# BladeCenter processor blades, I/O expansion adapters, and units

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
D. M. Desai

*This paper describes the electrical architecture and design of the IBM eServer™ BladeCenter® processor blades, expansion blades, and input/output (I/O) expansion adapters and units. The processor blades are independent, general-purpose servers containing processors, chipsets, main memory, hard drives, network interface controllers, power input control circuitry, and local systems management. The blade architecture is robust and flexible enough to enable the design of general-purpose Intel-processor-based two-way and four-way blades, IBM POWER™-processor-based blades, blades based on AMD Opteron™ processors, and common expansion blades. In addition, the processor blades, expansion adapters, and units use a serialized–deserialized (SerDes) interface for the internal I/O fabrics, thus giving blades the flexibility to support virtually any I/O protocol that supports SerDes. Support for I/O fabrics beyond the base Gigabit Ethernet is accomplished via optional I/O expansion adapters for Fibre Channel, Myricom Myrinet®, InfiniBand®, or additional Gigabit Ethernet. Additional storage or peripheral component interface capability can be added to the processor blades via expansion blades.*

## Introduction

The IBM eServer* BladeCenter* system is a high-density, rack-mounted system with the capability to support up to 14 independent 30-mm-wide central processor blades. An overview of its design concept and architecture is presented in [1]. The system has an integrated input/output (I/O) via switch module [2] and a management [3] infrastructure that interfaces with the processor blades via a midplane [4]. The midplane also provides power distribution from the power modules [5] to the blades. To achieve high availability and reliability, it is designed to provide complete redundancy for power, cooling, and midplane signaling for the processor blades.

The processor blade architecture is heterogeneous; blades can make use of processors from many manufacturers and support most operating systems. Blades have been developed that support IBM POWER* processors, Intel Xeon** processors—including those supporting extended memory 64-bit technology (EM64T) and AMD Opteron** processors. These blades use a common printed circuit board form factor, expansion interfaces for additional I/O and storage, and a midplane interface. In most cases, a blade is implemented as a single printed circuit board and occupies one blade slot in the chassis. However, some blades, such as those supporting four processors, may require multiple blade slots on the basis of the printed circuit board area and other features required by the design.

The blades can be general- or special-purpose servers that contain processors, memory, optional local integrated drive electronics (IDE) or Small Computer System Interface (SCSI) disk drives, Gigabit Ethernet controllers that support the two primary I/O fabrics, a local baseboard management controller (BMC), and power conversion circuitry to convert the redundant 12-V inputs to the various voltages required by the blade electronic components. **Table 1** presents a comparison of the processor blade features.

All blades are designed to comply with a common midplane interface, which consists of a redundant pair of 60-pin Teradyne VHDM** (Very High-Density Metric) [6] and Molex PowerPlus** [7] connectors. Redundant interconnects were used to ensure maximum availability and reduce the total cost of ownership. The blades can



**0018-8646/05/$5.00 © 2005 IBM**

**Table 1** Functional comparison of processor blade features.

| | *HS20 (8676)* | *HS20 (8832)* | *HS40 (8839)* | *JS20 (8842)* | *HS20 (8843)* | *LS20 (8850)* |
|---|---|---|---|---|---|---|
| Processors | 1–2 Intel Xeon | 1–2 Intel Xeon | 1–4 Intel Xeon MP | 2 PowerPC* 970 | 1–2 Intel Xeon | 1 AMD Opteron |
| Core chipset | GC-LE | GC-LE | GC-LE | CPC 925 | Intel E7520 | |
| FSB speed (MHz) | 400 | 533 | 400 | 1,100 | 800 | 1,000 |
| Dual inline memory modules (DIMMs) DDR = double data rate | 4 DDR200; 256–2,048 MB | 4 DDR266; 256–2,048 MB | 8 256–2,048 MB | 4 DDR PC2700 | 4 DDR2; 256–4,096 MB | 4 DDR PC3200 |
| Dual Ethernets | Yes | Yes | Yes | Yes | Yes | yes |
| Local disk drives | 2 in. × 2.5 in. IDE | 2 in. × 2.5 in. IDE | 2 in. × 2.5 in. IDE | 2 in. × 2.5 in. IDE | 2 in. × 2.5 in. SCSI | 2 in. × 2.5 in. SCSI |
| Supports standard I/O expansion card | Yes | Yes | Yes | Yes | Yes | Yes |
| Supports small-form-factor (SFF) I/O expansion adapter | No | No | No | No | Yes | Yes |
| Supports one SFF daughter card and two local disk drives | No | No | No | No | Yes | Yes |
| BladeCenter storage expansion (BSE) unit | BSE-1 | BSE-1 | BSE-1 | None | BSE-2 | None |
| PCI expansion unit (PEU) | Yes | Yes | Yes | None | Yes | No |
| Supports power management | No | No | No | No | Yes | Yes |
| Automatic BIOS recovery (ABR) | No | No | No | Yes | Yes | Yes |
| Serial over LAN (SOL) support | No | Yes | Yes | Yes | Yes | Yes |

interface with other subsystems in the chassis using the following redundant buses: RS-485 to the management module for systems management communication, Universal Serial Bus (USB) 1.1 [8] and analog video to the management module for keyboard, video, and mouse (KVM) support, USB 1.1 to the media tray for removable media function, and serialized–deserialized (SerDes) channels to the switch modules for the high-speed I/O fabrics. Redundant control signals are provided that communicate presence, bus control, and blade slot addresses.

There are two pairs of redundant power modules in the BladeCenter enterprise chassis that provide +12 V through the midplane to redundant power connectors on the blade interface and to the chassis infrastructure. (The BladeCenter T chassis for the telecommunications market has a different structure and is described in [9].) Power modules in bays 1 and 2 support power domain A, consisting of blades 1 through 6 and the chassis infrastructure, while power modules in bays 3 and 4 support power domain B, consisting of blades 7 through 14. Each blade must provide the voltage conversions required by the various processors, chipsets, and other electronic components. Given the multiple voltages required by any variation of chipsets, processors, and other electronic components that can be supported, providing a single voltage to the blades and the chassis infrastructure is the most efficient method of power distribution.

The blades provide an I/O expansion capability in two ways. The first way to expand the I/O beyond the two primary Gigabit Ethernet fabrics is through the use of optional I/O expansion adapters. I/O expansion adapters provide an interface between the 100-MHz or 133-MHz PCI-X (Peripheral Component Interface eXtended) 1.0 blade bus [10] and switch modules 3 and 4 through the SerDes buses in the midplane. The I/O expansion adapters offer the flexibility to expand the blade I/O by additional fabrics made of Gigabit Ethernet, Fibre Channel, Myricom Myrinet**, or InfiniBand**. Two I/O expansion adapter form factors were developed to maintain the ability of a blade to fit in a single slot: one standard size that allows the installation of one disk drive and a second, smaller form factor that allows two disk drives to be installed on some types of blades.

The second way in which blade I/O can be expanded is via expansion blades. These offer the capability to add disk drives or support standard-size PCI and PCI-X adapters by providing—depending on the primary chipset used—one of two interfaces in the architecture for the expansion blade option. The first- and second-generation HS20 and HS40 blades provide a Teradyne HDM** [7] connector through which a Broadcom ServerWorks** Inter Module Bus (IMB) [11] passes to the expansion blade option. The third-generation HS20 blade provides a high-performance FCI GIG-Array** connector [12] which passes PCI-Express** links and a parallel SCSI bus to the expansion blade option (the JS20 and LS20 blades do not support this expansion capability). Expansion blade options occupy the slot adjacent to the blades. They are designed to be added and removed without tools and pivot into place on the base blade, forming a multiple-slot blade. Hot-swappable SCSI disk drives can be added to blades by attaching a blade SCSI blade storage expansion (BSE) unit. Standard-size PCI and PCI-X adapters can be added to blades by attaching a PCI I/O expansion adapter.

Each blade must provide a BMC to support critical BladeCenter functions, which include communication with the management module to support the power management protocol, with the control panel and light-path diagnostic light-emitting diodes (LEDs),[1] with the control panel buttons for power on/off, with the KVM and media access control, and with inventory, error reporting, and environmental monitoring and reporting. The BMC also supports other functions such as serial over LAN (SOL) and wake on LAN (WOL).

[1]Light-path diagnostics is a serviceability architecture IBM has implemented across its server line. LEDs are placed next to major or replaceable components (processors, DIMMs, hard disk drives, etc.). If one of these components fails or is suspect, the blade BMC lights the LED; seeing the light makes it easy to locate the failing component.

Processor and chipset manufacturers are continually increasing the frequencies at which their components operate, resulting in increased power demands. As power demand increased, it became apparent that the BladeCenter design had to be changed to ensure that sufficient power could be supplied by each power domain in the system and to maintain redundant power. This technical challenge provided the motivation to develop the Power Executive power management protocol used in the third generation of the HS20 and LS20 blades. This protocol must be supported by any blade capable of driving a power demand that exceeds the nominal value of a single power module. When a redundant pair of power modules is operating nominally in a nonfault condition, the power module pair can sustain the power demand for its domain in the chassis. However, should one power module in the redundant pair fail, the demand from that domain can exceed the power capacity of the unaffected single power module in that domain. In that case, the power demand in that domain must be reduced to the limits of a single power module within a specified period of time. The third-generation HS20 processor can throttle the processor(s) when a power module fails and bring the power demand of that domain below the required limit. In this power module failure scenario, the blades continue to run uninterrupted and are fully functional in a degraded mode until the failed power module is replaced.

The body of this paper covers the blade electrical architecture and design, specifically the midplane I/O interface, the I/O expansion adapter interface, and the capability to expand storage and PCI. Following the expansion architecture, we discuss the BMC and the functions it provides, and describe the Power Executive power management protocol. Finally, a detailed description of the design of the various blade, Ethernet, Fibre Channel, and PCI I/O expansion adapters, and BSE is provided.

## Blade midplane interfaces

A primary goal of the BladeCenter architecture is to provide an electrical and mechanical platform that will accept and support a broad range of processor architectures. To achieve this, industry-standard connectors and signaling were used for the midplane interface to the blades. Three groups of signals were specified: management communication and control signals, I/O fabric signals, and KVM and media tray signals. Power in the form of +12 VDC is supplied to each blade, which converts it to other voltages as required. The processor blade midplane control and USB signal interfaces are shown in **Figure 1**.

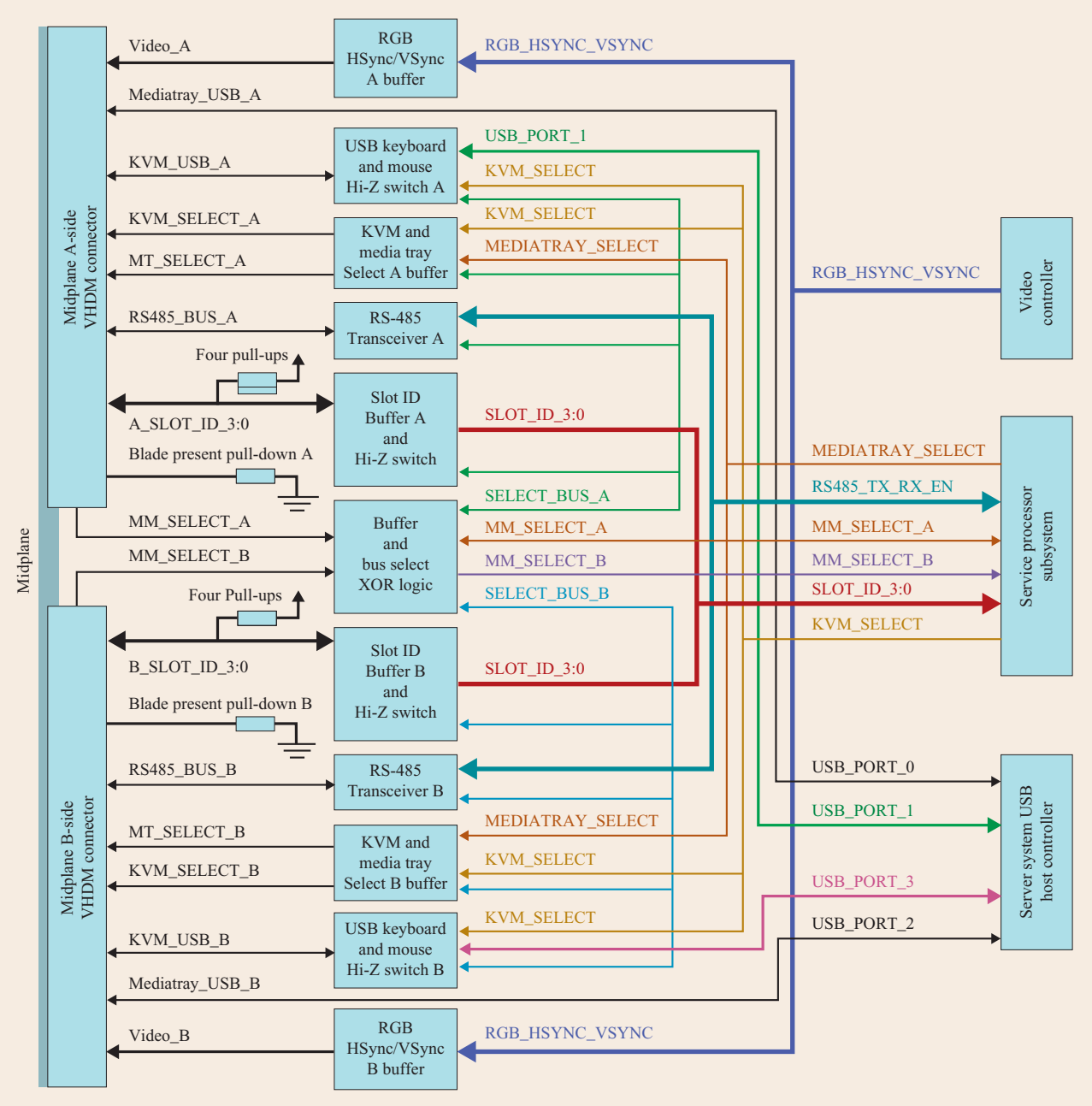


**Figure 1**

Processor blade to midplane USB, video, and control interface signals.

Another primary goal of the architecture is to ensure high reliability and availability; therefore, redundant signaling, connectors, and voltage distribution were specified, as was the capability for concurrent repair through the hot-swapping of replaceable components. In addition to specifying this common, redundant interface, a common printed circuit board form factor specified the maximum outline dimensions, keepout areas, and printed circuit board thickness.

### *Management communication and control signals*

The blade has redundant presence signals, one in each of the VHDM connectors. The blade signals its insertion or removal from the chassis to the management module

using these presence signals. There is a redundant pair of RS-485 buses (one in each VHDM connector) that are bidirectional, two-wire, and half-duplex. They connect all of the blades and both management module bays. These buses provide the communication link between the management module and each blade.

Four address signals (A3–A0) in each VHDM connector are used to establish a unique address for each blade slot. These signals are wired to the appropriate voltage and ground on the midplane. These unique slot addresses are used by the blade to determine whether it is being accessed by the management module on one of the two redundant RS-485 buses.

To determine which of the two redundant RS-485 buses is active, the blades decode the management module select A (`MM_Sel_A`) and management module select B (`MM_Sel_B`) signals from the management module. If `MM_Sel_A` is active, the A bus is used; if `MM_Sel_B` is active, the B bus is used. If both select signals are active or inactive, the blades assume an invalid state on both buses. A blade that occupies multiple blade slots communicates on the bus using the slot with the lowest address. For example, a blade occupying blade slots 2 and 3 must respond on the RS-485 bus using blade slot address 2.

#### *I/O fabric signals*

Four SerDes channels in the midplane connect each blade slot to switch module bays 1 through 4. Two differential pairs, one transmit and one receive, make up each SerDes channel. The first two channels, 1 and 2, are reserved for Gigabit Ethernet, while the remaining two channels, 3 and 4, are used by the optional I/O expansion adapters and provide support for most protocols that can transmit via SerDes; for example, Gigabit Ethernet, Fibre Channel, InfiniBand, and Myrinet.

#### *USB (KVM) and video signals*

The blades implement four independent USB buses, two through each VHDM connector for redundancy. The USB buses operate at 12 MB/s. The unique BladeCenter implementation of the USB buses allows the blades to share the keyboard, mouse, and media tray. Two of the buses are used for the keyboard and mouse, and two for the media tray. The blades use the `MM_Sel_A` and `MM_Sel_B` signals to determine which active set of buses to use for USB communication. The blade video subsystem, where applicable, provides redundant analog video output to the midplane video bus through the two VHDM connectors. The red, green, blue (RGB), horizontal sync (HSync), and vertical sync (VSync) signals are buffered on the blade.

#### *Power*

The blades accept +12 V from the midplane via redundant power connectors. The input circuit is designed to allow the blade to be hot-swapped in and out of an energized midplane without affecting the operation of other blades or modules. Upon insertion, auxiliary power is applied to a subset of the blade electronic components, particularly the BMC. Permission for power to be applied to the remainder of the blade is granted to the BMC by the management module.

#### *Printed circuit board and connector specification*

The maximum raw card size for blades is 394.2 mm × 226.99 mm (15.52 in. × 8.937 in.) with a thickness of 1.83 ± 0.02 mm (0.072 ± 0.007 in.). Each blade plugs into the midplane using a redundant pair of 60-pin VHDM and PowerPlus power connectors. See [13] for more detailed information.

### Blade expansion interfaces

To provide flexibility in expanding the I/O of a blade beyond the two base Gigabit Ethernet fabrics, the architecture specifies an interface for an optional I/O expansion adapter that can support a variety of protocols, including Gigabit Ethernet, Fibre Channel, Myrinet, or InfiniBand [2, 13]. An expansion capability for adding additional storage [13] or for adding standard PCI or PCI-X adapters is also specified.

#### *Blade I/O expansion adapter interface*

The blade provides two connectors for supporting I/O expansion adapters. One is a 64-bit PCI-X 1.0 electrical interface using a Molex 200-pin board-to-board stack connector. This interface supports 100 MHz or 133 MHz, depending on the I/O expansion adapter design. I/O expansion adapters with multiple controllers typically run at 100 MHz because of the additional capacitive load. The signaling levels are 3.3 V; otherwise, the I/O expansion adapter designs must conform to the wiring and timing specifications in the PCI-X 1.0 specification. The 200-pin connector provides power to the I/O expansion adapter and is limited to a maximum of 10 W. Auxiliary 3.3 V is provided to enable the BMC to read the I/O expansion adapter vital product data (VPD) EEPROM (electrically erasable programmable read-only memory) when the blade is not powered on (for example, when the blade is initially installed). This VPD is communicated to the management module, which verifies compatibility between the I/O expansion adapter and the type of switches installed in switch module bays 3 and 4. If there is a match, the management module grants permission to the blade to provide 12 V, 5 V, and 3.3 V to the I/O expansion adapter.

The second connector is a 12-pin Molex Plateau High-Speed Mezz** connector [7] which passes the four differential pairs of high-speed SerDes signals from the I/O expansion adapter to the blade. The blade then routes these signals to the VHDM connectors that plug into the midplane.

### Expansion blade interface

Two different blade storage and PCI expansion interfaces were provided for in the architecture. The BSE-1 interface specifies a 144-pin HDM connector through which a ServerWorks IMB bus passes. The IMB bus comprises two unidirectional 16-bit links running at 800 MHz, which provides a bandwidth of 1.6 GB/s in each direction. The BSE-1 interface also includes an Inter-Integrated Circuit ($I^2C$) Serial Bus, which allows the blade BMC to read VPD from the BSE-1 for inventory purposes, and +1.5 V to be used for proper termination of the IMB bus. The first- and second-generation HS20 blades and the HS40 blade used the BSE-1 interface because those blades used a ServerWorks chipset. The HS20 SCSI storage expansion unit, referred to as the BSE-1, connects to this interface.

The second interface in the design was the BSE-2 interface, supported by the third-generation HS20 blades. This interface passes an eight-lane PCI-Express bus through a 200-pin FCI GIG-Array connector along with an Ultra-320 SCSI bus. An $I^2C$ bus also passes through this connector, again to allow the BMC to access VPD. The BladeCenter SCSI storage expansion unit 2, referred to as the BSE-2, connects to this interface. The JS20 blade does not support this expansion capability.

### Blade control panel

Each blade has a control panel that contains blade status LEDs and push buttons that provide power on/off, KVM, and media access control. The power LED indicates the status of the blade power and is controlled by the BMC. When a blade is plugged into the chassis, the power LED blinks at a 4-Hz rate, indicating that the blade has standby power but has not yet been given permission to power up by the management module. Once permission is granted, the LED blinks at a 1-Hz rate, indicating that it can be powered on either by the control panel power button or by the management module.

The blue *location LED* is controlled by commands from the management module to the BMC and is used to assist in locating a specific blade in a chassis. The LED can be set to either blink or remain illuminated continuously. The *fault LED* indicates that the BMC has detected one of the following conditions: processor fault or mismatch, memory fault, one of the light-path diagnostic LEDs is on, or a critical temperature or voltage threshold has been exceeded. The *information LED* is used as a visible indicator that the blade may have a condition detected by the management module that may cause it to run in a degraded mode. The *activity LED* indicates that the blade has disk drive or network activity. If the *KVM* and *media tray LEDs* are on continuously, it means that the blade has ownership of these devices. The LEDs blink if ownership has been requested but permission has not yet been granted by the management module. The ownership of these devices is not locked together; this means that one blade can own the KVM while a different blade owns the media tray.

## Blade baseboard management controller (BMC)

Each blade must have a BMC, sometimes referred to as the *local service processor*. Depending on the particular blade, it will provide support for the following functions:

- Communication with the management module over the RS-485 bus.
- Power management functions.
- KVM and media tray ownership.
- Light-path diagnostics support.
- Automatic BIOS (basic I/O system) recovery (ABR).
- Automatic server restart (ASR).
- Predictive failure analysis (PFA).
- Serial over LAN (SOL).
- Wake on LAN (WOL).
- Inventory.
- Error logging.
- Environmental monitoring.

A major portion of the BMC function involves communicating with the management module over the RS-485 bus. Each blade has redundant RS-485 buses, one for each management module bay. If both management modules are installed for redundancy, only the RS-485 bus connected to the active management module is used. Each blade slot has a unique address determined by the four address signal lines from the midplane. The unique address allows the management module to communicate with one blade at a time.

The BMC coordinates with the management module for the power management capabilities of the blade. Upon insertion into the chassis, power domain 1 of the blade receives 12-V power from the midplane and converts this to auxiliary power for use by the BMC and other circuitry. The BMC begins its initialization sequence and places the power LED on the blade control panel in a fast-blink (4 Hz) mode. Once the BMC initialization completes, it reports VPD and status to the management module, then receives power-on permission from the management module. The BMC places the

power LED in a slow-blink (1 Hz) mode, indicating that initialization is complete, the blade has local power permission, the power button on the blade control panel is enabled, and the blade can be powered up by using the power button or by a management module command. If the power button is pressed, the BMC enables 12-V power to the remainder of the blade. There is also a remote capability for powering the blade on and off from the Web interface. This is accomplished by the management module sending the appropriate command to the BMC on the blade.

The BMC is involved with the selection of the KVM and the media tray. The control panel of each blade has a combination push button switch and LED for the KVM, and a similar one for media tray access. The BMC on each blade coordinates with the management module to control switching the KVM or media tray from one blade to another. The following list of steps illustrates the coordination between the local BMC and the management module when switching the KVM from one blade to another:

1. The KVM select switch on the control panel of blade A is pushed.
2. The BMC on blade A detects that the KVM button has been pushed and sends a request message to the management module over the RS-485 bus.
3. The management module checks to see whether another blade has control of the KVM.
4. The management module finds that blade B has control of the KVM and sends the blade B BMC a message to disable its KVM.
5. The BMC on blade B disables its KVM.
6. The BMC on blade B turns off the LED on its control panel and sends a message to the management module that it has disabled its KVM.
7. The management module sends a message to the BMC on blade A to enable its KVM.
8. The BMC on blade A enables the KVM, illuminates its KVM LED, and sends a message to the management module that it has control of the KVM.

The process for selecting the media tray is similar. Each KVM and media tray can also be assigned to a blade using the management module Web interface instead of the push buttons on the blade control panel.

The BMC provides light-path diagnostics support for the blade. The BMC can turn on a blade light-path diagnostic or control panel LED when it receives a command to do so from a source external to the BMC, such as the power-on self-test (POST) routine or the management module. It can also turn on the LEDs in response to its analysis of environmental (thermal and voltage) sensors. Light-path diagnostic LEDs controlled by the BMC are located on the blade control panel and the blade printed circuit board. The LEDs on the blade are strategically located next to removable subsystems, such as processors, dual inline memory modules (DIMMs), and disk drives, to assist in locating faulty components. The LEDs mounted directly on the printed circuit board have a backup capacitor controlled by a push button. This allows the LEDs to be illuminated when the blade is removed from the chassis, its cover removed, and the button pressed.

Some blades have automatic BIOS recovery (ABR) capability. In this case, the BMC will start a watchdog timer when the system is powered on. The POST routine executes code to the point at which it has ensured valid BIOS code by comparison with a checksum. If the code is determined to be good, the POST routine sends a message to the BMC telling it to turn off the watchdog timer, and the system continues booting. If the POST code does not send the message to the BMC before the watchdog timer expires, the primary BIOS code is assumed to be defective and the system is booted from a backup, or secondary, copy of BIOS.

The BMC also controls a watchdog timer for automatic server restart (ASR) in the operating system environment. The BMC periodically starts a watchdog timer, and an ASR device driver periodically sends a signal to the BMC to turn the timer off. If the operating system hangs up, the device driver is not able to turn off the watchdog timer. The watchdog timer then expires, and the BMC responds by rebooting the blade.

The BMC function aids in predictive failure analysis (PFA). In the case of memory PFA, for example, a systems management interrupt (SMI) handler routine keeps track of the number of single-bit (recoverable) errors within a given period of time. If this number is above a certain threshold, the SMI handler sends a signal to the BMC to label the memory as bad. The BMC logs the message to the management module and illuminates the light-path diagnostic LEDs in response. In this case, the specific LEDs illuminated would be the error LED on the blade control panel and the LED next to the defective memory. The management module would also illuminate the chassis control panel error LED.

Some blades support SOL capability. On those blades, the BMC and the network interface controller (NIC) route the serial data from the blade serial communications port to the network infrastructure of the chassis. The SOL traffic is routed through the Ethernet switch module located in switch module bay 1. The switch module routes the SOL traffic to and from the management module through the 100-Mb Ethernet connection between them. The SOL interfaces through a Telnet session on the management module. The

management module is also used to configure the components for SOL operation.

The WOL function can be used to power on a blade by sending a magic packet to either the blade's integrated NICs or the Ethernet I/O expansion adapter NICs. (A magic packet is an Ethernet packet that instructs the NIC to send a power-on signal to the power control circuitry of the blade.) The BMC can enable or disable the WOL function by command from the management module. Enabling or disabling this function is done at the blade level; it cannot be done at the individual port level for either the integrated NIC or the I/O expansion adapter NIC.

All major BladeCenter components, except the fans, have a VPD EEPROM that is used by the management module to compile a chassis inventory. For most of the components, the VPD EEPROM is accessed through an $I^2C$ bus connected directly to the management module; however, blades are an exception, since there are no direct $I^2C$ bus connections between the management module and the blades. All communication between the management module and blades regarding VPD is done through the RS-485 bus. The BMC receives requests from the management module to read the VPD from the local EEPROM on its internal $I^2C$ bus and transmits it to the management module. The management module can obtain data such as the type of blade, the code levels on the blade, the presence of an I/O expansion adapter, the presence of a expansion blade option, and the number of DIMMs installed.

The BMC accomplishes error logging by sending messages to the management module, which stores them in its log. The BMC passes messages from other sources, such as POST or diagnostics, and also generates its own messages when it determines that there is an error.

The BMC has environmental monitoring capability and monitors the temperatures and voltages of the blade, sending this information to the management module if a threshold is exceeded. For processor temperature monitoring, there are four levels of temperature, defined in increasing order: thermal warning reset, thermal warning, soft shutdown, and hard shutdown.

If the BMC reads a temperature higher than the thermal warning point, it sends a message to the management module, illuminates the error LED, and increases the fan speed but does not shut down the blade. If the temperature drops below the *thermal warning reset* level, which is lower than the *thermal warning* point, it sends a message to the management module that the blade is now operating in a normal temperature range and turns off the error LED. If the temperature continues to increase and reaches the *soft shutdown* threshold, the BMC attempts to signal the operating system to do a controlled shutdown. If the temperature reaches the *hard shutdown* threshold, the BMC shuts down the blade immediately.

## Power Executive function

A major BladeCenter design point is maximizing density, or processing power, in relation to its physical volume. As circuits become denser (higher circuit count per area) and faster, the power required by these devices increases. The processors in particular consume a large portion of the system power. As processor speed and the associated power consumption increased, design adjustments were needed to ensure that sufficient power could be supplied to the system. Power became a critical consideration of the design. One adjustment that was made was to use higher-capacity power modules. With the advent of higher-performance blades, such as the HS20 machine type 8843 blade, architecture and design extensions were made to accommodate their increased power consumption. These extensions take advantage of power-supply dynamic characteristics, redundant power system design, and power management capabilities of newer high-performance devices which, for example, have throttle capability that can suppress their processor speed to reduce power consumption.

To see how this increased power consumption is accommodated, we first describe the power architecture, then the pertinent power module and high-performance blade characteristics; finally, we describe how the advanced power management software—Power Executive—takes advantage of these design elements [3].

### *Power distribution*

The BladeCenter enterprise chassis has two independent power domains, each with two power modules for redundancy. Power domain A powers the first six blade slots and the chassis infrastructure, including the fans, switch modules, management modules, and media tray. Power domain B powers the remaining eight blade slots.

### *Power module characteristics and specifications*

Power supplies generally have specifications for the nominal and trip current values. The nominal value means that they can sustain that load indefinitely. The trip current value is a value that, if exceeded, will cause the power supply to shut down immediately. Additionally, a power module may have a specification for a short-term (e.g., one second) overload. In the BladeCenter design, when both (redundant) power supplies are operating, they can sustain a combined load greater than their nominal value. The maximum for this combined value can be up to 80% of double the power-supply nominal value. During operation above the nominal value of the power module, if a redundant module is lost, the aggregate power consumed by the

elements in the domain must be reduced to the nominal value or the remaining power module will also shut down. Currently, the highest-rated power supply is specified at 2,000 W for its nominal value. Its specified trip current is 2,623 W for 20 ms. Additionally, this power module is specified to maintain a load of 2,500 W for longer than one second.

#### *Power inventory*

The BladeCenter power management is under control of the management module. This management capability includes the power-on of a blade in conjunction with the BMC on each blade. There are power consumption numbers for all components in the BladeCenter chassis. Power is allocated to all common components first (fans, media tray, management modules, and switch modules); it is then allocated to installed blades. Every blade reports a maximum power consumption value. The high-performance blades, such as the HS20 machine type 8843, report a maximum power consumption value based on a dynamic inventory of options installed on the blade. The management module will allow complete power-on of a blade only if there is sufficient remaining power available for the blade. The power available within a domain is dependent on the capacity of the pair of power modules installed and the user-selectable power management policy.

#### *Blade advanced power management*

Beginning with third-generation blades (HS20 machine type 8843), not only are more accurate power consumption values reported, but they are also capable of dynamically reducing their power consumption by processor throttling. These blades report their power consumption reduction capability to the management module. Moreover, these blades have the capability to detect the loss of a redundant power module and then to reduce their power consumption immediately. If power reduction is required on loss of a power module, the amount of the reduction required by a blade is preset in the blade by the management module.

#### *Chassis-level power management*

When the user-selectable policy allows for the available domain power to be above the nominal value of the power module, the management module will allow power-on of blades to exceed this nominal value. When the nominal is exceeded, the management module will also set the amount of power reduction required in each of the blades. Thus, when a redundant power module is lost, each blade with power-throttling capability will automatically reduce its power consumption to at least the value preset by the management module. Thus, even though the power consumption is above the nominal value, power failure in a domain can be avoided by this rapid reduction of power demand when a single redundant power module is lost.

#### *User control of power management*

The behavior of the power management function is controlled by a policy that can be set by the user. This policy is given control if the nominal value of the power module can be exceeded and if the system can run in a nonredundant capacity. For example, setting this policy to "Redundant without performance impact" will cause the power management function to reject a blade that causes allocated power to be above the nominal value of the power module. Thus, in this case, the loss of a redundant power module would not cause any blades to throttle. With the policy set to allow power to be allocated above the nominal value, an event is set when the nominal is crossed and blades capable of throttling are set to reduce their power in the event of the loss of a redundant power module. Other customer-policy capabilities are being considered for future releases to allow customers to further define the throttle behavior of blades within a power domain.

#### *Continuing power management evolution*

High-performance blades have been designed with the capability of dynamically measuring their actual power consumption. The BladeCenter power management capabilities are evolving to further optimize the use of high-performance blade capabilities by taking advantage of their ability to measure power. On the basis of this measurement capability, both blade-level and chassis-level optimizations are being pursued for future releases.

### HS20 (machine types 8678 and 8832)

The first- and second-generation BladeCenter HS20s (machine types 8678 and 8832, respectively) are single-wide, dual-socket blades that use Intel Xeon processors and the ServerWorks Grand Champion** LE (GC-LE) chipset (**Figures 2** and **3**). Both designs support four DIMM sockets, dual Gigabit Ethernet ports to the midplane, expansion capability via the I/O expansion adapter for supporting two additional I/O fabrics, expansion capability for supporting a BSE-1 or PCI I/O expansion unit (PEU), a BMC, video and USB signals to the midplane for KVM and local media support, and two local integrated drive electronics (IDE) disk drives. A removable top cover provides access to install options such as additional processors, memory, disk drives, and I/O expansion adapters.

For both blades, the processors plug into two 604-pin zero-insertion-force (ZIF) sockets while supporting two

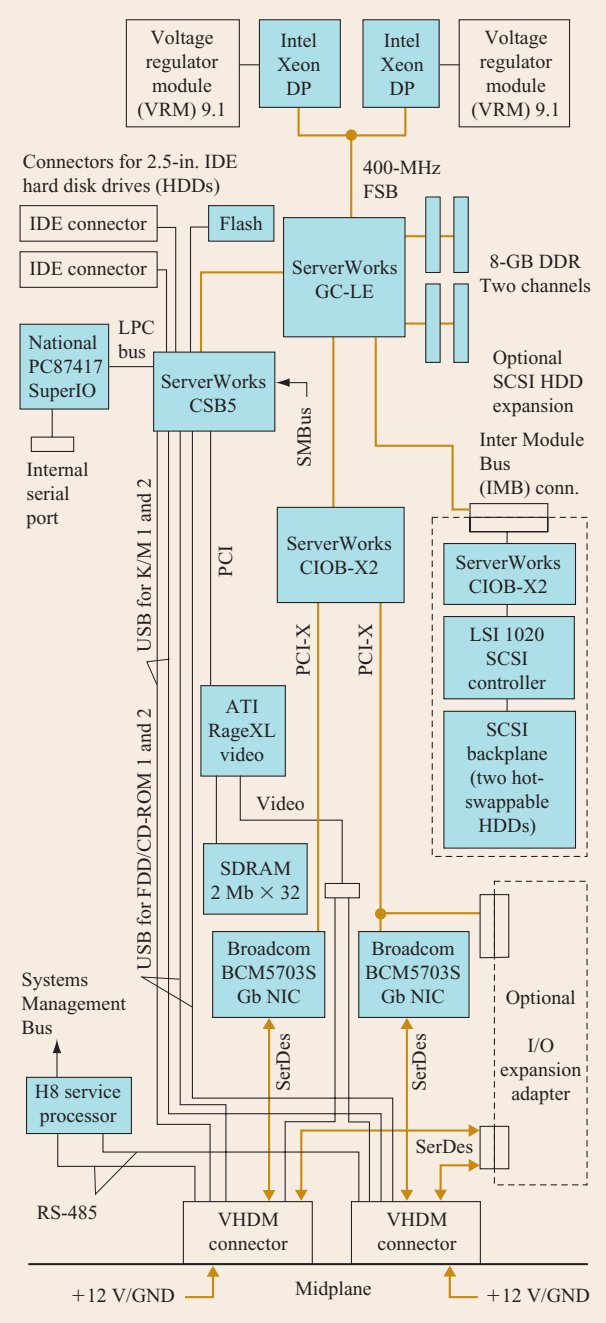


**Figure 2**

HS20 (machine type 8678) Intel Xeon DP processor blade.

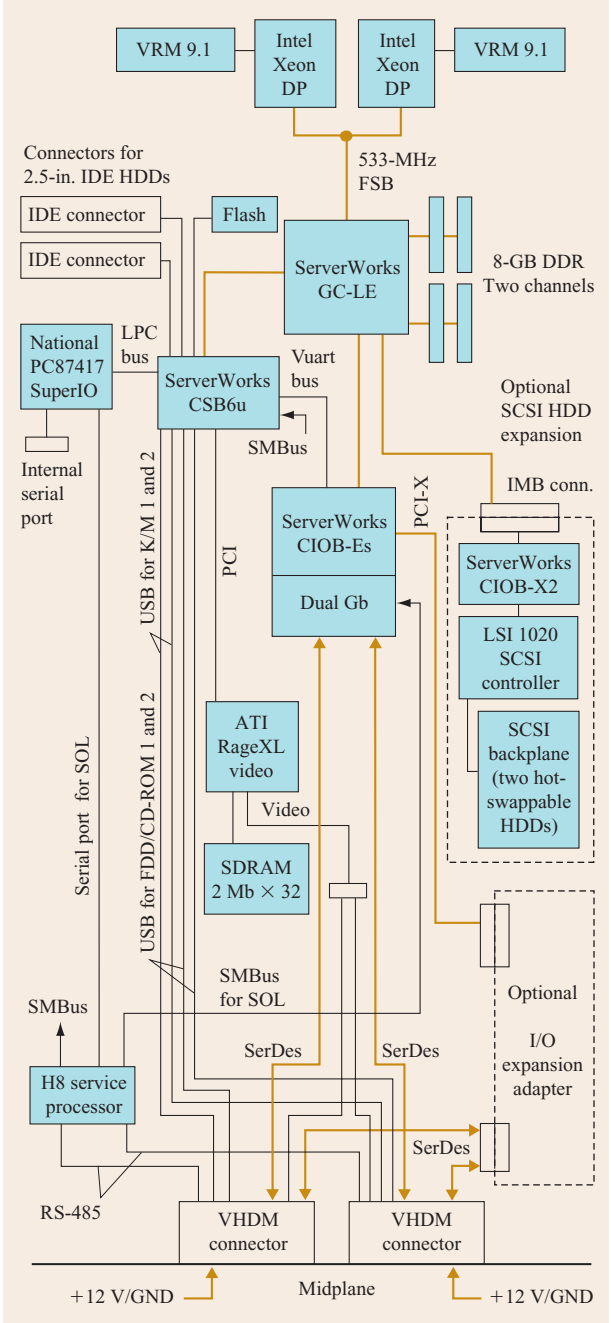


**Figure 3**

HS20 (machine type 8832) Intel Xeon DP processor blade.

different front-side-bus (FSB) frequencies. The first generation supports an FSB speed of 400 MHz, while the second generation supports an FSB of 533 MHz. The second-generation HS20 blade also supports the 533-MHz FSB version of Intel Xeon EM64T processors. Processor core speeds range from 2.0 GHz to 3.067 GHz.

Voltage to the processors is provided by two embedded voltage-regulator-down (VRD) modules designed using one Volterra VT1101 control module and six Volterra 1101 slave modules per processor. The slave modules use an aluminum heat sink riveted to the printed circuit board for cooling.

The ServerWorks chipset comprises the GC-LE and either the CIOB-X2 and CSB5 (on the 8678 blade) or the CIOB-ES and CSB6 (on the 8832 blade) modules. The GC-LE module serves as the interface between the processors, memory, the two IMB I/O interface units to interface with CIOB-X2 or CIOB-ES I/O bridges, and one Thin-IMB (so called because of fewer data signals) I/O interface unit to interface with the CSB5 or CSB6 Southbridge modules. This module provides an efficient interface between the processor bus and memory with data integrity features such as error correction code (ECC), chipkill, memory scrubbing, and hot-spare memory. The term *memory scrubbing* refers to a process that actively reads memory during idle periods to search for and correct errors in memory. *Hot-spare memory* involves setting aside a bank of memory as a spare, which is then activated if another bank of memory fails. ECC detects and corrects single-bit errors, and chipkill detects and corrects for a single failed dynamic random access memory (DRAM) module on a DIMM. The HS20 supports up to four pluggable DDR-200 (on the 8678 blade) or DDR-266 (on the 8832 blade) DIMMs. DIMM sizes of 256 MB, 512 MB, 1 GB, and 2 GB are supported, for a maximum capacity of 8 GB main memory. The memory subsystem is two-way interleaved, requiring the DIMMs to be installed in matched pairs.

The GC-LE, in conjunction with the CIOB-X2 (or CIOB-ES), provides a high-performance data path between the processor bus and the PCI-X-based I/O subsystem. The GC-LE filters all unnecessary bus traffic going to the processor bus and performs packing and unpacking of data for PCI-X accesses, which in turn reduces processor utilization and provides more processor bandwidth for server applications.

The CIOB-X2 is an integrated I/O bridge that provides a low-latency, high-performance data path between the IMB and the two independent external 64-bit PCI-X interfaces capable of running at 133 MHz. The CIOB-X2 interfaces with the GC-LE using an IMB bus, and its purpose is to bridge IMB bus transactions to PCI-X (used for the I/O expansion adapter interface) and to interface with the integrated Ethernet controllers. The bridge supports concurrent, peer-to-peer PCI-X transactions, deep I/O queues, write-through caching, and error-reporting functions. The IMB bus running at 800 MHz provides 1.6 GB/s bandwidth in each direction. It supports two unidirectional 16-bit links and is capable of supporting 32 pipelined transactions in each direction.

The CSB5 and CSB6 Southbridge modules integrate the following common I/O functions: bus mastering for supporting the dual IDE disk-drive interfaces, an enhanced 8237 direct memory access (DMA) controller, two cascaded 8259 interrupt controllers, support for four USB 1.1 ports compliant with the Open Host Controller Interface specification, a Low Pin Count (LPC) Interface for POST/BIOS flash memory, and General Purpose I/O (GPIO) and Advanced Configuration and Power Interface (ACPI) support. These Southbridge modules also provide a 32-bit 33-MHz PCI bus to which the video controller is attached. The graphics function is provided by an ATI RAGE** XL graphics and video accelerator chip and an external 2 Mb $\times$ 32 synchronous DRAM (SDRAM) module. The maximum resolution supported in the BladeCenter system is 1,024 $\times$ 768 pixels at a refresh rate of 75 Hz.

The Ethernet subsystem is another of the primary differences between the first- and second-generation HS20 blades. The first-generation 8678 uses one CIOB-X2 bridge paired with two Broadcom BCM5703S single-channel Gigabit Ethernet controllers, one connected to each of the PCI-X 133-MHz buses provided by the CIOB-X2. The second-generation 8832 uses one ServerWorks CIOB-ES bridge module in which the functions of the CIOB-X2 and a Broadcom BCM5704S dual-channel Gigabit Ethernet controller are integrated. Both of these Gigabit Ethernet media access controllers (MACs), the BCM5703S and CIOB-ES, provide redundant Gigabit SerDes interfaces, one to each of the VHDM connectors that plug into the midplane. The midplane then routes each channel to two different Gigabit Ethernet switch modules. To ensure consistent Ethernet port designation across all types of blades, the BIOS allows the port enumeration to be changed so that, for example, an Ethernet port routed to switch module 1 will be consistent with the port numbering of another blade.

The 8678 and 8832 blades provide the capability to expand their I/O through the support of optional I/O expansion adapters. These cards are connected to the third and fourth BladeCenter I/O fabrics supported by switch modules 3 and 4. These fabrics can be used to provide additional Gigabit Ethernet, Fibre Channel, Myrinet, or 1X InfiniBand capability (one transmit differential pair and one receive differential pair of wires). The I/O expansion adapter host interface is a PCI-X 133-MHz bus.

The 8678 and 8832 blades can be expanded to a double-wide blade to increase the storage capacity of the blade by attaching a BSE-1. The BSE-1 provides support for two hot-swap Ultra-320 3.5-in. slim disk drives. The HS20 blade has an HDM connector through which an 800-MHz IMB bus passes for connection to a CIOB-X2 bridge module on the BSE-1.

The BMC is an 8-bit Hitachi H8 microcontroller which supports the following functions: interfacing with the chassis management module via the RS-485 bus, local environmental monitoring for voltages and temperatures on the blade, local LED control for light-path diagnostics function, local power control, and ASR.

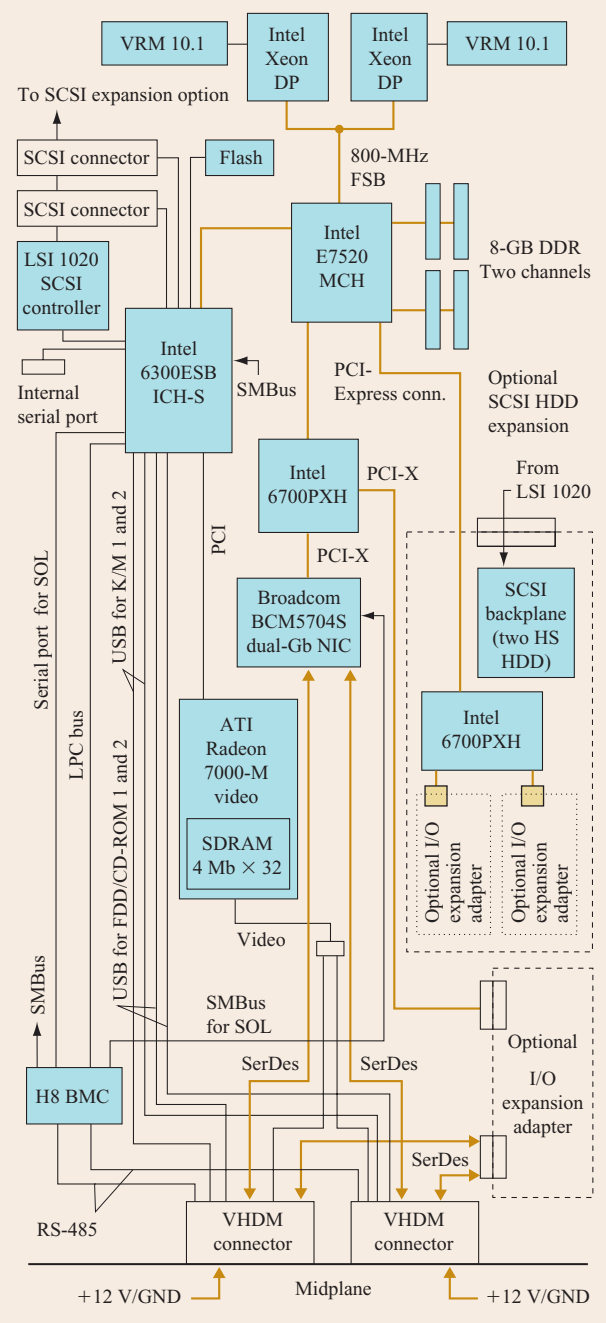


**Figure 4**

HS20 (machine type 8843) Intel Xeon DP processor blade.

## HS20 (machine type 8843)

The BladeCenter HS20 (machine type 8843) is the third generation of the HS20 blade. It extends the performance of the BladeCenter system with Intel Xeon EM64T processors, DDR-2 memory, small-form-factor (SFF) SCSI disk drives, and a new BMC (**Figure 4**). This new performance level is still contained in a single-wide blade with I/O expansion capability using either of two I/O expansion adapter form factors and the BSE-2 and PEU.

The 8843 blade is a single-wide blade with the same mechanical features as the HS20 blades discussed above. A removable top cover provides access to install options such as additional processors, memory, disk drives, and I/O expansion adapters. Due to the increased thermal load of Intel Xeon processors with 800-MHz FSB, a new heat-sink design was used that is larger than previous HS20 heat sinks and is made of extruded copper for greater heat transfer. Additional airflow to the SCSI disk drives is provided by a fresh air opening in the redesigned front bezel. The control panel was redesigned to be smaller than previous HS20 control panels and allows fresh air to bypass the processor heat sinks, directly cooling the SCSI disk drives.

In previous HS20 blades, one disk drive had to be removed to insert an I/O expansion adapter. In the 8843 blade, a new size of I/O expansion adapter was designed to fit onto the same I/O expansion adapter connectors, but still allow two disk drives to be installed on the blade. These new SFF I/O expansion adapters will fit onto only the 8843 blade and the BSE-2 option. The BSE-2 plugs into the FCI GIG-Array connector used for connecting expansion blade options.

The 8843 has two mPGA 604-pin ZIF sockets for the Intel Xeon processor with 800-MHz FSB. The base clock frequency is 200 MHz and is quad-pumped (i.e., every clock cycle is broken up into four sampling periods during which the data can change). This feature allows for up to 6.4 GB/s of bandwidth. The processor supports hyperthreading and new thermal management schemes called *thermal monitor 1* and *thermal monitor 2*. Processor speeds of 3.2 GHz up to 3.8 GHz are supported.

The processors in the 8843 require a split-plane power distribution design. Each processor has its own VRD, providing up to 110 W of power. This power is delivered to the processor by using one internal 2-oz copper plane and one external 1.5-oz copper plane per processor. The individual power planes are physically connected to only one processor each, but extend over both processors to provide a continuous reference plane for the front-side bus signals.

The processor VRDs are designed using one Volterra VT1103ML control module and eight Volterra 1103SC slave modules per processor. Using eight slaves per processor provides an adequate power margin to accommodate increased processor power requirements in the future. The slave modules use an aluminum heat sink riveted to the printed circuit board for cooling.

The 8843 uses the Intel E7520 memory controller hub (MCH) as the Northbridge chip that connects the processors to the main memory. The MCH is a 1,077-ball

flip-chip ball grid array (FC-BGA) package measuring 42.5 mm per side. The 8843 uses industry-standard PC3200 DDR-2 DIMMs arranged in an interleaved architecture. The MCH provides two DDR-2 channels operating synchronously with a maximum bandwidth of 3.2 GB/s per channel. The memory configuration of each channel must match in terms of technology, size, and speed. The maximum memory size supported using 4-GB DIMMs is 16 GB.

The MCH has two types of high-speed buses for connection to I/O devices. The MCH chip has three eight-lane PCI-Express buses operating at 2.5 Gb/s per lane, providing up to 12 GB/s of total I/O bandwidth. Each PCI-Express port can be configured as an eight-lane, four-lane, or two-lane port. The 8843 is designed with an eight-lane PCI-Express port connected to an Intel 6700 PXH 64-bit PCI hub (PXH) and an eight-lane PCI-Express port connected to an expansion socket for attaching a BSE-2. The second high-speed bus is the Hub Interface 1.5 (HI 1.5) bus, which is used for connection to the Intel I/O controller hub (ICH-S 6300ESB) Southbridge. The HI 1.5 operates at 133 MHz and is a 16-bit-wide interface. Because of the number of high-speed buses, the layer count of the printed circuit board was increased to ten layers. The MCH chip has several high-speed communication buses (FSB, PCI-Express, and HI 1.5) with exacting layout requirements. To minimize trace routing lengths for each of these buses, the MCH chip was rotated at a 45-degree angle on the board.

The MCH core is powered by a 1.8-V regulator and uses 1.5 V and 1.2 V as I/O voltages for the PCI-Express bus and FSB, respectively. The MCH chip has a total power consumption of approximately 10 W and receives most of its airflow preheated by the processors. To keep the MCH cooled below its 85°C limit, a 75-mm × 60-mm aluminum heat sink is mounted to the top and makes contact with the back of the die.

The Intel 6700PXH PCI-X bridge chip is a PCI-Express to PCI-X bridge. It generates two subordinate PCI-X buses running at 133 MHz which are connected to the blade I/O expansion adapter connector and to a BCM5704S dual Gigabit Ethernet controller. The two buses can operate asynchronously, but no peer-to-peer traffic is allowed between buses. The BCM5704S provides connection to the switch modules in switch bays 1 and 2 in the BladeCenter chassis.

The Intel 6300ESB Southbridge chip connects to the Northbridge MCH chip using the HI 1.5 bus. The ICH-S provides USB, real-time clock, and legacy I/O ports, and contains a PCI-X 1.0 bridge that operates at 66 MHz.

The 8843 is the first blade to use 2.5-in. small-form-factor SCSI disk drives as the primary disk storage. An LSI Logic** 1020 Ultra-320 SCSI controller generates an LVD (Low Voltage Differential) SCSI bus that connects to the two right-angle SCA2 connectors on the printed circuit board and to the GIG-Array connector. The LSI 1020 supports hardware RAID (redundant array of independent disks) 1 and enhanced RAID 1 (RAID 1E), which can be configured through the setup utility included in the system BIOS. When the optional BSE-2 is used, the two additional 3.5-in. disk drives in the expansion option are connected to the same SCSI bus as the SFF 2.5-in. disk drives, allowing RAID 1E and hot-spare configurations, which provide the blade with the capability to automatically replace a faulty hard drive with a spare and to reconstruct the lost data from the failing drive on the new replacement drive. The SCSI bus is terminated at each end of the bus. When the BSE-2 option is present, terminators at the GIG-Array connector are disabled, allowing the bus to continue on to the BSE-2 for connection to the 3.5-in. disk drives on the expansion option.

The 8843 uses a new BMC to provide a systems management interface to the chassis management module. This BMC module provides system environmental monitoring, local power control, SOL, and data reporting for the Power Executive feature in the management module. The BMC chip controls all communication between a blade and the management module. Once communication has been established between the BMC and management module, local power permissions that enable the blade to power on are passed to the BMC. During boot time, the BMC reports the expected power load to the management module Power Executive software for chassis power monitoring. The BMC responds to power-supply failures by managing power consumption to preset limits based on remaining power capacity.

### HS40 (machine type 8839)

The BladeCenter HS40 (machine type 8839) is a modular four-way blade server supporting up to four Intel Xeon multiprocessor (MP) processors operating at 2.0 GHz, 2.2 GHz, 2.5 GHz, 2.7 GHz, 2.8 GHz, and 3.0 GHz with L3 cache sizes of 1 MB, 2 MB, and 3 MB. It is a double-wide enclosure consisting of a processor printed circuit board and an I/O printed circuit board [**Figures 5(a)** and **5(b)**]. The chipset is the ServerWorks Grand Champion GC-LE comprising the CMIC-LE Northbridge, the CIOB-X2 PCI-X I/O bridge, and the CSB6 Southbridge.

The processor board mates with the I/O board through three high-speed flexible cables and a power cable. The processors plug into four 603-pin ZIF processor sockets with an FSB frequency of 400 MHz, yielding a bandwidth of 3.2 GB/s. In addition, the HS40 processor board supports a scalable memory subsystem. There are eight 184-pin PC2100 DDR DIMM sockets supporting a

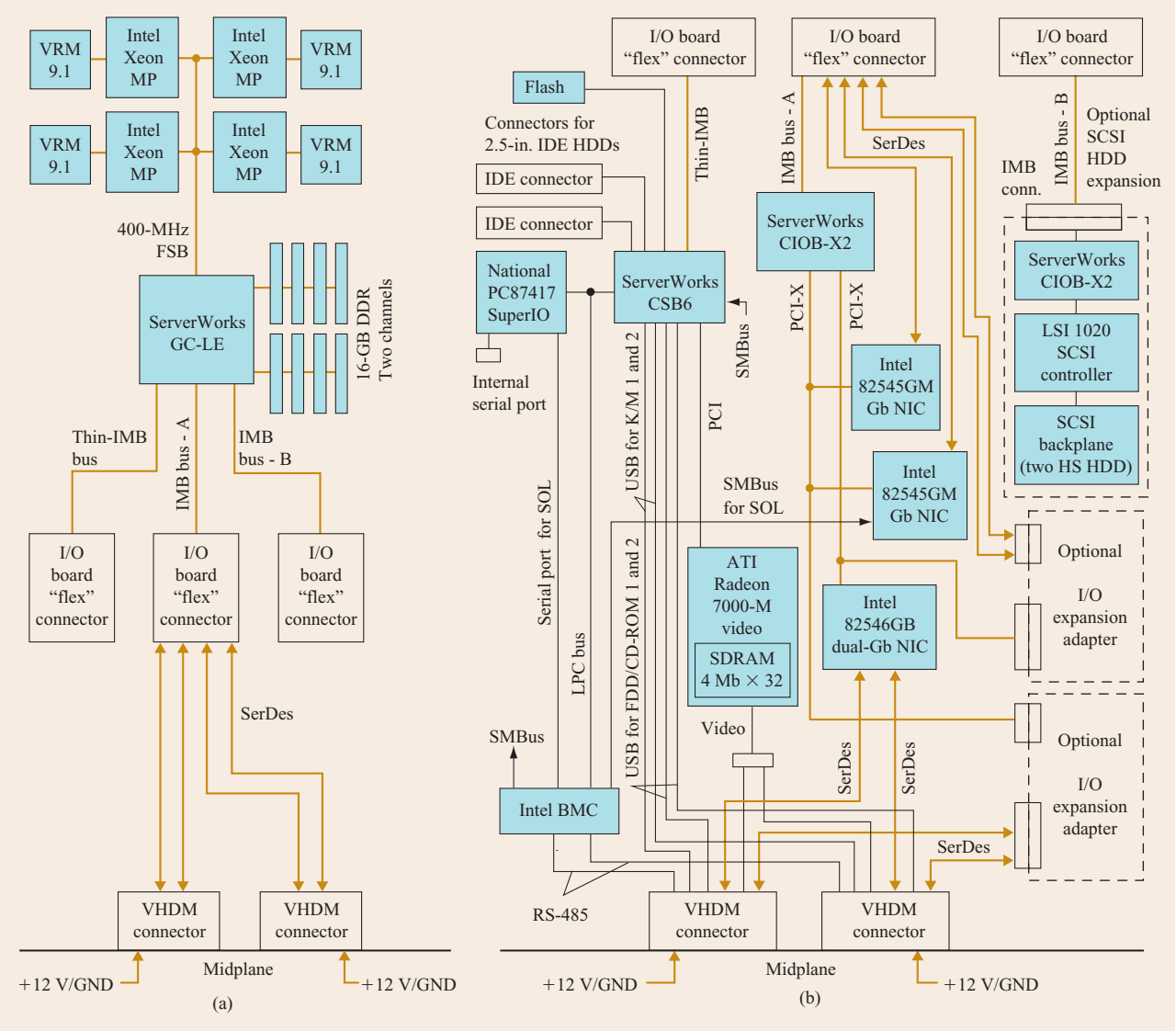


**Figure 5**

HS40 (machine type 8839) Intel Xeon MP processor blade. (a) Processor and memory printed circuit board. (b) I/O printed circuit board.

maximum of 16 GB of main memory. DIMM sizes supported are 256 MB, 512 MB, 1 GB, and 2 GB. These DIMM sockets are divided into two channels, providing a two-way interleaved memory architecture. DIMMs must be installed in matched pairs (one in each channel, with up to four DIMMs per channel). The dual-channel interleaved architecture provides 3.2 GB/s of memory bandwidth. Reliability, availability, and serviceability (RAS) features supported are ECC, chipkill, memory scrubbing, and hot-spare memory. The processor board construction is 16 layers of FR4 printed circuit board material.

The HS40 processor board has two primary interfaces with the I/O board. The first interface is the IMB bus from the GC-LE CMIC to the GC-LE CIOB and provides up to 3.2 GB/s of bandwidth between these two devices. The other primary interface—the Thin-IMB bus—links the GC-LE CMIC and the GC-LE CSB6 and provides 400 MB/s of bandwidth between them.

The I/O board provides the base for all I/O interfaces on the HS40 blade. It integrates four Gigabit Ethernet interfaces, redundant video and RS-485 interfaces, and four USB 1.1 interfaces. It is upgradable through several expansion interfaces including one GIG-Array expansion

connector, two pairs of 64-bit 100-MHz PCI-X and SerDes I/O expansion adapter connectors, and two IDE interface connectors. Two 64-bit PCI-X bus segments are supported via the CIOB-X2 PCI-X expander bridge component. The CIOB-X2 in the HS40 I/O design configuration provides up to 100 MHz PCI-X support. There are two PCI-X 3.3-V 64-bit I/O expansion adapter slots and three 3.3-V 64-bit integrated Ethernet controllers. Two integrated Intel 82545EB single Gigabit Ethernet controllers and one integrated Intel 82546GM dual Gigabit Ethernet controller provide the Ethernet I/O fabric support. The HS40 blade supports one or two I/O expansion adapters. These cards connect to the blade through the separate 100-MHz 64-bit PCI-X connectors. Integrated video is provided by an ATI Radeon** 7000 with 16 MB of memory.

BIOS flash storage, server management, video, and system I/O are all contained on the I/O board and interfaced through the CSB6 Southbridge component, which is connected by a Thin-IMB interconnect to the CMIC-LE. The IDE bus is controlled by the CSB6 and is routed to two SFF IDE disk drive connectors.

The I/O board contains both dc-to-dc (D2D) modules and embedded D2D converters, which are necessary to provide power to the electronic circuits, IDE disk drives, PCI-X I/O expansion adapters, and the processor board. The bulk power supply for the D2D modules comes from a system-wide 12-V bulk power distribution. The HS40 processor board contains embedded D2D converters necessary to provide power to the processor board circuits, two processors, and memory subsystems. Because of the connector current rating limitations, the HS40 processor board cannot generate all of its power from the processor board midplane connections. Hence, there is a requirement to pull 12-V power from the I/O board. This is accomplished through a pluggable four-wire cable between the two boards.

A BSE-1 can be attached to the HS40 to allow additional storage using one or two hot-pluggable SCSI Ultra-160 or Ultra-320 disk drives. When the BSE-1 is attached, the HS40 will use three slots in the chassis. Without the BSE-1, the blade can support one or two local IDE disk drives (2.5-in. form factor). These IDE disk drives are not hot-pluggable. The user loses an I/O expansion adapter slot for each local IDE disk drive that is installed.

Server management is provided by an Intel Sahalee[2]-based BMC and server management software. The BMC is located on the I/O board and provides power control, voltage sensing, temperature sensing, error sensing and logging, board identification via VPD, and D2D regulator fault detection. The BMC and management module communicate using the RS-485 interface.

[2]An Intel BMC chip that is IPMI-compliant (IPMI: Intelligent Platform Management Interface).

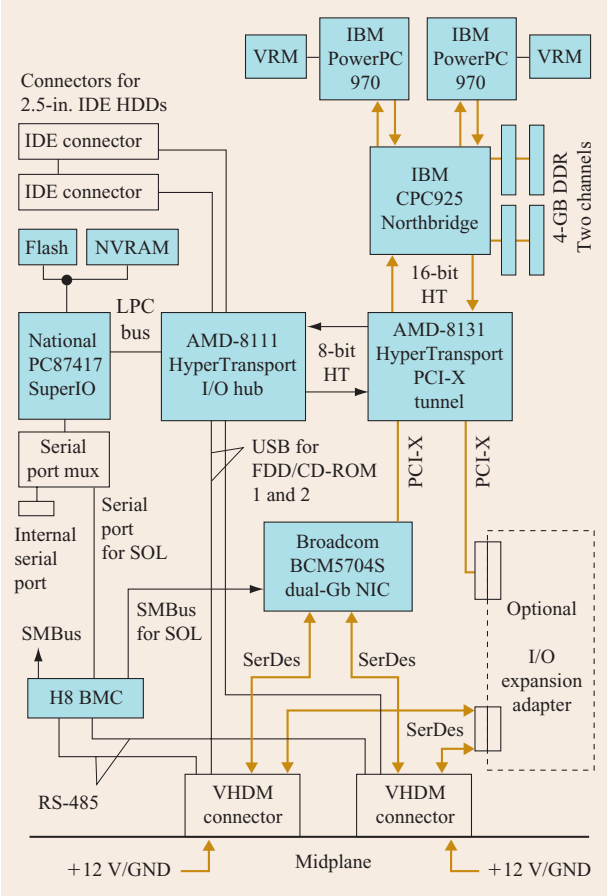


**Figure 6**

JS20 (machine type 8842) IBM PowerPC processor blade.

The HS40 processor board interfaces with the system management controller located on the I/O board through multiple $I^2C$ bus interfaces.

## JS20 (machine type 8842)

The BladeCenter JS20 (machine type 8842) is a two-way, single-wide blade server based on PowerPC* technology and offers a full 64-bit server platform for high-density, high-performance computing (**Figure 6**). Two 8842 models are available: the 8842-21X, which is based on the 64-bit POWER4*-based PPC970 processor, and the 8842-41X, which is based on the follow-on PPC970FX processor. The JS20 supports 64-bit Linux** and AIX*. It was a unique challenge during the design of the JS20 to reuse as many existing features and functions from HS20 blade servers as possible to ensure the seamless integration of the JS20 into the BladeCenter environment, and yet handle the specific requirements of the PowerPC chipset and architecture, including the different

initialization requirements and host-firmware stacks (BIOS compared with Open Firmware).

The processor subsystem comprises either two PPC970 or two PPC970FX processors which operate at internal clock frequencies of 1.6 GHz and 2.2 GHz, respectively. The processors are based on POWER4 Architecture* and provide a high-superscalar 64-bit design with dual fixed-point, dual floating-point, and dual load and store units with 64-bit data paths, along with a powerful single-instruction multiple-data (SIMD) unit with 128-bit data paths. Each processor includes a 32-KB L1 data cache and a 64-KB instruction cache, and a 512-MB L2 cache with ECC. They are manufactured in IBM 130-nm (970) and 90-nm (970FX) copper process technologies and use core voltages of 1.3 V and 1.1 V, respectively, and 1.3 V for the elastic interface. The processor bus operates at a 2:1 ratio to the processor frequency, i.e., at 0.8 GHz (970) or 1.1 GHz (970FX), and connects the processors to the CPC925, an integrated bridge and memory controller chip. The CPC925 provides bridging functionality among the processors, the memory subsystem, and an AMD HyperTransport** (HT) bus to the PCI-based I/O subsystem. An elastic interface to each processor supports a maximum aggregate bandwidth of either 6.4 GB/s for the PPC970 (800,000 transfers $\times$ 4 bytes $\times$ 2 directions) or 8.8 GB/s for the PPC970FX (1.1 million transfers $\times$ 4 bytes $\times$ 2 directions).

The memory subsystem of the JS20 consists of the memory controller function of the CPC925 and up to four DIMM modules. The memory controller supports ECC for single-bit error correction, double-bit error detection, and memory scrubbing to support PFA. The scrub function is used in *scrub immediate* mode for memory initialization during startup, and in *background scrub* mode during normal system operation. Chipkill or hot-spare memory is not supported. The DIMMs supported are 2.5-V, 84-pin (72 bit = 64 data + 8 bit ECC), DDR-1 333-MHz (PC2700) registered ECC modules with 512-Mb or 1-Gb chips, organized $\times$8 (8 data bits per DRAM module). Supported DIMM sizes are 256 MB, 512 MB, and 1 GB, up to a maximum system memory capacity of 4 GB. DIMMs must be installed in pairs of equal types, so that only 2-DIMM or 4-DIMM configurations can be used.

The I/O subsystem is basically a PCI-based I/O subsystem with some integrated features and some optional expansion features. It is based on a fast 16-bit-wide HyperTransport bus driven by the CPC925 bridge function and connecting to an AMD-8131** HT/PCI-X bridge (3.2 GB/s bandwidth). The AMD-8131 HT tunnel chip provides two PCI-X buses and a downstream 8-bit HyperTransport bus for ancillary I/O. One PCI-X bus is connected to an integrated BCM5704S dual Gigabit Ethernet controller, and the other PCI-X bus is connected to the I/O expansion adapter connector in which optional I/O expansion adapters can be installed. The I/O signals of the I/O expansion adapter are wired to the midplane VHDM connectors. The midplane routes these SerDes channels to switch modules 3 and 4. The downstream 8-bit HT connects to an AMD-8111** Southbridge, with an aggregate bandwidth of 800 MB/s, and provides USB, IDE, and LPC interfaces. The USB interface is wired directly to the midplane connectors, the IDE is wired onboard to two IDE connectors, and the LPC is wired to a PC87417 SuperIO chip for universal asynchronous receiver/transmitter (UART), real-time clock, and Xbus-attached boot flash and nonvolatile RAM (NVRAM). The JS20 blades do not have hardware RAID support for the IDE subsystem; they do not have an onboard video controller (switched KVM is not supported); and they do not have a GIG-Array connector for expansion blade options.

USB 1.1 and Gigabit Ethernet are integrated on the JS20 board. There are two USB 1.1 controllers, driven by the AMD-8111, with dual redundant connections to the midplane. The two Gigabit Ethernet ports use the BCM5704S and also have dual redundant connections to the midplane. There is also a single enhanced IDE (EIDE) controller with two channels, driven by an AMD-8111 and wired to two IDE connectors for optional IDE (40-GB, 5,400-rpm) drives. An onboard PCI-X I/O connector supports an optional I/O expansion adapter (e.g., Ethernet or Fibre Channel) common to all blades.

The JS20 uses a Hitachi H8 microcontroller for the BMC. The firmware running on this controller is almost identical to the non-PowerPC blades. The differences primarily handle the PowerPC chipset, including the initialization sequence. The non-platform-specific functions, such as SOL, environmental sensing, power control, communication to the management module, and LED handling, are the same.

The JS20 onboard VRDs generate all required voltages from the 12-V bulk power supply from the BladeCenter midplane connection. This power is sourced by two redundantly connected power supplies located in the rear of the BladeCenter chassis. Voltages required are 1.1 V, 1.3 V, and 1.5 V for the processor subsystem (including CPC925), 1.5 V and 2.5 V for the memory subsystem, and 1.8 V, 2.5 V, 3.3 V, and 5.0 V for the I/O and service subsystem.

The JS20 printed circuit board is an FR4 board in IBM FRED8 technology with 1-mm pitch, 3-mil line width, 4-mil line-to-line, and two lines per channel; it has 16 layers total with six signal planes, eight power planes, and two mounting planes.

## LS20 (machine type 8850)

The AMD Opteron LS20 for IBM BladeCenter (machine type 8850) is a single-wide blade which features a two-

socket Opteron processor and provides 64-bit computing capabilities that extend the x86 instruction set and maintain x86 32-bit binary compatibility (**Figure 7**). The Opteron architecture uses an integrated memory controller and innovative HyperTransport technology to provide memory and I/O bandwidth at the speed of the processor, reducing or eliminating bottlenecks.

The LS20 can be configured with up to two single-core (SC) or dual-core (DC) processors with currently supported operating frequencies of up to 2.6 GHz (SC) or 2.2 GHz (DC). The chipset consists of an AMD-8131 HyperTransport dual PCI-X tunnel, an AMD-8111 HyperTransport I/O hub, an LSI 1020 PCI-X-to-Ultra-320 SCSI controller, a BCM5704S dual Gigabit Ethernet controller, and an ATI7000M video controller. The memory is made up of four DDR PC3200s with ECC and chipkill support. The blade supports up to 16 GB of memory using DIMM densities of 512 MB, 1 GB, and 2 GB.

The blade contains one 133-MHz, 64-bit I/O expansion adapter using a PCI-X host interface and a SerDes interface through the chassis midplane. The first-generation I/O expansion adapter displaces one of the optional local disk drives. There is also support for a second-generation SFF I/O expansion adapter, which allows installation of both local disk drives. The LS20 has four USB 1.1 ports, one pair for the keyboard and mouse and one pair for a CD-ROM (compact disk, read-only memory), floppy disk drive, and USB. It also has two integrated Gigabit Ethernet ports, one or two optional 2.5-in. Ultra-320 SCSI 36-GB or 72-GB disk drives, integrated hardware RAID 1 support with the two-drive configuration (LSI 1020 SCSI controller), and video using the ATI Radeon 7000 with 16 MB of memory. Systems management is through a system management processor compliant with Intelligent Platform Management Interface (IPMI) 1.5 or through an SOL management interface.

The internal design of the LS20 is different from the other blade designs. The memory is not located behind the processors, as in the classical layout; instead, the DIMMs are located at the front of the blade with the processors. The Opteron processor requires the physical location of the memory to co-reside with the processors. The LS20 configuration with the four DIMMs and two processors across the front of the blade was designed to eliminate impact to the heat-sink cooling capabilities. This required that the DIMMs stand vertically. The vertical height allowance from the top of the printed circuit board to the bottom side of the top lid of the blade is approximately 1 in., and the DIMM total height had to be reduced to 0.72 in. after accounting for the socket height above the printed circuit board. Maintaining this blade in a single-wide slot required the use of a unique

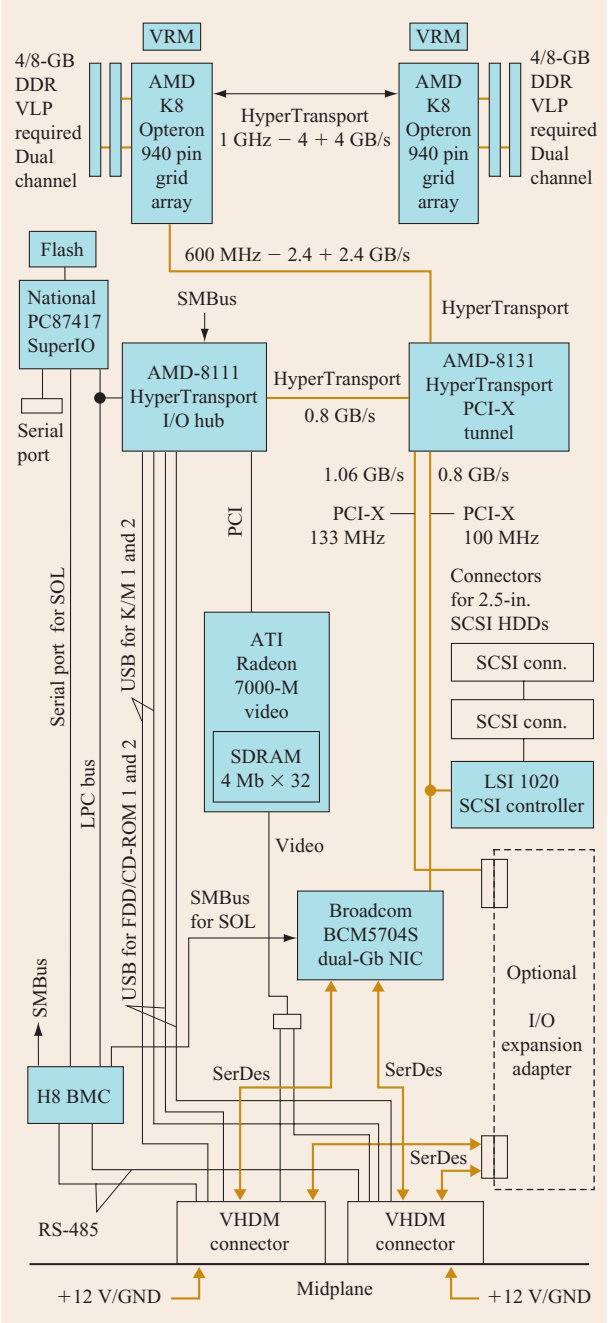


**Figure 7**

LS20 (machine type 8850) AMD Opteron processor blade.

DIMM package, called *very low profile*, with a maximum height of 0.72 in. This unique DIMM package is now in the process of being adopted by the Joint Electronic Device Engineering Council (JEDEC).

The memory is spread across the two processors or internal memory controllers of the processor to provide sufficient bandwidth for memory-intensive applications.

The DDR is PC3200-interleaved to provide 6.4 GB/s (total memory bandwidth capability of 12.8 GB/s). The interconnect between the processors and downstream I/O uses a HyperTransport bus. The interprocessor link capabilities can achieve 1 GHz or up to 4 GB/s in each direction (total bandwidth of 8 GB/s per link). However, for the downstream I/O link, the bandwidth is limited to 600 MHz or 2.4 GB/s in each direction (total bandwidth of 4.8 GB/s). Still, the HyperTransport link to the Southbridge is 200 MHz or 800 MB/s (total bandwidth of 1.6 GB/s).

The I/O subsystem includes the standard redundant midplane interconnect of a blade with Ethernet, RS-485, USB, and video. The Ethernet is provided by a BCM5704S dual Gigabit Ethernet controller, which supports WOL, virtual LAN (VLAN) tagging, teaming, and failover capabilities. The RS-485 interfaces the BMC with the BladeCenter management module. Systems management features include voltage sensing, temperature sensing, board identification, error logging, system configuration (processors and memory), SOL, and D2D fault detection. The four USB 1.1 ports from the AMD-8111 Southbridge interface with the midplane to provide KVM support. The video is the same controller used on the other blade designs discussed in this paper.

## HS20 SCSI storage expansion unit

The BladeCenter HS20 SCSI storage expansion unit, referred to as the BSE-1, increases the disk storage capability of the HS20 (first- and second-generation) and HS40 blades by providing two 3.5-in. hot-swappable Ultra-320 SCSI disk drives. This first-generation storage expansion option works with the blade to provide a local disk subsystem and cannot be used in the BladeCenter chassis as a standalone blade. The BSE-1 is listed in the system information area of the management module as a multiblade configuration along with the blade to which it is attached. The BSE-1 reports any environmental information to the management module under the blade information area.

Installation of the BSE-1 option was designed to be a tool-less operation. The blade top cover is removed, exposing a Molex 144-pin HDM connector. This connector provides the signaling path between the blade and the BSE-1 option. The BSE-1 is mounted on the HS20 blade in place of the top cover and is rotated into place and locked together, forming a double-wide blade set. In the case of the HS40, which is already a double-wide blade, the BSE-1 option forms a triple-wide blade. The blade and BSE-1 are then installed in the chassis as an integral unit. The latching handles that install the blade in the chassis and remove it also lock and install the BSE-1 in the chassis.

The front bezel of the BSE-1 option provides access to the two 3.5-in. hot-pluggable Ultra-320 SCSI disk drives. The option ships to customers with fillers installed in the disk drive bays to prevent airflow bypass when the BSE-1 is installed in the chassis without disk drives present. Once the disk-drive fillers are removed, two 3.5-in. hot-pluggable disk drives can be inserted into the BSE-1 and mated to the Molex 80-pin SCA2 connectors mounted on the BSE-1 printed circuit board. Plastic 90-degree-angled light pipes on the BSE-1 align the status LEDs mounted on the board with light pipes mounted on the disk drives and provide fault and activity status to the customer.

The BSE-1 option has two sets of PowerPlus connectors providing 12-V power as the only connection between the midplane and the BSE-1. There are no signal connections between the BSE-1 option and the midplane. All signal connections to the BSE-1 option are through the blade to which it is attached. The BSE-1 is responsible for creating the voltages it requires by converting the 12-V power supplied from the midplane. Redundant power is supplied from the midplane through an ORing circuit on the BSE-1; failure of a power module results in continued operation from the redundant half of the midplane. Regulators on the BSE-1 provide 1.8-V and 3.3-V standby, and 5 V, 3.3 V, and 2.5 V. For signal quality reasons, the blade provides 1.5 V to the BSE-1, which is used as a reference for one of the high-speed buses between the blade and the BSE-1.

The Molex 144-pin HDM connector provides an impedance-controlled high-speed path between the blade and the BSE-1, as well as a low-speed path for $I^2C$ bus communication, presence detect, and power-on sequence handshaking. A ServerWorks IMB connects the CMIC memory controller on the blade to the ServerWorks CIOB-X2 PCI-X bridge on the BSE-1 through this HDM connector. The IMB bus supports two unidirectional 16-bit links operating at 800 MHz and provides a bandwidth of 1.6 GB/s in each direction.

The ServerWorks CIOB-X2 module is a PCI-X bridge chip that provides two PCI-X buses capable of running at 133 MHz. Only one of the secondary PCI-X buses is used to connect to the LSI 1020 SCSI controller. The CIOBX2 chip also provides fault-alert functions to the CMIC chip on the blade for faults such as PERR/SERR on the secondary PCI bus segments and link training failure of the IMB bus link. (PERR is *parity error*, used for reporting data parity errors during all PCI bus transactions except for "special cycle" transactions. SERR is *system error*, used for reporting address parity errors during all PCI bus transactions or data parity errors during "special cycle" transactions or other fatal system errors.)

The LSI 1020 SCSI controller generates a SCSI bus that connects to the two hot-pluggable disk drive bays

and a QLogic Gem318 enclosure monitor. The SCSI bus is capable of Ultra-320 speeds, but will downshift to the slowest common speed of any disk drives installed. The LSI 1020 chip has embedded hardware RAID 1E function, which allows the disk drives to operate in a mirrored arrangement; disk drive mirroring is transparent to the operating system. If a hard drive fails, mirroring allows the data to continue to be retrieved from the remaining disk drive until a replacement is installed. The LSI 1020 then automatically rebuilds the replacement disk drive upon insertion. Fault LEDs on the blade control panel signal fault conditions and identify a failing disk drive in a mirrored configuration. The mirrored disk-drive pair appears as just one logical disk drive to the operating system, and disk-drive capacity is reported as the total capacity of just one disk drive.

An Analog Devices ADM1024 supervisory chip monitors temperatures and voltages inside the BSE-1 and reports these to the blade through the $I^2C$ bus. The management module Web interface then reports these environmental readings and can send alerts for over- and under-voltage values and for temperatures falling outside the optimal operating range.

The BSE-1 printed circuit board is constructed of eight layers (six signals, two power) of FR4 material and provides controlled impedance layers for the IMB and PCI-X buses and sufficient power distribution for the Ultra-320 SCSI disk drives, CIOBX2 module, and LSI SCSI controller. Controlled impedances of 50 Ω and 60 Ω single-ended and 120 Ω differential-ended are attained in the same stackup by having dedicated layers for these impedances.

## SCSI storage expansion unit 2

The BladeCenter SCSI storage expansion unit 2, or BSE-2, is the second-generation storage expansion option for blades. BSE-2 takes advantage of higher-bandwidth interconnect buses on the third-generation HS20 blade (machine type 8843). It shares many mechanical characteristics of the BSE-1. Like the BSE-1, the BSE-2 option provides for installation of two 3.5-in. hot-pluggable Ultra-320 SCSI disk drives. Unlike the BSE-1, the BSE-2 allows up to two I/O expansion adapters of either form factor to be installed, for a total of eight I/O fabric connections to the midplane for the blade and BSE-2 combination.

Installation of the BSE-2 option was designed to be a tool-less operation and is similar to the installation of the first-generation BSE-1, although the BSE-1 and BSE-2 are not interchangeable. The top cover of the 8843 blade is removed, exposing an FCI 200-pin GIG-Array connector. This connector provides the signaling path between the blade and the BSE-2 option. The BSE-2 is mounted on the 8843 blade in place of the top cover and pivots into place to form a double-wide blade. The top cover of the BSE-2 has an access hatch for the I/O expansion adapter slots. As with the BSE-1, the blade and BSE-2 are installed in the chassis as an integral unit.

The BSE-2 option is designed to enhance the I/O capability of the new 8843 blade and to provide increased disk storage capability and enhanced RAID configurations. The BSE-2 design extends the SCSI bus from the blade and connects the two 3.5-in. hot-swappable disk drives to the same SCSI bus as the local SCSI disk drives on the blade. This layout allows the four disk drives to be configured in different RAID modes depending on customer requirements. The LSI 1020 SCSI controller on the 8843 blade supports RAID 1 and Raid 1E with embedded hardware mirroring and hot-spare disk-drive capability. When the BSE-2 is attached, the SCSI bus active terminators on the blade are disabled, and the bus terminates at the last disk drive on the extended BSE-2 bus. The two local SCSI disk drives on the 8843 blade are identified as 0 and 1. The two 3.5-in. hot-swappable disk drives in the BSE-2 become disk drives 2 and 3 when present. All SCSI control is provided by the LSI 1020 on the blade.

The 8843 uses PCI-Express buses as the high-speed interconnect between components in the chipset. One eight-lane PCI-Express bus is routed through the GIG-Array connector and connects to an Intel PXH PCI-X bridge on the BSE-2. This bridge provides two secondary PCI-X 1.0 buses operating at 133 MHz that connect to two I/O expansion adapter slots. Each PCI-Express lane operates at 2.5 Gb/s each direction and is 8/10-bit encoded. The eight-lane PCI-Express bus provides a total of 2 GB/s of bandwidth in each direction. This is more than adequate considering that each PCI-X bus operating at 133 MHz requires only 1 GB/s.

The two I/O expansion adapter slots provide connection to either the first-generation standard-form-factor I/O expansion adapters (Ethernet, Fibre Channel, Myrinet, or InfiniBand) or the second-generation SFF I/O expansion adapters (Ethernet or Fibre Channel). The I/O expansion adapter in slot 1 must be compatible with the switches installed in switch module bays 1 and 2 of the BladeCenter chassis, and any I/O expansion adapter installed in I/O expansion adapter slot 2 must be compatible with the switches installed in switch module bays 3 and 4. The management module uses the $I^2C$ bus connected to the BSE-2 for presence detection of I/O expansion adapters and will not power on the 8843 blade if incompatible I/O expansion adapters are detected. With two I/O expansion adapters installed in the BSE-2 option, the available I/O fabric connections now total eight: four on the base blade and four in the BSE-2. This provides for a maximum I/O bandwidth of 3.5 GB/s.

The BSE-2 provides connection to the BladeCenter chassis midplane for all I/O fabrics through two Molex VHDM connectors. The BSE-2 is provided with 12-V power from the midplane through two Power Plus 3 connectors, and the power is distributed by switching and linear regulators to the components on the BSE-2 printed circuit board. The BSE-2 is responsible for creating any voltage it requires and does not draw power from the blade to which it is connected. The BladeCenter management module displays power requirements for each blade and lists the BSE-2 as a separate blade.

An ADM1024 supervisory chip monitors temperatures and voltages inside the BSE-2 and reports these to the blade via the $I^2C$ bus. The management module Web interface then reports these environmental readings and can send alerts for over- and under-voltage values and for temperatures falling outside the optimal operating range.

The BSE-2 printed circuit board is constructed of eight layers (six signals, two power) of FR4 material and provides controlled-impedance layers for the PCI-Express and PCI-X buses and sufficient power distribution for the Ultra-320 SCSI disk drives, Intel PXH chip, and LSI SCSI controller.

Signal connections between the blade and the BSE-2 go through the GIG-Array connector. These signals include both the differential-ended PCI-Express and Ultra-320 buses and the single-ended $I^2C$ buses, presence-detect bits, and power-sequence handshake signals. The GIG-Array connector was chosen because of its high pin density and its ability to provide impedance control for both differential- and single-ended signals. This connector also had to support a zipper insertion to allow the BSE-2 to pivot into position onto the base blade. The GIG-Array is a 496-pin connector that can be configured for up to 200 differential ground-shielded signals. Low-speed single-ended signals can be routed onto the high-speed signal pins or onto the shield pins for higher-pin-count applications. The BSE-2 uses both configurations by segregating the pin field into a high-speed section with interstitial ground-shield pins and a low-speed section in which the interstitial ground pins are used as additional signal pins.

## PCI I/O expansion unit

The PCI I/O expansion unit (PEU) is an option available for certain blades to provide a mechanism to support standard PCI or PCI-X adapters in a chassis. The PEU utilizes the host expansion interface connector similarly to the BSE, along with an IMB-to-PCI-X bridge to develop two independent PCI or PCI-X buses that can accommodate 32-bit or 64-bit adapters at frequencies of 33 MHz, 66 MHz, 100 MHz, or 133 MHz. Bus size, protocol, and frequency (with the exception of 100 MHz) are automatically detected by the PEU on the basis of the adapter requirements. To aid in the transition from legacy to next-generation networks, the PEU is ideal for applications requiring connections to existing communications infrastructure, such as Signaling System 7 (SS7), asynchronous transfer mode (ATM), or T1/E1/J1 interfaces. Both PEU slots can handle full-sized, full-power (25 W) PCI or PCI-X adapters, thus enabling solutions that use intelligent adapters for Internet Protocol Security (IPSec) acceleration, voice recognition, and other offload engine technologies. IBM has partnered with various industry-leading suppliers to certify their adapters as part of the IBM ServerProven* program for both enterprise and telecommunications applications.

## Gigabit expansion adapter

The Gigabit expansion adapter provides the blade with two additional Gigabit Ethernet SerDes interfaces which are routed to compatible switches in the BladeCenter chassis. This I/O expansion adapter is electrically connected to the blade through two connectors. One is a 64-bit PCI-X 1.0 electrical interface using a Molex 200-pin board-to-board stack connector. The I/O expansion adapter also obtains power (12 V, 5 V, and 3 V, and 3-V standby), $I^2C$ bus, and other miscellaneous signals from this 200-pin electrical interface. The other connector is a Molex high-speed mezzanine connector used to connect the SerDes interface to the blade through four differential high-speed signal pairs.

The I/O expansion adapter supports only 3.3-V PCI signaling environments and conforms to the wiring and timing specifications of PCI-X 1.0. This card is designed to run up to 133 MHz.

The design of the Gigabit Ethernet I/O expansion adapter is based on the BCM5704S dual-port Gigabit Ethernet controller. The BCM5704S combines in a single device two IEEE Standard 802.3-compliant MACs with two integrated SerDes interfaces, one shared PCI/PCI-X bus interface, and on-chip buffer memory. Support for the following is featured in the MAC: VLAN tagging, Layer 2 priority encoding, link aggregation, and 1,000-Mb/s full-duplex flow control. Failover, WOL, and Preboot eXecution Environment (PXE) booting are supported on this controller (PXE can be enabled or disabled). The controller also provides large on-chip buffer memory for standalone operation.

This card draws a maximum of approximately 5 W of power, primarily 3.3 V standby, which is provided to power the I/O expansion adapter EEPROM and enough of the BCM5704S circuitry to operate while the blade is powered off. VPD is extracted from the I/O expansion adapter EEPROM and communicated to the management module. The management module then verifies compatibility between the I/O expansion adapter and the switch modules before granting power-up

permission to the blade. Once the verification is successful, the I/O expansion adapter receives 12 V, 5 V, and 3.3 V power from the blade.

The Gigabit Ethernet I/O expansion adapter was designed using two different printed circuit board form factors. The first-generation I/O expansion adapter was designed with the standard I/O expansion adapter form factor, a slightly larger card that limits the blade to one local disk drive. The second generation of this design is a smaller form factor that enables coexistence of the I/O expansion adapter with both local disk hard drives installed on the blade.

## Fibre Channel expansion adapter

The Fibre Channel expansion adapter provides the blade with two optional Fibre Channel interfaces routed to compatible switches in the chassis. Like the Gigabit Ethernet expansion adapter, this I/O expansion adapter is electrically connected to the blade through two connectors. One is a 64-bit PCI-X 1.0 electrical interface using a Molex 200-pin board-to-board stack connector. The I/O expansion adapter also obtains power (12 V, 5 V, and 3 V, and 3-V standby), $I^2C$ bus, and other miscellaneous signals from this 200-pin electrical interface. The other connector is a Molex high-speed mezzanine connector used to connect the SerDes interface to the blade through four differential high-speed signal pairs.

The expansion adapter supports only 3.3-V PCI signaling environments and conforms to the wiring and timing specifications of PCI-X 1.0. This card is designed to run up to 133 MHz.

The design of the Fibre Channel expansion adapter is based on the QLogic ISP2312 dual Fibre Channel controller. The ISP2312 combines two independent 2-GB Fibre Channel ports with two integrated SerDes interfaces, one shared PCI/PCI-X bus interface, and flash and NVRAM interfaces. Each channel consists of an internal reduced instruction set computing (RISC) processor (with an external memory interface), a receive DMA sequencer, frame buffer, internal transceivers, and DMA channels. Although the controller supports two synchronous burst static random access memory chips (SRAMs), this expansion adapter supports only one device per channel.

The expansion adapter draws a maximum of 6 W of power. To power the serial EEPROM, 3.3-V standby is provided. VPD is extracted from the serial EEPROM and communicated to the management module; the management module then verifies compatibility between the expansion adapter and the switch modules before granting power-up permission to the blade. Once the verification is successful, the I/O expansion adapter receives 12 V, 5 V, and 3.3 V power from the blade. This expansion adapter is primarily powered by 5 V, which is supplied to an onboard dual-switching regulator that creates 2.5 V at 1.75 A and 3.3 V at 1.75 A.

The Fibre Channel expansion adapter was designed using two different card form factors. The first-generation I/O expansion adapter was designed with the standard I/O expansion adapter form factor, a slightly larger card that limits the blade to one local disk drive. The second generation of this design is a smaller form factor that enables the coexistence of the I/O expansion adapter with both local disk hard drives installed on the blade.

## Summary

In this paper we describe the processor blade, expansion blade, and I/O expansion adapter architecture for the IBM eServer BladeCenter system. The BladeCenter architecture is not limited to a single manufacturer's processor, a single processor architecture, or a single operating system. Processor blades that support IBM POWER, Intel Xeon, and Opteron processors are described in detail. In most cases, blades fit in a single-wide slot; however, the architecture allows the development of multi-slot blades in order to accommodate designs requiring more features, such as additional processors or memory.

One of the key attributes of the blade design is the redundant design of the midplane signal and power interface. The redundant signaling is described, along with the high-speed connector requirements and selection. In addition to redundancy, expandability is another requirement of the BladeCenter architecture. The blade I/O bandwidth can be increased by two additional Gigabit Ethernet ports or by an additional fabric protocol, such as Fibre Channel or InfiniBand. This is done through the addition of optional I/O expansion adapters. The blade storage capacity can also be increased by attaching a storage expansion blade. Similarly, standard-size PCI-X adapters can be supported with the addition of a PCI expansion unit.



## References

1. 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).
2. 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).

3. T. Brey, B. E. Bigelow, J. E. Bolan, H. Cheselka, Z. Dayar, J. M. Franke, D. E. Johnson, R. N. Kantesaria, E. J. Klodnicki, S. Kochar, S. M. Lardinois, C. M. Morrell, M. S. Rollins, R. R. Wolford, and D. R. Woodham, "BladeCenter Chassis Management," *IBM J. Res. & Dev.* **49**, No. 6, 941–961 (2005, this issue).
4. 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).
5. 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).
6. VHDM Backplane Connector System; see *http://www.molex.com/cmc_upload/0/000/0-8/388/tab01vhdm.pdf*.
7. Molex Connector Systems; see *http://www.molex.com/cgi-bin/bv/molex/index_login.jsp*.
8. Universal Serial Bus Specification, Revision 2.0, pp. 15–24 and 119–170; see *http://www.usb.org/developers/docs/*.
9. S. L. Vanderlinden, B. O. Anthony, G. D. Batalden, B. K. Gorti, J. Lloyd, J. Macon, Jr., G. Pruett, and B. A. Smith, "BladeCenter T System for the Telecommunications Industry," *IBM J. Res. & Dev.* **49**, No. 6, 873–886 (2005, this issue).
10. PCI-X 1.0 Bus Specification, pp. 15–150; see *http://www.pcisig.com/home*.
11. ServerWorks IMB and Chipsets; see *http://www.broadcom.com/products/Enterprise-Small+Office/SystemI-O+Chips*.
12. FCI GIG-Array Connectors; see *http://www.fciconnect.com/highspeed/highspeed_04.asp*.
13. W. G. Holland, P. L. 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).

*Received December 16, 2004; accepted for publication February 22, 2005; Internet publication October 7, 2005*

**James E. Hughes** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (jehughes@us.ibm.com)*. Mr. Hughes is a Senior Technical Staff Member working in BladeCenter Architecture and Hardware Development. He received a B.S. degree in electrical engineering from Pennsylvania State University in 1980 and joined IBM Endicott that same year. Between 1980 and 1995, he worked on several application-specific integrated circuit (ASIC), board, and system-level designs in the development of IBM System/370* and digital video products. In 1995 he joined the Personal Computer (PC) Server team in Research Triangle Park, where he was the lead engineer for several PC server and xSeries* server products. Mr. Hughes has been part of the BladeCenter architecture team since 2000 and is responsible for system and electrical architecture and design.

**Michael L. Scollard** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (scollard@us.ibm.com)*. Mr. Scollard received a B.S. degree in electrical engineering from Auburn University in 1988, joining IBM that same year to work in the Personal Systems Division failure analysis department. In 1992 he moved to the ThinkPad* Development Group in Research Triangle Park to design graphics systems for the ThinkPad 350C and 701C. He moved to the Workstation Development team in 1995 and worked on system board design for various Intel processor-based workstations and servers. He joined the BladeCenter development team in 2001 and currently works on future Intel-based blade servers.

**Rudolf Land** *IBM Deutschland Entwicklung GmbH, Schoenaicherstrasse 220, 71032 Boeblingen, Germany (land@de.ibm.com)*. Mr. Land is a Senior Technical Staff Member in hardware development responsible for platform architecture and system design of IBM eServers. He studied mathematics at Ohio State University and graduated from the University of Cologne in 1979 with a thesis on computational algebra. In 1981 he joined IBM in the server hardware development organization, where he held various positions in S/390* microcode development, server system design, and system integration. Since 2002 Mr. Land has been working on concepts and system design to integrate PowerPC technology into the BladeCenter environment.

**James Parsonese** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (jparsone@us.ibm.com)*. Mr. Parsonese is a Senior Engineer working in the xSeries BladeCenter Design Group. He graduated from Marquette University in 1993 with a B.S. degree in electrical engineering and received an M.S. degree in computer engineering from George Washington University in 1996. Since he joined IBM in 1998, his work has included ASIC development efforts on a memory controller for memory compression, including development of the xSeries server which incorporated the memory compression. Mr. Paronese joined the blade development effort on JS20 (PowerPC) and LS20 (AMD Opteron) blades in 2001.

**Christopher C. West** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (ccwest@us.ibm.com)*. Mr. West received a B.S. degree in electrical engineering from the University of Florida in 1998. He joined IBM in December 2000 as a development engineer in the hardware development department for xSeries 220 servers.

This group subsequently transitioned into the BladeCenter Development Group, which was responsible for the development of the original eServer BladeCenter system and xSeries HS20 blade. Mr. West is a Staff Engineer currently working on further BladeCenter development; he is one of the key hardware designers of xSeries HS20 blade servers.

**Victor A. Stankevich** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (vics@us.ibm.com)*. Mr. Stankevich is a Senior Engineer and certified Project Manager in BladeCenter product development. He received a B.S.E.E. degree from the University of Tennessee and an M.S.C.E. degree from Florida Atlantic University, joining IBM in 1983. He was the project manager for the original BladeCenter family of products after managing several xSeries server projects. He was previously the Development Engineering Manager for the Project Test Leads and System Integrators in the PC Server organization. Mr. Stankevich has received several IBM awards and published several papers.

**Challis L. Purrington** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (challisp@us.ibm.com)*. Mr. Purrington received a B.S. degree in electrical engineering from Norfolk State University in 1974, joining IBM Endicott that same year. He is currently one of the lead engineers on the blade development team and has had key roles in all four Intel-based blade designs. He was team leader on various projects and holds patents in the areas of pico-processor buffer management and token ring fault recovery. Mr. Purrington has numerous publications and technical reports.

**Danny Q. Hoang** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (dqhoang@us.ibm.com)*. Mr. Hoang is an Advisory Engineer working in the Blade Server Hardware Development Group. He received a B.S. degree in electrical engineering from the University of Washington in 1992 and joined IBM at Boca Raton, Florida, that same year. In 1997 he joined the PC server team in Research Triangle Park, where he was the lead ASIC simulation engineer for several PC server products. Mr. Hoang joined the blade server team in 2001; he is a systems engineer responsible for four-way blade server products for xSeries.

**Gary R. Shippy** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (gshippy@us.ibm.com)*. Mr. Shippy is a member of the BladeCenter development team. He received a B.S. degree in electrical engineering (1966) and an M.S. degree in computer science (1974), both from Ohio State University. When he joined IBM, he initially worked on diagnostic packages for the communications multiplexers. He has since worked in various hardware and software design and architecture areas, including data telecommunications, private branch exchange (PBX), multimedia, and systems management. Mr. Shippy joined the xSeries team in 2001 and is currently focusing on BladeCenter management functions.

**Mitchell L. Loeb** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (loeb@us.ibm.com)*. Dr. Loeb received a B.A. degree in business administration from Franklin and Marshall College in 1971, an M.S. degree in electrical engineering from Duke University in 1976, and a Ph.D. degree in electrical engineering from North Carolina State University in 1985. He joined IBM in 1980, working on the holographic scanner. He is currently a Senior Engineer responsible for the Open Specifications for the BladeCenter products. Dr. Loeb has taught courses in computer science and electrical engineering for 18 years as an Adjunct Associate Professor at both Duke University and North Carolina State University; he was a Visiting Assistant Professor in electrical engineering at Duke in 1987. Mr. Loeb is also a licensed professional engineer in North Carolina.

**Mark W. Williams** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (mw@us.ibm.com)*. Mr. Williams received a B.S degree in electrical engineering from Lafayette College in 1978 and an M.S. degree in computer engineering from Florida Atlantic University in 1994. Since joining IBM in 1982, he has worked on many desktop and server systems in product engineering and product development.

**Bruce A. Smith** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (bruces@us.ibm.com)*. Mr. Smith received a B.S.E.E. degree from the University of Pittsburgh in 1976. Since joining IBM in 1977, he has worked primarily on Intel architecture systems, beginning with DataMaster, a PC predecessor product. He was the system leader on seven microchannel desktop systems and is currently working on BladeCenter development within the xSeries organization. Mr. Smith is the inventor or coinventor of nine patents.

**Dhruv M. Desai** *IBM Systems and Technology Group, 3039 Cornwallis Road, Research Triangle Park, North Carolina 27709 (ddesai@us.ibm.com)*. Mr. Desai is a Distinguished Engineer in IBM eServer xSeries development, working as a chief architect and strategist on BladeCenter systems. He holds an M.S. degree in computer engineering from Nova Southwestern University and an M.S.E.E. degree from Texas A and M University. Mr. Desai has 24 years of experience in systems design and architecture of Microsoft Windows**/Intel-based systems. He holds 33 patents.