Distrib Parallel Databases
DOI 10.1007/s10619-007-7010-3

General-purpose blade infrastructure for configurable system architectures

Kevin Leigh · Parthasarathy Ranganathan · Jaspal Subhlok

Abstract Bladed servers are increasingly being adopted in high-density enterprise datacenters by virtue of the improved benefits they offer in form factor density, modularity, and more robust management for control and maintenance with respect to rack-optimized servers. In the future, such servers are likely to form the key foundational blocks for a variety of system architectures in data centers. However, designing a commodity blade system environment that can serve as a general-purpose infrastructure platform for a wide variety of future system architectures poses several challenges. This paper discusses these challenges and presents specific system architecture solutions, along with application examples to illustrate the general-purpose nature of the infrastructure for parallel and distributed applications.

Keywords Blade servers · Enclosure · Power · Networking · Data centers

1 Introduction

Several recent trends are likely to impact the design of future enterprise servers. These include the move towards large consolidated data centers, commoditization of high-performance hardware, increasing adoption of virtualization, and greater convergence

Recommended by: Monem Beitelmal.

K. Leigh (✉)
Hewlett-Packard (HP), 20555 SH249, Houston, TX 77070, USA
e-mail: kevin.leigh@hp.com

P. Ranganathan
Hewlett-Packard (HP), 1501 Page Mill, MS 1177, Palo Alto, CA 94304, USA
e-mail: partha.ranganathan@hp.com

J. Subhlok
University of Houston (UH), Houston, TX, USA
e-mail: jaspal@cs.uh.edu

between different networking protocols. At the same time, end-user system requirements are increasingly focusing beyond performance to also include higher levels of manageability, availability, scalability, power, etc. The system-on-a-card approach represented by blade servers is emerging to be an interesting architectural platform to address these trends.

Consider for example, the focus on better manageability and lower costs. Although datacenter capital expenses (CapEx) to procure hardware/software are non-trivial, over 80% of the total datacenter costs are in the operational expenses (OpEx). Blade systems lower server costs, dramatically reduce labor costs on cable management, and eliminate expensive transceivers and cables between the server blades and the edge switches (due to the use of backplane traces). They also have lower electricity costs, provide a lower labor cost environment with ease and speed of service/upgrade, and more efficient interfaces to datacenter configuration and automation tools.

From a consolidation point of view, server blades epitomize how dense high-performance server systems' form factors can be implemented. The power and associated thermal densities are directly proportional to the performance density and inversely proportional to volume density. Typical datacenters can enjoy the benefits of small datacenter footprint requirements of dense servers, but they can no longer sustain the required growth of power delivery and heat extraction. The good news is that blades are more efficient in power consumption and cooling, compared to stand-alone rack-optimized dense servers, because the pooled power supplies and fans within a blade enclosure can be designed and managed more efficiently. In addition, fluctuating utilization profiles of server blades for many datacenter applications can be exploited to manage the total power consumption of an enclosure to be within an affordable threshold for a deployment.

Similarly, consider availability and flexibility. Service availability is the bottom-line for the users of the datacenter resources, and hardware resources need to be agile enough to support fluctuating service demands. A key requirement for most businesses is a top-to-bottom well orchestrated software and hardware solution set that will help them significantly reduce the total cost of ownership, while addressing their ever changing business challenges (including fluctuating demands, merger/acquisition, etc.). Blades provide an environment where applications can be easily migrated across blades, for fail-over recovery, load balancing, or even plant disaster recovery, under the control of datacenter automation tools.

In addition, bladed environments offer unprecedented modularity in building different higher-level system architectures. For example, the HP BladeSystem c-Class enclosure includes the following elements: server blades, storage blades, interconnect modules (switches and pass-through modules), a signal midplane that connects blades to the interconnect modules, a shared power backplane, shared power supplies, shared fans, and enclosure management controllers. Most of these elements are hot-pluggable and all of these elements are field-replaceable.

The modularity is further strengthened by recent trends in network protocols. From a bandwidth point of view, the local IO interface PCI has evolved from PCI 32-bit/33 MHz at 1 Gbps to PCIe $\times 16$ (gen1) at 40 Gbps within one and half decades. Ethernet also has evolved from 10 Mbps to 1 Gbps, and will soon be at 10 Gbps. InfiniBand has been evolving for several years, and bandwidth for IB $4\times$ has gone from SDR 10 Gbps to DDR 20 Gbps, and soon to QDR 40 Gbps. The bandwidth

of these fabrics have converged at 10 Gbps. Additionally, there is a lot of similarity in high-speed backplane signaling rate and physical layer across different protocols including Backplane Ethernet, Fiber Channel (FC), InfiniBand (IB) and PCI Express (PCIe).

From a historical perspective for modern mainstream data centers, the first generation blades were dense blades [1, 2] that were low power and correspondingly limited in functionality. These were followed by higher-performance blades such as HP BladeSystem p-Class [3], introduced in the early 2000, and later followed by Egenera BladeFrame [4], IBM BladeCenter [5] and those from a few other system OEMs. Given the need to interoperate with then-existing IT practices, most of the server blades were designed as repackaged rack-optimized servers simply interconnecting traditional server blades and network switches. Egenera made an attempt towards interconnect virtualization but their method lacked in cost efficiency, space efficiency, node scalability and interconnect flexibility. However, the next generation blade infrastructure [6] and future blade designs should and are likely to break free from these constraints.

As an extension of these trends, we argue that, in the future, blade servers are likely to be used as key foundational blocks for future enterprise systems, and consequently, future blade environments need to be designed as a *general-purpose infrastructure platform* on which other architectures can be layered. However, this approach poses several interesting challenges. This paper describes these challenges and solutions.

The rest of the paper is organized as follows. Section 2 provides a broad overview of the issues with architecting and engineering a general-purpose blade infrastructure platform along the various dimensions of cost, performance, power, availability, manageability, and flexibility. Section 3 then discuss three key solutions—better power and cooling, improved networking abstraction, and better management and automation—that enable it to provide a general-purpose platform for different end-user scenarios. Sections 4 and 5 illustrate how the general-purpose blade infrastructure designed can address traditional scale-out applications as well as distributed parallel applications. Section 6 concludes the paper.

2 Designing blades to be a general-purpose infrastructure

Modern day general-purpose computers are constructed with commodity hardware components and interconnect protocols based on open-standards, and can be configured with off-the-shelf software for special or general-purpose use. We define a general-purpose infrastructure within a blade enclosure to have similar attributes as a general-purpose computer. The differences are that a general-purpose infrastructure can accommodate different functional modules (e.g., general-purpose server blades, storage blades, network protocol switches and IO fabrics), and it can be configured to function as an ensemble of interconnected systems of varying capabilities or one system.

Examples of ensembles of interconnected systems (or *scale-out* systems) in an enclosure are a group of web servers, a group of database application-layer servers, and a cluster of HPC (High-Performance Computing) nodes. In those ensembles, each

Higher density ➔ better cost amortization
Higher density ➔ lower volume space ➔ small modules
Small modules ➔ lower performance blades/switches or more expensive components
Higher density ➔ more complex backplane
Higher performance ➔ more complex backplane
Higher complexity ➔ higher cost
Higher performance ➔ higher blade power consumption
Higher density ➔ higher enclosure power consumption
Higher power consumption ➔ more cooling ➔ higher power consumption
Higher power consumption ➔ Higher thermal environment ➔ lower reliability
More complex design ➔ lower reliability
Lower reliability ➔ lower availability
Lower reliability ➔ more redundancy needed for higher availability ➔ higher cost

Fig. 1 Blade enclosure design trade-off parameters

blade is a server system and they are interconnected with various protocol switches such as Ethernet, Fiber Channel or InfiniBand.

Another example of a system in an enclosure is a backend database server consisting of multiple processor/memory blades with a coherent interconnect that ties them together to make up a *scale-up* system. It is well understood by hardware system designers that coherent links interconnecting the processor/memory subsystems are significantly more complex than communication network interconnects.

These scale-out and scale-up systems are examples of the spectrum of flexibility that a general-purpose infrastructure has to address. We will describe the challenges and rationale behind the blade enclosure we designed as a general-purpose infrastructure, and will illustrate application examples to address a variety of system architectures.

In this section, we describe the key dimensions in designing a blade enclosure to be an optimal general-purpose infrastructure. Specifically, we will discuss optimizing the six key parameters—cost, performance, power, availability, manageability and flexibility. One major challenge is that each of these parameters cannot be optimized independently, as they are inter-related as illustrated in Fig. 1. Another significant challenge is that, the optimized solutions should still be valid to support technologies during the infrastructure life-span of about five to ten years after its first deployment in the market, since longevity is an implied requirement for a general-purpose infrastructure.

2.1 Cost

We will first address the costs for blades, switches and enclosure infrastructure. Balancing an optimal point of maximum enclosure density and simplest enclosure design will minimize per-blade total cost which is a combination of a blade cost plus the amortized cost of the blade infrastructure. The enclosure density means the maximum number of blades installable in a blade enclosure, and it depends on the form factors of the blades and the enclosure.

In practice, popular commodity server configurations require a set of components (such as processors, memory, core IO devices, disk drives and network interface devices) to be contained within a blade form factor. The main components are

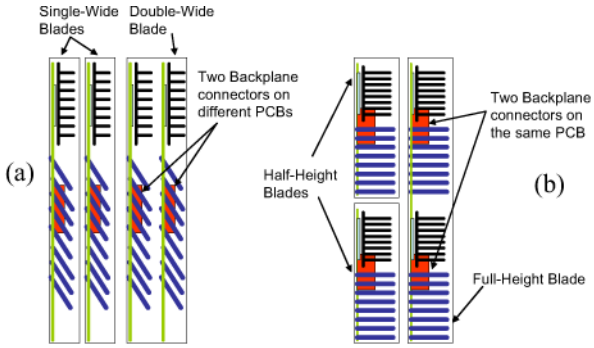

Fig. 2 **a** Side-by-side vs. **b** over-under blade form factor scaling

processors with associated memory modules (DIMMs) and IO devices. Historically, 2-processor commodity servers with varying memory and IO choices are the most dominant deployment in the enterprise data centers. 4-processor servers are the next popular configuration for the mainstream high-end applications, such as database. Here, we are using "processor" to refer to processor sockets. A 4-processor blade will need twice the number of processor sockets, DIMMs and power budget than a 2-processor blade. Therefore, there are at least two blade form factors that need to be supported—one optimized for a 2-socket blade and the other for a 4-socket blade configuration.

Simplifying the designs is clearly important for lowering implementation costs. As we discussed, blades need to be scalable in form factor to be implementable for different configurations of processor, memory and I/O. A general approach is to have one or more connectors for the smallest form factor blade, and have twice of these connectors for a two times larger blade. Blade form factor can be scaled by using two side-by-side blades for a larger blade as shown in Fig. 2(a), or over-under as shown in Fig. 2(b).

As the blades are scaled in the direction of the PCB plane, the system's main PCB (also commonly known as motherboard) is typically a single plane for a larger blade in Fig. 2(b). Figure 2(b) also shows the benefit of blade form factor to be thick, to accommodate tall heat sinks for the processors and tall DIMMs. If the side-by-side blade form factor (as shown in Fig. 2(a)) is too thin, then it might limit a blade design to low-profile DIMMs instead of standard height DIMMs, which will limit cost, capacity, performance choices, or they might require the DIMM connectors to be angular which will require more real estate (fewer DIMMs) and create signal integrity challenges. We prefer the over-under form factor scaling of half-height blades and full-height blades as shown in Fig. 2(b). We designed the volume space of the single-wide half-height blade to accommodate the most popular 2-socket systems, and the single-wide full-height blade to accommodate a 4-socket system with a fair amount of memory (e.g., 16 DIMMs), plus disk drives and IO adapter cards. We were aware that scaling the full-height blades to be double-wide full-height will require different PCB planes as in Fig. 2(a), but we chose to be cost efficient for the most popular 2-socket and 4-socket systems.

It is important to note that the cost of the DIMMs installed in a server can overwhelm the cost of the original system. Typically, the per-byte price of top capacity DIMMs is much higher than their lower capacity counterparts. For example, today's

prices of server-class DIMMs are linear with respect to density for up to 2 GB, start going up above the linear curve for 4 GB, and goes exponentially higher for 8 GB and 16 GB. This DIMM cost curve with respect to the top capacity bins looks the same over time as the costs on the DRAMs get lower and the capacity per DIMM doubles every 12 to 18 months.

For each memory controller design, the numbers of DIMM slots for a memory channel are limited. However in blades, volume space and power budget limitations within a blade may impose bigger challenges before the electrical capacitance limit is reached. Therefore in blades from real-estate and cost efficiency perspectives, vertical-mount DIMMs as shown in Fig. 2(b) are preferred to angular-mount DIMMs as shown in Fig. 2(a). In general, more DIMM slots in a blade provide better memory choices for users in terms of capacity vs. cost.

To control the cost of the backplane, its construction needs to be simple. In the following paragraphs, we discuss the cost impact of the backplane as well as its performance and availability attributes.

2.2 Performance

In the previous section, we discussed optimization of blade form factor to be scalable, to accommodate different performance blades, such as a half-height blade supporting two processors while a scaled-up higher performance full-height blade supporting four processors and more DIMMs. In this section, we discuss performance optimization of blades, switches and backplane.

Before we discuss the backplane connectivity for blades and switches, it is important to understand the physical layer of the fabrics that are to be supported. The popular fabrics for blades connectivity described earlier are backplane Ethernet, FC, IB 4× and PCIe ×4. There are also three backplane Ethernet standards emerging under IEEE 802.3ap workgroup [7], which are 1000-Base-KX, 10G-Base-KX4 and 10G-Base-KR. Table 1 lists the number of wires or traces required for these fabrics, and their corresponding bandwidths. The "Aggregate BW" column shows the "labeled bandwidth" for all the lanes for simplicity, rather than the actual aggregated bandwidth.

Table 1 Physical layer signal traces and bandwidths of fabric protocols

Interconnect	Lanes	# Wires	BW Per Lane	Aggregate BW
GbE (1000-Base-KX)	1×	4	1.2 Gbps	1 Gbps
l0GbE (10G-Base-KX4)	4×	16	3.125 Gbps	10 Gbps
l0GbE (l0G-Base-KR)	1×	4	l0 Gbps	10 Gbps
FC (1, 2, 4, 8 Gb)	1×	4	1, 2, 4, 8 Gbps	1, 2, 4, 8 Gbps
SAS	1×	4	3 Gbps	3 Gbps
IB	1×–4×	4–16	2.5 Gbps	2.5–l0 Gbps
IB DDR	1×–4×	4–16	5 Gbps	5–20 Gbps
IB QDR	1×–4×	4–16	l0 Gbps	10–40 Gbps
PCI Express	1×–4×	4–16	2.5 Gbps	2.5–l0 Gbps
PCI Express (gen2)	1×–4×	4–16	5 Gbps	5–20 Gbps

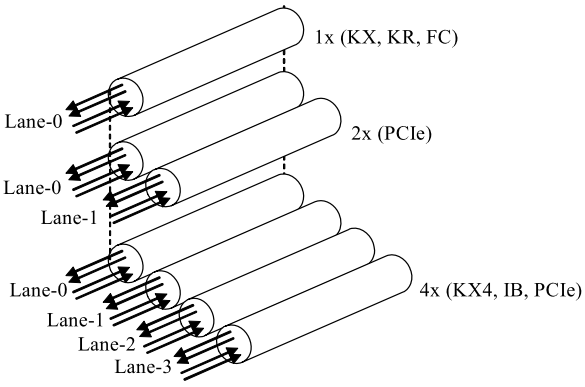

Fig. 3 Physical layer similarities for different fabric protocols

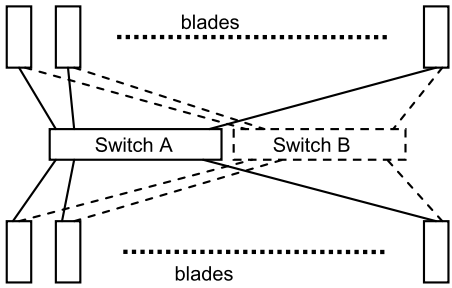

Fig. 4 Dual-star topology to interconnect blades

Figure 3 illustrates how these popular fabrics' physical lanes can be "overlaid" on a set of traces.

A 4-trace signal group (also referred to as a lane or $\times 1$ or $1\times$) consists of a differential transmit and a differential receive signal pair. KX, KR and FC each require $1\times$. Additional traces are needed for wider $4\times$ lane interfaces such as KX4, IB and PCIe. This signal lane reuse is achieved by arranging the interconnect module bays' positions. If two smaller (single-wide) interconnect bays are positioned side-by-side then they can be used together as a larger (double-wide) interconnect bay. This interconnect bay layout in conjunction with the backplane traces overlaying enables an interconnect module to support traditional network switch modules with different lane widths, as well as different fabric modules, as depicted in Fig. 3. Consequently, a set of backplane traces support network-semantic traffic (over Ethernet, FC, IB) or memory-semantic traffic (over PCIe) depending on the modules installed in the interconnect bays.

A single-wide interconnect module can connect to all the blades, and will provide a connectivity with a "star" topology. Therefore, there will be a dual-star topology with two single-wide interconnect switch modules (e.g., Switch-A and -B in Fig. 4). And if Switch-A and -B are used in combination then there will be one star topology (with wider lanes to all the blades).

When a $1\times$ lane supports 10 Gbps data rate, an IB QDR $4\times$ port from a blade connecting to a double-wide interconnect module will yield 40 Gbps in one direction. For both direction, the aggregate bandwidth of a double-wide interconnect module will be 80 Gbps. The cross-sectional bandwidth of a blade backplane is the product

of this number and the maximum number of blades and the maximum numbers of double-wide interconnect modules within an enclosure.

In this design, the fabric connectivity choices for the blades will dictate the interconnect module form factor to be single-wide or double-wide. The size of the interconnect module can be determined by the amount of connectors on the switch faceplate, which can be derived from the switch over-subscription ratio, i.e., the downlinks to the blades vs. the uplinks to the external switches. For example, for 16 blades and 4 external connectors on the faceplate, the switch's over-subscription ratio will be 4:1.

Signal integrity challenges are not trivial for a pair of differential signals on a blade backplane at 10 Gbps, particularly when the backplane supports several blades and switches. The challenges include minimizing the signal losses along the signal path (or channel) consisting of multiple connectors and long traces on a PCB, while minimizing the cost of the backplane. These can be addressed through general signal integrity best practices such as carefully defining the signal pin assignments (such as grouping same-direction and isolating different-direction high-speed signals), keeping the traces short, keeping the traces within the PCB layers, keeping the through-hole via stubs short (by design or by back-drilling), etc. Although modern high-speed transmitters and receivers are capable of controlling the transmit signal waveform and adaptively filtering out the noise at the receivers, respectively, the end-to-end channel losses and noises (such as cross-talks) need to be minimized. A transmitter's signal waveform can be shaped by selecting the signal *emphasis* settings [8]. The purpose is to anticipate the high frequency losses in a way that after the signal travels through a channel the waveform will still have enough energy in the leading edges. Relatively higher amplitude at the leading portion of a positive and a negative waveform at the transmitter can give a wider and taller signal "eye" pattern for the receiver to discern the signal.

Figure 5(a) shows a hypothetical original signal, and (b) shows the signal after going through a channel where most of the high frequency components have been attenuated in the channel. Figure 5(c) shows a simple de-emphasized version of the signal of (a), where the first bit has relatively higher amplitude than the trailing bits of the same polarity. The signal at the receiver (d) is a much improved version compared to (b). Alternately, the signal can be pre-emphasized, i.e., the leading portion(s) of a wave forms have higher amplitudes than the original amplitude. There can also be multiple pre-/de-emphasis levels that can vary the amplitude levels within a bit time. A caveat is that the emphasis settings of a transmitter may depend on the channel topology, and thus it is a challenge to optimally set them when the channel topology changes for a transmitter, e.g., when a blade is inserted in a different position in an enclosure. This problem can be addressed during the configuration phase of the enclosure, which will be discussed in the manageability section.

As shown in Table 1, the IO or communication interconnect bandwidth are in the 10–40 Gbps, at the top end. In addition, depending on the usage these interconnects are used for the distances of a meter to hundreds of meters—about a meter for PCIe, less than a few meters for SAS, about 10–30 meters for IB, 10's to 100's meters for FC, and 100's of meters for Ethernet. Consequently, the protocols are designed to be serial and require only a few signal pins to conserve the number of wires within a cable. In contrast, coherent protocols (such as HyperTransport) are by nature very

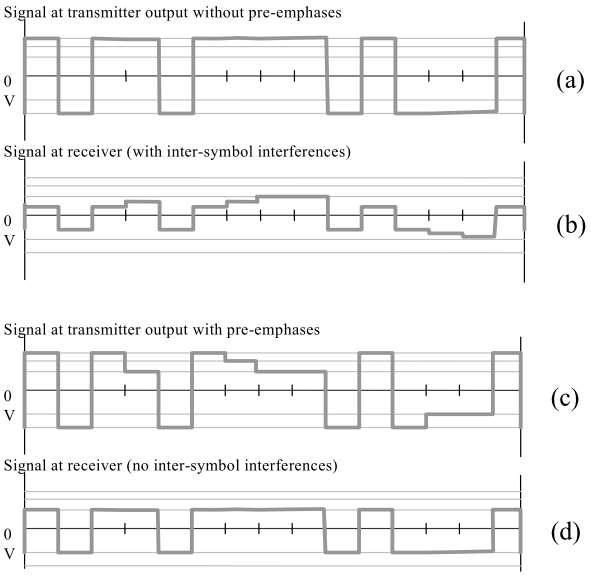

Fig. 5 High-speed signal transmitter emphasis

latency sensitive and the interconnections are traditionally on-chip, on-board, or between boards. Therefore, the coherent links typically have source-synchronous clocks and several pins are used for the protocol to be latency-efficient. In addition, multiple of these links are used to minimize the hop count. In summary, an IO or a network interconnect can require 4–16 wires per port, where a coherent interconnect can require about 80–100 wires. Obviously, the complexity of a backplane in a blade enclosure cannot be practical for implementation and economic reasons to support IO links, communication networks and coherent links.

With the number of cores per processor chip increasing to two, four and more, it is relatively easy for a modern server blade to have 16 cores. Using this example blade, conjoining two of them yields 32 cores. Conjoining two blades can be achieved by means of connectors or a PCB with connectors for one or more coherent links. This concept can be extended to more than two blades if there are enough applications to justify building the products. The over-under scalable model illustrated in Fig. 2(b) allows a blade to be tall and have enough space for a tall connector to support coherent links for inter-blade connectivity.

2.3 Power

A blade enclosure connects to facility power by interfacing directly to power cable feeds routed to rack cabinets, or indirectly to in-rack power distribution units which are in-turn connected to facility power feeds. Regardless, it makes sense to design an enclosure power budget to be some multiples of the facility power lines. Table 2 lists the most commonly used facility power feeds.

An enclosure power budget needs to be designed to accept some multiples of facility power feeds to support a number of blades with certain power envelope. As discussed earlier, although a maximum number of blades will help on the infrastructure amortization to lower the cost per blade, the power budget per blade limits the

Table 2 Commonly used datacenter facility power feeds

	Region	Line voltage	AC breaker [Cord]	Current (derated)	AC Power (derated)
Single-phase	NA	208 V	20 A	16 A	3328 VA
Single-phase			30 A	24 A	4992 VA
3-phase 30 A			30 A	24 A	8640 VA
3-phase 60 A			60 A	48 A	17 292 VA
Single-phase	International	230 V nom.	16 A		3680 VA
Single-phase			32 A		7360 VA
3-phase			16 A		11 040 VA
3-phase			32 A		22 080 VA

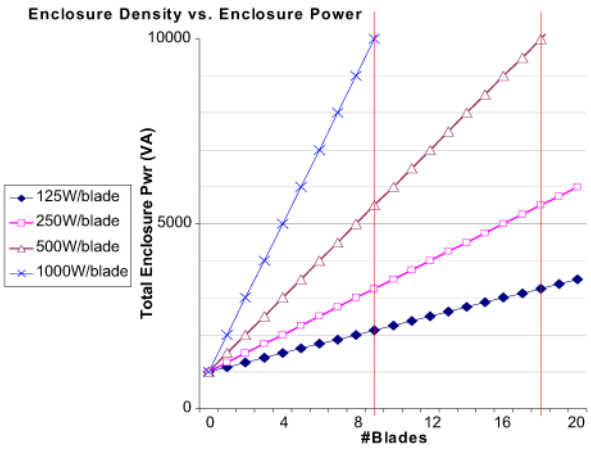

Fig. 6 Blades power consumption within an enclosure

number of blades that the enclosure can support given a limited power budget for the enclosure.

Figure 6 illustrates the amount of enclosure power required for generic blades with varying power budgets of 125 W, 250 W, 500 W and 1000 W per blade. Also, as discussed earlier on how the form factor of blades are designed to be scalable for performance, the power budget for the smaller and larger blades should be sized within the power budget of the enclosure. For example, Fig. 6 shows that if an enclosure has 5000 W for the blades, then there can be 16 250 W blades or 8 500 W blades.

Power is a scarce resource in datacenters. Multiple stages of power conversion are done within a blade enclosure and within blade and switch modules for different components' power requirements at different levels and tolerances.

For maximum power utilization efficiency, the following needs to be optimized:

- High efficiency voltage conversion at every stage.
- Minimized losses through the power distribution paths by minimizing the DC resistance along the path. Higher power losses will be converted to heat, which will translate to more cooling requirements, i.e., more power consumption.

- Minimize power consumption of the cooling fans, by using high pressure power efficient fans where the RPM can be adjusted according to the equipment cooling requirement. Another way to lower power consumption of the fans is to optimize the airflow paths in the entire enclosure to use less total airflow.
- Operate power supplies in their highest efficiency modes, i.e., operate at high utilizations. For example, if multiple AC-to-DC conversion power supplies are not utilized high enough, then shed the load to fewer power supplies to run them at higher utilization, if possible.

In addition, power management methods should be extensively implemented including capping power budgets at module and component levels, monitoring actual power consumptions, power budget profiling according to the application utilizations and processor utilization levels, etc.

2.4 Availability

In a blade system enclosure there are multiple servers, network equipment and infrastructure support elements (such as power supplies and fans). It is important that there shall be no catastrophic failure of the enclosure caused by any single failure of a component or module within the enclosure. There are several ways to define availability. Below, we qualitatively describe some general methods to maximize availability in our blade systems.

2.4.1 Minimize single point of failure (SPoF)

- Provide redundant modules such as redundant power supplies, fans, switches, enclosure managers, etc. There can be multiple redundant models, such as $N + m$, where $m = 1, \ldots, N$. For example, a $3 + 3$ redundancy for power supplies means 1 to 3 power supplies can fail and service will not be interrupted. $3 + 1$ redundant power supplies means only one power supply can fail for service to be uninterrupted if the load requires all 3 power supplies.
- Provide redundant paths such as facility power feed connectivity, power delivery to modules within an enclosure, blades to interconnect bay connectivity, and blades to enclosure manager connectivity. There are choices for implementing redundant paths for a blade in connecting to the backplane. There can be one connector with redundant pin paths, or multiple connectors. There are other considerations that should be noted in making this choice, on single connector or multiple connectors. In the example of combining two smaller blades to form a larger blade in scaling the blade form factor, if there are multiple connectors on a smaller blade, then the number of connectors for a larger blade will be potentially doubled. This increase in connector count can be counter productive such as mechanical mating tolerances which can affect the failure rate of a blade, e.g., during blade insertions, blade handling outside of the enclosure, etc.

2.4.2 Maximize mean time to failure (MTTF) of modules

- This is especially true for a critical component that would be a single point of failure (SPoF). If there is only one backplane PCB within an enclosure, it is important to make the backplane with a high MTTF, such as minimizing the number

of active components and minimizing the connector count. Ideally, a backplane is completely "passive," i.e., no electronic components at all. The next level to relax this constraint is to make the backplane having only passive devices, such as resistors and capacitors. Yet another level to relax is to have minimum active components, but with high mean times between failures, and ensure that they will not cause critical failure.
- Minimize the operating temperature of the components. First, deliver fresh cool air to every critical module that requires cooling (servers, switches and power supplies). Also, strategically place hot components in the best airflow paths while providing ample volume space for heat extraction mechanisms, e.g., heat sinks.
- Minimize connector failure by maximizing mechanical robustness, such as using connectors with rigid enough body and alignment pins. For heavy modules, such as server blades, we prefer press-fit type contacts to surface-mount type to prevent solder joint failures.
- Minimize the number and types of backplane connectors on each blade or interconnect module for most consistent mechanical alignment such as initial mating, connector contact-wipe, and mated pair bottom-out.

2.4.3 Maximize fault isolation

- Ideally, any failure within a component will not affect the functionality of other components. A relaxed requirement for blade systems is "Any failure within a FRU will not affect functionality of other FRUs". For example, servicing a failed fan should not require another fan (or any other FRU) in operation to be removed.

2.4.4 Minimize the mean time to repair (MTTR)

- Blade systems inherently provide field replaceable units (FRU) within a blade enclosure for ease of installation and replacement.
- Detection and reconfiguration are further discussed in the manageability discussion. The key point is that when a failure occurs on a blade, the down time is minimized by migrating the service from the failed blade to another functional blade in shortest time possible.

We will address availability again at the end of the Manageability section.

2.5 Manageability

Each blade has a management controller commonly known as a blade management controller (BMC). A blade enclosure commonly has one or two enclosure management (EM) controllers.

The BMC monitors thermal and operational conditions within each blade, in a way where the statuses can be monitored by the EM. The BMC also handles other tasks, such as providing remote console access to users, remote peripheral attachments (to floppy and CD of a remote console client system), programmatic interface to EM as well as to external software environment such as datacenter management console or automation software. The BMC on a blade can also operate under stand-by power,

before the blade is allowed to be powered on. The BMC allows users and management tools to completely manage a server using the same method regardless of the physical location, such as in front of the server, across the rack (room, building or world), truly enabling lights-out management of a server.

The EM monitors thermal and operational conditions within an enclosure, in a way where the statuses can be monitored by external datacenter management software. The EM also handles other tasks, such as providing remote console access to users and external software. There can be redundant EM pairs in an enclosure, since it is a critical module within an enclosure and it should not be a single-point-of-failure. How the redundant EM pairs intercommunicate to maintain coherent state, and how they communicate to detect a failure condition and fail-over from the active EM to the stand-by EM is implementation dependent. The EM's are operational as soon as the enclosure is supplied power. The following paragraphs describe significant advantages for having the EM's in an enclosure to manage blades and switches:

2.5.1 Hardware configuration management

- Blades installed in an enclosure can be in different form factors, of different types and have different configurations with network interface devices installed to connect to network switches. There can also be multiple different network switch modules installed in the same blade enclosure. The EM has to ensure each blade has the correct devices installed to interface to the network switches. If so, the EM will continue to turn on the blades per the power management policy. If not, the EM can choose to not power the blade or not turn on just the network ports that are not compatible, depending on an implementation.
- If the network ports are compatible then the EM discovers the connectivity of the devices on both ends of the backplane traces, and sets up any necessary equalization parameters, as discussed in the Performance section.

2.5.2 Power/thermal management

- For the blades that pass the hardware configuration verification, the EM will verify whether each blade can be allowed to power up provided that the blade's BMC has requested power, and there is enough power and cooling budget by querying the power supplies and fans installed.
- If not, the EM negotiates with each blade for lower power budgets predefined by the system administrators.
- Modern processors are capable of setting "power states" to operate in certain operating voltage and frequency. Using these, the power consumption of a blade can be easier to manage by the EM.
- Once blades are operational, the EM continues to monitor the blades for power consumption, power supplies' status, thermal conditions throughout the enclosure, fans' status, and enclosure configuration changes (e.g., new blades installed, blades removed). The EM then makes necessary adjustment such as power budget for each blade and communicates with blades' BMC to control the blades' power modes.

2.5.3 Availability management

- Since the EM has access to each blade's BMC and their respective interconnect modules, it is possible for the EM to detect failure conditions including component failures, thermal conditions and software malfunctions.
- The EM can then take actions on the to-be-failed or already-failed blade, such as migrating or redeploying applications on another blade and reconfiguring the interconnect modules accordingly.
- Failure detection algorithms and fail-over policies can be defined within EM, or at a higher management software level with direct communication to the EM, to improve the service availability of the blades and the interconnect modules within an enclosure.

2.6 Flexibility

We have discussed methods to optimize an enclosure design for generic blade enclosures. Traditional blade enclosures are primarily designed to support traditional general-purpose server blades and traditional switch modules.

For a blade enclosure to be an optimal general-purpose infrastructure, it has to be a lot more flexible than a traditional blade enclosure. Some of the elements from the previous discussions that make the blade enclosure more flexible, and therefore a more general-purpose infrastructure, include:

- Scalable blade form factors for blades to be general-purpose scale-out and scale-up servers, application-specific processors, storage, IO, etc.
- Scalable interconnect module form factors and the backplane infrastructure supporting network-semantic and memory-semantic interfaces on the same set of traces.
- The EM to enable the connectivity of compatible blades and interconnect modules.
- The EM to allocate power depending on the types of blades and available power budgets.

3 Bladesystem™ C-class case study

In the previous section, we explained how we optimized and suggested solutions for each of the six key parameters, namely cost, performance, power, availability, manageability and flexibility. In this section, we will use HP BladeSystem c-Class architecture as a case-study in designing a general-purpose infrastructure leveraging the solutions suggested in the last section, and further defined the implementation specifics.

The first instantiation of that architecture is the c7000 enclosure. This 10U enclosure form factor was derived from several directions. It is to hold 16 modern blades that can accommodate system components equivalent to the most popular server model in datacenters—the 2-socket, 8-DIMM, 2 hot-plug drive blade and two optional IO cards (primarily for fabric connectivity). The 42U rack is the most commonly used rack cabinet form factor in datacenters. The 42U rack height should be

Table 3 Enclosure sizing in a 42U rack

Enclosure size (height)	Max. # enclosures in a 42U rack	Worst-case rack space wasted [U, % of 42U]	Min. # of blades to be competitive (with respect to 1U rack-optimized)
4U	10	2U, 5%	5
5U	8	2U, 5%	6
6U	7	0U, 0%	7
7U	6	0U, 0%	8
8U	5	2U, 5%	9
9U	4	6U, 14%	11
10U	**4**	**2U, 5%**	**11**
11U	3	9U, 21%	15

evenly divisible by the blade enclosure height, and even if it cannot, there should be minimum waste on the left-over rack space. Table 3 lists how well different enclosure sizes fit within a 42U rack.

The 4U and the 5U are too small to accommodate modern high-performance server electronics and still provide space for the minimum number of blades to be competitive (listed in the last column). The 6U and 7U enclosures are optimal in rack space utilization, but they are still too small to accommodate high-performance blades and switches, and the number of blades do not allow for efficient amortization. The 8U and 10U are very similar in rack space wastage. Although the 8U gives one more enclosure than the 10U, per blade form factor is still too limited and thus not enough number of blades to justify the infrastructure. The 9U wastes too much rack space at the same enclosure count as the 10U in a 42U rack.

The last column is the minimum number of blades needed for a 42U rack to have a higher density than 1U rack-optimized servers, as many users compare blade density with the 1U rack-optimized server. In other words, fewer blades than this number will not be attractive from density perspective. For the 11U, there will be one enclosure fewer in the 42U rack, but the amount of space gain is not justifiable at the expense of an entire enclosure. As the enclosure size gets larger, it becomes impractical to handle from size and weight perspectives, and therefore larger enclosure sizes are not discussed further here.

The 10U seems to be an optimal enclosure size balancing the trade-offs on enclosure blade density, per-blade volume size, the number of switches, power supplies, fans, rack density and 42U rack space wastage.

Figure 7(a) shows the front view of the c7000 enclosure. It has 16 *half-height* server blade bays organized as 8×2 over-under form factor, 8 *full-height* blade bays or a mix of half-height and full-height blade bays. This scalable configuration allows 64 blades in a 42U rack since there are 16 blades per enclosure and there can be four 10U enclosures in a 42U rack with 2U left over for miscellaneous use such as aggregating switches, a laptop/KVM (keyboard/video/mouse) tray or Power Distribution Units (PDU). 64 blades in a rack means 50% more servers compared to 1U rack-optimized servers in a 42U rack. The half-height blade form factor is also optimized to accommodate six 2.5″ hot-pluggable disk drives.

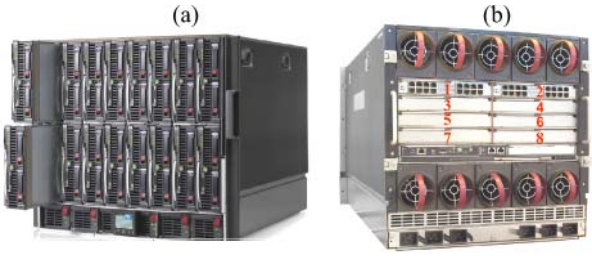

Fig. 7 BladeSystem c7000 enclosure (**a**) front view, (**b**) rear view

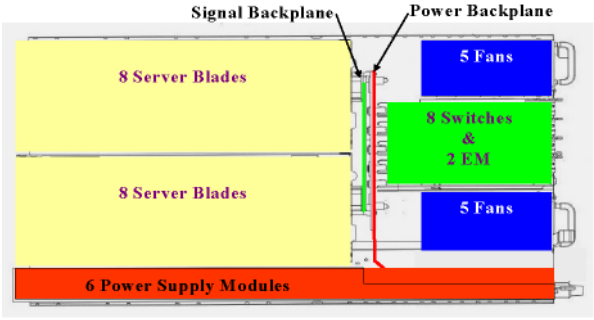

Fig. 8 BladeSystem c7000 enclosure side view

In addition to the server blades, other modules accessible at the front are 6 power supplies and a LCD called Insight Display for enclosure and blade configurations as well as for status reports. The six power supplies can be configured to be not redundant, $N + N$ (e.g., $3 + 3$) redundant, or $N + 1$ (e.g., $5 + 1$) redundant. As shown in Fig. 7(b), the c7000 enclosure rear supports 10 fans, 8 interconnect modules, 2 redundant enclosure managers (also known as OA—Onboard Administrator), and power source connectors. Each half-height and full-height blade can consume up to 450 W and 900 W, respectively.

Figure 8 illustrates the side view of the c7000 enclosure, where the 16 half-height blades on the left and the 8 switches on the right are connecting to the same signal backplane. The power backplane is totally independent from the signal backplane, to simplify both the power backplane and the signal backplane construction. The power backplane is a solid metal construction with no components, making it a very reliable power distributor. The signal backplane is also a passive backplane board. The design of the signal backplane followed high-speed signal design best practices, including impedance control, skew control, back-drill, etc.

The form factors for the switches are also scalable to be either single-wide or double-wide. The single-wide form factor is optimized to support 16 RJ45 for Ethernet or 16 SFP connectors for FC modules.

Figure 9 illustrates the 8 interconnect bays 1 through 8 also already shown in Fig. 7(b), where 1 and 2 (1/2) can be used as two single-wide redundant switches 1A/1B, respectively. Similarly, the interconnect bays 3/4, 5/6, and 7/8 are three redundant pairs. For the double-wide switches, the interconnect bays 1 and 2 are combined to form 1AA, 3 and 4 are combined to form 1BB, allowing 1AA and 1BB to form a redundant pair. Similarly, 2AA and 2BB are redundant pair made up of the interconnect bays $5 + 6$ and $7 + 8$, respectively.

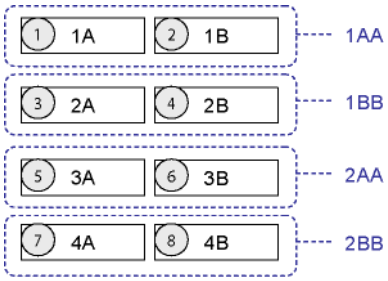

Fig. 9 Scalable interconnect bays

Each double-wide interconnect bay can support 4× interface and the backplane is capable to support 10 Gbps per 1× interface, and therefore 40 Gbps for a 4× interface. With connectivity to four double-wide interconnect bays at the back of the enclosure, a half-height blade can have a one-way bandwidth of 160 Gbps and bidirectional bandwidth of 320 Gbps. For 16 half-height blades at the front of the enclosure, the backplane "front-to-back" cross-sectional bandwidth can be up to 5.12 Tbps.

3.1 Bottom-up design for power and cooling

The power source connectivity for the c7000 enclosure is optimized for the most popular power feeds in enterprise datacenters. The initial implementation offers either six single-phase power cords or two 3-phase power cords. The six power supplies are sized for the most popular power sources. Each power supply module is rated at 2250 W output. When the six power supplies are configured to be in $3+3$ redundant, the power consumption load within an enclosure can be up to 6750 W.

The following methods are used to maximize the total power efficiency within an enclosure:

(1) Maximize the power supply modules' conversion efficiency
(2) Regulate the available power budget for blades
(3) Maximize the fans' power consumption efficiencies

3.1.1 Maximize power supply efficiency

With Dynamic Power Saver, fewest number of power supplies within an enclosure are turned on to support the load with $N+N$ power supply redundancy, so that all the power supplies can operate at high efficiency. Power supplies operate at higher efficiency levels when their utilizations are high.

Figure 10 shows the enclosure power supplies output requirements in three ranges with relative power supply efficiencies, where the number of power supplies is varied: two (in $1+1$ configuration) at 2250 W per supply; four (in $2+2$ configuration) at 4500 W per 2-supplies; and six (in $3+3$ configuration) at 6750 W per 3-supplies. The highest efficiency range is 85% to 90%. The efficiency of the power supplies drop dramatically when their load is not high enough. For example, when all the six power supplies are used and when the load is about 33% the efficiency drops to 80%, i.e., this load can be handled by just two power supplies with $1+1$ configuration (see the "$3+3$ not-managed" curve and the vertical marker line in Fig. 10). By managing

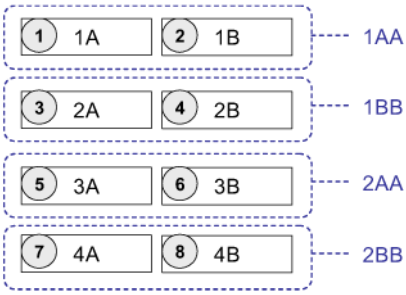

Fig. 10 Power supply efficiencies vs. load

Table 4 Power and cost savings by power supplies load balancing

PS output	#PS	Watt/PS	PS eff%	PS input	Power waste
1800 W	$3+3=6$	300 W	75%	2400 W	600 W
1800 W	$1+1=2$	900 W	89%	2023 W	223 W
Power savings for an enclosure					377 W
Power savings for 20 enclosures					7540 W
Power saving costs per year (assuming ~$0.10/KWh)					~$6600

the six power supplies in a way that only the minimum number of power supplies are active to support the load allows the active power supplies to operate at their peak efficiency. Therefore, the overall power supply efficiency can be dramatically improved. Note the power supply sharing effect (small dips of the "3 + 3 Managed" curve in Fig. 10) when the power supplies are activated from 1 + 1 to 2 + 2, and from 2 + 2 to 3 + 3.

Table 4 illustrates an example of the benefits of Dynamic Power Saver in terms of lower loss in power conversion and lower utility cost. In this example, all the modules within an enclosure draw 1800 W of power from the power supplies. If all the six power supplies (3 + 3) are used then each power supply will be supporting 300 W at 75% efficiency. Therefore the AC input to the six power supplies will be 2400 W, with 600 W wasted. However, if only two power supplies (1 + 1) are used then each will be supporting 900 W at 89% efficiency. Therefore, the AC input to the two power supplies will be 2023 W, with 223 W wasted. That means the power savings due to higher conversion efficiency is 377 W per enclosure. This lower waste in power conversion directly translates to utility saving. For 5 racks with 4 enclosures in each rack, there will be 20 enclosures and the power saving will be 7540 W. Note that, when only two power supplies are used, the remaining four supplies will be in stand-by, and are available if the power draw is increased by the blades.

3.1.2 Regulate the blades' power budgets

Modern processors are inherently much more power efficient than their predecessors because of advances in silicon processes and chip designs. In addition, modern processors are also designed to operate in different performance states (p-states), where their operating voltage and frequency can be stepped down and up dynamically. Processors consume less power in lower p-states. One notable characteristic of

the p-states is that some processors' throughputs are not affected at lower p-states when the processor utilization is not near its peak [9]. Typically, the throughput is not affected at all by lowering the power when the utilization is less than 80%, and is not significantly different even at 90% utilization. By dynamically adjusting the p-states, the system can operate at full performance level for the full range of workload while reducing power consumptions for lower workloads. Generally, server processor utilizations in enterprise datacenters are below 80% most of the time. (This is due to various reasons—e.g., the processor outperforming other subsystems within servers, servers' resources over-provisioned to handle potential peak loads, workload capping at 50% to handle spikes, etc.)

HP named its BladeSystem blades' p-states control mechanism HP Power Regulator. The power consumption and temperatures within a blade are monitored by each blade's baseboard management controller called iLO (Integrated Lights-Out controller), and the p-states of the processors within the blade are adjusted accordingly by the system firmware in real-time. The iLO also sets the system firmware to not allow processors to exceed certain power consumption level by capping the highest p-states the firmware can set on the processors.

Each blade within an enclosure reports its corresponding power consumption levels for the OA to regularly manage each blade's power requirement to be optimal. For example, if the actual power consumption of a blade is constantly above a certain watermark level, then its maximum power level can be incremented, if its iLO requests.

In addition to the blade- and enclosure-level power management, datacenter management tools can spread the load across different groups of servers to further balance power consumption and cooling requirements across the datacenter facility. Server virtualization methods based on VMM [9] can also be used to migrate applications across blades to save power while maximizing the ratio of performance/watt.

3.1.3 Maximize the cooling efficiency

The BladeSystem c7000 enclosure is designed for the ambient cool air to be drawn from the front and for the extracted heated air to be exhausted at the rear of the enclosure. The server blades and the interconnect modules are at the front and rear portions of the enclosure, respectively. Therefore, the blades and the interconnect modules interface to the signal and the power backplanes from the front and from the rear, respectively, as shown in Fig. 11. Figure 11 also shows the air plenum in the center region of the enclosure, where the signal and the power backplanes are.

The 10 fans extract the hot air from the center plenum to the rear of the enclosure. There are no fans in the blades and switches. The power supplies pull fresh cool air from the front and exhaust directly to the rear of the enclosure, independently from the blades and switches.

Since the server blades' faceplates are exposed at the front of the enclosure, the fresh cool air from the front gets pulled into the blades and the heated air gets extracted into the center plenum by the enclosure fans. There are "air scoops" on the extreme sides of the enclosure that allow the fans to draw the fresh cool air from the front of the enclosure through these side air scoops via the center air plenum and the

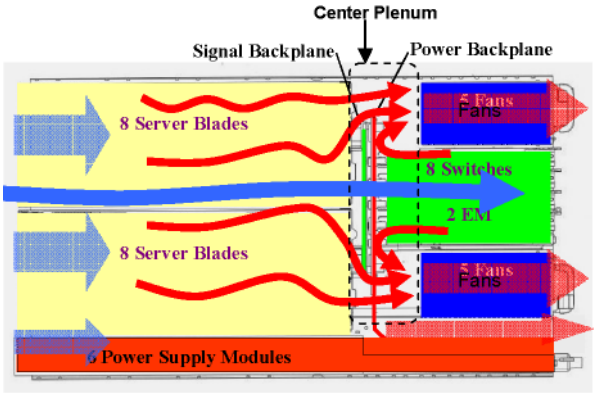

Fig. 11 BladeSystem c7000 airflow paths

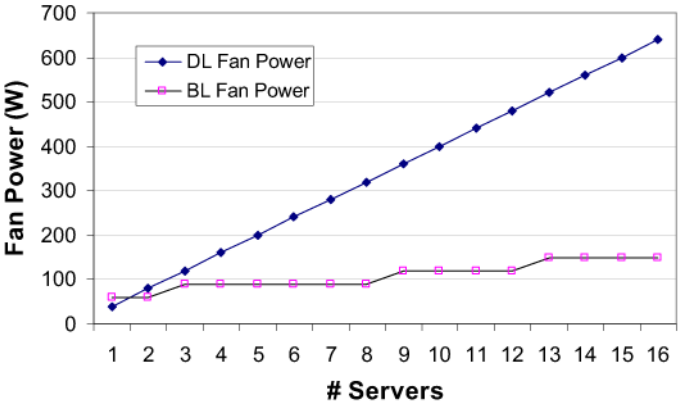

Fig. 12 Fan power consumptions for blade vs. rack-optimized servers

interconnect modules. There are air ingress holes on the sides and rear portion of the interconnect modules for the cool air from the scoops to be pulled in.

The airflow through the center plenum is also directed by means of air louvers and mechanical trap doors, which are actuated only when fans are running and a module is inserted, respectively. In addition, when a blade or an interconnect module is inserted it is seated close to the backplane assembly and the perimeter of the module is sealed to prevent air leakage.

HP called the c-Class enclosure fans the Active Cool Fans, which can move more air at lower power than traditional fans. The ambient temperature in cool aisles in datacenter ranges from 22 °C to 30 °C, with a typical value of 25 °C. The Active Cool Fans can move the same amount of air at lower RPM and thus lower power consumption, due to their efficiency [10]. Figure 12 compares the cooling fans power consumption for sever blades vs. rack-optimized servers.

Understandably, the power consumption of fans of rack-optimized servers scales linearly with the number of servers. For the c-Class, the numbers of fans required in an enclosure are 4, 6, 8 and 10 for 2, 8, 12 and 16 blades, respectively, and therefore the power consumption of fans in an enclosure increases at a lower rate. On average,

the power consumption for cooling fans per server blade in c-Class is about 10 W vs. 40 W per rack-optimized server at similar system configurations.

The Active Cool Fans' RPM can be lowered to consume even lower power in the most common datacenter ambient temperature range of 22 °C to 28 °C. Note that the fans run at different RPM for the same ambient temperature for different processors' performance (which is directly related to processors' power consumption).

The fan control logic synchronizes with the OA to manage the thermal requirements throughout the enclosure, and optimizes the amount of airflow, the power consumption, and the acoustic noise of the fans.

3.2 Network abstractions

Despite all the advantages of switches inside blade enclosures that reduce the cable management complexity and costs, these switches in blade enclosures added significant switch count for the network administrators to manage. Not using switches to avoid that problem, by means of pass-through modules, would bring back one of the key problems that blades solved—cable management.

The goal is to aggregate the physical ports from the blades to fewer physical ports by using a switch, and make the switch be "transparent" to the network administrators' management domain. Figure 13 shows two hypothetical blades with each having a FC host bus adapter (HBA) connecting to the FC switch across the backplane. The HBA-1 (in Blade-1) and the HBA-2 (in Blade-2) have the hardware port addresses of WWN1 and WWN2, respectively.

A traditional FC HBA's port have the port type called N-port (Node-port), which can connect to another N-port for a point-to-point interface, or to an F-port (Fabric-port) of a FC switch for a fabric interface. Therefore, a FC HBA's N-port in a blade will interface to an F-port of a FC switch across the backplane in an enclosure as shown in Fig. 13. For a FC switch in a blade enclosure, its external uplink port connecting to the FC core switch is typically an E-port (Expansion-port), and therefore the FC switch will be managed by the storage administrators to be part of a SAN fabric, as it will be "seen" by the core switches as a *switch*. With a Virtual Connect FC module supporting N-port identifier virtualization (NPIV) [11], the external FC port illustrated in Fig. 13 is an N-port. A FC core switch will then "see" this N-port the FC module similar to a FC port directly off a FC HBA in a server. In other words, the FC ports on the blades have virtual connectivity to the external switches via fewer physical ports on an interconnect module HP called Virtual Connect FC module.

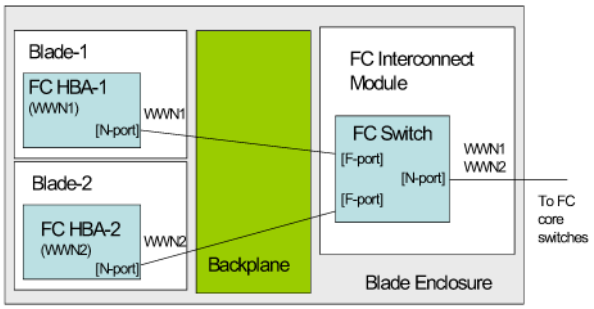

Fig. 13 Fiber Channel port types using NPIV

A Virtual Connect FC module essentially aggregates the FC ports of the blades and presents them with fewer physical ports to the external switches as Node-ports, rather than as a FC switch participating in a FC SAN fabric.

In other words, from a port management perspective, the FC ports are now logically moved from the back of the server blades to the back of the enclosure, solving the problem of FC switch count explosion in datacenters. In common FC SAN fabric designs, there are limited number of switches that can be incorporated in a SAN. This number varies depending on the vendor (McData, Cisco and Brocade allow 24, 40 and 56 FC switches in a SAN fabric, respectively). Virtual Connect allows port aggregation without introducing a (managed) switch in the SAN and therefore Virtual Connect can be used as many times as needed without affecting the switch count in a SAN fabric. Multiple Virtual Connect modules can be connected (or stacked) together to create a single Virtual Connect domain, so that only one Virtual Connect manager (VCM) is needed. A second VCM can be used as an option for redundancy.

3.3 Support for data center automation

We will use the application of NPIV by virtual machine monitors (VMM) [10] to illustrate an example of how the hardware addresses are migrated along with applications to different physical servers. VMM can keep a pool of locally administered hardware address WWN's (globally unique worldwide names) to be assigned to the virtual machine (VM) instances. VMM can also migrate a VM instance from one physical server to another, for hardware fail-over, hardware upgrade for the application running on the VM, or other reasons. When a VM is migrated to another platform, it is important that the VM continues to have the same network accesses without noticeable service interruption, e.g., same connectivity to a FC target SAN without any changes required in the SAN switches and target (which can take weeks). VMMs achieved this by migrating the locally administered WWN (associated with the previous VM) to the new VM along with the application during the migration.

A method similar to how the VMM manages a pool of hardware addresses, can be applied to blades where a management controller could assign temporary hardware address(es) to each network interface device, and help migrate them when application on that blade is migrated to another blade. For the Virtual Connect modules, the hardware addresses (WWN for FC and MAC addresses for Ethernet) are managed by the Virtual Connect Manager and are assigned to the network interface devices' ports in a manner transparent to the operating systems.

4 Application examples

In this section, we discuss traditional and emerging application categories in dense data centers, where blade servers are most suitable. We will describe application characteristics and how they can be mapped onto hardware systems to validate the flexibility of the general-purpose infrastructure.

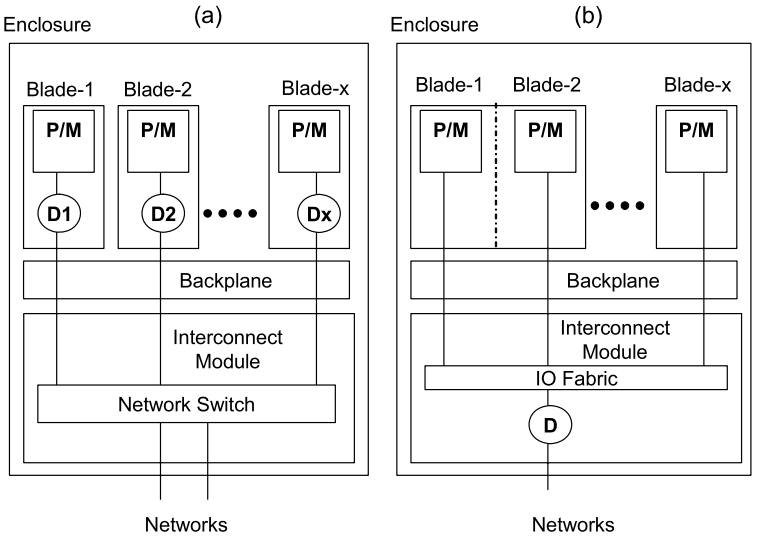

Fig. 14 Network connectivity using (**a**) network switches, (**b**) IO fabrics

4.1 Traditional enterprise scale-out servers

Since blade servers evolved from dense rack-optimized servers, blades are inherently suitable for supporting scale-out applications such as web server farms and terminal servers. For these scale-out server farms, blades are interconnected with traditional switches such as GbE switches for data networking and FC switches for storage networking. Figure 14(a) illustrates a simplified model with each blade having an IO device (D) which interfaces to a network switch via the backplane. In a typical data center, the "edge" Ethernet switches are over-subscribed at about 6:1, i.e., the downlink bandwidth from the servers side of an Ethernet switch is six times the uplink bandwidth to the core network side. Popular network bandwidth capabilities per port have also grown—10 GbE and 4 Gb FC are not uncommon. To address applications that do not require blades to have high IO bandwidth, IO fabrics can be used to reduce or eliminate the IO devices in each blade, and let all the blades share fewer IO devices via an IO fabric, as suggested in Fig. 14(b). The general-purpose infrastructure does not preclude the implementation of IO device sharing such as the methods developed under the PCI SIG [12].

Therefore the general-purpose infrastructure can accommodate traditional methods where multiple protocol interfaces in blades and corresponding switches in interconnect bays are used, as well as blades sharing IO via IO fabrics.

4.2 Database

Historically, servers with a relatively high number of processors (e.g., 16-way, 32-way) tightly coupled with shared memory subsystems were used in scale-up systems to achieve multiple threads for database applications. Large scale-up systems

require complex core logic to interconnect processors, memory and IO to achieve high bandwidth and low-latency performance. Due to long development time and specialized software requirements, large scale-up systems are neither economic nor competitive compared to today's commodity servers with multiple-core processors and large memory subsystems. In today's fast-paced technology era, one major disadvantage of long development time (e.g., >2 years) of traditional scale-up systems is that the system will potentially be out of date by the time it is ready to be shipped.

To overcome these problems associated with traditional scale-up systems, there are two methods to achieve high-performance systems for database applications using commodity server components.

4.2.1 Modular scale-up

One method is *modular scale-up*, where processor/memory pairs are interconnected via cache coherent links to form a CC-NUMA architecture. The number of processor/memory pairs and the interconnect topology can vary by implementation to trade-off cost vs. performance. Although, an enclosure can accommodate 16 half-height blades, a half-height blade is not designed to be able to accommodate more than two processor sockets, and the cost burden for each half-height blade to be a part of a modular scale-up environment will not be justifiable. A full-height blade can accommodate four modern processor/memory pairs. It can be extrapolated here that a double-wide full-height volume space can accommodate eight processor/memory pairs, as illustrated in Fig. 15(b). Details on how the processor/memory pairs are interconnected within the volume space of multiple widths of full-height blades is implementation-dependent. An enclosure designed for eight full-height blades can support up to 32 processors/memory pairs. By using the quad-core processors, an enclosure can support up to 128 cores. With this modular scale-up approach using server blades based on CC-NUMA architecture in a GPI, a scale-up system can be realized using commodity components at an economic price point due to cost efficient components and relatively fast development time.

In a more traditional blade environment, each blade contains interface controllers (e.g., D1 in Fig. 15(a)) to connect to the networks via the backplane. In this case, the processor/memory complex within each blade is the "root complex" for all the IO devices in that blade.

Alternatively, IO devices can be implemented within an interconnect module as shown in Fig. 15(b) for more flexibility in associating IO devices to the root complexes. Recall that we have discussed in the previous section how the PCIe signals and network protocol signals have been "overlaid" on the backplane traces. Therefore, a blade can still be interfacing to a network controller when it is relocated from the blade to an interconnect module. When two processor/memory complexes (Blade-1 and Blade-2) are attached together via CC-NUMA links to form a system as shown in Fig. 15(b), there is only one root complex for the IO devices D1 and D2. Note that, an implementation can choose to use only D1 or D2, and yet both Blade-1 and Blade-2 will have access to that device.

An extension of this concept is to use an IO fabric, as hinted with dashed line box in the interconnect module in Fig. 15. With an IO fabric similar to the one discussed in

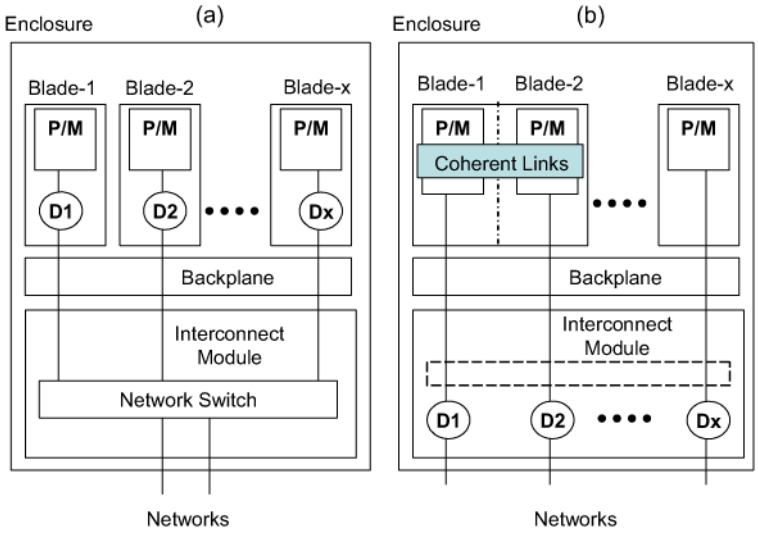

Fig. 15 Flexible connectivity for (**a**) scale-out blades, (**b**) scale-up blades

Fig. 14, potentially fewer IO devices can be used (compared to dedicated IO devices in each blade) within the enclosure.

4.2.2 Scale-out clusters

The other method to scale for performance is *scale-out clusters*, where independent servers are interconnected via high-speed low-latency switches, such as InfiniBand or 10 GbE. RDMA (Remote Direct Memory Access) methods are used in InfiniBand or in Ethernet infrastructure, for a server to directly read or write data, to or from another server, respectively. Scale-out clusters have been already used for parallel distributed database applications. The general-purpose infrastructure allows server blades to be configured as building blocks for scale-out clusters, by each enclosure supporting up to 16 scale-out nodes, and four InfiniBand or eight 10 GbE high-speed low-latency switches in the interconnect bays. With these high-speed low-latency switches, scale-out clusters can be used for the back-end database engines that interface to a shared database. Multiple enclosures can be interconnected to expand the cluster size to several nodes.

The scale-out blades can have different protocol switches such as Ethernet for data networking, Fiber Channel for storage networking and InfiniBand for cluster networking, which is figuratively illustrated in Fig. 16 with A-, B- and C-switches in the enclosures. Core switches are also illustrated as Protocol-A, -B and -C switches, and they interconnect multiple enclosures to form a larger cluster. Alternatively, one protocol switch can be used for data, storage and cluster networking, by using Ethernet or InfiniBand switches, which is illustrated with the dashed-line boxes in Fig. 16. The general-purpose infrastructure is flexible to support either individual protocol switches, or single-protocol switches, by employing the appropriate interfaces in the blades and switches in the interconnect bays.

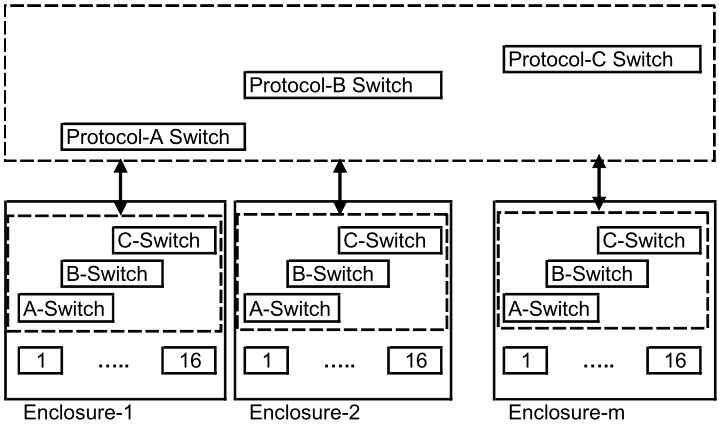

Fig. 16 Network connectiving using dedicated or consolidated networks

For both modular scale-up and scale-out cluster methods, there will be multiple threads within each environment. The difference is that, the scale-out cluster environment provides low latencies (e.g., 1–10 us) at lower costs, and the modular scale-up environment provides lower latencies (e.g., in ns) across the blades within an enclosure at higher costs. In addition, the scale-out cluster methods provides much higher scalability.

4.3 HPC (high-performance computing)

Another area of application that the blade infrastructure addresses well is High Performance Computing (HPC). HPC applications are inherently parallel, and therefore an HPC environment commonly consists of a large number of scale-out nodes interconnected with high-speed fabrics. There is a wide range of HPC applications. Depending on the application, system requirements can vary significantly within nodes as well as in the interconnects, Within a node, applications behave differently—some perform better with processor floating point operations per second capabilities (flops), some perform better with larger processor caches, some perform better with higher processor memory bandwidth. Across the nodes, some applications scale better with the interconnect link bandwidth, some require large bisectional interconnect bandwidth, and some scale better with low latencies.

The message passing link bandwidth requirement ranges from 0.01 Bpf (Byte/flop) to 1.0 Bpf per core [13]. For a modern system capable of 25 Gflops, the IO requirement at the low-end will be 250 MBps (2 Gbps), which is the bi-directional bandwidth of a GbE. For the applications that require a high-end of 1 Bpf, the same system will require 25 GBps (20 Gbps) per core, which is the bi-directional bandwidth of a 10 GbE or an InfiniBand (single data rate) 4×. For a 2-socket dual-core system, the former example will require 4 GbE NICs, and the latter example will require 4 10 GbE NICs or 4 IB SDR 4×, or 2 IB DDR 4×, or one IB QDR 4×. The blade infrastructure we designed supports 16 2-socket blades that can interface to four Ethernet switches and two IB 4× switches.

High bisectional bandwidth is important for applications such as for FFT domain conversions requiring n^2 communications. Latency is important for electromagnetic simulations and FFT operations, where small messages (64 B to 512 B) are exchanged among many nodes. Therefore, these applications scale better with NICs having low-latency high message/sec rates (being able to handle high number of messages in flight) and work well with small messages.

For large-scale scientific simulations, it is not uncommon for the HPC clusters to have hundreds or thousands of nodes in computing centers. In addition to the issues discussed for data centers, HPC system design involves additional considerations discussed as follows.

HPC systems are primarily employed to run large scale data dependent parallel applications. In this case a failure of a single process can result in a catastrophic application failure potentially involving 1000s of nodes. Hence, a scheme to checkpoint application state routinely and restart on failure, with failed processors replaced by spare processors, is increasingly the norm for large-scale application execution. The Virtual Connect method (explained in the previous sections) enables configuration of a spare blade in its pre-boot state and therefore allows fast fail-over or migration of the failed blade's application states to the new blade either within an enclosure or across a different enclosure.

An HPC application can involve direct communication between any pair of processors executing an application. Clearly, a scenario where each of the 1000s processors needs to exchange messages with all the others can be challenging for any network infrastructure. Fortunately, this is a rare scenario. The dominant communication pattern in most HPC codes is a *stencil* where the application processes are organized in a topology, typically a grid, torus, hypercube or a tree, and communication occurs primarily between neighbors within that topology.

A study at Los Alamos National Lab [14] with a representative set of codes of interest to Department of Energy and Department of Defense, combined with our own analysis of NAS Benchmarks from NASA, discovered the following:

- Of the 17 combined benchmarks/applications, 13 codes primarily or exclusively had stencil communication with two to eight communicating neighbors per process.
- One code was dominated by a non-stencil collective communication pattern, while the remaining three exhibited a combination of stencil and other patterns.

The point is that the communication capability between a small set of logical neighbors within a topology is of critical importance in HPC applications although all communication patterns must be supported well.

In a blade infrastructure we designed, the optimal server blade granularity is 8 or 16 in an enclosure. The blades within an enclosure can be connected across one or more high-speed low-latency protocol switches (e.g., InfiniBand, 10 GbE) via the backplane to form an 8-node or 16-node cluster. Multiple blade enclosures can also be interconnected via a network of switches to form a larger cluster size.

As illustrated in Fig. 16, each enclosure consists of 16 blades interconnected via an intra-enclosure switch, and all the enclosure switches are interconnected with an external switch. Figure 17 shows nine enclosures, each consisting of 16 blades (or nodes). It also illustrates multiple node examples, where the neighboring nodes are

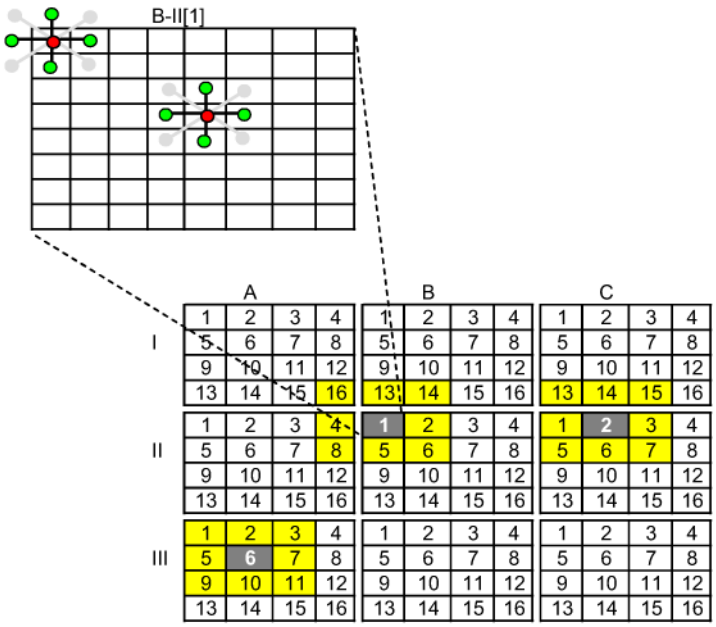

Fig. 17 An example of a communication pattern in a cluster of blades

within the same enclosure or across the enclosures. We consider a blade to contain a node (processor/memory complex) and assume that each node holds a sub-matrix of data. The stencil pattern is fundamentally within the elements of a large matrix as illustrated in Fig. 17. Most of the processing in a stencil communication requires no inter-node communication as large contiguous blocks of matrices can be assigned to the same processor within a node. When a matrix stencil is projected to a processor array, the communication pattern changes in a subtle manner. For example, in a 2D 8-point stencil illustrated in Fig. 17, the vast majority of the communication of a node-level stencil will be with its North, South, East and West neighbors. In general, for a 2D layout of a *nxn* matrix within a blade, the intra-node communication will be $O(n^2)$ and the inter-node communication will be *O(n)*. Similarly, with a cluster within an enclosure with *nxn* blade nodes, the total intra-enclosure communication will be $O(n^2)$, while the inter-enclosure traffic will be *O(n)*. For a 1D layout, the inter-enclosure communication will be a constant irrespective of the number of nodes in an enclosure, because there is approximately the same amount of communication between a pair of nodes and a pair of enclosures.

Smaller stencil sizes will reduce inter-enclosure communication traffic compared to intra-enclosure traffic. The lower the ratio of inter-enclosure traffic to intra-enclosure traffic the lesser the bandwidth required for the switch uplinks.

The number of switch uplink ports to interconnect multiple enclosures will depend on the implementation requirements. In the enclosure we designed, each switch can support up to 16 ports and there can be several switches per blade, allowing flexible inter-enclosure connectivity that can be easily customized to the requirements of the above scenarios.

5 Discussion

The c7000 enclosure, of course, supports traditional blades and network switches. In addition, as a general-purpose infrastructure the c7000 enclosure also has the following attributes:

- The signal backplane of the c7000 enclosure can support up to 5.12 Tbps of cross-sectional bandwidth and allows both network-semantic and memory-semantic traffic across the backplane, which opens up opportunities to reconsider how a system is defined within an enclosure. A server system boundary is no longer limited to rigid physical boundaries within a blade form factor.
- The blade bays are scalable in form factor (for scale-out or scale-up blades), power budget and connectivity bandwidth, which enables different types of blades to be used in the enclosure. A blade can be an IO blade (e.g., storage blade) or a traditional server blade of different sizes.
- The interconnect bays are scalable in form factor, power budget and connectivity bandwidth, which enables different types of interconnect modules to be used in the enclosure. An interconnect module can be a traditional network protocol switch, port aggregator (such as Virtual Connect module), simple traditional protocol pass-through module, or an IO fabric module with pooled IO devices.
- Flexible and scalable power and cooling resources to support different facility power requirements and enclosure power/cooling capabilities. The power source connectivity can be interchangeable to support different facility power feeds. The power distribution within the c7000 enclosure is hefty enough to scale to the power envelope of the enclosure. The Active Cool fans can be scaled in conjunction with the power source scaling.

Server blades can save datacenter costs in several areas. The followings are cost saving examples of the c-Class blade environment compared to rack-optimized servers [15]: 36% less capital equipment cost, 90% savings in deployment expenses, 69% reduction in energy consumption over a 3-year period, and 25% facility expenses on power, cooling and space.

6 Conclusions

Blades represent one of the fastest-growing segments of the server market, with most major computing vendors adopting this approach. Blades offer increased compaction, consolidation and modularity, with better management and maintenance. In this paper, we argue that blades provide a key foundational block for enterprise systems in future data centers.

We introduced the concept of architecting the next generation blade environment to be a *general-purpose infrastructure*, where the infrastructure will foster different system architectures, enabled by high bandwidth interconnects, interconnect flexibility and intelligent management controllers. We discussed in detail the key attributes and trade-off's in designing an optimal general-purpose infrastructure, and explained an instantiation of the HP BladeSystem c-Class infrastructure with scalable blades

and interconnect bays connected across a high bandwidth backplane, along with specific methods in the c-Class pertaining to management of power, network connections and fail-over automation. Finally, we described example application classes' interface characteristics and demonstrated the flexibility of the c-Class enclosure as a general-purpose infrastructure.

In the future, enterprise systems will have a common fabric for computation where users will be able to "blade everything", including storage, PC's, workstations, servers, and networking, in a variety of configurations—from scale-out to scale-up—in a simple, modular, and integrated way. Similarly, at a communication level, recent trends show promise for a common fabric for data communication, storage networking, and cluster networking. At the same time, these environments will use a rich layer of virtualization—to pool and share key resources including power, cooling, interconnect, compute and storage—and automation—to streamline processes from monitoring and patching to deploying, provisioning, and recovery—to provide enterprise environments customized and optimized for future end-user requirements. The generality, efficiencies and robustness of the general-purpose blade environment discussed in the paper is a key to such a future and we believe that this area offers a rich opportunity for more innovation for the broader community.

Acknowledgements We would like to thank the reviewers, especially Monem Beitelmal and Richard Kaufmann, for their feedback on the paper. We would like to acknowledge the HP BladeSystem design team for various insightful conversations that were valuable in the development of the c-class architecture. We would also like to thank Vanish Talwar, John Sontag, Gene Freeman, Dwight Barron and Gary Thome (all at HP) for their comments and support of the work. Subhlok was supported by the National Science Foundation under Grant No. ACI-0234328 and Grant No. CNS-0410797.

References

1. RLX Technologies: RLX System 300ex Hardware Guide, v4.0 (2002) (note: All RLX blade hardware products had been discontinued just before HP acquired RLX in late 2004)
2. HP BladeSystem e-Class Overview and Features (2004)
3. HP: HP BladeSystem (p-Class) Technology, HP Tech Brief (2005)
4. Egenera®, BladeFrame® System Specification (2006)
5. Desai, D., et al.: IBM BladeCenter system overview. IBM J. Res. Dev. **49**(6) (2005)
6. HP BladeSystem c-Class Architecture, Technology Brief (2006)
7. IEEE Draft 802.3ap, Ethernet Operation over Electrical Backplanes (2006–2007)
8. Liu, J., Lin, X.: Equalization in high-speed communication systems. IEEE Circuits Syst (second quarter 2004)
9. Herrod, S.A.: The Future of Virtualization Technologies, ISCA (2006)
10. Vinson, W.: Turning Blade Density to a Power and Cooling Advantage. Presentation slides at LinuxWorld (2006)
11. ANSI INCITS T11, FC N-port Identifier Virtualization (NPIV) standard
12. PCI-SIG IO Virtualization. http://www.pcisig.com/specifications/iov/
13. Kaufmann, R.: Multi-Core Technologies for High Performance Computing. http://www.scimag.com/multicore (2006)
14. Kerbyson, D. Barker, K.: Automatic identification of application communication patterns via templates. In: Proc. 18th International Conference on Parallel and Distributed Computing Systems (PDCS) (2005)
15. Quinn, K., et al.: Forecasting Total Cost of Ownership for Initial Deployments of Server Blades, IDC (2006)
16. HP: Power Regulator for ProLiant Servers, Tech Brief, 2nd edn. (2006)