storage virtualization can be based on SAN volume controller software and a SAN File System. The SAN File System provides a centrally managed SAN-wide file structure that accessed by servers running heterogeneous operating systems. The SAN File System essentially converts the SAN into a highly scalable file sharing network. Clustered file systems and SAN File Systems are just two examples of how cluster and Grid computing technologies are beginning to be exploited to provide a greater level of integration between the NAS and SAN approaches to networked storage.

**Unified Fabrics**

As computational cluster computing becomes a mainstream data center technology, there is the possibility that data center servers would be attached to three distinct switching fabrics: Ethernet for IP communications, a SAN for shared access to storage resources, and a cluster fabric for IPC between cluster nodes. Although each fabric might be well optimized for its function, deploying three separate fabrics, each with its own cabling, host adapters, and management systems, involves too much cost and complexity for most enterprise IT departments to accept. The ideal alternative would be to use a single switching fabric (a "unified" fabric) for all three functions. Settling on a single fabric involves some degree of sub-optimization at any particular point in time, but in the long run could contribute significantly toward optimizing the cost-effectiveness of the data center through consolidation of disparate technologies and the resulting reduction or control of TCO.

The primary candidate for a unified fabric is Ethernet in conjunction with TCP/IP. Ethernet has been adopted almost universally as the network technology of choice for enterprise LANs, and is beginning to extend its reach into the MAN and WAN. Although there are fabric technologies that currently offer better end-to-end performance benchmarks for SANs and IPC cluster interconnect, Ethernet/TCP/IP has done a creditable job of providing very flexible and cost-effective solutions in both of these areas. As the next generation of iWARP GbE and 10 GbE NICs become available and prices of 10 GbE NICs and switch ports continue to fall, the end-to-end performance gap of Ethernet will either disappear completely (for SANs) or be reduced to relative insignificance (for IPC). Figure 6 shows a data center with Ethernet as the unified fabric providing IPC, SAN, NAS, and LAN/MAN/WAN connectivity. In addition, a FC gateway or storage bridge/router can be deployed to provide backward compatibility to protect the prior investment in a FC SAN.

![Figure 6. Ethernet as a unified data center switching fabric](image-url)
InfiniBand is a second possible candidate for a unified fabric. IB has performance characteristics that have been optimized for use as an IPC or SAN fabric. However, IB is still in the early stages of adoption as a cluster interconnect and has not yet gained support from storage device manufacturers for IB-attached storage. Therefore, IB can provide a unified fabric only in the sense that servers in the data center would not need three separate types of adapters for different networks. A single IB server adapter can provide indirect connections to SAN and LAN switch fabrics via the IB switch, as shown in Figure 7. Connectivity from the IB fabric to Ethernet and FC SANs is provided by a gateway function that may be performed by a standalone device or a gateway card installed in the IB switch. These gateways are potential bottlenecks that can limit the aggregate LAN or SAN bandwidth available to the server cluster, unless multiple gateways are deployed in parallel. The main drawback, however, is that with the IB-based fabric unification, there are still three separate fabrics to provision and manage.

Although the goal of a unified switching fabric is worthwhile pursuing, it is fairly certain that multiple switching fabrics will co-exist in many high end data centers for some time to come. When a data center upgrade is required to support the current business plan, the most cost-effective currently available technology must be used. Waiting for next year’s technology improvements is generally not a viable option. However, the decision to add a new fabric to the data center has a fairly significant impact and should be based on a careful analysis that balances significant application benefits against all the related cost factors contributing to TCO, including the training and staffing aspects of introducing a complex new technology into the heart of the IT infrastructure.

**Ethernet Switch/Routers for the Data Center Network**

As data centers undergo location consolidation and more business critical information assets are centralized in a smaller number of data center locations, the issue of non-stop accessibility to business-critical data and computing will become even more important. The requirements of business continuity call for data centers with highly redundant architectures designed to eliminate as many single points of failure as possible. A multi-level approach to security is also required to protect the data center from external or even internal attack.

In order to guarantee business continuity even if an entire data center is lost due to a catastrophic event, critical data must be replicated in remote data centers that have the computing resources available to continue to support key business operations. Network level redundancy and data center replication/redundancy are even more effective when the servers and networking devices are designed with the robustness and high availability features that minimize the number of device failures that require failovers to redundant backup devices while simultaneously minimizing the impact of any failover that does occur.

Data center consolidation has additional implications for the network that provides connectivity within the data center and between the data center and the rest of the enterprise network. With location consolidation, there will be fewer data centers with each remaining data center housing more computing and storage resources. This means that each data center will need expanded internal networking capacity and increased bandwidth connectivity to the enterprise network.

As a result of consolidation, a less predictable mix of traffic will be funneled into each data center raising the possibility of congestion that could lead to increased response time in accessing critical data center resources. In addition, with data center consolidation and the introduction of cluster computing, the Ethernet network will be called upon to carry more different types of mission-critical traffic than ever before (possibly including IPC and/or SAN traffic in addition to the normal client/server and NAS traffic typical today).

"Data center optimized Ethernet" networks that are deployed to support server consolidation should be selected based on their capabilities to support both near term consolidation and the future steps toward the next generation Virtual Data Center. As a result, the emphasis should be on families of switch/routers
that offer the right combination of high availability/robustness features, scalable performance, security, and intelligent traffic control.

**Robustness, High Availability, and Security:** The switch/router should provide non-stop operations even in the face of the full gamut of software or hardware errors that can possibly occur. Basic high availability features include full hardware redundancy for all major subsystems, including control plane/management modules, switch fabric, power, and cooling. Additional robustness is provided by software resiliency features that allow protocol upgrades and restarts without interrupting traffic flow. Network resiliency features include support for 802.3ad Layer 2 Link Aggregation for trunking and link redundancy, and 802.1w Rapid Spanning Tree Protocol or Layer 3 Equal Cost Multi-Path routing to allow the implementation of highly fault tolerant networks. Additional software robustness can be added by modularization of the control plane to isolate the effects of soft failures. Security features should contribute to a multi-tiered security solution by being available to help to protect the switching system from unauthorized access or denial of service (DoS) attacks. Wire-speed access control lists (ACLs) should be available to restrict access to sensitive data and resources and should also be able to filter on Layer 2 to Layer 4 information regardless of the forwarding mode of the interface.

**Interface Density and Scalability:** Switch/routers should be capable of supporting large numbers of GbE and 10 GbE interfaces in compact enclosures. High density switching simplifies network design, makes efficient use of scarce rack space, and minimizes the cost and complexity of interconnecting and managing multiple switch chassis in the data center or across the enterprise LAN/MAN. Scalable switch/routers are engineered to accommodate subsequent generations of technology (such as 40 Gbps or 100 Gbps Ethernet) without radical architectural changes or forced obsolescence of earlier generation systems. Highly scalable, high port-count switches can reduce the total number of switches required in a traditional two-tier Ethernet data center network based on a distribution (aggregation) tier of switches plus a separate tier of access switches that provides server connectivity, as shown in the left hand side of Figure 8. Even greater consolidation can be achieved by using high density Ethernet switch/routers to collapse the two tiers of data center switches into a single aggregation/access tier of Layer 2/Layer 3 switches, as shown on the right hand side of Figure 8.

**Non-blocking Switch Architectures:** Switch/routers should support full line-rate, non-blocking packet forwarding performance even when the chassis is fully populated with the densest interface cards available. For data center applications, including computational clusters, switch/router configurations that over-subscribe the uplink switch ports have very limited applicability because of the additional latency that can result from port congestion. Non-blocking architectures also provide the most consistent and predictable performance possible, irrespective of traffic patterns. Furthermore, non-blocking line-rate packet forwarding performance should not be compromised even with all Layer 3 services enabled, including the full range of traffic management/control and QoS services supported by the device.

**Integrated L2 Switching and L3 Routing:** Integrating Layer 2 switching and Layer 3 routing in a single switch/router device simplifies the network topology by reducing the number of systems and network interfaces that must be deployed to implement multi-tiered network designs. For example, in the collapsed aggregation/access tier shown in Figure 8, the switch/router would play the dual role of a Layer 2 switch for server access connections and a Layer 3 VRRP router for distribution function and connections to the core of the enterprise LAN. In addition to providing robustness for large network designs, Layer 3 routing allows optimum load balancing using equal cost multi-path (ECMP) routing of traffic over the redundant paths. Redundant paths are prevalent in highly meshed networks including very large data centers designed for both highest performance and maximum availability.

**QoS:** As the Ethernet network is used to support more applications that are highly sensitive to throughput and latency, such as IPC traffic and iSCSI SAN traffic, it will be highly desirable to take steps to assure predictable performance for all applications. One possibility would
be to use separate Ethernet switch fabrics dedicated to IPC or the iSCSI SAN. Another approach is to use a single fabric and VLANs to isolate IPC and SAN segments from general purpose traffic. The right choice would depend on the application mix and scale of the data center network. In any case, QoS will be required in order to provide the highest level of priority for critical traffic. QoS may be based on packet identification/classification using TCP/UDP port number or end system packet marking using the IEEE 802.1p or IP DiffServ specifications. QoS may be enforced by providing strict priority to IPC, SAN and other critical traffic, in conjunction with enforcing policies for rate limiting or policing non-critical sources of traffic.

Building the Foundation for the Virtual Data Center with Ethernet
The basic network design for a consolidated data center is shown in Figure 9 where a single Aggregation/Access tier of Ethernet Layer 2/Layer 3 switches provides the connectivity for servers and NAS, as well as connections to the Core of the network. This particular data center is based on a 3-Tier server architecture with Web presentation servers in the first tier, Application Servers implementing business logic applications in the second tier, and the Database and other back end resources in the third tier. Each tier of servers is configured as a separate subnet/VLAN. The web servers and applications servers may be implemented as load-balancing clusters, and the database servers as computational cluster. Storage for the web server tier is typically provided by NAS with the database tier using a redundant SAN. Even in this relatively simple design with only two Aggregation/Access switch/routers, the data center shown could be scaled to support over 300 dual-ported servers by taking advantage of the current generation of high end GbE/10 GbE switch/routers such as the Force10 E1200.

The dual redundant Ethernet switch/routers are configured for Layer 2 switching between nodes in each cluster VLAN/subnet and Layer 3 switching (IP routing) among the cluster subnets and between the data center and the remainder of the enterprise network. Servers are connected with dual port NICs to each of the Ethernet switches using Gigabit Ethernet (GbE), GbE trunks, or 10 GbE depending on the performance level required.

For server connectivity, one of the switches is configured as the 802.1D primary root bridge for the Layer 2 connectivity within each group of clustered servers and as the Virtual Router Redundancy Protocol (VRRP) primary router for connectivity outside the cluster subnet. The second switch would be configured to act as the secondary root bridge and the secondary VRRP router. In this way, each aggregation/access switch plays a consistent role relative to the servers in each subnet at both Layer 2 and Layer 3.

VRRP creates a virtual router interface that can be shared by two or more routers. The interface consists of a virtual MAC address and a virtual IP address. Each database server is configured with the VRRP virtual interface as the default gateway. Each of the routers participating in the virtual interface is assigned a priority that determines which is the primary router. The VRRP routers multicast periodic "hello" messages that include the priority of the sending router. In the event the primary router fails, the secondary router detects the missing hello, determines that it has the highest remaining priority, and begins processing the traffic addressed to the virtual interface. VRRP, in conjunction with STP and dual ported server NICs, assures that server connectivity will be maintained even if one of the switch/routers fails.

Layer 3 routing between the clusters and the remainder of the network maximizes the degree of control over entering the data center and within the server tiers of the network.
data center. Layer 3 Extended ACLs, rate limiting, and policing can all be enabled on the Layer 3 interfaces to implement tight control over the traffic entering each subnet.

In order to provide the highest level of service for critical data center traffic, such as latency sensitive IPC traffic in the database subnet, the basic goal should be to give high priority to general data center traffic, while ensuring that IPC traffic and other more critical traffic does not incur any added latency by spending time queued up in a buffer behind less critical traffic.

The first consideration would be to ensure that the network has sufficient capacity for the anticipated traffic. This may involve careful analysis of traffic patterns to minimize the possibility of congestion on the links between the data center and the core of the network. One of the advantages of the 3-Tier server architecture is that it minimizes the bandwidth consumed on the data center's core links because of the localization of presentation and business logic within the data center. As a result, significantly more bandwidth is consumed within the data center than is required for transferring results to client systems throughout the network. Therefore, the data center's uplinks to the core may be theoretically over-subscribed, but the probability of congestion occurring on these links is very low.

Traffic analysis of this sort can allow more of the switch capacity to be safely allocated to intra-server traffic within the data center. Once traffic patterns are well-understood, the network designer can exploit the QoS capabilities of the modern Ethernet switch to prioritize critical data center traffic. Some of the tools available include:

**Priority queuing:** With priority queuing, the forwarding capacity of a congested port is immediately allocated to any high priority traffic that enters the queue. Therefore, with priority queuing, the only traffic that can delay the most critical data center traffic is other critical data center traffic already in the queue.

**Rate limiting and policing:** One way to help control traffic is to use ACLs, rate-limiting, and policing to limit the amount of lower priority traffic that can consume bandwidth of the critical data center subnets.

**Weighted Random Early Detection (WRED):** Packet loss can be essentially eliminated if buffers are never allowed to fill to capacity with resulting overflows. Overflows can be avoided by applying WRED to the lower priority traffic as the buffer fills beyond specified levels, which will begin to drop packets randomly in order to throttle the sender. This eliminates the possibility of high priority packets arriving at a buffer that is already overflowing with lower priority packets.

Dual redundant storage subsystems are deployed within the database tier to provide data integrity and data replication services. Storage replication is commonly deployed to protect against failure of an entire storage subsystem. Ideally, the replication can be performed at the storage subsystem/SAN level without consuming server capacity or bandwidth of the Ethernet/IP network. Data may also be backed up to a local tape archive attached to the SAN. In addition, data replication to a remote site may be desirable to protect against the possibility of a disaster that could threaten an entire data center or site. In this case, the backup data could traverse the site Ethernet LAN and a high speed enterprise wide area network. Another alternative would be a dedicated long haul Fibre Channel link from the data center SAN switch to another SAN switch in a remote data center located at a hardened disaster recovery site. A third option is to use direct Ethernet connectivity from the data center switch/routers to remote disaster recovery sites and other data centers over the MAN and WAN by 10 GbE over dark fiber or CWDM/DWDM, or by telecommunications services that support 10 GbE interfaces. Using Ethernet for direct connections among data centers offers the advantage of supporting different types of traffic over a single link (e.g., data backup and recovery, SAN interconnection, and Grid computing).

The basic data center design shown in Figure 9 can be scaled in a number of ways. The simplest approach is to add more Aggregation/Access switches to the VRRP group in order to support a larger number of servers among the clusters. A second approach to scaling is shown in Figure 10 where each tier of the server hierarchy has its own set of Aggregation/Access switches. This allows the different server tiers to scale independently and allows for the choice of switch model (or even type of switching fabric) to be optimized at each server tier. For example, the application server tier and database tiers could use either Ethernet or InfiniBand switches for IPC cluster switching depending on application requirements and performance tradeoffs at the time of deployment.
Conclusion

As enterprise data centers move through consolidation phases toward next generation architectures that increasingly leverage virtualization technologies, the importance of very high performance Ethernet switch/routers will continue to grow. The high volume of Ethernet products continues to spur rapidly declining prices and a constant stream of enhancements/innovations including 10 GbE WAN PHY, Ethernet MAN/WAN services, iWARP RDMA/TOE NICs, 10GBase-T for 10 GbE over twisted pair, and the next generation of the Ethernet bandwidth hierarchy at 40 or 100 Gbps. These developments testify to the unparalleled flexibility and resilience of Ethernet and TCP/IP to respond to challenges from competing technologies. As Ethernet evolves over the next two to three years, it is likely to significantly enhance its standing as the switching fabric of choice for data center IPC and as a highly competitive SAN fabric. As these developments unfold, data center architectures based on a unified switching fabric can be implemented without any of the performance compromises that exist today. Fabric unification at the highest level of performance for all data center applications will serve as the basis for future waves of data center consolidation and architectural evolution.

In today’s high performance data center, where information transformed into answers is the currency, information technology is regaining its role as a strategic asset that enables competitive advantages. Continued data center evolution will ensure businesses are prepared to meet the new challenges and opportunities of the emerging answer economy.