BladeCenter networking

S. W. Hunter
N. C. Strole
D. W. Cosby
D. M. Green

The IBM eServer™ BladeCenter® system physically consolidates the server and network into a common chassis. It was introduced as a new server architecture that provides many benefits over the traditional data center model of clustered independent systems linked by a network fabric. This paper describes the BladeCenter networking architecture and relates it to user requirements for multi-tier servers, scale-out models, networking technology advances, and industry trends. Design decisions and challenges, the switch subsystem and input/output technology options, services that are currently supported by the architecture, and future enhancements and extensions are addressed.

Introduction

Over the past decade, a universal model for designing data centers has been the scaling of computing performance by aggregating large numbers of low-profile servers with one or more networking fabrics in a clustered or multi-tiered framework, as shown in **Figure 1** and described in [1–3]. The grouping of servers in this manner is sometimes referred to as a *Web cluster* [3], such that a *cluster* is a parallel or distributed system consisting of a collection of interconnected whole computers used as a single, unified computing resource [4]. In this type of configuration, tier 1 servers are typically dedicated to a specific function, such as load balancing, security, or caching and are sometimes referred to as *appliance servers*. Tier 2 servers may be referred to as *application servers* and typically have the ability to host a variety of applications or be dynamically provisioned with an application when more resources of that type are needed. If present, *database servers* may be located in tier 3 and typically have the most stringent performance and dependability requirements.

While there are many advantages with the multi-tiered framework, the complexity of managing the large number of independent systems and the networking fabrics has become cumbersome because of issues such as cabling and the number of independent control points. To reduce this complexity, the concept of using network technology to physically consolidate servers as blades in a common chassis was introduced to help overcome these issues and to provide additional benefits over those of the traditional data center model.

The *blade* concept is not new. In fact, networking blades have been used in networking products since the early 1990s for the same reasons as server blades—reduction of cables, ease of scalability, built-in redundancy, and a single point of control. The additional benefits of combining server blades with networking technology include improved density compared with independent servers and networking systems and a tighter coupling of server and networking technology for advanced workload management. Additional details can be found on past and current networking blade products in [5–7]. For a general overview of the BladeCenter architecture, motivation, and design tradeoffs, see [8].

Network architecture

The BladeCenter network architecture comprises multiple independent network subsystems for the interconnection of a collection of blade servers consolidated within a common chassis. For example, the initial deployment of the BladeCenter chassis supports up to 14 server blades interconnected with one or two Ethernet network switches and up to two optional switches. Each server blade slot has up to four high-speed network interfaces, with each interface connected to a switch module bay in such a way that the 14 blades have point-to-point connections to each of the four integrated network switch module bays. In addition, the switch modules provide multiple external uplinks for connectivity to the external network infrastructure.

The networking links across the midplane are independent of the technology being used. Ethernet, Fibre Channel, InfiniBand** [9], and Myricom Myrinet**

0018-8646/05/$5.00 © 2005 IBM

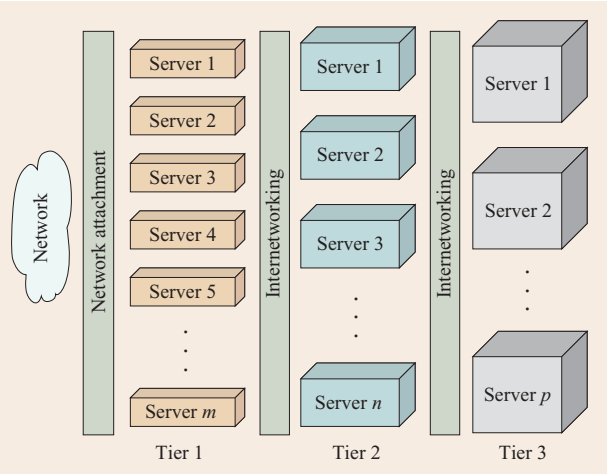


Figure 1

Multi-tier framework.

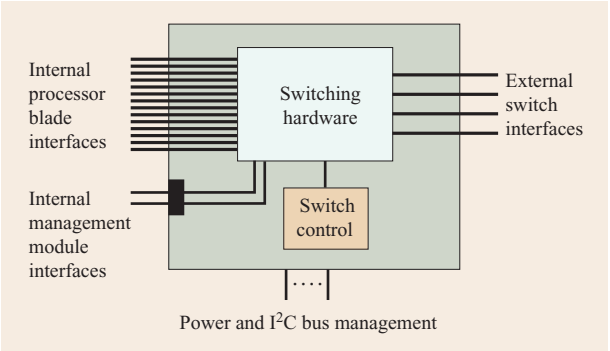


Figure 2

Switch module interfaces.

technologies are currently supported. While the midplane is independent of the networking technology used, current server blades implement two input/output (I/O) interfaces that use Gigabit Ethernet networking application-specific integrated circuits (ASICs) on the system board, thus fixing these links to Ethernet only. Therefore, the corresponding switch modules in bays 1 and 2 must also have compatible interfaces. The technology for the other two blade I/O interfaces is optional and is determined by the choice of blade I/O expansion adapter on the system board. The technology chosen for the I/O expansion adapter must match the technology in switch module bays 3 and 4. For details, see switch packaging, midplane interconnection, and blade I/O expansion adapter [10–12].

To leverage industry technology and standards, the Open Systems Interconnection (OSI) model is adhered to for the BladeCenter internal and external interfaces, so that it provides a framework for the exchange of data and network information from applications on one server, through the network media, to an application on another client or server. The OSI model categorizes the various processes needed in a communications session into seven distinct functional layers [13, 14].

Adhering to network industry standards on both sides of the networking link provides the flexibility of using industry-available I/O and switch technology. For example, an Ethernet switch may provide Layer 2, 3, 4, and/or 7 functions, and the blade I/O may provide basic media access control (MAC), Transmission Control Protocol/Internet Protocol (TCP/IP) offload, Internet Small Computer Serial Interface (iSCSI), and/or remote direct memory access (RDMA) [15] functions, with the level of function being offered on each end of the link being mutually exclusive.

The switch external interfaces are the primary communication paths in and out of the BladeCenter chassis. Multiple media types and protocols may be offered to provide cost, connectivity, and/or performance tradeoffs. Some of the options for external interfaces include multispeed Ethernet (copper or fiber), Fibre Channel, and InfiniBand.

Each of the two management module bays has point-to-point Ethernet network connections to each of the four switch module bays to provide redundancy and to eliminate single points of failure in the networking and management domains. Management modules provide a single point of control for the chassis and act as proxy devices to connect to the switch control point on each of the switch modules, as shown in **Figure 2**. The Ethernet connection between the management module and the switch modules provides management access to each of the network switches via common interfaces such as Telnet, Simple Network Management Protocol (SNMP), and Web browser. In addition to the Ethernet interface, an Inter-Integrated Circuit (I^2C) Serial Bus Interface also exists between each of the management modules and switch modules. The I^2C bus interface is a low-level management interface used for collecting information, performing initial configuration, monitoring, and controlling the switch modules.

In summary, each networking switch module has four types of interfaces with internal (within chassis) or external devices. These are external interfaces, internal serializer/deserializer (SerDes) interfaces with the server blades, internal SerDes interfaces with management modules, and internal I^2C bus interfaces with management modules. As an example, the initial BladeCenter Ethernet switch module implemented four

external 1000BaseT Ethernet links, 14 internal point-to-point 1-Gb/s SerDes links to the processor blades, two point-to-point internal 100-Mb/s SerDes links to the management modules, and an I^2C bus interface with each of the two management modules for I^2C bus vital product data and register access. Other signals for the Ethernet switch module include power, switch bay identifier, and presence indicator.

Scalable performance and dependability

One of the benefits of the multi-tier framework is its ability to easily scale performance by attaching additional servers. This scheme is sometimes referred to as a *scale-out architecture* to highlight its ability to add independent compute nodes—with each blade having an independent memory mapping—around a networking interconnect. The scalability of a system can be affected by many areas [16, 17], one of them being the networking interconnect. One requirement is for the underlying network fabric to be nonblocking. For example, to design a nonblocking networking configuration of 14 internal 1-Gb/s blade links and four external 1000BaseT links, the design must support up to 18 Gb/s of total throughput. This configuration allows full 1-Gb/s blade-to-blade communication and makes it easy to increase external bandwidth capability.

Another reason for the popularity of the multi-tier framework is its potential to improve the dependability of a system by using redundancy and device failover in a cost-effective manner. By defining the internal networks as independent fabrics and by maintaining standard interfaces, each device can be viewed as an autonomous unit with well-defined boundaries. This approach makes it possible to leverage industry-standard technology and techniques to provide redundancy and appropriate failover support within the chassis. For example, since a Layer 2 switch collects very little networking *state* in comparison with a Layer 4 or 7 switch, the approach taken to achieve minimal failover time is quite different. Similarly for the blade I/O, an Ethernet MAC device collects very little networking state in comparison with an Ethernet device capable of TCP/IP offload. Industry approaches for achieving dynamic failover for Layer 3 switches and network interface card (NIC) I/O are the Virtual Router Redundant Protocol (VRRP) and NIC teaming, respectively. Layer 2 switches have relied upon the Spanning Tree Protocol (STP) to provide redundant links and alternate paths.

Additional areas related to dependability include performability [18, 19] and security [20–22]. For example, techniques for improving performance and dependability of Web sites that receive a large number of requests include redundant hardware, load balancing, Web server acceleration, and efficient management of dynamic data [1, 2]. Examples for enhancing network security are the use of virtual local area networks (VLANs) for management purposes to protect against malicious access, such as denial-of-service attacks, and the use of configurable access control lists to filter ingress traffic at the switch ports. The BladeCenter architecture and design must continue providing the same level of scalability and dependability benefits as described above for the multi-tier framework. In addition, as described in the following sections, decisions were also made to further enhance these areas.

Design challenges and considerations

The design of the BladeCenter chassis is optimized around space, power, and cooling, and this applies to the networking components as well. The switch module mechanical enclosure is based on a single-wide, single-high InfiniBand module with the following requirements:

- Width: 29 mm; height: 112 mm; depth: 259.8 mm (i.e., longer than InfiniBand).
- Card-to-card pitch: 30 mm.
- Power: 45 W maximum.
- Airflow: top-to-bottom or bottom-to-top, depending on the slot in the chassis.
- Retention: single-cam lever with snap latch.
- Connector: Molex VHDM** signal and power with an alignment pin.

With a limited amount of space and power being allocated for the networking switch, a significant amount of effort was required to conform to these constraints. The first and second design challenges below were direct results of some of the above limitations; the fourth was an indirect result, since a separate physical network for management purposes was not feasible.

IBM (Layer 2) switch: power, packaging, and thermal constraints

The first challenge of integrating off-the-shelf network technology was the packaging of the 20-port Layer 2 Ethernet switch into the switch enclosure. Using Figure 2 as a reference, the initial Ethernet switch module consisted of two Broadcom BCM5632 switching ASICs interconnected via the Broadcom HiGig** interface to achieve the necessary port density, with 18 1-Gb/s ports (14 blade and four external) and two 100-Mb/s ports to each of the management modules. Since switching ASICs at that time did not provide integrated SerDes, four quad SerDes devices were required for internal interfaces and a quad 1000BaseT transceiver was required for the external interfaces. Finally, a small processor complex was required to provide the switch control point, power conversion, and control circuitry. To reduce the amount

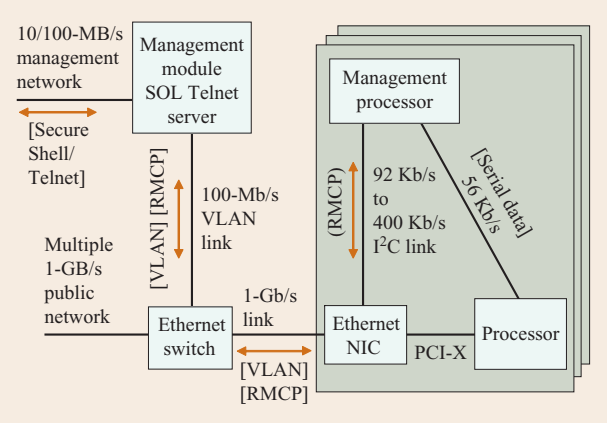


Figure 3

BladeCenter serial-over-LAN overview.

of space and power required, the decision was made to use SerDes interfaces across the midplane instead of 1000BaseT transceivers, which required approximately 1 W per port. Through the use of SerDes instead of 1000BaseT, the ability to automatically negotiate transmission speeds across the backplane was lost.

Nortel (Layer 4–7) switch: power, packaging, and thermal constraints

A second challenge was to meet space, power, and thermal limitations in the development of the second-generation Ethernet switch, which also included the integration of Layer 4–7 capabilities. IBM partnered with the Nortel Alteon group—a proven leader in providing load-balancing and content-switching functions—to provide this advanced switch.

The design consisted of two BCM5690 switching ASICs and additional functions in a configuration similar to that of the first-generation switch described above, with the exception that the SerDes interfaces were integrated into the switch ASICs, thus providing additional space for other components. To provide the Layer 4–7 networking function, a Broadcom SiByte** network processor, a field-programmable gate array (FPGA), and external static random access memory (SRAM) were required to handle the deep-packet processing and connection state. While the integration of SerDes into the switching ASICs helped to free up space, excellent engineering practices were also followed by the Nortel team to meet space, power, thermal, and stringent electromagnetic interference (EMI) requirements.

EMI considerations

Another related challenge was determining the appropriate design constraints for EMI, because of its effect on signal integrity and radiated emissions. Advanced EMI modeling and analysis techniques were necessary to determine methods of meeting packaging requirements and to consider the superposition effects of EMI when up to four switch modules are installed in the chassis. Results of this work led to specific solutions, such as increasing the number of printed circuit board layers and selection of specific parts from vendors.

Secure management and control

Another challenge was the design of the network initialization, configuration, and secure services into management modules, switches, and blade NICs to provide a flexible and secure environment for chassis management. For example, an internal management VLAN is defined on the physical fabric, with the management module and switch module control point(s) being default members. This approach is taken to keep management traffic secure on its own private (logical) network. As the default, this network is accessed externally only through the management module and is hidden from the external uplink ports within the switch, blocking malicious users (e.g., denial-of-service attacks) from accessing devices on the internal management network. Additional VLANs may also be configured as *private* in order to contain other management traffic internally. As described below, the serial-over-LAN design uses this approach.

Serial over LAN

One of the major server blade design decisions was to remove the legacy physical serial port. However, serial connectivity is useful on virtually all rack-mounted equipment, particularly with Linux** servers, where it is required for effective system configuration and administration. Serial concentrators are used for several reasons: to limit the distance serial connections must be run, to provide more complex terminal server function, and to enable multi-protocol routing between serial connections and LAN-based terminal consoles.

The chassis and blade design and the goal of centralizing control within the chassis management module led to the decision to implement the serial interface on each server blade as a logical—rather than physical—interface. This design leverages the internal Ethernet subsystem in the chassis and provides a more flexible option for use in future applications.

The BladeCenter serial-over-LAN (SOL) solution (**Figure 3**) preserves the simplicity of serial-connected LAN management of servers within the cabling constraints of dense, rack-mounted blade servers. Logical serial connectivity to LAN-based terminal applications is preserved without dedicated serial cabling. Serial data can be transparently forwarded to remote terminal

applications via the existing Ethernet network by routing it over the internal Ethernet fabric within the chassis between the server local service processor or baseboard management controller (BMC) and chassis management modules. In the chassis, basic terminal server functions are implemented in the management module via an imbedded SOL Telnet server.

The Distributed Management Task Force defines Remote Management and Control Protocol (RMCP) in the Alert Standard Format Specification. When implemented to take advantage of all features, this simple protocol provides for client control functions in pre-operating-system (OS) and OS-absent states. The protocols are intentionally simple so that firmware can easily generate and parse messages. These concepts have been extended by the Intelligent Platform Management Interface (IPMI) Consortium and are now incorporated in the IPMI specification. Both of these schemes are based on the encapsulation of serial data within User Datagram Protocol (UDP) frames to provide a sufficiently flexible and functional communication protocol foundation for handling serial data streams over LAN systems.

The BladeCenter management module provides remote access to a Telnet proxy server application by the external 10/100-MB/s Ethernet port. Serial connectivity to each blade is initiated by the remote client via a Telnet or Secure Shell session and command-line interface. Sessions can be established simultaneously with all 14 blades for OS console management. Internally, UDP frames encapsulating serial data pass between the service processor or BMC and the management module using a dedicated, internal IEEE Standard 802.1q-tagged VLAN. The Ethernet NIC firmware on each blade provides VLAN filtering along with the MAC and Internet Protocol address-compare function to capture those frames intended for the management processor on the blade.

Redundant switch failover with NIC teaming

The dual Ethernet interfaces on the server blades are capable of supporting interface teaming, which is used to group internal blade interfaces and external switch interfaces into a "team" to provide fault tolerance and load balancing. Teaming drivers also offer an option for IEEE Standard 802.3 link aggregation, but this can be supported only for blades where two or more Ethernet I/O interfaces on the blade connect to the same switch.

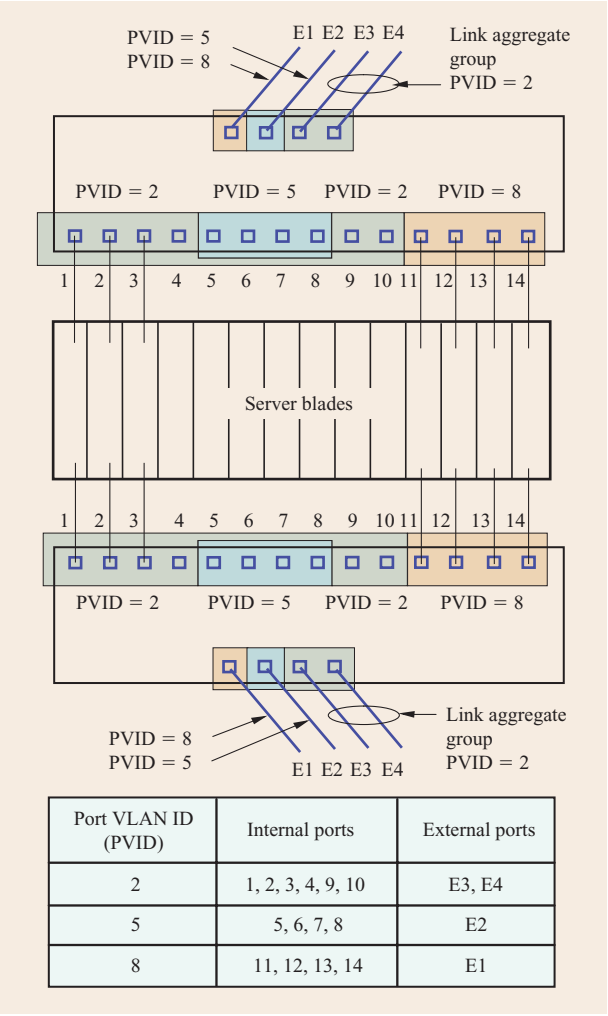
A teaming driver can react to the loss of a link to one or more of the teamed NICs by redirecting traffic to the remaining available links. This normally occurs whenever an Ethernet switch module (ESM) is physically removed from the chassis. However, loss of the external uplinks is not detected by the teaming driver because of the intervening switch fabric within the ESM. Current industry practice relies on either the Layer 2 STP or Layer 3 VRRP protocols to activate an alternate external link or network path. IEEE Standard 802.1 defines STP operation and associated timings to prevent data loops from disrupting operation of a Layer 2 switched network. For a Layer 2 Ethernet network to function properly, only one active path can exist between any two stations. STP operation is transparent to end stations, which cannot detect whether they are connected to a single LAN segment or a switched LAN of multiple segments. However, these protocols may take several seconds and result in disruption of the end-to-end server sessions.

A design innovation unique to BladeCenter ESMs, known as *redundant switch failover* or *trunk failover*, allows the ESM to deactivate the internal links whenever the associated external links are lost, thus enabling the teaming driver to react to the external link loss. Recovery times with this scheme are shorter and less disruptive than with STP.

The Broadcom Advanced Server Program (BASP) [23], Intel PROSet, or Linux TG3 driver software can create multiple IP interfaces on unique VLANs using teamed NICs. Essentially, multiple IP interfaces are created by the OS, while the software drivers then insert the configured VLAN tag to create a separate virtual connection for each interface, but share the physical network interface connections. The VLAN identifications (IDs) within the 802.1q tags maintain separation of the flows while transversing intermediate Layer 2 network nodes and common physical connections. IEEE Standard 802.1p prioritization within the 802.1q tag and round-robin scheduling can be configured to police the bandwidth utilization between the VLANs. This flexibility potentially provides a more economical (and higher-bandwidth) approach to providing fully redundant connections to multiple network interfaces.

Figure 4 illustrates a pair of ESMs, each partitioned with the same port VLAN configuration to provide redundancy to the blade servers. The NIC teaming drivers typically allow a pair of server blade NICs to be configured in either active/active or active/backup modes. If supported by the OS, the active/active mode attempts to balance traffic between the two ESMs during normal operation to maximize use of the available network capacity. The active/backup scheme directs all traffic to the designated active NIC and ESM during normal operation. With either scheme, if one of the ESMs is removed from the chassis or if the ESM reacts to a loss of the uplinks via the trunk failover option, all traffic is automatically forwarded through the other ESM to reach the external network.

The external ports in this example, including the dual-port aggregate group, are dedicated to selected VLANs.



Port VLAN ID (PVID)	Internal ports	External ports
2	1, 2, 3, 4, 9, 10	E3, E4
5	5, 6, 7, 8	E2
8	11, 12, 13, 14	E1

Figure 4

Redundant blade connections via dual Ethernet switch modules (ESMs).

Thus, the physical separation within the chassis can be extended to the external network infrastructure to maintain data separation if required.

The dual NICs allow a blade server to be attached to separate physical networks via the two ESMs within a chassis. Since the two NICs are attached to separate ESMs via point-to-point internal links, total separation of the physical networks can be maintained. While not required, the applications within the blade server may be associated with one network or the other. This type of configuration also allows a blade server to act as a router or security firewall between the two networks while maintaining physical (Layer 2) separation between the networks.

Networking technology and options

As described above, the architecture and chassis midplane design allow a single-slot blade to have up to four networking interfaces, each connecting to one of the internal switch modules. For the dual-processor blades, two of the I/O interfaces are defined as Ethernet since these are placed directly on the system board. The other two interfaces are determined by the type of I/O expansion adapter attached to the blade. The current I/O technologies supported on the expansion adapter are Ethernet, Fibre Channel, InfiniBand, and Myrinet. The current switch technologies supported are Ethernet, Fibre Channel, and InfiniBand. Copper and optical passthrough modules are also available. With the exception of Fibre Channel, each of these interconnect technologies is described below. For a detailed description of Fibre Channel, see [24].

Ethernet

Ethernet standards and products have continued to evolve for the past two decades since the initial IEEE Standard 802.3 was introduced in the early 1980s. Current Ethernet technology provides high-speed, 1-Gb/s full-duplex, point-to-point links in both NIC and multiport Layer 2 switch components over a wide variety of copper and fiber media. Ethernet infrastructure products and management tools are pervasive in worldwide commercial, industrial, and telecommunications systems.

Consequently, Ethernet was determined to be the natural choice for low-cost, high-volume server blade internetworking within the BladeCenter chassis and for connection to the external network infrastructure. Therefore, as described in the previous sections, Ethernet technology is inherent in the server blades and forms the basis for many of the underlying networking subsystem designs.

IBM is partnered with recognized leaders in the networking industry to integrate the latest advances in Ethernet technology. Nortel Networks has incorporated advanced load balancing and content-based applications in its first product developed especially for the BladeCenter chassis—the Nortel Layer 2–7 Gigabit Ethernet Switch Module. This switch subsystem is particularly applicable to advanced workload-management applications, as described below. A more recent pair of products—Nortel Layer 2/3 Gigabit Ethernet Switch Modules with both 1-Gb/s copper and fiber uplinks—addresses the combination of Layer 2 switching and Layer 3 routing.

Cisco Systems networking products are recognized throughout the industry for their advanced feature set and the broad range of applications they support. The Cisco Intelligent Gigabit Ethernet Switch Module (Cisco

IGESM) was specifically developed by Cisco Systems as a BladeCenter network subsystem and brings many of the Cisco advanced Layer 2 switching and filtering functions into the chassis. In addition, the broad range of Cisco management tools and protocols are supported to facilitate the integration with existing networking infrastructures.

The next generation of Ethernet components to support 10 Gb/s and higher speeds over fiber and copper media are also emerging. The server blade design allows this technology to be incorporated on the blade via the I/O expansion adapter or on the ESMs as high-speed uplinks. IBM is also involved in industry activities to advance Ethernet as a midplane interconnect [25] and to overcome some of its deficiencies in areas such as congestion management [26]. In addition, next-generation Ethernet I/O technology will incorporate advanced packet-processing schemes to relieve some of the processing bottlenecks that occur in today's end-to-end protocols. These are discussed in more detail below.

InfiniBand

InfiniBand, Myrinet, and Quadrics Ltd. QsNet are all interconnects generally available on the market today with high-bandwidth (greater than 1 Gb/s) and low-latency (less than 10 μs) capabilities. To provide an industry standard for clustering and generalized I/O, Compaq, Dell, HP, IBM, Intel, Microsoft, and Sun worked together to develop the InfiniBand standard. As described in the specification [9], the InfiniBand architecture provides a first-order interconnect technology for interconnecting processor nodes and I/O nodes to form a system area network. The architecture is independent of the host OS and processor platform and is designed around a point-to-point switched fabric with end-node devices (host computers, I/O devices, Ethernet adapters, etc.) interconnected by cascaded switch devices. The focus of the InfiniBand interconnect is on two environments by making the appropriate bandwidth, distance, and cost optimizations: *module-to-module*, as typified by computer systems that support I/O module add-in slots, and *chassis-to-chassis*, as typified by interconnecting computers, external storage systems, and external LAN/WAN access devices (e.g., routers, gateways) in a data center environment.

The InfiniBand switched fabric provides a reliable transport mechanism to enqueue messages for delivery between end nodes. In general, message content and meaning is not specified by InfiniBand, but is left to the designers of end-node devices and the processes that are hosted on end-node devices. The architecture defines hardware transport protocols sufficient to support both reliable messaging (send/receive) and memory manipulation semantics, such as remote direct memory access (RDMA), without software intervention in the data movement path. Specialized protocols, such as RDMA, will achieve one of the overall objectives of increased processor utilization and decreased latency. While InfiniBand supports implementations as simple as a single computer system, it is also capable of supporting replication of components for increased system dependability, cascaded switched fabric components, additional I/O units for scalable I/O capacity and performance, additional host node computing elements for scalable computing—or any combination of these.

Since it is designed to be a first-order network, InfiniBand focuses on moving data in and out of node memory and is optimized for separate control and memory interfaces. This permits hardware to be closely coupled or even integrated with the memory complex of a node, removing any performance barriers. However, InfiniBand is flexible enough to be implemented as a secondary network that permits legacy networks and migration while still permitting maximum available bandwidth use and increased processor efficiency. The components that make up an InfiniBand fabric may include an InfiniBand switch, host channel adapters (HCAs), target channel adapters (TCAs), an Ethernet gateway, a Fibre Channel gateway, and an I/O expansion unit. In contrast to the decentralized topology management approach used by Ethernet (e.g., STP), the InfiniBand architecture describes a subnet manager for defining the topology and controlling the fabric. In a possible InfiniBand configuration (**Figure 5**), one or more switches can make up the InfiniBand fabric, with end-node connectivity being achieved with HCAs or TCAs. HCAs are used to connect server systems to the fabric in a clustered configuration. TCAs are typically used to connect the InfiniBand fabric to some other networking media, such as Ethernet or Fibre Channel, via respective gateway devices. InfiniBand switches operate at Layer 2 in the OSI model, and local IDs (LIDs) are used for addressability at this level.

The BladeCenter chassis was originally designed with the integration of a generic clustering and I/O fabric, such as InfiniBand, in mind. The support of InfiniBand enables BladeCenter servers for high-performance clustering applications that require low latency and for applications that require scalable I/O. The initial InfiniBand I/O expansion adapter and switch are provided through a partnership with TopSpin Communications. The adapter has two IB-1x (i.e., a 2-Gb/s data rate and 2.5-Gb/s baud rate) point-to-point connections to a corresponding InfiniBand switch in switch module bays 3 and 4. The TopSpin switch provides multiple IB-4x (i.e., 8-Gb/s data rate and 10-Gb/s baud rate) external links; however, the architecture also supports networking interfaces such as Ethernet and Fibre Channel to be brought out as external

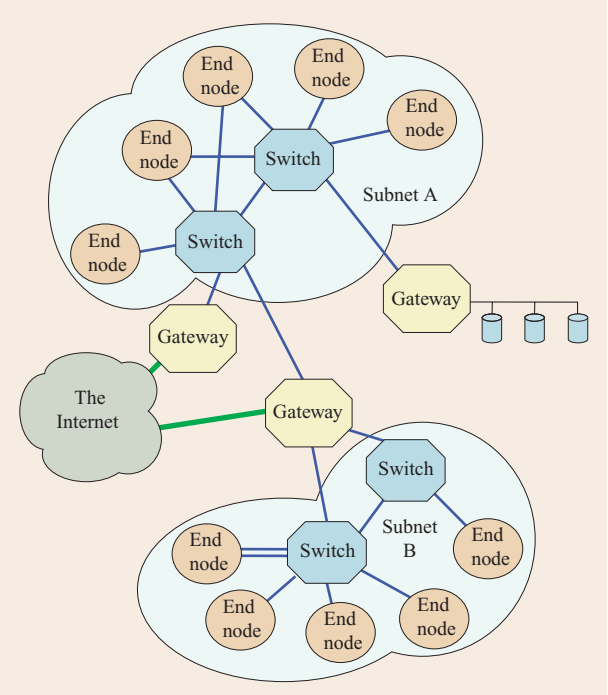


Figure 5

InfiniBand configurations.

links via an InfiniBand-to-Fibre Channel or InfiniBand-to-Ethernet gateway. It is also possible to scale the I/O beyond that provided by the BladeCenter IB switch by connecting to one or more external InfiniBand I/O chassis.

Myrinet

Myrinet** is a high-performance clustering interconnect option for servers created by Myricom. It is used predominantly for computationally demanding scientific and engineering applications and for data-intensive Web and database applications. Since complex technical problems that were once reserved for supercomputers are now being addressed with a combination of programming techniques, commodity computers, and high-performance, low-latency networks such as Myrinet, Myricom has been able to establish itself as a leader in this area.

Myrinet interconnect provides a BladeCenter implementation with the M3S-PCIXD-2-I interface, which is functionally identical to the M3F-PCIXD-2 version of the "D card" except that the Myrinet link outgoing from the blade is electrical signaling rather than fiber. The Myrinet links are carried from the blades across the BladeCenter midplane to the switch module bay. By use of a Myrinet I/O expansion adapter and the optical passthrough module (OPM), blades have access to a full-duplex 2.0-Gb/s data rate (2.5-Gb/s baud rate) link. The OPM converts the Myrinet signaling from the 14 blades to four quad-fiber links. Each fiber ribbon cable is terminated on the opposite end with four LC fiber pairs that can be plugged into a Myrinet switch. Myrinet software support for the "D card" interfaces is based on the Myrinet GM-2, with its usual suite of MPICH-GM, VI-GM, and Sockets-GM middleware.

Passthrough modules

While the integration of the switch technology into the chassis provides several benefits, direct blade connectivity to the external network infrastructure may be preferred for some applications. For cases such as these, the system architecture allows for copper and optical passthrough modules that perform no internal switching. Using Figure 2 as a reference, this is accomplished by passing the 14 internal processor blade I/O interfaces through the switch module complex and performing the necessary conversion to the appropriate external media (copper or optical fiber).

The copper passthrough module (CPM) is integrated into the chassis as a switch module; however, it performs no switching function within the chassis. Instead, a single CPM passes all 14 1-Gb/s connections (one from each blade) to interfaces on the module that are external to the chassis. This solution may fit well for those cases in which sufficient external 1-Gb/s connections are already available on an external switch, thus providing no need for internal switching. The CPM supports only Gigabit Ethernet connectivity.

The OPM is similar to the CPM in that it provides 14 connections (one from each blade) to external interfaces. Two differences between the CPM and OPM are that the CPM supports only Ethernet, while the OPM supports Ethernet, InfiniBand, and Myrinet, and that the OPM supports multiple speeds in relation to the media. On this basis, the OPM creates fiber connections for Ethernet connectivity *only* in switch module bays 1 and 2, since the blade I/O is fixed. Similarly, the OPM handles Ethernet, Myrinet, and Fibre Channel in switch module positions 3 and 4, depending on the I/O expansion adapter installed on the blades. That is, for switch module bays 3 and 4, the protocol used for each blade depends solely on the installed blade I/O expansion adapter and the external switch connection. The OPM is the only option that will allow the blades within the chassis to implement a mix of I/O expansion adapters.

It should be noted that the CPM and OPM contribute little to the cable simplification and reduction or to the designed tight integration of servers and networking. Also, internal chassis operations that require Ethernet switching in switch module bays 1 and/or 2, such as SOL,

are disrupted by the passthrough modules. However, they do provide flexibility to mix and match the connectivity of protocols within a BladeCenter system and connectivity to the external environment, with absolutely no inherent oversubscription.

Services supported by the BladeCenter architecture

To illustrate the flexibility of the BladeCenter architecture and how the technology described above may be applied, examples are given below covering workload management, application and security of VLANs, virtualization, virtual infrastructure, and grid.

Workload management

Workload management is often deployed to proactively shift workload on the basis of the current state of the system, server, and/or networking metrics. Recognizing the importance of workload management technology, Cisco, F5 Networks, Foundry Networks, Nortel, and others have applied this technology to Web clusters by enabling networking products with load-balancing and content-switching functions. The BladeCenter chassis supports these switching functions with two models. The first model is with an integrated Nortel Layer 2–7 Gigabit Ethernet Switch Module, which allows a chassis to appear as a virtual service with a single virtual IP (VIP) address. The second model is for one or more chassis to be connected to an external Cisco switch (e.g., Catalyst** 6500) containing a Cisco Content Switching Module.

Using information in the network packet header, networking products of this type dynamically redirect work requests to servers on the basis of the server performance, health, power, or other aspects of the system. That is, Layer 4 switches use information up through the Layer 4 packet header to recognize an IP flow based on its 5-tuple (e.g., source IP address and port, destination IP address and port, protocol type) to determine how it should be directed. Similarly, Layer 7 switches use information up through the Layer 7 packet header to determine how it should be directed. Thus, the application of workload management to vendor networking product has enabled them to be an important component of the data center by providing a means for directing traffic to appropriate servers.

As mentioned, load balancing or Layer 4 switching requires that a switch be TCP-aware (i.e., Layer 4 in the OSI model) in the sense that it must be able to identify a new TCP connection, assign a TCP connection to a real server, and ensure that all ensuing packets related to the TCP connection continue to be sent to the same real server. TCP connection dynamics between two end stations may change with the introduction of intermediate Layer 4 switching, such that Layer 4

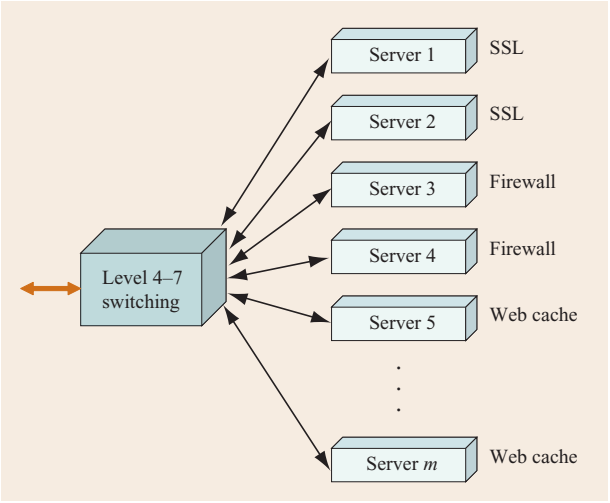


Figure 6

Load balancing.

switching implements the concept of *virtual services* that are indexed with VIP addresses. By presenting a VIP for a service provided by multiple servers or blades, metrics are used to select which server or blade in a group will receive the next client connection. Some options for these metrics include minimum misses, hash, least connections, round robin, and response time [14]. Perfectly equal load balancing between server blades in a virtual service pool is generally difficult to achieve, but is also generally irrelevant. The main purpose of load balancing is to maintain application availability at whatever level of performance the currently active resources will allow. In practice, this generally means keeping all servers working at relatively equal but efficient operating points. To accommodate changes in the workload, adding servers to a virtual service pool or removing them from it may occur when the workload is increased or decreased. **Figure 6** shows how m servers are physically connected to a Layer 4 and/or 7 switch. By recognizing the application type (e.g., HTTP with TCP port number equal to "80"), a Layer 4 switch could balance the load across the two Secure Sockets Layer (SSL) devices, the two firewall servers, or the multiple Web cache servers.

For example, consider a service whose Domain Name Server (DNS) name is *Service_A.com*. Normally, there would be a server set up somewhere with that host name and IP address. With Layer 4 switching, the switch module itself takes ownership of the IP address as a VIP and has multiple "real" server blades behind it capable of delivering the service *Service_A.com*, whose addresses can be arbitrarily assigned, since they are of only local significance. Using this approach, a chassis containing

14 physical blades can be viewed as a single service connected to the network. Additional examples of load-balancing techniques can be found in [1, 2, 14], with a comparison of approaches given in [27]. The ability to provide a virtual service also enables autonomic functions for power management [28] and improving dependability from an end-user perspective [29].

As described earlier, to perform content or Layer 7 switching, the switch must be aware of information—for example, Uniform Resource Locators (URLs), session identifications, and cookies—in the application header of a packet to direct requests to appropriate servers. Content switches also perform load-balancing functions when multiple physical servers are required. For example, URL-based server load balancing allows optimization of resource access and server performance so that content dispersion can be optimized by making load-balancing decisions on the entire path and file name of each URL. URL requests are load-balanced among multiple servers matching the URL according to the load-balancing metric configured for the real server group (e.g., *least connections*). Using Figure 5 as a reference and assuming that m is greater than 6, servers 5 and 6 could be load-balanced for a specific URL, while the remaining Web cache servers are load-balanced for HTTP applications in general on the basis of Layer 4 information (i.e., TCP port number is "80").

Similarly, cookies can be used to provide preferential services for customers, ensuring that certain users are offered better access to resources than other users when site resources are scarce. A Web server could authenticate a user via a password and then set cookies to classify a particular customer by groups or priorities. Using cookies, the switch can distinguish traffic by individuals or groups of users and place them in groups or communities that are redirected to better resources and receive better services than all other users. A detailed example covering content switching can be found in [30].

Virtual LANs

Good network design practices limit the number of stations that share a common broadcast domain to a few hundred. This is sometimes referred to as a Layer 2 broadcast domain and serves to limit the flow of broadcast traffic within a network by controlling the number of stations that can be the source of broadcast traffic. Separation of stations can be accomplished by creating networks that are physically separated. However, a more common practice today is to create VLANs, which maintain a virtual separation among stations that share a common physical network infrastructure.

VLANs are usually implemented within a Layer 2 switch to provide an additional level of control with regard to packet handling. The IEEE Standard 802.1q defines an optional 4-byte field, known as the *VLAN tag*, which may be included in the packet header. A 12-bit VLAN ID uniquely delineates up to 4,094 VLANs. The switch uses the tag information along with the Layer 2 destination address to determine which ports are allowed to receive the packet.

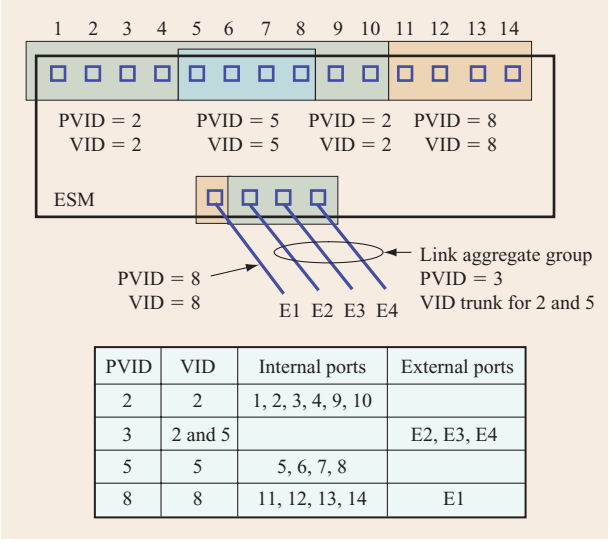
VLANs can be used to partition the BladeCenter switch ports into logical groups to prevent data traffic for one group from being seen by the other groups. The two most common approaches for assigning VLANs allows administrators to choose either physical or logical separation. Physical separation is maintained through the use of port-based VLANs. With this scheme, a switch port can be assigned to only one port VLAN. Ports that share a common port VLAN ID (PVID) are thus within the same VLAN or broadcast domain. The internal switch fabric prevents packets from crossing between ports that are assigned different PVIDs.

The use of port VLANs and physical separation at the external switch ports enables servers within the same chassis to participate in both secure and nonsecure networks while maintaining the data traffic separation that is required for server blades to have secure access to either network. This is quite different from standalone server systems, in which the networking elements are separated from the servers and are interconnected via cables. This separation allows servers to implement multiple NICs for simultaneous connection to both secure (trusted networks) and nonsecure (nontrusted) networks.

As shown in Figure 4, port VLANs allow selected blade server ports to be logically associated with one or more external ports in the same VLAN. The external ports provide the interconnect links to the external network infrastructure, with one or more external ports sharing a PVID with a subset of blades in each VLAN. As shown in the example illustrated by Figure 4, blade servers 1–4 and 9–10 could be attached to the secure network, while blade servers 5–8 and 11–14 are attached to separate nonsecure networks.

The ESM maintains total isolation among these three networks. The use of port VLANs is the recommended VLAN scheme to ensure that no data packets flow between blades within separate VLANs within the ESM. Since a blade can belong to only one port-based VLAN, the switch fabric will maintain traffic separation among the ports and cannot be subverted without compromising the network configuration.

As mentioned above, a port can be a member of multiple 802.1q VLANs in addition to the single port-based VLAN. The ESM must rely upon the 4-byte VLAN tag to associate a packet with the appropriate subset of member ports. An example of overlapped port VLANs and q-tagged VLANs is shown in **Figure 7**. External ports



PVID	VID	Internal ports	External ports
2	2	1, 2, 3, 4, 9, 10	
3	2 and 5		E2, E3, E4
5	5	5, 6, 7, 8	
8	8	11, 12, 13, 14	E1

Figure 7

VLAN example.

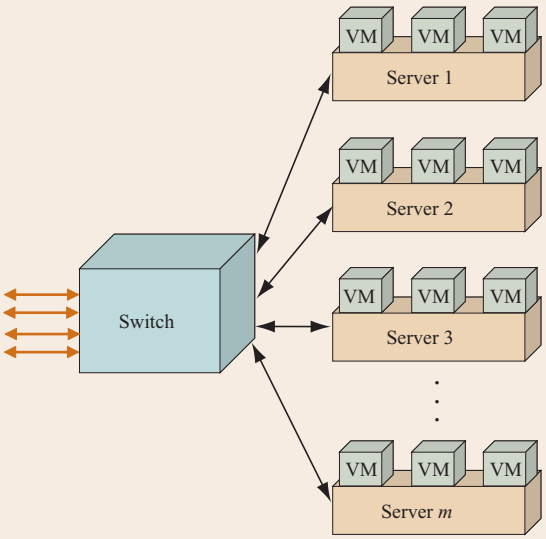


Figure 8

Virtual machine technology.

E2, E3, and E4 are configured as a three-port link aggregate group and as a VLAN trunk in order to transport packets associated with VLANs 2 and 5.

A VLAN trunk permits data traffic from multiple tagged VLANs to share the same physical media, and thus the VLAN separation can be extended throughout the network infrastructure. The 802.1q VLAN scheme may be vulnerable to attempts to subvert the mechanisms within a switch to maintain separation [31–33]. In general, security exposures are typically attributed to systems and VLAN trunks that are not configured properly. Also, improvements in the underlying switch fabric components have reduced the security exposure associated with earlier designs.

Virtualization

As shown in **Figure 8**, each of the server blades can support virtual machine (VM) technology, such as VMware** virtual infrastructure [34–36], in order to share the blade physical resources by hosting multiple instances of OS images. In addition to the blade being shared, the networking infrastructure can also be shared with the use of VLAN technology, and security can be maintained between VMs. For example, each VM shown in Figure 8 can be logically associated with an independent VLAN configured on the switch, so that with three VMs per blade, there could be a total of $3 \times m$ total VLANs configured internally and trunked out to the uplinks of the switch.

VMware virtual infrastructure also supports NIC teaming so that the two Ethernet NICs on each blade can be grouped to appear as a single networking device called a *bond*. The VMware NIC teaming function has the same benefits as described for NIC teaming in the section on redundant switch failover above, but with the added benefits of VM technology. For example, when a BladeCenter system is used for server consolidation, intelligent workload management can be achieved by moving VMs along with their respective applications to different blades to optimize around server and network performance without disruption to the user.

A component of VMware intelligent workload management is accomplished with VMware VMotion** technology [34]. VMotion is capable of transferring the entire system/blade and memory state of a running VM from one VMware ESX Server** to another if all of the systems disk information is located on a shared storage infrastructure, such as a storage area network (SAN). Ongoing memory transactions are transferred in a bitmap to the other system, and when all system state has been transferred, VMotion suspends the source VM, transfers the bitmap, then resumes the VM on the target ESX blade. Since the bitmap transferred is small, the process takes less than two seconds on a Gigabit Ethernet network and appears as no more than a temporary network loss to the application, service, and user. This is accomplished by leveraging the windowing operation of the TCP protocol for guaranteed delivery of lost packets. In addition to workload management in general, this approach is useful for achieving highly dependable

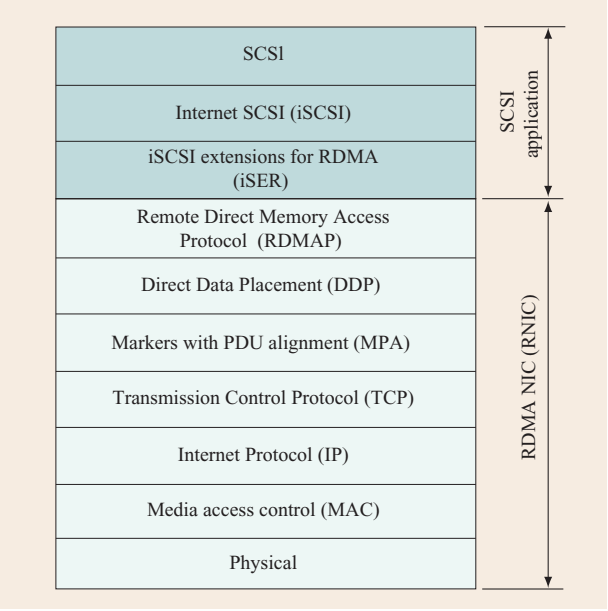


Figure 9

RNIC with SCSI application layers. (PDU: protocol data unit.)

applications by providing security between virtual domains, minimizing failover time, and limiting disruption to the end user. A BladeCenter system enhances dependability by enabling a single control domain for both the physical and virtual environments.

While the focus here is on VMware, virtualization technology offered by others on industry platforms includes Microsoft Virtual Server [37] and the Xen project [38].

Grid

In conjunction with Figure 5, a brief description of the clustering and I/O capabilities of InfiniBand technology was presented. With the subnet being a local cluster and the aggregation of clusters being formed by interconnecting these over the Internet or a private network, it can also be used for describing an underlying grid fabric. As defined in [39], this topology qualifies as a grid fabric layer, with a *resource* being a logical entity such as a distributed file system, computer cluster, or distributed computer pool. Additional details on the expected purpose and architecture of future grid systems are presented in [40].

Solutions resulting from IBM partnerships with Cisco (Layer 2 switching) and Nortel (Layer 2–7 switches), TopSpin (InfiniBand), and Myricom (Myrinet) can be used to enable BladeCenter systems for a variety of grid applications by enabling the clustering of up to 14 blades in a single chassis on the appropriate interconnect fabric. Multiple BladeCenter systems can then be interconnected externally with one or more high-end backbone switches to create a much larger cluster. It has been shown that some grid applications will suffice with large numbers of servers interconnected with 1-Gb/s Ethernet bandwidth [41], while other applications will require the bandwidth and performance guarantees of InfiniBand.

Future enhancements and extensions

The following networking enhancements are planned to further advance areas of InfiniBand, Ethernet, and workload management.

InfiniBand

An area of focus with InfiniBand 2.0 is the application of double-data-rate (DDR) technology for communications. That is, the initial baud rate of 2.5 Gb/s per channel (2.0-Gb/s data rate) for an IB-1x link will be doubled to a 5.0-Gb/s baud rate per channel (4.0-Gb/s data rate). As with InfiniBand 1.0, this speed will also apply to IB-4x and IB-12x cabling, with IB-8x being newly supported. This increase in bandwidth will be significant for data center computing and database clustering as the scale-out model continues to evolve in parallel with the convergence of communication, storage, and clustering traffic onto a common fabric.

InfiniBand will also provide increased flexibility for supporting the scale-out model. Depending on the application, there may be a need to increase compute power by adding servers or blades, or there may be a need to increase the I/O capability by adding adapters or attaching one or more I/O chassis. For example, a single BladeCenter system could be connected to multiple I/O chassis via a common InfiniBand fabric to provide a large number of InfiniBand, Ethernet, Fibre Channel, or other types of external connectivity.

Ethernet and networking offload

Like InfiniBand, Ethernet is slowly evolving through standards to become a system area network fabric through extensions to support clustering and storage applications. The support of TCP/IP offload technology provided by network adapters and OS providers is a primary step for enabling networking convergence of this type. With the support of TCP/IP offload, the TCP/IP bottleneck can be overcome, and additional services (e.g., iSCSI [42] and RDMA [43]) provide clustering and I/O capabilities similar to those of InfiniBand. That is, using Ethernet as the common fabric, the RDMA protocol sits on top of the MPA [44] and DDP [45] layers, as shown in **Figure 9**, to provide applications with direct memory access to the memory of another computer system with significantly less latency.

The initial definition of the iSCSI protocol describes a means to transport SCSI commands over TCP. However, more recently a second option was defined to leverage the direct memory access provided by the RDMA service, as shown in Figure 9. This approach is known as iSCSI extensions for RDMA (iSER) [46]. These technologies will be supported within the BladeCenter system as the network adapter and OS support become enabled.

10-Gb/s Ethernet

Ethernet standards and product support for 10 Gb/s over both copper and fiber have already been introduced. However, the network and I/O component costs, and subsequent end-to-end connection costs, are much higher than for 1-Gb/s interfaces. Also, the drive distances over copper and low-end fiber media are limited. These design challenges will be addressed by the industry as a whole and will soon result in 10-Gb/s Ethernet as a viable technology option for internetworking servers within a BladeCenter chassis or among servers within a data center.

Enterprise workload management

The IBM Enterprise Workload Manager (EWLM) [47] is part of the IBM Virtualization Engine (VE) offering [48] and can dynamically monitor and manage distributed heterogeneous workloads to optimize the operation of applications. Infrastructure simplification is a key area of focus for BladeCenter products, and the alignment with EWLM and VE is consistent with this direction. EWLM simplifies the management of IT resources by creating a consolidated logical view of resources across a processor complex, cluster, or distributed network through automation and virtualization.

As described in the Introduction, the multi-tiered framework shown in Figure 1 is very common. The servers could be physically located in one data center or could be spread across sites in different cities or countries. EWLM is an implementation of policy-based performance management, with the scope being a set of servers logically grouped into what is called an enterprise workload *management domain*. The set of servers included in the domain have some type of relationship, such as a group supporting a particular line of business (which may consist of multiple business processes spread across a few servers or a thousand servers). On each server OS instance in the domain, a thin layer of enterprise workload management logic, called the *managed server*, is installed. The managed server layer is positioned between applications and the OS to gather resource usage and delay statistics known to the OS.

A second role of the managed server layer is to gather relevant transaction-related statistics from middleware applications. The application middleware implementations, such as the IBM WebSphere* Application Server, are notified when a piece of work starts and stops, and the middleware is notified when a piece of work has been routed to another server for processing (for example when a Web server routes a servlet request to a WebSphere Application Server). The managed server layer dynamically constructs a server-level view describing relationships between transaction segments (known by the applications) with resource consumption data (known by the OS). A summary of this information is periodically sent to the *domain manager*, where the information is gathered together from all of the servers in the management domain to form a global view.

When incoming requests arrive at a load balancer, the load balancer may have limited information about the status of an application or the performance of the servers to which it is routing. EWLM does not route the work itself, but provides recommendations to the routing entity using the IBM Server/Application State Protocol (SASP). Through SASP messages, a load balancer can notify the domain manager which systems and applications can be load-balanced, and the domain manager can make recommendations to load balancers regarding how to distribute work based on the statistics gathered and policies set. However, it is up to the load balancer to actually make use of EWLM recommendations to route incoming requests to the members.

In relation to BladeCenter servers, the external Cisco Content Switching Module switch and the integrated Nortel Layer 2–7 Gigabit Ethernet Switch Module both support the EWLM function and SASP for enhanced workload management. As previously described, the Cisco switch can be placed among multiple BladeCenter chassis, while one or more Nortel switches can be integrated into each BladeCenter chassis. With this function, the combining of EWLM with BladeCenter servers and integrated switches brings an even tighter coupling between the server, network, and application layers.

Conclusion

This paper describes how the BladeCenter networking architecture and technology is converged with server and management technology to consolidate and simplify infrastructures built around the scale-out model. Areas of focus include showing incorporation to networking standards, discussing some of the switch and I/O technology options, and sharing example applications and some future directions. Recent advances in networking technology that further enable scalable and dependable computing were discussed in the context of how these technologies could be leveraged within BladeCenter products.

Blade server architectures will have a definite place in meeting customer requirements of the future, but it will be important to continue leveraging industry and customer trends, developing appropriate technology, and establishing new standards. As described, the integration of servers and network subsystems within the BladeCenter chassis helps to simplify systems management in the data center. In the future, this management may extend to include global workload management and virtual organizations enabled by grid technology.

Acknowledgments

The authors gratefully acknowledge constructive comments from Dr. Satish Gupta of the IBM Systems and Technology Group, Dr. Rick Harper of IBM Research, and the reviewers, who provided their comments anonymously.

References

1. A. Iyengar, J. Challenger, D. Dias, and P. Dantzig, "High-Performance Web Site Design Techniques," *IEEE Internet Computing* **4**, No. 2, 17–26 (March 2000).
2. D. Oppenheimer and D. A. Patterson, "Architecture and Dependability of Large-Scale Internet Services," *IEEE Internet Computing* **6**, No. 5, 41–49 (September 2002).
3. D. A. Menasce, "Trade-Offs in Designing Web Clusters," *IEEE Internet Computing* **6**, No. 5, 76–80 (September 2002).
4. G. Pfister, *In Search of Clusters: The Ongoing Battle of Lowly Parallel Computers*, Prentice-Hall, Inc., Upper Saddle River, NJ, 1998.
5. IBM Corporation, The IBM 8260 Multiprotocol Switching Hub; see *http://www.rmav.arauc.br/equipamentos/8260/8260.html*.
6. J. Parker, M. Koskinen, B. Kuppers, and J. Reichel, "The IBM 2220 Nways Switch: Concepts and Products," *IBM Redbooks,* December 10, 1998; see *http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/sg244307.html*.
7. Cisco Systems, Inc., The Cisco Catalyst 6500 Series Switches; see *http://www.cisco.com/en/US/products/hw/switches/ps708/*.
8. D. M. Desai, T. M. Bradicich, D. Champion, W. G. Holland, and B. M. Kreuz, "BladeCenter System Overview," *IBM J. Res. & Dev.* **49**, No. 6, 809–821 (2005, this issue).
9. InfiniBand Trade Association; see *http://www.infinibandta.org/home*.
10. M. J. Crippen, R. K. Alo, D. Champion, R. M. Clemo, C. M. Grosser, N. J. Gruendler, M. S. Mansuria, J. A. Matteson, M. S. Miller, and B. A. Trumbo, "BladeCenter Packaging, Power, and Cooling," *IBM J. Res. & Dev.* **49**, No. 6, 887–904 (2005, this issue).
11. J. E. Hughes, P. S. Patel, I. R. Zapata, T. D. Pahel, Jr., J. P. Wong, D. M. Desai, and B. D. Herrman, "BladeCenter Midplane and Media Interface Card," *IBM J. Res. & Dev.* **49**, No. 6, 823–836 (2005, this issue).
12. J. E. Hughes, M. L. Scollard, R. Land, J. Parsonese, C. C. West, V. A. Stankevich, C. L. Purrington, D. Q. Hoang, G. R. Shippy, M. L. Loeb, M. W. Williams, B. A. Smith, and D. M. Desai, "BladeCenter Processor Blades, I/O Expansion Adapters, and Units," *IBM J. Res. & Dev.* **49**, No. 6, 837–859 (2005, this issue).
13. J. F. Kurose and K. W. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet*, Addison-Wesley Publishing Co., Boston, 2005.
14. S. Hochstetler, D. Green, E. Johanson, and N. Strole, "IBM eServer BladeCenter Layer 2–7 Network Switching," *IBM Redpaper,* REDP-3755-00, January 2004; see *http://www.redbooks.ibm.com/redpapers/pdfs/redp3755.pdf*.
15. RDMA Consortium; see *http://www.rdmaconsortium.org/home*.
16. K. Hwang and Z. Xu, *Scalable Parallel Computing: Technology, Architecture, Programming*, McGraw-Hill Book Co., Inc., San Francisco, 1998.
17. K. Hwang, *Advanced Computer Architecture: Parallelism, Scalability, Programmability*, McGraw-Hill Book Co., Inc., New York, 1993.
18. R. A. Sahner, K. S. Trivedi, and A. Puliafito, *Performance and Reliability Analysis of Computer Systems,* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.
19. S. W. Hunter, Performance and Dependability Modeling of Some Switched Network Applications, Ph.D. dissertation, Duke University, Durham, NC, March 1997.
20. R. K. Iyer, "Introduction to the IEEE Transactions on Dependable and Secure Computing," *IEEE Trans. Dependable & Secure Computing* **1**, No. 1, 2–3 (January–March 2004).
21. A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic Concepts and Taxonomy of Dependable and Secure Computing," *IEEE Trans. Dependable & Secure Computing* **1**, No. 1, 11–33 (January–March 2004).
22. D. M. Nicole, W. H. Sanders, and K. S. Trivedi, "Model-Based Evaluation: From Dependability to Security," *IEEE Trans. Dependable & Secure Computing* **1**, No. 1, 48–65 (January–March 2004).
23. Broadcom Corporation, Broadcom NetExtreme Gigabit Ethernet Teaming, Broadcom; see *http://www.broadcom.com/collateral/wp/570X-WP100-R.pdf*.
24. W. G. Holland, P. Caporale, D. S. Keener, A. B. McNeill, and T. B. Vojnovich, "BladeCenter Storage," *IBM J. Res. & Dev.* **49**, No. 6, 921–939 (2005, this issue).
25. IEEE 802.3ap Backplane Ethernet Task Force; see *http://www.ieee802.org/3/ap/*.
26. IEEE 802.3ar Congestion Management Task Force; see *http://www.ieee802.org/3/ar/*.
27. V. Cardellini, M. Colajanni, and P. S. Yu, "Dynamic Load Balancing on Web Server Systems," *IEEE Internet Computing* **3**, No. 3, 28–39 (June 1999).
28. D. J. Bradley, R. E. Harper, and S. W. Hunter, "Workload-Based Power Management for Parallel Computer Systems," *IBM J. Res. & Dev.* **47**, No. 5/6, 703–718 (2003).
29. V. Castelli, R. E. Harper, P. Heidelberger, S. W. Hunter, K. S. Trivedi, K. Vaidyanathan, and W. P. Zeggert, "Proactive Management of Software Aging," *IBM J. Res. & Dev.* **45**, No. 2, 311–332 (2001).
30. G. Apostolopoulos, D. Aubespin, V. Peris, P. Pradhan, and D. Saha, "Design, Implementation and Performance of a Content-Based Switch," *Proceedings of Infocom 2000*, 2000, pp. 1117–1126.
31. D. Taylor, "VLAN Security Test Report," The SANS Institute, July 12, 2000; see *http://www.sans.org/resources/idfaq/vlan.php*.
32. C. Hoffman, "VLAN Security in the LAN and MAN Environment," The SANS Institute, April 27, 2003; see *http://www.giac.org/certified_professionals/practicals/gsec/2850.php*.
33. Cisco Systems Inc., "Virtual LAN Security Best Practices," VLAN security white paper; see *http://*

www.cisco.com/en/US/products/hw/switches/ps708/products_white_paper09186a008013159f.shtml.
34. VMware, Inc.; see *http://www.vmware.com/products/vmanage/vc_features.html*.
35. VMware, Inc., Configuring and Installing IBM BladeCenter; see *http://www.vmware.com/pdf/esx21_IBM_blade.pdf*.
36. VMware, Inc., VMware ESX 2.1 NIC Bonding and VLANs on BladeCenter HS20; see *http://www.vmware.com/pdf/esx21_IBM_NIC_VLAN.pdf*.
37. Microsoft Corporation, Microsoft Virtual Server 2005; see *http://www.microsoft.com/windowsserversystem/virtualserver/default.mspx*.
38. University of Cambridge Computer Laboratory, Xen: The Xen Virtual Machine Monitor; see *http://www.cl.cam.ac.uk/Research/SRG/netos/xen/*.
39. I. Foster, C. Kesselman, and S. Tueke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *Intl. J. Supercomputer Appl.* **15**, No. 3, 200–222 (2001); see *http://www.globus.org/alliance/publications/papers/anatomy.pdf*.
40. I. Foster and C. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*, "Computational Grids," Chapter 2, Morgan Kaufmann Publishing Co., Los Altos, CA, 1999.
41. Force10 Networks, Inc., Building Scalable, High Performance Cluster/Grid Networks: The Role of Ethernet; see *http://www.force10networks.com/applications/roe.asp*.
42. J. Satran, K. Meth, C. Sapuntzakis, M. Chadalapaka, and E. Zeidner, "Internet Small Computer Systems Interface (iSCSI)," RFC 3720, April 2004; see *http://www.alliedtelesyn.co.nz/resources/rfc37xx.html*.
43. R. Recio, P. Culley, D. Garcia, and J. Hilland, "An RDMA Protocol Specification," Internet Draft draft-ietf-iwarp-rdma-01.txt, The Internet Society, February 2003.
44. P. Culley, U. Elzur, R. Reico, S. Bailey, and J. Carrier, "Marker PDU Aligned Framing for TCP Specification," Internet Draft draft-ietf-iwarp-mpa-02.txt, The Internet Society, February 2003.
45. H. Shah, J. Pinkerton, R. Reico, and P. Culley, "Direct Data Placement over Reliable Transports," Internet Draft draft-ietf-iwarp-ddp-01.txt, The Internet Society, February 2003.
46. M. Chadalapaka, H. Shah, U. Elzur, P. Thaler, and M. Ko, "A Study of iSCSI Extensions for RDMA (iSER)," *Proceedings of the Association for Computing Machinery (ACM) Special Interest Group on Data Communication (SIGCOMM) Workshop on Network-I/O*, August 2003, pp. 209–219.
47. P. Bari, C. Covill, K. Majewski, C. Perzl, M. Radford, K. Satoh, D. Tonelli, and L. Winkelbauer, "IBM Enterprise Workload Manager (EWLM)," *IBM Redbooks,* August 27, 2004; see *http://www.redbooks.ibm.com/abstracts/sg246350.html?Open*.
48. S. E. Bach, M. Cathcart, M. Ferrier, C. Matthys, and J. Schuneman, "Virtualization and the On Demand Business," *IBM Redpaper,* August 18, 2004; see *http://www.redbooks.ibm.com/abstracts/redp9115.html?Open*.

Received December 16, 2004; accepted for publication March 8, 2005; Internet publication October 13, 2005

Steven W. Hunter *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (hunters@us.ibm.com)*. Dr. Hunter is an IBM Distinguished Engineer in the xSeries/BladeCenter Architecture and Technology Department. He received a B.S.E.E. degree from Auburn University in 1984, an M.S.E.E. degree from North Carolina State University (NCSU) in 1988, and a Ph.D. degree from Duke University in 1997. Since 1997 he has worked in the xSeries organization, where he has been an architect and designer for network and clustering technology, the InfiniBand Architecture, and BladeCenter systems. He is a licensed Professional Engineer, a Senior Member of the IEEE, and an adjunct professor at NCSU, where he teaches networking courses. Dr. Hunter holds patents spanning both hardware and software, has published numerous papers, and has presented at a variety of conferences and symposiums.

Norman C. Strole *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (ncstrole@us.ibm.com)*. Dr. Strole is a Senior Technical Staff Member. He received a B.S.E.E. degree from North Carolina State University in 1973 and M.S.E.E. and Ph.D. degrees from Duke University in 1975 and 1980, respectively. He is currently involved with the development of next-generation BladeCenter network subsystems. He is an IBM Master Inventor, with more than 25 patients issued or pending, and has published several technical papers in the field of networking. Dr. Strole taught computer network architecture and switching theory courses as an Adjunct Professor at Duke University for 16 years. He is a Senior Member of the IEEE.

David W. Cosby *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (cosbydw@us.ibm.com)*. Mr. Cosby works in BladeCenter ecosystem development. He received dual B.S. degrees in electrical and computer engineering from North Carolina State University in 1991, followed by an M.S. degree in computer engineering in 1993. He joined the IBM Networking Hardware Division in 1992, focusing on high-performance device drivers. Before joining the BladeCenter team, Mr. Cosby worked in the IBM Microelectronics Division as the technical leader for network processor system integration.

David M. Green *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (greendm@us.ibm.com)*. Mr. Green is a Technical Project Manager for the BladeCenter Options Development team. He received a B.S. degree in computer systems from the University of North Carolina at Greensboro in 1995. He joined IBM in 1997. Mr. Green is a contributing author on two previous Redpapers for networking on BladeCenter systems.