BladeCenter storage

W. G. Holland
P. L. Caporale
D. S. Keener
A. B. McNeill
T. B. Vojnovich

This paper describes the available storage options for the IBM eServer™ BladeCenter® system—local hard disk drives on each processor blade, network-attached storage accessed over Ethernet, and storage area network (SAN)-attached storage accessed over Fibre Channel—highlighting where each provides a compelling or unique storage solution. The basic selection criteria are presented, focusing on the most significant attributes of each option. Additional information is provided for emerging technologies, iSCSI (SAN-attached storage accessed over Ethernet) and InfiniBand®, and the technologies and attributes that have unique importance in a blade-server environment.

Introduction

When the IBM eServer* BladeCenter* system was first conceived in the late summer of 1999, storage solutions were a critical design aspect of the overall architecture. Current and anticipated storage technology and market transitions were factored into the architecture and subsequent product design. These factors included the following:

- Most servers at that time had between one and six local Small Computer Systems Interface (SCSI) drives.
- Servers requiring more storage would add a PCI RAID (Peripheral Component Interface redundant array of independent disks) SCSI adapter connected to an external storage JBOD (just-a-bunch-of-disks) array.
- Servers provided both a compact disk read-only memory (CD-ROM) drive for installing operating system and large application software packages and a 3.5-in. floppy diskette drive for installing smaller software packages, basic input/output system (BIOS) updates, and patches.
- Network attached storage (NAS) file-level access to remote storage had become very common for end-user access to data on servers. But NAS had only limited applications as a server storage solution because of its performance, which was limited by the available 100-Mb/s Ethernet local area network (LAN), and the processing overhead of the network protocols.
- Fibre Channel storage area networking (SAN) providing access to remote block-level storage was becoming more common.
- The anticipated trend was that remote storage would surpass local storage for both capacity and performance. The PCI RAID adapter was simply not going to be able to provide the space, power, and cooling necessary to deliver the performance and connectivity available from a larger external storage controller.
- Beyond Fibre Channel SANs, Internet Protocol (IP)-based SAN storage had become the next wave, anticipated to compete directly with Fibre Channel storage within just a few years. Although IBM did announce an early Internet SCSI (iSCSI) product [1]—the 200i iSCSI storage controller—in February 2001, iSCSI did not capture much of the SAN market with the first wave of products. With broader support and a wider range of products in the market, IBM reentered the iSCSI storage controller market with the release of the TotalStorage* DS300 [2] in late 2004.
- In the late 1990s, high-bandwidth, low-latency communications over a single converged network was the objective for two competing interconnect standards efforts—Future I/O (FIO) and Next Generation I/O (NGIO). The two standards efforts merged in August 1999, and the combined standard was renamed *InfiniBand*** in October of that year.

Thus, the storage solution portfolio that the BladeCenter design would require in order to be a competitive server included local internal SCSI disks, Ethernet NAS connectivity, and Fibre Channel SAN connectivity. Future trends indicated that it would also

0018-8646/05/$5.00 © 2005 IBM

Table 1 Storage solution attributes.

Capacity	Sufficient disk space and quantity. *Capacity* = (*storage capacity per drive*) × (*number of drives*). - Drives allocated for RAID redundancy. - Drives assigned as standby hot spares.
Performance	Sufficient access latency and data transfer bandwidth. *Latency*: Key for reads that stall the server (server waits for response) and writes that must be confirmed before further processing. *Bandwidth*: Pipelined or prefetched reads (data arrives before needed). Most write operations can be pipelined.
Scalability	Quick and simple capacity and/or performance changes. Minimize future constraints by planning for growth from the start. - Capacity and performance may grow predictably as business grows. - New applications may drive abrupt growth or alter the predicted growth. - Seasonal or sporadic events may cause abrupt or temporary growth.
RAS	Reliability, availability, and serviceability. *Reliability*: Likelihood that no component will fail. *Availability*: Likelihood that the application is up, inclusive of failures. *Serviceability*: Minimal time, risk, and cost for repair and maintenance.
Management	Visibility and control of the prior four metrics. *Visibility*: Identifying, tracking, diagnosing, or predicting problems. *Control*: Enabling new functions and resolving or preventing problems.

have to accommodate the emerging iSCSI Ethernet SAN technology and InfiniBand high-bandwidth, low-latency converged networking technology.

Storage solution criteria

A storage solution can be defined through a few attributes, as shown in **Table 1**. The BladeCenter design provides a wide range of storage choices simply because applications and customers differ in the required storage attributes and the time, space, people, and money available to meet those requirements.

Storage solution rollout

When the system was in development, two different deployment models were anticipated. One was that a customer would move an existing application onto a processor blade, keeping other changes to a minimum. Thus, it would be valuable if BladeCenter capabilities were the same as those already relied on by customers' applications. The other deployment model was that new applications would be developed and deployed specifically for a blade-server environment, incorporating more automation, autonomic and remote management, and consolidated storage and networking. All of these capabilities were anticipated to be of growing interest and value to large data centers. It seemed reasonable to plan that early deployments would be very conservative, using existing technologies, while customized applications that would fully exploit the blade-server environment would take longer to develop.

The first BladeCenter storage offerings supported existing applications by providing the following capabilities for connecting storage to the processor blades:

- Local integrated drive electronics (IDE) disk storage in the processor blade for boot, operating system (OS), applications, and small data sets.
- Local hot-swappable SCSI disk storage with RAID mirroring in the BladeCenter storage expansion unit (BSE) for added performance, capacity, availability, and serviceability.
- Remote NAS storage access through Ethernet for larger data repositories and data that would be read simultaneously by multiple servers.
- Support for installation of an optional processor blade I/O expansion card, henceforth referred to as the I/O expansion adapter.
- Chassis-level shared floppy disk and CD-ROM drives to enable local software installation, BIOS updates, and patch applications.

Within six months after that initial offering, the first complete high-performance remote storage option for a

blade server was released. It offered the option of a Fibre Channel I/O expansion adapter and Fibre Channel switch module to provide connectivity to the additional capacity, performance, scalability, and RAS features of Fibre Channel SANs. Over time, additional storage options were released:

- A two-port Ethernet I/O expansion adapter to provide a processor blade with additional LAN bandwidth and/or connectivity to additional LAN segments.
- A PCI expansion unit (PEU), allowing processor blades to use standard IBM ServerProven* PCI adapters.
- Local IDE RAID mirroring for additional availability.
- IDE flash disks, high-reliability local storage for harsh environments.
- Optical passthrough module (OPM) to provide direct external fiber-optic connectivity from each processor blade (without any added switching function) for Ethernet, Fibre Channel, and Myricom Myrinet** I/O.
- Copper passthrough module (CPM) to provide direct external connectivity from each processor blade for Ethernet on RJ45 connectors.
- Local SCSI disk storage with RAID mirroring in the processor blade for added performance, capacity, and availability.
- InfiniBand I/O expansion adapter and InfiniBand switch module options for low-latency clustering and LAN and SAN network convergence onto a single communications fabric.

Each solution pushes the storage solution envelope a little bit wider. Application and storage architects have been increasingly comfortable with this growing family of storage offerings. As expected, applications customized to exploit the best features of blade servers began to emerge. Diskless processor blades that booted directly from Fibre Channel storage became a fully supported offering in response to customers who were deploying such configurations in their production environments in 2003.

At the same time, it was obvious that the marketplace was not just going to shift from legacy storage to some new blade-storage paradigm. Rather, different customers and different applications would settle on the most suitable storage solution and stick with it for stability. In addition to pushing the envelope to enable new storage paradigms, storage solutions were released to provide choice within existing storage paradigms:

- Additional Ethernet switch modules, offering additional networking capabilities [3].
- Additional Fibre Channel switch modules from Brocade Communications Systems that provided full interoperability with existing Brocade SANs. Two models addressed the needs of the smallest one-switch and two-switch Fibre Channel applications at a low price point and the needs of larger-center environments, respectively.
- A Fibre Channel switch module from QLogic provided six external 2-Gb/s Fibre Channel ports. This tripled the previous Fibre Channel switch module available external connection bandwidth for high-storage bandwidth applications.
- Lower-cost remote Fibre Channel and iSCSI disk storage controllers to extend the remote storage options to smaller and more cost-sensitive applications.

Local storage for processor blades

Drive types

When the BladeCenter system was introduced, drives were easily classified into three groups: enterprise class server drives, desktop class drives, and mobile class drives (**Table 2**). Because processor blades are too small to accommodate the 3.5-in. disk drives used in larger servers and desktops, small-form-factor (SFF) 2.5-in. drives were selected. (The BSE section below describes how full-sized disks can be added to processor blades.) The SFF drives also use less power than larger drives. However, the SFF drives used in the original BladeCenter processor blades were custom-modified to remove their laptop-inspired feature of powering off the motor during periods of inactivity. Server applications do not perform well when the disk drives occasionally go to sleep. Additional qualification and testing was done to deliver the best reliability available in a SFF drive.

With the introduction of SFF SCSI drives, followed by SFF serial-attached SCSI (SAS) drives, the higher performance and reliability aspects of enterprise-class drives became available in standard SFF drives (Table 2). SAS [4] is the next-generation SCSI interface. It uses a high-performance serial interface rather than the parallel SCSI bus. The SAS architecture is designed to be compatible with and accept the attachment of desktop serial-attached AT attachment (SATA) disk drives. SATA [5] is the corresponding next-generation ATA [6] interface for IDE drives. This will allow a processor blade to offer a unified disk attachment interface capable of attaching lower-cost SATA drives or more expensive but higher-performance SAS drives.

RAID functionality

RAID [7] uses multiple disks in a coordinated fashion to provide enhanced availability or performance. With two

Table 2 Disk drive options.

	(2002)			(2004–2005)	
	Enterprise	*Desktop*	*Mobile*	*SFF SCSI*	*SFF SAS*
Seek time	Fast	Moderate	Slow	Moderate	Moderate
Rotation speed (rpm)	Fast (15K, 10K)	Moderate (7,200)	Slow (5,400, 4,200)	Moderately fast (10K)	Moderately fast (10K)
Form factor (in.)	3.5	3.5	2.5	2.5	2.5
Interface	SCSI or Fibre Channel	IDE/ATA	IDE/ATA	SCSI	SAS
Key attributes	Performance, reliability	Cost, capacity	Size, power, impact tolerance	Performance, reliability, size	Performance, reliability, size
Relative cost per GB	High	Low	Moderate	High	High

drives, as offered on processor blades, only two RAID levels are available. Both combine the two physical disks into one logical disk: RAID 0 spreads the data across both drives to deliver higher performance, and RAID 1 puts a copy of the same data on each drive to deliver higher availability.

The availability improvement from RAID 1, also called *drive mirroring*, is very good. With almost any failure condition on one drive, the system will continue to operate from the remaining drive. The local disk drives are the only moving parts on the processor blade, and most customers would prefer to be protected from a disk failure, no matter how remote the chance. Providing RAID 1 for local disks complements the BladeCenter high-availability design, in which redundancy is used extensively to improve overall system availability.

RAID support comes in three levels: OS-based RAID, basic RAID, and hardware RAID. Standard storage (OS-based RAID) provides no RAID capability in the local disk subsystem. RAID can be provided via the OS if desired. OS-based RAID solutions are the least expensive. Basic RAID consists of a RAID-aware BIOS for the storage controller and a device driver that contains the RAID algorithms. Basic RAID is superior to the OS-based RAID solution in that it provides redundancy and recovery during the boot process, before the OS is loaded far enough to provide redundancy or recovery. Hardware RAID implements RAID outside the host processor environment. It typically comprises a special storage controller containing a dedicated microprocessor, memory, and firmware to implement RAID functionality.

Local storage implementations

Since the BladeCenter system was introduced in late 2002, there have been three different two-way Intel-based processor blade designs, a four-way Intel processor blade, and a two-way IBM PowerPC*-based processor blade (**Table 3**).

With the release of the two-way Intel processor blade (HS20 model 8843), hardware RAID and SCSI drives are both available on the processor blade. The HS20 model 8843 implements an entry hardware RAID solution via the LSI 1020 single-channel SCSI controller. The 1020 is an industry-unique SCSI controller that can optionally act as an entry RAID controller. Additionally, the 1020 RAID solution is totally integrated into the IBM xSeries* ServeRAID* manager toolset.

BladeCenter storage expansion

BladeCenter storage expansion (BSE) provides the ability to expand the storage capability of the HS20 (8678, 8832) and HS40 processor blades with the attachment of one or two IBM xSeries 3.5-in. Ultra 320 hot-swap drive options (10K or 15K rpm). The BSE contains an LSI 1020 hardware RAID controller. Full-size 3.5-in. drives provide higher capacity and performance than the drives contained on the base processor blade (**Table 4**).

Ethernet NAS

The base Ethernet connectivity consists of at least two 1-Gb/s Ethernet ports on each processor blade. With the support built into all supported operating systems, these can be used to connect to file systems or remote disks on other servers or NAS devices through NAS protocols. Microsoft Windows** generally uses the Common Internet File System (CIFS) protocol to access a disk that another server has chosen to make available over the network. UNIX** and Linux** more commonly use network file system (NFS) for similar network file access. Additional third-party software is available for each of these operating systems to use either or both file access protocols in either environment. The BladeCenter system

Table 3 Processor blade storage implementations.

	Processor	*Local disk*	*Local disk RAID*	*Blade storage expansion*	*BSE RAID*
HS20–8678	Two-way Intel	SFF IDE	Standard	BSE	Hardware
HS20–8832	Two-way Intel	SFF IDE	Basic	BSE	Hardware
HS40–8839	Four-way Intel	SFF IDE	Standard	BSE	Hardware
JS20–8842	Two-way PowerPC	SFF IDE	Standard	(none)	(n/a)
HS20–8843	Two-way Intel	SFF SCSI	Hardware	BSE-2	Hardware

Table 4 Drive details.

	SFF IDE	*SFF IDE flash*	*SFF SCSI*	*Full-size SCSI*
Capacity (GB)	40, 60	1, 2, 4	36, 73	36, 73, 147
Rotation speed (rpm)	5,400	(n/a)	10K	10K or 15K
Interface speed	ATA-100	ATA-100	U320	U320 or U160
Cost	Low	Very high	Low to high	Low to high
Supported	HS20–8678 HS20–8832 HS40 JS20	HS20–8678 HS20–8832 HS40 JS20	HS20–8843	BSE BSE-2

provides the basic 1-Gb/s Ethernet connectivity to enable each processor blade to access NAS storage or to provide NAS storage for use by other devices.

The addition of an Ethernet I/O expansion adapter can provide additional network and NAS bandwidth to the server processor blade through two additional 1-Gb/s Ethernet links. Providing these additional network links can also allow separation of storage and networking traffic or isolation of the public (or external) network from the private (or internal) network. LAN segment isolation can provide a guarantee that performance or functional problems on one network will not automatically disrupt other networks.

Additional discussion of the BladeCenter networking capabilities can be found in [3]. Described in that paper are the technology advances for improving the performance or functionality of the Ethernet connection that will provide the same benefits to NAS communications. Incorporation of advancing Ethernet functions into both the processor blade network interface card (TCP/IP offload and remote direct memory access–RDMA [8]) and the switch module (Layer 3 switch, L4–7 switching, virtual LANs, priority) contribute to increasing NAS performance and functionality.

Specific NAS acceleration technology is also being developed that goes beyond benefiting the wide range of network communications and directly aims to enhance storage access over the network. Standards work on the NFS protocol includes the addition of RDMA mechanisms to the file system itself. This would increase the efficiency of networked storage by reducing the data handling on both ends of the link. With RDMA, one system can directly read from, or write to, specific memory locations on another system in order to transfer data. One of the two systems does not even have to be involved in processing the data transfer. For NFS, this could allow requested data to simply appear in the right data buffers. Or one system could access data directly from the data cache of another system to avoid going out to the physical disk for the same data.

Enhancements such as this are expected throughout the next few years. The technology- and network-independent BladeCenter midplane and modular I/O structure will allow the incorporation of valuable new technologies as they become available.

Fibre Channel SAN storage

Over the past five to six years, Fibre Channel SAN-attached storage has become widely deployed to meet customer requirements for increasing storage capacity, performance, and availability for all server types. SAN storage attachment provides consolidation of disk and tape storage into a manageable resource for shared access by multiple servers (including BladeCenter processor blades) via high-speed Fibre Channel connections. SANs are most often deployed in a high-availability

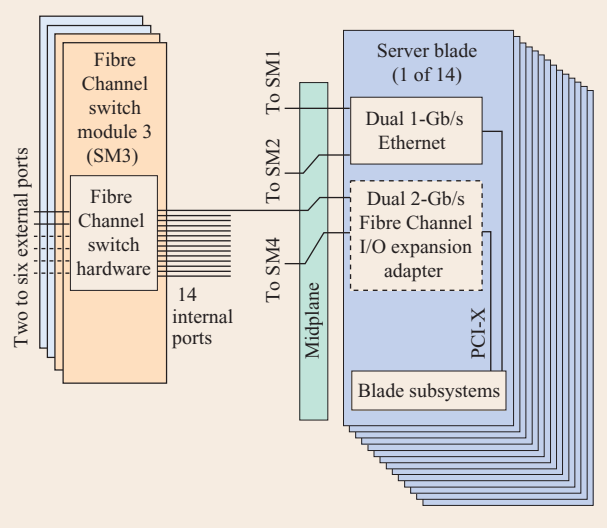

Figure 1

BladeCenter internal midplane wiring.

environment. SAN technology satisfies the most critical requirements through the use of high-reliability redundant Fibre Channel links and SAN switches, RAID storage controllers, and hot-swappable components to provide a highly fault-tolerant and serviceable storage solution. SAN-based storage consolidation can provide reductions in total cost of storage ownership via improvements in storage management efficiency and provisioning and by increasing overall utilization of storage resources.

The BladeCenter design supports a number of options for attaching disk and/or tape storage using SAN technology. SAN-attached storage can be used to augment local disks installed on the processor blades for additional capacity and performance, or SAN-attached storage can completely replace the local disks, providing remote boot support for the processor blades. Remote boot support allows BladeCenter processor blades to boot OS software and load application programs from external SAN-attached storage.

Fibre Channel SAN options

To support the attachment of Fibre Channel SAN storage, a dual-ported Fibre Channel host bus adapter (HBA) I/O expansion adapter can be added to each BladeCenter processor blade that must connect to a SAN. The BladeCenter Fibre Channel I/O expansion adapter is a repackaged version of the standard Fibre Channel HBA PCI adapter used in xSeries nonblade servers. The Fibre Channel I/O expansion adapter contains a two-port HBA that provides the processor blade with connectivity to two 2-Gb/s Fibre Channel ports. The high-speed serialized/deserialized (SerDes) electrical signals that carry the Fibre Channel port connections are routed from the I/O expansion adapter through a high-speed connector to the processor blade and then through a second high-speed connector into the midplane, where each port is routed to one of the two associated switch modules. Switch modules 3 and 4 (SM3, SM4) are wired to the two ports on the Fibre Channel I/O expansion adapter. This is shown pictorially in **Figure 1**.

Several different BladeCenter Fibre Channel switch module options and a passive optical passthrough module (OPM) option are supported and can be installed in the SM3 and SM4 slots in the rear of the chassis. These options provide the connection between the internal Fibre Channel signals and the external SAN, thereby completing the connection between the BladeCenter processor blade, its I/O expansion adapter, and the external SAN.

Either one or two Fibre Channel switch modules or OPMs can be installed in a BladeCenter chassis depending on the performance required and/or the requirement for redundant path connections to the SAN. If two Fibre Channel switch modules or OPMs are installed and connected to an external dual-fabric SAN, connections to external storage that supports redundant controllers and SAN connections (along with fault-tolerant RAID storage technology) can be provided to eliminate any single points of failure in the storage subsystem. This redundant path SAN connectivity, coupled with storage multipathing device drivers and server clustering technology, can be used to meet the needs of a wide variety of high-performance, fault-tolerant computing and data storage solutions.

Choosing SAN connectivity options

The choice of whether or not to connect a BladeCenter system to a SAN and which switch module option to use depends on a number of factors. Reasons for connecting a BladeCenter system to a SAN are typically the same as for other types of servers. For example, if applications that run on the processor blades require more capacity, performance, or fault-tolerance from the storage than can be provided from local disks, SAN-attached storage may provide a superior solution. In another situation, if an existing SAN is already in place, it may be desirable to connect to the existing SAN. SAN connectivity can also be used to support connection to external tape drives and tape libraries to allow multiple servers or blades to share those resources. Mixing disk and tape storage together on a common SAN can require additional considerations to ensure proper operation and optimum performance. Consolidation of boot disks to lower overall costs and aid in automated storage and

server provisioning techniques is yet another reason why it may be desirable to connect a BladeCenter system to a SAN. The subject of booting processor blades from SAN storage is revisited below.

Once the decision has been made to attach a system to a SAN, the next choice is which of the available internal Fibre Channel switch module or OPM options to use. If only a few processor blades have to be connected to the SAN, it may be more cost-effective to use one OPM (or two for redundancy) to route the signals to the external SAN. If the use of Fibre Channel in the chassis is expected to increase or the majority of blades already require connectivity to the SAN, using one or two internal Fibre Channel switch modules may be a more cost-effective alternative. A pair of switch modules will aggregate the processor blade HBAs into fewer cables to connect into a SAN through fewer ports. The OPM provides an electrical-to-optical conversion similar to existing small-form-factor pluggable (SFP) modules, with no aggregation. Each processor blade HBA port to be connected to the SAN requires a cable and an open Fibre Channel port in the SAN. Switch modules provide two to six external Fibre Channel ports. For applications that require very high data bandwidth to external storage, the OPM provides all 14 Fibre Channel ports for connection to the SAN. For connectivity to a SAN that uses proprietary protocols or otherwise precludes the use of a switch module, the OPM may be an ideal solution, although the overall solution cost of the OPM, cables, and additional SAN ports may be higher than when internal switch modules are used.

Booting from SAN-attached storage

If SAN connectivity has been added to a BladeCenter system, it is possible to boot the OS and load application programs from SAN-attached disk storage instead of local disks. In fact, if SAN connectivity has been provided, processor blades do not require any local disks.

Automated server and storage provisioning software such as IBM Remote Deployment Manager can be used to automate the process of configuring selected combinations of SAN-attached storage, internal and external SAN switches, processor blades, and Fibre Channel I/O expansion adapters to support booting from SAN storage. In many cases, booting from SAN storage can provide better performance and fault-tolerance than local disks. Before installing a BladeCenter system to boot from SAN, it is advisable to check with IBM and/or the SAN fabric and storage device vendors to make sure that booting from SAN storage is supported for the particular configuration desired. Most storage and SAN vendors provide this type of information as part of interoperability guides or other documents.

Fibre Channel SAN

BladeCenter SAN connectivity options provide the ability to connect BladeCenter servers to a wide variety of SANs and storage devices. IBM offers a variety of SAN-attached disk and tape products that have been tested with BladeCenter servers to meet a wide range of price, performance, capacity, and fault-tolerance characteristics. From the entry-level DS400 through the mid-range DS4000 series (formerly FAStT family), the recently announced DS6000 series, and up to the enterprise-level DS8000 series, IBM offers BladeCenter-compatible SAN-attached disk storage to meet the needs of any type of application. Many kinds of non-IBM storage are also supported. Through the IBM ServerProven and BladeCenter Alliance programs, vendors can provide interoperable products.

IBM offers a variety of SAN-attached tape drives and libraries, as well as internal and external Fibre Channel SAN switches and directors to create cost-effective SAN fabrics that can scale from one BladeCenter chassis to hundreds of SAN-attached processor blades. The choice of SAN fabric type to use typically depends on a number of factors, such as interoperability with an existing SAN, the number of Fibre Channel ports required, the number of hops (switches in the datapath), required advanced SAN fabric features, and specific SAN management requirements. These topics are discussed in more detail below.

Fibre Channel switch interoperability

With the rapid growth of Fibre Channel SANs, ensuring that various products can work together is extremely important. The BladeCenter Fibre Channel switch modules support direct storage attachment to external ports (E-ports) and switch-to-switch connections to external SAN switches. By utilizing E-port connections, a SAN can grow from a small, single-switch environment to a very large network of switches. *Interoperability* is the general term used to describe how two switches work together when they are connected. When two switches are connected, a number of parameters must be exchanged to ensure that the SAN fabric will expand correctly. Zoning configurations, buffer credits, and time-out values are among the parameters exchanged during the E-port creation, and these must be consistent among switches to ensure that the SAN fabrics will merge correctly. Switch products provide various solutions to ensure Fibre Channel interoperability, and these topics, along with the way in which the BladeCenter design incorporates Fibre Channel interoperability, are discussed below.

The FC-SW2 standard provides the ground rules that must be followed to ensure that Fibre Channel switches from various vendors are able to work with each other. The FC-SW2 standard provides an open environment

that does not restrict further expansion of the SAN to a single vendor. Three levels of communication are defined in the standard. They detail how switches should interoperate in the following ways: link level and fabric zoning, path selection via fabric shortest path first, and management information (for example, the name server table). BladeCenter switch modules conform to the FC-SW2 standard for interoperable multivendor (heterogeneous) environments, allowing them to be installed in currently existing Fibre Channel fabrics independently of the current SAN vendor.

Early deployments of BladeCenter systems into existing single-vendor (homogeneous) SANs consisting of Brocade or McData switches highlighted the criticality of E-port interoperability and BladeCenter solutions for a number of interoperability issues. Many SAN users were not aware that the switches they had already deployed came with an FC-SW2 standards-based mode of operation. They were traditionally using the vendor's proprietary mode default. By showing how to configure open SAN fabrics, BladeCenter systems have demonstrated that stable interoperable environments can be established with the FC-SW2 standard.

In addition to FC-SW2 interoperability, some SAN vendors use a proprietary mode of interoperability that enables an advanced set of functions available within a homogeneous fabric. This mode of operation is limited to switches from the same vendor; however, the features provided within this proprietary mode can improve SAN management, performance, and security because the product may deliver functions beyond those in the current Fibre Channel standards. Prior to the BladeCenter introduction, many SAN fabrics were homogeneous fabrics using a proprietary mode of interoperability. With the introduction of compatible Brocade switch modules, current Brocade SAN environments are able to retain the proprietary mode of interoperability and any advanced features that are in use, and incorporate a BladeCenter chassis with integrated switching into the SAN. This benefits users who require a Brocade homogeneous SAN, since a BladeCenter system provides seamless installation of up to 14 additional processor blades in the current SAN.

A third choice for Fibre Channel interoperability is available from select vendors whose switches can operate in the "native" mode of another vendor's switch products. This feature allows the switch running the special native-mode code to attach to a homogeneous SAN currently operating in a non-FC-SW2 or proprietary mode. This interoperability feature has the advantage of not requiring changes to the current SAN configuration. For example, if a homogeneous SAN is operating in a proprietary mode, a configuration change is required to make it operate in the open FC-SW2 mode. The benefit of the native-mode operation is that the configuration change is not required.

For environments in which it is simply undesirable to have an additional E-port connection to manage, there is a need to provide direct access from the processor blades to the SAN. The BladeCenter OPM provides up to 14 external connections directly to the individual processor blades. As discussed in the section on SAN connectivity above, the OPM provides an external fiber cable for each of the 14 processor blade positions. When a Fibre Channel expansion adapter is installed on the processor blade, the OPM provides a cable capable of connecting into the SAN fabric, thereby allowing a direct link between the expansion adapter and the external SAN switch. This connection will never be an E-port because there is no switch-to-switch connection from the BladeCenter system to the SAN. Therefore, traditional interoperability complexities are avoided with the use of the OPM.

Fibre Channel switch module options

The BladeCenter design offers a number of Fibre Channel switch module options that provide varying levels of SAN features and external bandwidth capabilities. The switch modules contain full nonblocking switch technology that provides 14 internal 2-Gb/s ports accessible from the processor blades and either two or six external ports that accept SFP modules. The external ports provide autodetect capabilities to determine the speed and port type required. The external port speed is autodetected to support the current 2-Gb/s Fibre Channel devices and older 1-Gb/s devices. The available SFP modules are capable of running at 1-Gb/s or 2-Gb/s speeds, so a user does not have to change SFP modules when supporting different speeds. The external port type can be configured for E-port, fabric port (F-port), or fabric loop port (FL-port). The E-port capabilities were described above in the interoperability section. The F- or FL-port types allow the switch modules to be connected directly to storage, tape, or another SAN device that is not a switch. F-ports are created for the majority of direct-attached devices. This port provides a direct point-to-point connection that does not require an arbitration procedure to communicate. FL-ports are required for direct connection of Fibre Channel disk drives and a small number of other storage devices and tape drives that support only Fibre Channel loop attachment. The primary difference between an F-port and an FL-port is that the FL-port can consist of multiple devices attached to a single port. All devices on the loop must arbitrate for access in order to begin a transfer with another device.

Fibre Channel switch modules provide the additional benefit of port aggregation via nonblocking switching that requires fewer external available SAN ports to attach a chassis. The same number of traditional standalone

servers would require more available ports in the SAN, as a 1:1 connection ratio would be required. Another advantage of port aggregation is that the number of fiber cables required is reduced by using the switch modules. The traditional requirement for two redundant fiber cables per server is replaced with a minimum of two cables for the entire BladeCenter chassis. The two cables provide redundant SAN connectivity to all 14 processor blades. Redundant power, cooling, and Ethernet connections are provided for the switch modules by the chassis. The Fibre Channel switch module is a hot-pluggable modular enclosure.

Two switch module choices that provide standards-based Fibre Channel connectivity are the IBM two-port switch module and the QLogic six-port Enterprise switch module for the BladeCenter system. These two switch modules support the FC-SW2 standards for interoperable E-port connections in heterogeneous SAN fabrics. IBM provides an interoperability guide [9] for these two switch choices that details how to configure the E-port connections to a variety of external SAN devices. Because users may not be familiar with the FC-SW2 mode of operation, the guide provides a configuration walkthrough with detailed steps and pictures that show exactly how these switch modules can be integrated into the SAN. Instructions for integration into Brocade, McData, CNT, Cisco, and QLogic SAN environments are provided in this guide, making it one of the most comprehensive switch interoperability guides in the industry. The availability of two or six external ports allows the user to select a switch module that meets the bandwidth and connectivity requirements of the environment. For applications that demand high bandwidth, the six-port switch module provides almost a 2:1 bandwidth ratio to the internal processor blades, for a total of 24 Gb/s full-duplex capacity at 2 Gb/s.

The other switch module choices are two Brocade Fibre Channel models that provide Brocade switching and advanced SAN features within the BladeCenter chassis. The Brocade Entry and Enterprise SAN switch modules provide two external ports and support all Brocade advanced features, such as interswitch link trunking for increased E-port performance between switches, advanced performance monitoring for optimizing performance of the SAN fabric, and secure fabric for increased security policies within the SAN. The advanced feature products from Brocade are available as license key features for the switches, allowing for immediate upgrade of the switch module without requiring new firmware. These switch modules provide the proprietary mode of operation for homogeneous Brocade fabrics and allow for seamless integration into a Brocade SAN. The Brocade switch modules, as with all BladeCenter switch modules, also support the FC-SW2 standard. Brocade provides configuration guidelines for attaching to McData SAN fabrics via the FC-SW2 mode of operation. The difference between the two Brocade models is in the number of switches that are permissible within the SAN fabric. The Brocade entry switch module contains a two-domain license. This license restricts the BladeCenter switch module to a SAN fabric that contains no more than two switches in total. It is important to remember that each one of the entry switch modules counts as one switch. Therefore, the entry switch module is an excellent choice for environments in which the storage is directly connected to the BladeCenter system or where the entry switch module will be attached to only one additional switch, with no other switches contained in the fabric.

For environments in which there are more than two switches in the SAN fabric, the Brocade enterprise switch module is required. This switch module has no preset limitation on the size of the SAN fabric, since it can support up to the Fibre Channel architectural switch limit of 239. The two modules provide the user with choices based on the size of the deployed SAN fabric. For installations that will grow over time, one can install the entry switch module today and upgrade later. When the SAN fabric exceeds the two-switch limit, an upgrade license can be installed on the entry switch module, which will then provide the full 239-switch capability.

iSCSI SANs for the BladeCenter system

Among the several methodologies to provide storage for processor blades, iSCSI-based SAN technology is emerging as a viable storage solution for blades. iSCSI SANs provide storage resources over any IP network, from a small Ethernet LAN up to the Internet.

iSCSI technology provides a low-cost remote storage solution. iSCSI initiators for the processor blade are available in a range of price/performance ratios. A software initiator provides the lowest cost, with a reasonably small tradeoff in peak server performance. Hardware initiators are better suited for more demanding workloads. And, for the most storage-intensive workloads, Fibre Channel SANs continue to provide the maximum performance, scalability, and robustness (see the section on Fibre Channel SAN above).

iSCSI storage target implementations provide a range of storage solutions that differ in capacity, performance, and level of physical integration. Storage blades can be ideal for situations in which integration within a BladeCenter system is desired. External storage controllers offer additional performance and, through add-on expansion disk drawers, additional scalability for larger and more complex solutions.

Building a SAN solution on an Ethernet infrastructure provides a value proposition to customers from several perspectives. First is the reduced solution cost in terms of

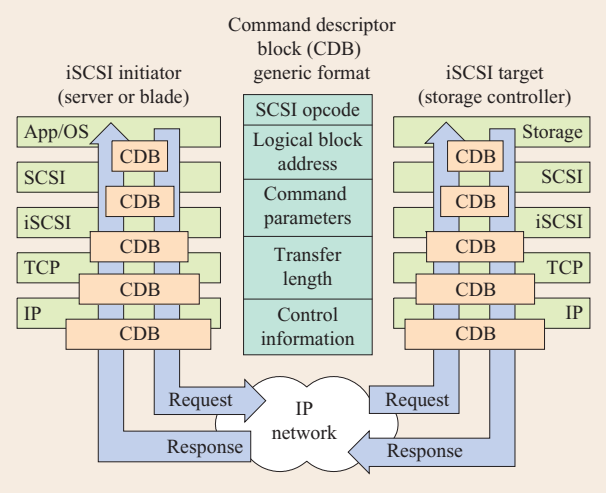

Figure 2

iSCSI protocol overview.

acquisition cost (the cost of the components and services) and the total cost of ownership (involving administration and expertise). Second, with the pervasiveness of Ethernet and IP, access to storage resources that are geographically local or remote is viable. Third, because iSCSI combines accessibility with traditional storage and content management, robust solutions—which address rich content available to a wider set of users and applications—can open new ways to provide value to customers.

The iSCSI protocol makes possible the transportation of SCSI transactions over TCP/IP networks by encapsulating SCSI requests and responses within IP packets. Since IP packets can be transported over Ethernet fabrics, the iSCSI protocol is supported transparently over Ethernet fabrics, and, similarly, over other fabrics supporting IP services. The SCSI protocol itself is a request/response protocol, with which computer resources, called *initiators*, can read data from or write data to storage resources, called *targets*, which manage, retrieve, and store the data on hard disks. By enabling SCSI transactions to travel over Ethernet networks, many initiators and targets can be supported in a wide variety of topologies without requiring special hardware or expertise. With the network paradigm, capabilities such as discovery, security, and load balancing can be provided natively or provided via value-add applications. In **Figure 2**, an overview of the iSCSI protocol is presented.

The next several sections present the architectural highlights of initiators and targets. Subsequently, several applications ideal for blade-server environments, such as booting diskless processor blades, are discussed.

iSCSI initiators for processor blades

An iSCSI initiator provides the processor blade the capability to make storage requests over an IP network. The iSCSI initiator must comply with the Internet Engineering Task Force [10] request for comments (RFC) 3720, which defines such iSCSI processes as transactions, security, discovery, and error handling.

In simple terms, the iSCSI initiator logs into the iSCSI target and negotiates session parameters. Once logged in, the iSCSI initiator makes read or write requests to the iSCSI target, with the target responding appropriately. When the storage resource is no longer needed, the iSCSI initiator logs out of the target and the session is terminated.

There are two basic types of initiators that can issue iSCSI transactions on behalf of an OS or application. A software initiator leverages a standard Ethernet network interface card (NIC) to provide a network connection, while the iSCSI protocol is handled in software on the main processor on the processor blade. A hardware initiator is a custom-built iSCSI hardware adapter focused on offloading the iSCSI protocol processing from the main processor. A software initiator is the most cost-effective solution, since no additional hardware is required. A hardware initiator, by offloading the network and iSCSI processing to an adapter, can free up the main processor(s) for other tasks, while the hardware initiator provides superior transaction performance. While a hardware iSCSI initiator will lower the processing load on the host, it may increase or decrease the performance of the storage transactions, depending on the design and processing capabilities of the specific iSCSI initiator hardware. **Figure 3** presents the respective iSCSI initiator architectures.

Software initiators provide an iSCSI processing layer that resides in the OS storage protocol stack. Above this layer is a SCSI layer; below is a TCP/IP layer that provides access to the TCP/IP transport services. These services may reside within the storage stack or in the OS network stack. The iSCSI layer makes network calls to the TCP/IP set of services and depends on those services to transport the packets out of the processor blade. These software initiators use resources already present on the processor blade: the central processing unit (CPU), memory, and NICs.

Hardware initiators provide a software interface similar to a Fibre Channel HBA where an adapter, or set of dedicated hardware, provides the complete set of functions associated with iSCSI, including the TCP/IP functionality. With this implementation, the SCSI layer in the OS storage stack can, in essence, make direct calls to

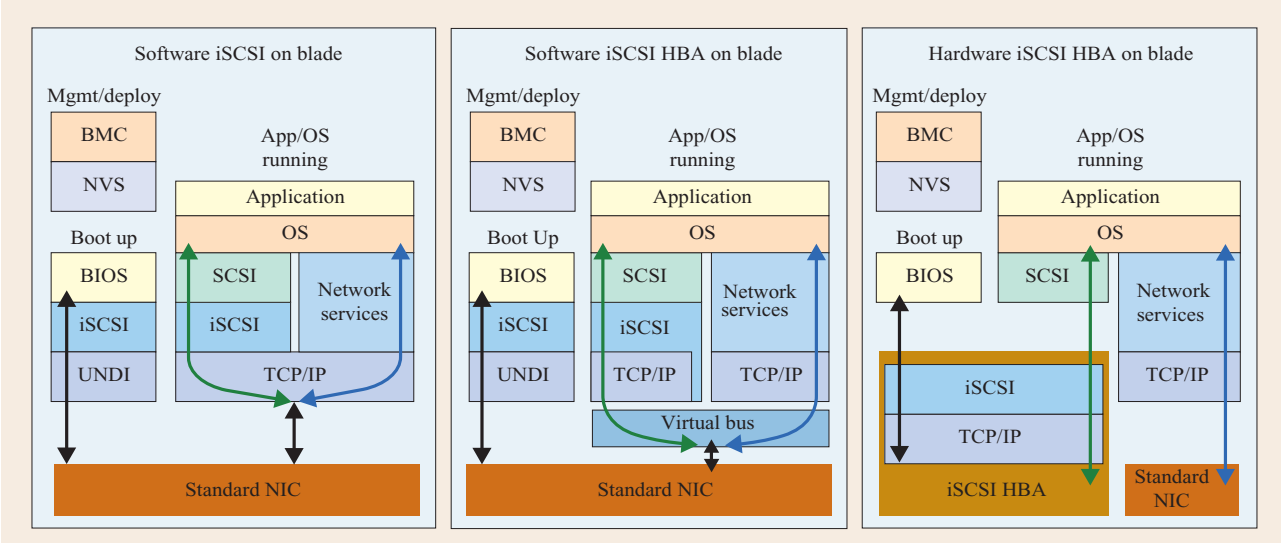

Figure 3

iSCSI initiator architectural approaches on BladeCenter blades. (BMC: baseboard management controller; UNDI: Universal Network Device Interface; NVS: nonvolatile storage.)

the dedicated hardware to initiate a storage transaction. This dedicated hardware may or may not be visible to the OS network stack. These hardware initiators are dedicated adapters that provide iSCSI functionality through the use of their own processing engines, buffer memory, and Ethernet interface.

The value of software initiators is in their cost-effectiveness, since all of the hardware involved in a software initiator approach is already present on the processor blade, and thus is effectively free. The tradeoff for the lower cost is that software initiators consume significantly more of the CPU than a hardware initiator under the same circumstances. This tradeoff is a reasonable one for many processor blade applications that do not heavily stress storage communications or the CPUs. Conversely, the strength of hardware initiators is in lowering CPU utilization, and in some cases, improving storage performance, while their weakness is the cost of the hardware.

To put the tradeoff in some perspective, a fictional software initiator may cost nothing and consume 25% of the CPU, while an equally fictional hardware initiator may cost $500 and consume 5% of the CPU. The decision is to determine whether freeing up 20% of the CPU capacity is worth $500. Continuing this example, the storage performance requirements of the intended application and the projected processor utilization of the application must both be evaluated. A Web application, such as hosting Web pages or an online catalog, would generate little demand for storage, since the hosted content is infrequently updated and the most commonly accessed portions are cached in memory by the Web server application. A large e-mail or database application, on the other hand, would make a significant demand on the storage, since each transaction must be saved to disk. After determining the value of reducing CPU utilization or improving storage performance, attention should be given to the investment alternatives available to the customer. An investment could be made to obtain an iSCSI hardware initiator for a certain level of improvement, or an alternate investment could be made instead to obtain an additional processor for the processor blade, change to a faster processor blade, or add a processor blade to take some of the load off the first processor blade. Depending on specific application requirements, an iSCSI software initator may provide the optimal solution, or an iSCSI hardware initiator may be required in order to achieve the appropriate solution.

iSCSI targets for the BladeCenter system

An iSCSI target responds to requests by iSCSI initiators by performing the data writes and the data reads to and from the physical disk. At a minimum, the target consists of a host network interface that provides the iSCSI target protocol support, a RAID controller function, and a pool of physical disks.

The host interface on the target provides the iSCSI functionality by responding to the iSCSI initiator log in, responding to the specific iSCSI initiator read and write requests, and supporting the iSCSI initiator log out. The

host interface also supports all of the TCP/IP processing and IP address management. Similarly to the iSCSI initiators, the iSCSI target host interface can be based on a software or hardware target, with the software target executing on the storage controller processor and the hardware target executing on dedicated hardware designed to unburden the storage controller processor. A given iSCSI target can leverage the best host interface approach for the intended market. For example, an entry-level target supporting a small number of disk drives may choose to implement the iSCSI host interface using a software approach, while a higher-end iSCSI target supporting a large number of disks may choose to implement the iSCSI host interface using a hardware-based approach. The choice is decided by the capacity and, in turn, the solution performance desired compared with the cost of such a solution.

The RAID controller function performs the coordination of disk reads and writes to disk rather than to the local storage controller cache. The coordination involves logical unit number (LUN) creation and management, logical block address mapping to the specific disks, and the scheduling of operations to minimize platter rotations and armature movement required to satisfy all pending transactions.

The actual pool of disks provides the storage capacity of the storage controller. The topology and quantity of disks provide support for the RAID configurations and the overall storage performance of the storage controller. The pool of disks can be based on Fibre Channel, SCSI, SAS, or SATA disks.

iSCSI targets can be implemented as a standalone external enclosure, either with or without redundant controller subsystems, power, and cooling. They can manage a pool of disks that can range from a handful to hundreds. iSCSI targets can also be implemented as a BladeCenter storage blade. There are similar tradeoffs in selecting an iSCSI target. The capacity and performance of an external iSCSI storage controller may be suitable for certain applications, while the integration and compactness of an iSCSI storage blade target within the system may be ideal for another application.

iSCSI solution for the BladeCenter system

Leveraging the BladeCenter architecture and iSCSI provides flexibility and integration that leads to many interesting and useful topologies and capabilities. Flexibility is provided by the rich network capabilities within the chassis and across multiple interconnected chassis. By using iSCSI and the BladeCenter design as the foundation, some interesting applications (including diskless processor blades, processor blade provisioning, content provisioning, and content management) can be provided. Diskless processor blades provide a path to better and tighter server, storage, and networking integration over time, while provisioning and content management provide new dimensions in terms of solution flexibility and optimization.

Putting the iSCSI initiators and iSCSI targets together on an Ethernet network creates an iSCSI solution. The three general steps to creating an iSCSI solution are deployment of subsystems and components, internal configuration, and definition of initiator and target relationships.

Briefly reviewing the architecture, each processor blade slot in a chassis connects to four independent switch fabrics, each implemented via an independent switch module residing in the chassis. A processor blade typically dedicates the first two fabrics for 1-Gb/s Ethernet by embedding two Ethernet NICs in the processor blade. A processor blade can use the remaining two switch fabrics via an optional I/O daughter card. The daughter card can be for use on a 1-Gb/s Ethernet fabric, a Fibre Channel fabric, or another protocol. Of course, the switch module must provide the same fabric type as the daughter card. An iSCSI hardware initiator can be implemented as a daughter card for use with a 1-Gb/s Ethernet switch module. Therefore, the customer has the choice of using software iSCSI on the base NICs, software iSCSI on a 1-Gb/s Ethernet NIC daughter card, or a hardware iSCSI daughter card—all with any standard processor blade and Ethernet switch module. The chassis modularity offers the customer an opportunity to select the best balance of cost and performance tradeoffs. Given this architecture, the implementation of iSCSI initiators is clear, while the iSCSI targets have a choice of using the same paradigm as processor blades or using an external storage controller.

The first step of a BladeCenter iSCSI solution is to determine and deploy the key elements. On the iSCSI initiator side, the deployment of the appropriate initiator elements on each processor blade provides iSCSI capabilities. On the target side, the deployment of the appropriate target provides iSCSI storage resources available to initiators.

With the vision of the iSCSI solution established, the iSCSI targets must be configured in terms of LUN creation for holding the actual data and security settings to ensure that only the proper iSCSI initiators can gain access. The configuration of an iSCSI target varies from vendor to vendor, but usually can be done by using a graphical user interface suite or a command-line interface (CLI). Ultimately, the method for configuring iSCSI targets will standardize.

Once the iSCSI initiators and iSCSI targets are in place, the next step is to define the relationship between iSCSI initiators and iSCSI targets. For example, in a given BladeCenter system, up to 14 processor blades

using the iSCSI initiator subsystem must be related to one or more iSCSI targets representing one or more LUNs. This relationship can be achieved via several methods, including using network services such as Dynamic Host Configuration Protocol (DHCP) [11] to acquire iSCSI configuration information, using manual configuration entry to explicitly define the relationship, or using a management tool suite to remotely configure iSCSI initiators and targets. The combination of iSCSI technology and BladeCenter architecture provides interesting solution capabilities. In **Figure 4**, several storage configurations are presented.

iSCSI applications

The individual components and complete solutions provide the foundation for several intriguing applications of iSCSI: supporting processor blade boot-up, provisioning processor blades, and managing content.

Processor blade boot-up using iSCSI

Since iSCSI provides the ability to read disk blocks, the idea of moving general storage and boot storage to an iSCSI target introduces new flexibility by detaching the application, OS, and user content from the physical processor blade. The boot process involves the processor blade BIOS to verify and configure processor blade hardware, retrieve the OS image from a disk, and hand off processor execution to that image. With iSCSI, the retrieval of the OS image can be done from a remote iSCSI target instead of a local embedded disk.

To integrate iSCSI into the BIOS boot process, a software- or hardware-based iSCSI initiator can be integrated into the BIOS so that it can utilize the remote iSCSI target as if it were a local disk. With no OS or protocol stacks available before boot, the integrated iSCSI function must include its own discovery, address resolution, TCP/IP protocol stack, network services, boot process, and standard iSCSI initiator functions. With this integration, the BIOS will issue a request to read the master boot record (MBR) by tasking the iSCSI initiator to retrieve the MBR from the remote target.

Configuring a BladeCenter system to support iSCSI booting involves informing the initiator and remote target of the specific iSCSI and IP parameters needed for each to find and acknowledge the other. For the initiator, the configuration parameters can be found in three ways: The first is to be acquired from the network via network services such as DHCP, Service Location Protocol (SLP), or iSCSI Server Name Service (iSNS); second, accepted via manual entry on user interface panels and subsequently placed in the processor blade nonvolatile storage (NVS); and third, accepted from a remote deployment wizard that has placed the parameters in the processor blade NVS through a different path. For the target, the configuration parameters can be configured via storage management interfaces such as a Web interface or a CLI.

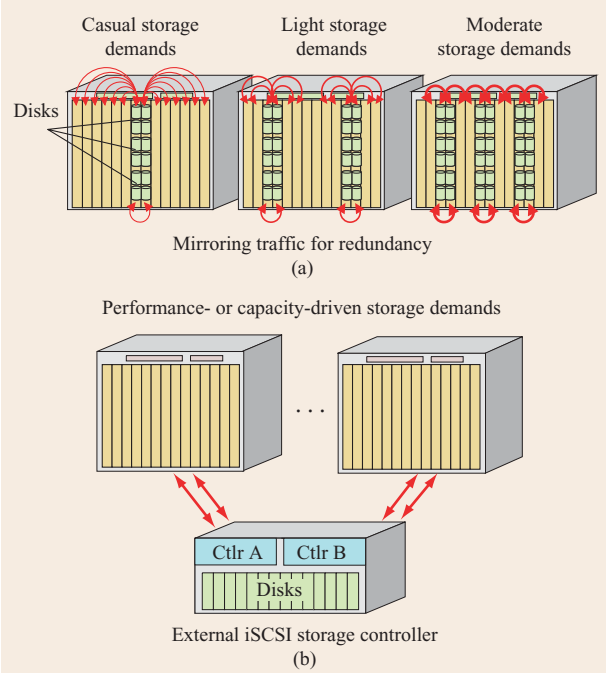

Figure 4

BladeCenter use across the storage-demand spectrum: (a) BladeCenter systems populated with a mix of application and iSCSI storage blades. (b) BladeCenter systems populated with application blades using external iSCSI storage controllers.

With the detachment of OS, applications, and content from the processor blade, local disk drives can be eliminated. Eliminating local disk drives frees up board space, reduces power consumption, and reduces heat generation. This space, power, and heat capacity that is saved can be used for other functions on the processor blade. Another benefit of diskless processor blades is the ability to replace the processor blade without affecting the data stored on the iSCSI target.

Application or processor blade provisioning

With the OS and/or application content residing on an iSCSI target, whether on the same or different LUNs, tools at the solution management level can assign the content to any given processor blade. This assignment can be accomplished in any one of several ways. One way is to set processor blade BIOS configuration parameters to reflect which iSCSI target and LUNs to use and then initiate a reboot of the given processor blade. Another way is to reconfigure an application available on a running processor blade with the new iSCSI target and

LUN parameters and then start the desired application via an agent running on the processor blade.

With the ability to flexibly provision processor blades, the decision process becomes of interest. When and why would a server have to be reprovisioned? Provisioning parameters can be manually entered each time a server has to be changed from one application to another. If regular identifiable events trigger the need to reprovision a server, the process can be automated. A provisioning policy would define the conditions that trigger a provisioning change and would define how the new parameters are determined. For example, for tasks that must to run during a specific time period or on specific days, the time of day could be a trigger to provision servers to perform that task. For workload balancing, the application workload or hardware performance data could be a trigger to reprovision a higher-performance server to take over that application. Similarly, if the application supports a scale-out processing model in which many processor blades can work in concert, additional processor blades could be provisioned as the workload increases, and processor blades could be reprovisioned to a different task when the workload subsides. This level of automated server provisioning is possible only with the dynamic flexibility of SAN storage.

Content management

With the content and processor blade no longer integrated, tools for managing the content, such as backup, remote replication, local mirroring, and capacity planning, can operate directly on the iSCSI targets without burdening the processor blade. iSCSI targets can provide backups by taking snapshots of a LUN, freezing an image of that LUN, and copying it to a "safe" LUN. The targets can make these safe LUNs available to the management tools to extract the critical content to back up devices such as tapes or near-line storage. Near-line storage is an alternative to tape storage for backup. It uses inexpensive disks in a subsystem that is optimized to provide a low-performance, high-capacity storage target. With these backups, the content is protected and retrievable as needed.

A unique feature of iSCSI is that the same Ethernet connection that connects the processor blade to the storage can also connect the storage controller to a peer controller anywhere on the LAN or, with an Internet connection, anywhere in the world. When an iSCSI target is configured to mirror content to a peer, updates to the primary target LUNs are copied to a remote LUN residing on a remote target; the content is safe from any physical site disasters such as fire, flood, or major power outage. Depending on the criticality of keeping the mirrored data perfectly synchronized, the mirroring can be synchronous, with each LUN update verified on the primary and remote LUNs prior to acknowledging back to the initiator, or it can be asynchronous, with LUN updates to the primary mirrored to the remote LUN in a later, less time-constricted manner.

The mirroring capability can be extended; several branch offices can mirror to one central location, where a content view across all locations can be built. This type of scenario is ideal for building a consolidated view of the content (for example, usage billing) across the entire business.

InfiniBand

InfiniBand has established itself as an industry-standard high-bandwidth, low-latency interconnect for high-performance computing (HPC) clusters in scientific and technical computing applications. The InfiniBand standard [12] defines a basic building block—a four-wire serial link comprising one 2.5-Gb/s differential pair for transmitting data and one 2.5-Gb/s differential pair for receiving data. Accounting for the ten-bit data encoding on the wire, this provides a peak bandwidth of 250 MB/s in each direction. Putting this in context with other industry-standard I/O links, the 250-MB/s peak bandwidth of an InfiniBand link, referred to as a *1x link*, is twice that of 1-Gb/s Ethernet (125 MB/s peak), and slightly ahead of 2-Gb/s Fibre Channel (200 MB/s peak).

The 1x link building block is scalable through two highly efficient mechanisms defined in the standard: multilane links and higher-bit-rate links. Using multiple lanes of the 1x link building block in a coordinated fashion provides higher bandwidth in a single logical link. Multilane InfiniBand spreads data traffic evenly across the individual lanes at the byte level with no additional load-balancing overhead, ensuring full utilization of the entire link [13]. During 2004, InfiniBand products were available supporting the first three defined InfiniBand lane widths: 1x, 4x, and 12x. These respectively provide single-direction peak bandwidth of 250 MB/s, 1 GB/s, and 3 GB/s. The second scaling mechanism in the InfiniBand standard, higher link bit rate, specifies the double-data-rate (DDR) 5-Gb/s link rate and the quad-data-rate (QDR) 10-Gb/s link. Prototypes of components capable of 4x DDR and 12x DDR became available near the end of 2004. By starting with a 1x building block that is comparable to existing link technology and defining link scalability to 48 times that starting point (12x link width combined with QDR link rate), InfiniBand provides a compelling solution for bandwidth-intensive applications.

An additional lane width of 8x was added to the InfiniBand standard in 2004 specifically to provide a bandwidth increment above 4x DDR that could be wired in a blade-server chassis. In designing the BladeCenter system, we identified physical limitations and cost factors

that precluded the implementation of 12x links. The BladeCenter design, with its 14 processor blades and 1x links, required a switch module connector with 80 active pins. Growing the switch to support 4x links and other future capabilities would increase the pin count to 280 pins per switch module. While this is a challenge, it is quite feasible with the technology available in 2005. However, looking forward to supporting IB-12x and the anticipated 792 pins per switch module, it seemed that the connector itself would cause the switch module to cost more and consume more chassis space than was reasonable. It was determined that IB-8x, requiring 536 pins, is a more practical design to implement.

InfiniBand for network communications

In addition to high bandwidth, InfiniBand provides highly efficient software interfaces that deliver much lower end-to-end latency than existing industry-standard networking protocols. The protocol processing associated with the most common networking interconnect, TCP/IP over Ethernet, rapidly consumes even the fastest processing speed available as the networking bandwidth approaches the 1-Gb/s or 2-Gb/s level. The TCP/IP protocol relies on the server processor to manipulate individual data buffers multiple times as transmit processing progresses from the application through TCP protocol processing, TCP checksum generation, TCP segmentation, IP protocol processing, IP segmentation, and finally the adapter hardware transmit queue. The receive process reverses these steps, with the additional requirement that incoming packets, which arrive interspersed with packets from other ongoing communications, must be reassembled and sometimes reordered and placed in the proper application buffer. Recent technology innovations to reduce the overhead of TCP/IP communications and make networking more efficient have focused on moving the protocol processing from the host processor down to the adapter hardware. This can provide significant benefits for bulk data transfers, but at an increase in the complexity and cost of the adapter. An alternate approach is to change the underlying protocol such that the processing burden is reduced rather than moved.

InfiniBand provides four server communications protocols that provide the scalable high-bandwidth benefit of InfiniBand, combined with different levels of software compatibility, increased efficiency, and decreased latency (**Figure 5**):

- *Internet Protocol over InfiniBand (IPoIB):* Maximum software compatibility for all IP communications. Offers increased bandwidth, minimal efficiency benefit, and small latency improvement.

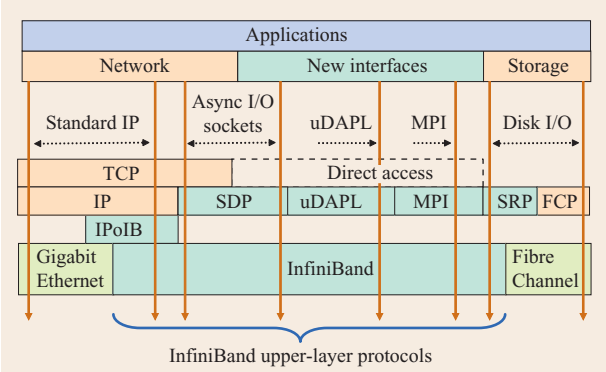

Figure 5

InfiniBand upper-layer protocols. (FCP: Fibre Channel Protocol.)

- *Sockets Direct Protocol (SDP):* Software compatibility for sockets-based communications. Offers additional bandwidth, increased efficiency, and additional latency improvement.
- *User-level direct access transport APIs (user-level programming layer, uDAPL):* Requires a new application software interface or a new interface library. Offers maximum bandwidth, maximum efficiency, and minimum latency.
- *Message Passing Interface (MPI):* A unique application software interface that is common in scientific and technical HPC. Offers maximum bandwidth, maximum efficiency, and minimum latency.

For storage, InfiniBand provides for both file-level remote NAS storage access and block-level remote SAN storage access. For file-level storage, the available networking protocols listed above are used to transport TCP/IP, User Datagram Protocol over Internet Protocol (UDP/IP), or sockets communications between servers and storage. File servers for UNIX and Linux environments use NFS to access files on the network. In Windows environments, CIFS provides a comparable file access method. InfiniBand supports both of these access methods layered on top of IP.

InfiniBand for storage communications

For block-level storage, InfiniBand follows the same path as other SAN technologies, exploiting the capabilities of the underlying network to deliver standard SCSI communications between servers and storage devices. Fibre Channel is the most common SAN fabric today, transporting SCSI communications throughout the Fibre Channel network with minimal overhead and latency. Currently, InfiniBand provides one native storage

protocol, SCSI RDMA Protocol (SRP), which provides the bandwidth of InfiniBand while maintaining efficiency and latency similar to that of Fibre Channel. It offers maximum software compatibility for SCSI storage communications, maximum bandwidth, and similar efficiency and latency.

Additional storage communications methods are available over InfiniBand through combinations of the supported protocols, such as iSCSI on top of SDP or IPoIB. iSCSI provides a SCSI transport over TCP/IP Ethernet networks, although with the added overhead of TCP/IP protocol processing discussed above. While it is a block-level protocol, iSCSI can be transported over InfiniBand through either SDP or IPoIB networking protocols, as with other TCP/IP-based protocols. Future InfiniBand enablement is expected that will combine further advances in IP storage protocols with the InfiniBand RDMA and high-bandwidth capabilities. iSCSI extensions for RDMA (iSER), Direct Access Transport (DAT), and proposed RDMA enablement of NFS are all efforts intended to provide additional performance and efficiency benefits through increased exploitation of RDMA-enabled fabrics such as InfiniBand.

InfiniBand for diskless servers

As part of providing networking and storage connectivity, InfiniBand also provides the capability for a server to boot directly from storage accessible through the InfiniBand fabric. Utilizing IPoIB allows for file-level network booting through either the preboot execution environment or Bootstrap Protocol (BOOTP) methods. Both of these methods use DHCP so that a server can discover the proper network address (TCP/IP address) to be used for network communications and the proper destination address and file name where it should find its boot disk. With this information, the server then reads in the boot image across the network using Trivial File Transport Protocol (TFTP) over UDP/IP and begins execution of the boot record as though it were being read from a directly attached local disk. On top of the required network capabilities, InfiniBand also provides a level of abstraction between the logical connectivity among multiple TCP/IP or UDP/IP end nodes and the physical InfiniBand network. Regardless of the complexity and scope of the physical InfiniBand network, the server attempting to boot can be presented with a view of the network containing only those devices with which it is required or is authorized to connect. In addition to the high bandwidth described above, InfiniBand provides multiple virtual lanes that allow different network traffic flows to be assigned different priorities through the network.

TFTP and UDP are used for network boot because they provide a very basic file transfer mechanism, but they provide only minimal error handling and retry mechanisms. TFTP is generally used for data transfers, such as remote booting, when there is limited space for storing the executable code or limited processing performance for executing the file transfer. For remote boot support, limited-option ROM space is the primary reason why TFTP is used rather than the more elaborate and more robust File Transfer Protocol (FTP). The result of using the simpler TFTP protocol is that network boot implementations can suffer from reliability problems if the network and boot server are not able to meet the bandwidth and response-time requirements of the servers that are booting. Increasing the number of booting servers, altering the sequence with which servers boot, or even unrelated increases in nonboot network traffic can lead to a situation in which one or more of the booting servers does not receive responses to its storage requests quickly enough, TFTP times out, and the server boot process terminates. The increased bandwidth provided by InfiniBand offers a more scalable network for remote booting. The additional functional capabilities provided by InfiniBand to isolate, prioritize, and simplify the network boot environment can further increase the robustness and performance of remote booting to a level that would otherwise require a dedicated boot network to achieve.

Remote booting can also be done through block-level access methods that even more closely mimic accessing a directly attached local disk. Using SRP, an InfiniBand-connected server can issue a SCSI read command to a storage device, asking to read the data blocks that comprise the disk boot record. This is fundamentally the same as Fibre Channel remote boot. InfiniBand can use the same process of manually entering into each diskless server the address of the remote disk controller and LUN to be used (which specifies which disk image on the controller) as its boot volume. However, InfiniBand remote boot implementations provide an additional level of abstraction. The InfiniBand hardware provided for use with a BladeCenter system is factory-set to read a specific LUN at a specific well-known address when it is the boot device in a server. The BladeCenter InfiniBand switch module is set to intercept any requests to this well-known address and redirect them to the real boot disk for that server. The mapping of servers to boot disks is managed within the InfiniBand fabric itself. These mappings can be preallocated and configured even before the servers are deployed. They can be dynamically or manually allocated as each server is deployed, and they can even be dynamically remapped as servers are put into use and removed from use for specific applications.

InfiniBand for I/O virtualization

The added flexibility that InfiniBand provides to the boot process is just one aspect of the I/O virtualization capabilities provided by the InfiniBand standard and enhanced by the BladeCenter InfiniBand solutions being delivered through the development partnership between IBM and Cisco Communications.

The BladeCenter InfiniBand hardware consists of an InfiniBand expansion adapter for each processor blade and an InfiniBand switch module to interconnect the processor blades to one another and to external InfiniBand devices. The card adds two 1x ports to the processor blade. The switch module connects one 1x port from each processor blade and four external 4x connectors for connecting to other InfiniBand equipment. The aggregate bandwidth available through the switch module to the 14 blades is 35 Gb/s (14×2.5 Gb/s) of peak inbound traffic and another 35 Gb/s of peak outbound traffic. The four connections out of the switch module add up to a total of 40 Gb/s in each direction so that no bandwidth restriction, or pinch point, is created. With the use of two switch modules, the total bandwidth available to the chassis is 70 Gb/s in each direction—a tremendous bandwidth, far beyond what is required for most data center applications.

This level of I/O performance is valuable primarily for tightly connected cluster systems. High-bandwidth data movement and low-latency messaging between nodes allows for higher overall throughput across a large cluster. This level of performance becomes valuable in the data center when it is combined with the ability to abstract, or virtualize, the server connections to storage and networks.

As described above, InfiniBand has the defined protocols to transport storage and network traffic among connected servers and other InfiniBand devices. In addition, the key protocols can also be converted through bridges to allow connection into other existing networks. An InfiniBand–Fibre Channel bridge allows an InfiniBand-connected server to communicate directly with Fibre Channel-attached devices as though it were natively connected to the SAN. An InfiniBand–Ethernet bridge provides communication between the InfiniBand servers and connected Ethernet devices, as though all were on a single LAN.

The Cisco InfiniBand–Fibre Channel bridge provides the boot virtualization function described in the previous section. It also provides this same I/O virtualization for each InfiniBand device and its connection to Fibre Channel SAN devices. This starts with the allocation of one or more physical Fibre Channel ports to provide connectivity for each server. Each server is provided one or more unique Fibre Channel World Wide Names (WWNs) so that Fibre Channel SAN functions, such as storage controller LUN masking and Fibre Channel switch zoning, operate across all Fibre Channel- and InfiniBand-connected members of the defined SAN. This extends the SAN, with all of its storage and network management capabilities, across the InfiniBand fabric to every InfiniBand device configured to connect to that SAN. The virtualization goes one step further. Multiple isolated SANs can be defined across the InfiniBand fabric, with different physical SAN connections through multiple InfiniBand–Fibre Channel bridges. This allows the InfiniBand devices to be selectively connected with specific Fibre Channel SAN devices on any of the connected SANs. The isolation provided to each configured SAN environment prevents unauthorized accesses to or from any device that is not configured to be a member of that SAN.

The Cisco InfiniBand–Ethernet bridge provides a similar environment connecting together Ethernet devices and InfiniBand-connected devices into easily defined LAN segments. InfiniBand allows the creation of multiple logical networks and the explicit control of determining which devices can communicate with other devices. All of these virtualized LAN segments can span the connected InfiniBand and Ethernet fabrics.

The SAN and LAN virtualization provided by InfiniBand addresses one of the limitations shared by all blade servers—finite connectivity through the chassis. Once all of the networks available in the chassis are put to use, there is simply no way to add another. Using InfiniBand I/O virtualization, each server can have as many virtual Ethernet NICs or virtual Fibre Channel HBAs as needed. There is no longer a physical constraint due to the connectors or wires inside the box. The available connectivity scales with the available InfiniBand bandwidth and the number and type of InfiniBand-connected gateways.

When all I/O connectivity is provided through virtual connections, whether to other InfiniBand-connected devices or through bridges to Ethernet- and Fibre Channel-connected devices, the reconfiguration of those same connections becomes as simple as a few clicks on the InfiniBand configuration screens. Servers and I/O can now be dynamically allocated to different applications as workload demands rise and fall. Deploying a server for an application might consist of selecting a boot disk for the application, identifying the necessary data storage and networking connectivity, then setting an available server to reboot with the selected resources properly configured for its use. The newly deployed server would run that application as though it had always been connected to those resources and committed to this specific application. When the extra server was no longer required, it could be drained of work, shut down, and all connections removed. The application data and the boot

Table 5 BladeCenter storage options.

Storage option	*Performance*	*Capacity*	*RAS*	*Convergence*
SCSI drives in processor blade	High	Moderate	Simple redundancy	
SCSI drives in BSE	High	Moderate	Hot-swap redundancy	
Ethernet, NAS	Low to moderate	Scalable	External options	Network and storage
Ethernet iSCSI SAN	Moderate	Scalable	External options	Network and storage
Fibre Channel SAN	High	Scalable	External options	
InfiniBand SAN/NAS	High	Scalable	External options	Cluster, network, and storage

disk would remain as they were until needed again, when they would be provided with some other server to do the processing. In the case of purely stateless processing, where the boot disk and connected data disks would not be preserved, they would be refreshed prior to the next deployment.

Summary

Today's customers can choose from a wide range of BladeCenter storage options (**Table 5**). Choosing which option or options to use requires the same understanding of the application storage requirements as any other storage solution. With an understanding of the performance, capacity, and RAS requirements and an eye to the broader value of network convergence, the optimal solution can readily be discerned. By incorporating the server connectivity into the BladeCenter chassis, the cost for even the most elaborate storage solutions can be made significantly less expensive than a comparable solution with individual server boxes. Over the next year, IBM and other vendors will bring additional products to market that will further integrate and simplify storage and storage management, continue the delivery of new interconnect and storage technologies to improve performance, capacity, and RAS, and further enable network convergence—and do so at competitive price points.

References

1. J. Satran, K. Meth, C. Sapuntzakis, M. Chadalapaka, and E. Zeidner, "Internet Small Computer Systems Interface (iSCSI)," IETF Request for Comments 3720, The Internet Society (April 2004); see *http://www.ietf.org/rfc/rfc3720.txt*.
2. IBM TotalStorage Products; see *http://www-1.ibm.com/servers/storage/disk/index.html*.
3. S. W. Hunter, N. C. Strole, D. W. Cosby, and D. M. Green, "BladeCenter Networking," *IBM J. Res. & Dev.* **49**, No. 6, 905–919 (2005, this issue).
4. Technical Committee T10, SCSI Storage Interfaces; see *http://www.t10.org*.
5. The Serial-ATA International Organization; see *http://www.serialata.org*.
6. Technical Committee T13, AT Attachment; see *http://www.t13.org*.
7. IBM Corporation, "IBM eServer xSeries ServeRAID Technology," white paper; see *ftp://ftp.software.ibm.com/pc/pccbbs/pc_servers_pdf/raidwppr.pdf*.
8. RDMA Consortium, Architectural Specifications for RDMA over TCP/IP; see *http://www.rdmaconsortium.org/home*.
9. IBM Corporation, "IBM eServer BladeCenter Fibre Channel Switch Interoperability Guide;" see *http://www-1.ibm.com/servers/eserver/bladecenter/literature/solutions_lit.html*.
10. The Internet Engineering Task Force (IETF); see *http://www.ietf.org/*.
11. R. Doms, "Dynamic Host Configuration Protocol," The Internet Engineering Task Force (IETF), Request for Comments 2131; see *http://www.ietf.org/rfc/rfc2131.txt*.
12. InfiniBand Trade Association, InfiniBand Architecture Specification, Volume 1. Release 2.1, October 2004; see *http://www.InfiniBandTA.org*.
13. InfiniBand Trade Association, InfiniBand Architecture Specification, Volume 2, Release 2.1, Chapter 5.1, pp. 79–80, October 2004.

Received December 16, 2004; accepted for publication March 4, 2005; Internet publication October 13, 2005

William G. Holland *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (wholland@us.ibm.com)*. Mr. Holland is a Senior Technical Staff Member working in BladeCenter development. Having provided technical guidance for the iSCSI, InfiniBand, and Fibre Channel BladeCenter options, he has ongoing responsibility for the BladeCenter I/O architecture and overall storage solution strategy. He received a B.S. degree in electrical engineering from Worcester Polytechnic Institute in 1984. Mr. Holland has worked in a number of roles at IBM, including circuit board tools development, S/390* processor logic design, worldwide product engineering manager for S/390, PCI network adapter design, network router architecture and performance, and xSeries performance analysis. With this diverse experience base, he was an original member of the BladeCenter team that created and refined the design from 1999 until it first shipped in 2002. Mr. Holland has been awarded 11 patents.

Patrick L. Caporale *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (caporale@us.ibm.com)*. Mr. Caporale joined IBM in 1995 in the IBM Research Division. He received a B.S. degree in electrical engineering from Manhattan College in 1996 and an M.S. degree from Columbia University in 2003. He spent the first half of his career in IBM working on networking solutions in IBM Research and the IBM Network Hardware Division. Mr. Caporale works on the BladeCenter development team as a technical leader for Fibre Channel options.

Don S. Keener *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (keener@us.ibm.com)*. Mr. Keener is an IBM Senior Technical Staff Member. He joined IBM after receiving a B.S. degree in electrical engineering from West Virginia University in 1978. Most of his career at IBM has been spent developing storage hardware products, such as ServeRAID and other storage adapter subsystems used with xSeries servers and systems. He currently works on the architecture and design of future RAID products and entry-level storage products. Mr. Keener holds 16 patents, most related to storage technology.

Andrew B. McNeill *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (amcneill@us.ibm.com)*. Mr. McNeill is an IBM Distinguished Engineer. He joined IBM after receiving a B.S. degree in electrical engineering from Clemson University in 1979. Most of his career at IBM has been spent developing storage hardware and software products, such as ServeRAID and FAStT subsystems used with xSeries servers. He worked on the BladeCenter storage attachment strategy from the inception of the project, including local disk, Fibre Channel SAN, and, most recently, iSCSI options. He currently works on the architecture and design of future entry and midrange storage products and management software. Mr. McNeill is a member of the IBM Academy of Technology and holds 27 patents, most related to storage technology.

Theodore B. Vojnovich *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (tbvojnov@us.ibm.com)*. Mr. Vojnovich joined IBM in 1984 to work on networking and communication products after receiving a B.S. degree in electrical engineering from Pennsylvania State University. He received an M.S. degree in computer engineering from North Carolina State University in 1989. He has worked on a variety of products, including video conferencing systems, Internet routers, and, most recently, network-attached storage systems. Mr. Vojnovich currently works on BladeCenter storage subsystems, focused on iSCSI at a product and solution level.