Blades as a General-Purpose Infrastructure for Future System Architectures: Challenges and Solutions

Kevin Leigh, Parthasarathy Ranganathan
Enterprise Systems and Software Laboratory
HP Laboratories Palo Alto
HPL-2006-182
January 4, 2007*

blades, c-Class, power

Bladed servers are increasingly being adopted in enterprise data centers 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 future 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 some specific solutions in the context of the HP BladeSystem™ c-Class products.

* Internal Accession Date Only
Approved for External Publication

Blades as a General-Purpose Infrastructure for Future System Architectures: Challenges and Solutions

Kevin Leigh and Parthasarathy Ranganathan
Hewlett Packard
kevin.leigh@hp.com; partha.ranganathan@hp.com

ABSTRACT

Bladed servers are increasingly being adopted in 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 future 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 some specific solutions in the context of the HP BladeSystem™ c-Class products.

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 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 optical 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.

Similarly, 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 standalone 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 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/33MHz at 1Gbps to PCIe x16 (gen1) at 40Gbps within one and half decades. Ethernet also has evolved from 10Mbps to 1Gbps, and will soon be at 10Gbps. InfiniBand has been evolving for about 5 years, and bandwidth for IB 4x has gone from SDR 10Gbps to DDR 20Gbps, and soon to QDR 40Gbps. The bandwidth of these fabrics have converged at 10Gbps. 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, the very first blades were dense blades [2][3] that were low power and correspondingly limited in functionality. These were followed by higher-performance blades such as HP BladeSystem p-Class [1], introduced in the early 2000, and later followed by Egenera BladeFrame [5], IBM BladeCenter [4] and 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. However, future blade designs should and are likely to break free from these constraints. Egenera made an attempt towards interconnect virtualization but their method lacked in

cost efficiency, space efficiency, node scalability and interconnect flexibility.

As an extension of these trends, we argue that, in the future, blade servers are likely to be used as key foundational blocks for most 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. Sections 3 and 4 then discuss three key solutions from the recently-announced HP BladeSystem c-Class -- 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. Section 5 concludes the paper.

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. One difference is that a general-purpose infrastructure can be configured with blades and switches with different functions including general-purpose server blades, storage blades, network protocol switches, and IO fabrics.

Below, we describe the key dimensions in designing a blade enclosure to be an optimal general-purpose infrastructure. Specifically, we will highlight cost, performance, power, availability, flexibility, and manageability. Note that while we discuss these separately, these are interrelated in several ways (as shown by some of the examples in Figure-1).

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

Figure-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 most popular main components are two or four processors with associated memory modules (DIMMs) and IO devices. 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 (a), or over-under as shown in (b).

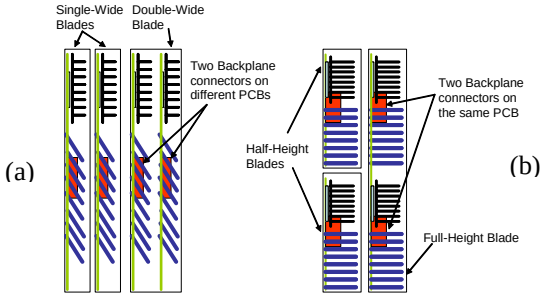

Figure-2

As the blades are scaled in the direction of the PCB plane, the system's main PCB (also known as motherboard) is typically a single plane for a larger blade in Figure-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. Skinnier blades might limit the use of low-profile DIMMs that 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 Figure-2(b).

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 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 2GB, start going up above the linear curve for 4GB, and goes exponential for 8GB and 16GB. This DIMM cost curve with respect to the top capacity bins changes over

time as the costs on the DRAMs get lower and the capacity per DIMM goes up every year.

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 maximum DIMM slots allowed for each memory channel is reached. Therefore in blades from real-estate and cost efficiency perspectives, vertical-mount DIMMs as shown in Figure-2(b) are preferred to angular-mount DIMMs as shown in Figure-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 will address the cost impact of the backplane when we discuss the performance and the availability attributes.

Performance

We will next discuss the performance of blades, the performance of the backplane for the blades to connect to the switches within an enclosure, and the performance of the switches.

In the previous section, we discussed the blade form factor to be scalable to support 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.

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 4x and PCIe x4. There are three backplane Ethernet standards under development [6], 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.

Table-1

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

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 1x) consists of a differential transmit and a differential receive signal pair. KX, KR and FC each require 1x. Additional traces are needed for wider 4x 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 Figure-3. Consequently, a set of backplane traces support network-semantic traffics (over Ethernet, FC, IB) or memory-semantic traffic (over PCIe) depending on the modules installed in the interconnect bays.

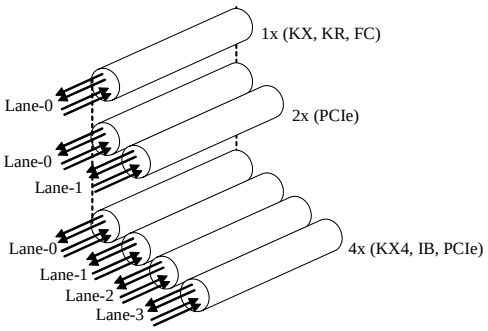

Figure-3

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 Figure-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).

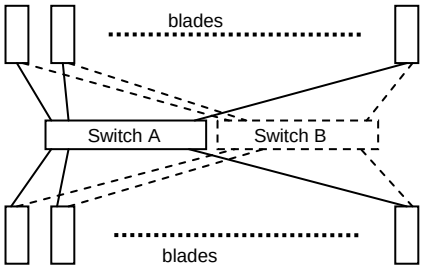

Figure-4

When a 1x lane supports 10Gbps data rate, an IB QDR 4x port from a blade connecting to a double-wide interconnect module will yield 40Gbps in one direction. For both direction, the aggregate bandwidth of a double-wide interconnect module will be 80Gbps. 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 with 10Gbps signaling, 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. 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.

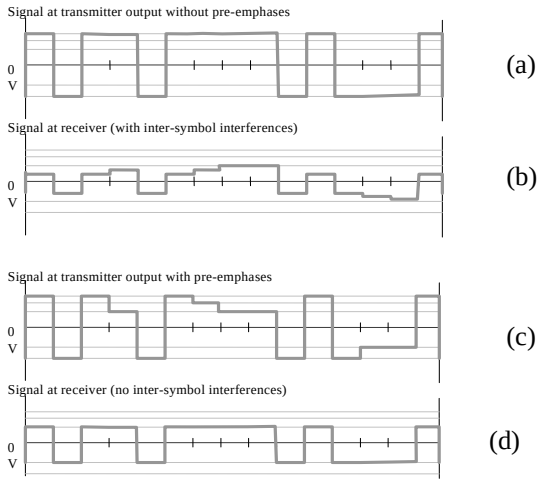

Figure-5

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.

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.

Table-2

	Region	Line Voltage	AC Breaker [Cord]	Current (Derated)	AC Power (Derated)
Single-Phase	NA	208V	20A	16A	3328VA
Single-Phase			30A	24A	4992VA
3-Phase 30A			30A	24A	8640VA
3-Phase 60A			60A	48A	17292VA
Single-Phase	International	230V nom.	16A		3680VA
Single-Phase			32A		7360VA
3-Phase			16A		11040VA
3-Phase			32A		22080VA

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 maximum number of blades will help on the infrastructure amortization to lower per-blade cost, power budget per blades limits the 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 125W, 250W, 500W and 1000W per blade.

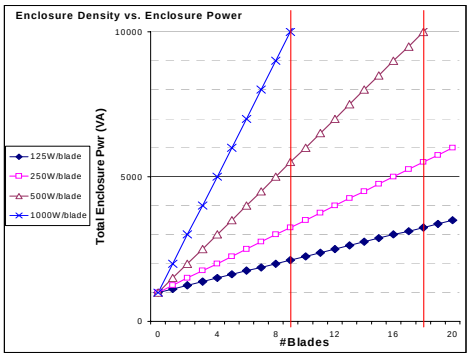

Figure-6

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, Figure-6 shows that if an enclosure has 5000W for the blades, then there can be 16 250W blades or 8 500W blades.

Power is a scarce resource in datacenters. Multiple stages of power conversions 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 followings need to be optimized:

- High efficiency voltage conversion at every stage.
- Minimize losses through the power distribution paths, by minimizing the DC resistance along the path. Power losses will be converted as heat, which will translate to more cooling requirement, i.e., more power consumption.
- Minimize power consumption of the cooling fans, by using high pressure power efficient fans where their 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.

Availability

In blade systems, there are multiple servers, network equipment and infrastructure support elements (such as power supplies and fans) within an enclosure. 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.

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...N. For example, a 3+3 redundancy for power supplies means 1 to 3 power supplies can be failed and service will not be interrupted. 3+1 redundant power supplies means only up to one power supplies can be failed 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.

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 to have 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 to have 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.

Maximize fault isolation

- Ideally, any failure within a component will not affect 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.

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.

Manageability

Each blade has a management controller commonly known as a base-board management controller (or the integrated Lights-Out controller – iLO – in hp servers). A blade enclosure commonly has one or two enclosure management controller (also known as the Onboard Administrator – OA – in hp servers).

The iLO monitors thermal and operational conditions within each blade, where these statuses can be accessible by the OA. The iLO 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 OA as well as to external software environment such as datacenter management console or automation software. The iLO on a blade can also operate under stand-by power, before the blade is allowed to be powered on. The iLO 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 OA monitors thermal and operational conditions within an enclosure, where these statuses can be accessible by external datacenter management software. The OA also handles other tasks, such as providing remote console access to users and external software. There can be redundant OA 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 OA's communicate to maintain coherent states is implementation dependent. OA's are operational as soon as the enclosure is supplied power. There are significant advantages for having OA in an enclosure managing blades and switches:

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 OA has to ensure each blade has the correct devices installed to interface to the network switches. If so, the OA will continue to turn on the blades per the power management policy. If not, the OA 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 OA also can discover the connectivity of the devices on both ends of the backplane traces and sets up the any necessary equalization parameters, as discussed in the Performance section.

Power/thermal management

- For the blades that passes the hardware configuration verification, the OA will verify whether each blade can be allowed to power up provided that the blade's iLO has requested power, and there is enough power and cooling budgets by querying the power supplies and fans installed.
- If not, the OA negotiates with each blade for lower power budgets predefined by users.
- 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 OA.
- Once blades are operational, the OA can continue to monitor the blades' power consumptions, power supplies status, thermal conditions throughout the enclosure, fans' status, enclosure configuration changes (e.g., new blades installed, blades removed) and make necessary adjustment such as power budgets for each blade and communicate with blades' iLO to control the blades' power modes.

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 to 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.
- OA to enable the connectivity of compatible blades and interconnect modules.
- OA to allocate power depending on the types of blades and available power budgets.

BLADESYSTEM™ C-CLASS CASE STUDY

HP BladeSystem c-Class enclosure architecture is designed to enable general-purpose infrastructure. The first instantiation of that architecture is the c7000 enclosure which is 10U tall. The 10U enclosure form factor was derived from several directions. It is to hold 16 modern blade model category 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 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.

Table-3

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

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-off's 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 8x2 over-under form factor scalable to 8 *full-height* blade bays. This 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 or a laptop/KVM (keyboard/video/mouse) tray. 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.

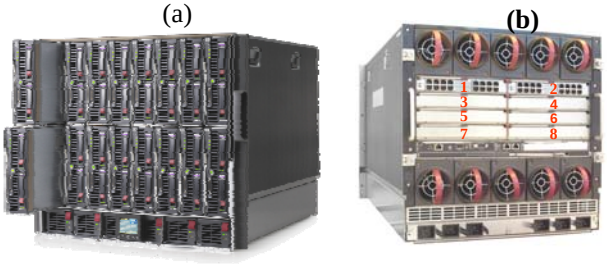

Figure-7

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 Figure-7(b), the c7000 enclosure rear supports 10 fans, 8 interconnect modules, 2 redundant OA, and power source connectors. Each half-height and full-height blade can consume up to 450W and 900W, 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, which got significant attention on high-speed signal design best practices, including impedance control, skew control, back-drill, etc.

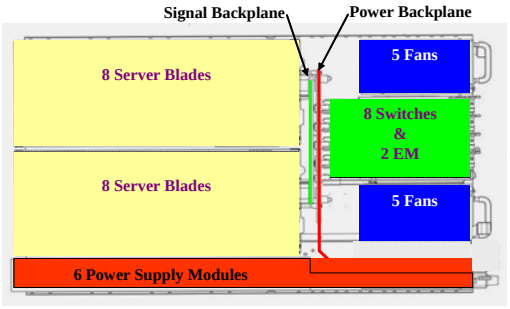

Figure-8

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

Figure-9 illustrates the 8 interconnect bays 1 through 8 also already shown in Figure-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.

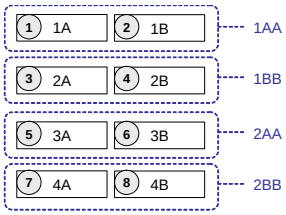

Figure-9

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

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 2250W 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 6750W.

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

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 (1+1) at 2250W per 1-supply; four (2+2) at 4500W per 2-supplies; and six (3+3) at 6750W per 3-supplies. The shaded region is the highest efficiency range (close to 90%) where the efficiencies of the power supplies are lower when the load is not high enough. For example, the efficiency is about 80% when all the six power supplies are used (3+3 not-managed in Figure-10), while the load is about 33%, i.e., the load can be handled by two power supplies (1+1). By managing the six power supplies in a way that only the minimum number of power supplies are active to support the load so that the active power supplies operate at their high enough load, i.e., at their peak efficiency range, the overall power supply efficiency can be dramatically improved. Note the power supply sharing effect when the power supplies are activated from 1+1 to 2+2, and to 3+3.

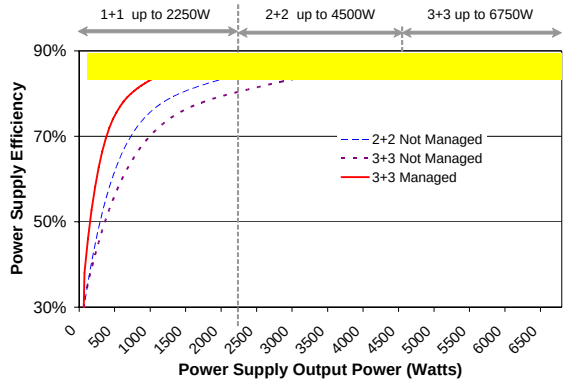

Figure-10

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 1800W of power from the power supplies. If all the six power supplies (3+3) are used then each power supply will be supporting 300W at 75% efficiency. Therefore the AC input to the six power supplies will be 2400W, with 600W wasted. However, if only two power supplies (1+1) are used then each will be supporting 900W at 89% efficiency. Therefore, the AC input to the two power supplies will be 2023W, with 223W wasted. That means the power savings due to higher conversion efficiency is 377W 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 7540W. 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.

Table-4

PS Output	#PS	Watt/PS	PS Eff%	PS Input	Power Waste
1800W	3+3=6	300W	75%	2400W	600W
1800W	1+1=2	900W	89%	2023W	223W
Power savings for an enclosure					377W
Power savings for 20 enclosures					7540W
Power saving costs per year (assuming ~$0.10/KWhr)					~$6,600

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 [10]. Typically, the throughput is not affected by lowering the power when the utilization is less than 80%, and not significantly different even for 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 overprovisioned to handle potential peak loads, workload capping at 50% to handle spikes, etc.)

HP named its BladeSystem blades' p-states control mechanism the Power Regulator. The power consumption and temperatures within a blade are monitored by the iLO, 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 not to 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 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 loads 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.

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 Figure-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 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.

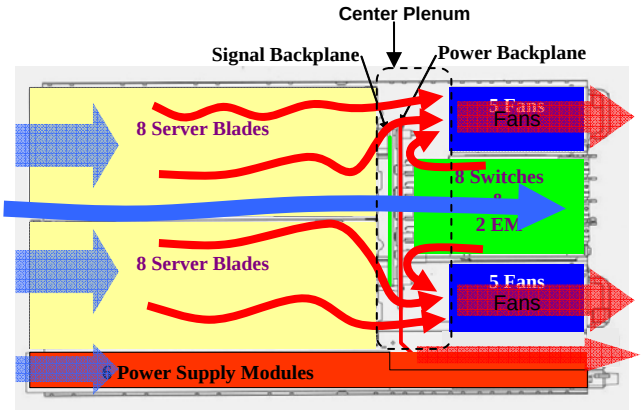

Figure-11

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 [8]. Figure-12 compares the cooling fans power consumption for sever blades vs. rack-optimized servers.

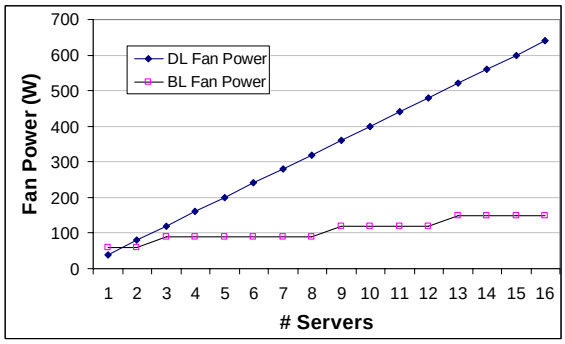

Figure-12

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 were 4, 6, 8 and 10 for 2, 8, 12 and 16 blades, respectively, and therefore the power consumption of fans in an enclosure increases whenever more fans are added. On average, the power consumption for cooling fans per server blade in c-Class is about 10W vs. 40W per rack-optimized server at similar system configurations.

The Active Cool Fans' RPM can be lowered to consume even lower power (by the fans) 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.

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 BladeSystem c-Class introduced an agile network port connectivity management method called Virtual Connect as part of the infrastructure within the c-Class enclosure, for businesses to be change-ready, while benefiting from other advantages of using server blades such as cable management.

The goal is to "hide" the switches, and move the network touch point from the back of the server blade to back of the enclosure on fewer physical ports. For example, 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 HBA-2 in Blade-2

have the hardware port addresses of WWN1 and WWN2, respectively.

A traditional FC switch in the blade enclosure will have the F-ports interfacing to the HBA's across the backplane, and the E-ports or NL-ports (on its faceplate external ports) to connect to the FC core switches, and therefore the switch has to be managed by the storage administrator to be part of a SAN fabric, as it is "seen" by the core switches as a *switch*. The Virtual Connect FC module supporting N-port identifier virtualization (NPIV) [7], the external FC port is now an N-port, where the FC core switches will "see" this port similar to the FC ports directly off a FC HBA in a server. The Virtual Connect FC module then is a FC port aggregator rather than a FC switch participating in a FC SAN fabric.

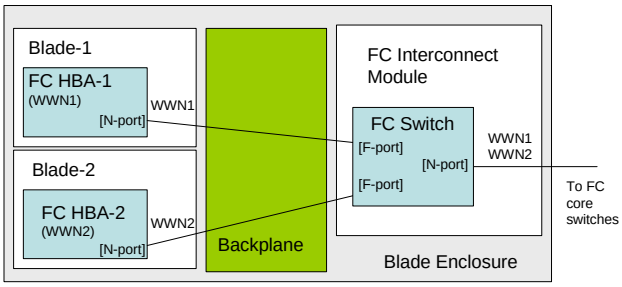

Figure-13

In other words, from 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 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.

Support for data center automation

We will use the application of NPIV by virtual machine monitors (VMM) [9] 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 transparent to the operating systems.

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 5Tbps of cross-sectional bandwidth and allows both network-semantic and memory-semantic traffics 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 a storage blade, an IO blade, etc. in addition to be different size traditional server blades.
- 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 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 [11]:

- 36% less capital equipment cost, [Note: The cost saving will vary depending on the network connectivity configuration, such as number of switches and the network cable types used for the rack-optimized servers. For example, FC optical cables between rack-optimized servers and the edge FC switches, and the associated optical transceivers will be entirely eliminated in a blade enclosure with FC switches because of the backplane.]

- 90% savings in deployment expenses,
- 69% reduction in energy consumption over a 3-year period, and
- 25% facility expenses on power, cooling and space.

CONCLUSIONS

Blades represent one of the fastest-growing segments of the computer market, with most major computing vendors adopting this approach for the benefits it offers with increased compaction, consolidation and modularity, and better management and maintenance. In this paper, we argue that blades can provide a key foundational block for various 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 (such as iLO and OA). We discussed in detail the key attributes and trade-offs in designing an optimal general-purpose infrastructure, and explained an instantiation of the HP BladeSystem c-Class infrastructure with scalable blade and interconnect bays connected across a high bandwidth backplane, along with intelligent controllers, as a general-purpose infrastructure. We also discussed specific technologies in the c-Class pertaining to power management, virtual connections, and automation and management.

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 key to such a future and we believe that this area offers a rich opportunity for more innovation for the broader architecture community.

REFERENCES

1. HP, HP BladeSystem (p-Class) Technology, HP Tech Brief, 2005.
2. 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.]
3. HP BladeSystem e-Class Overview and Features, 2004.
4. D. Desai et. al., IBM BladeCenter System Overview, IBM Journal of Research and Development, Volume 49 Number 6, 2005.
5. Egenera®, BladeFrame® System Specification, 2006.
6. IEEE *Draft* 802.3ap, Ethernet Operation over Electrical Backplanes, 20xx.
7. ANSI INCITS T11, FC N-port Identifier Virtualization (NPIV) standard
8. W. Vinson, Turning Blade Density to a Power and Cooling Advantage, presentation slides at LinuxWorld, 2006.
9. S.A. Herrod, The Future of Virtualization Technologies, ISCA, 2006.
10. HP, Power Regulator for ProLiant Servers, Tech Brief, 2nd Edition, 2006.
11. K. Quinn et. al., Forecasting Total Cost of Ownership for Initial Deployments of Server Blades, IDC, June 2006.