



Sun Constellation System: To Infinity and Beyond

Insight

Gordon Haff

14 April 2009

High performance computing today is clustered computing. Clusters can deliver performance to most types of HPC problems less expensively than monolithic specialized systems like the vector supercomputers of yore. Clusters also leverage the volumes and the pace of development associated with mainstream commercial computing. This is no small thing in an industry defined in no small part by the self-fulfilling observation known as Moore's Law.

The question today isn't really whether your supercomputer is a cluster or not. With rare exceptions, that's a given. It is. Rather, the question is whether the connections between the cluster nodes are "commodity" Gigabit Ethernet, specialized proprietary interconnects such as those found in IBM's Blue Gene/L,¹ or something in between that's faster than Ethernet while still being largely off-the-shelf. That middle tier is now dominated by InfiniBand, which has largely supplanted a fragmented landscape of performance-oriented clustering technologies such as Myricom's Myrinet, Quadrics' QsNet, and Dolphin's SCI.



Compute nodes span the gamut too. There are fat nodes (large SMP servers, often running Unix), skinny nodes (such as the pair of sub-GHz PowerPCs in Blue Gene/L), and—more commonly—nodes consisting of a basic dual-socket x86 server. However, one of the big trends, especially in larger HPC installations, is the use of blade servers. Blades buy density and general physical simplification—useful attributes when you're talking hundreds or thousands of servers.

The latest iteration of the Sun Constellation System (which is based on Sun's blade server product line) has a strong HPC bent that doubles down on what came before. The new Sun Blade X6275 server module packs two nodes—that is, two dual-socket servers—into a single server module. And it not only supports the new InfiniBand quad data rate (QDR)—which doubles the bandwidth of the prior DDR generation—but it lays QDR InfiniBand right onto the node itself. Constellation aims to be not just an HPC building block, but one that that's high-performance (in both compute and I/O domains) even by HPC standards.

Licensed to Sun Microsystems, Inc. for web posting. Do not reproduce. All opinions and conclusions herein represent the independent perspective of Illuminata and its analysts.

¹ See our [Blue Gene's Teraflop Attack](#). Although often categorized as Massively Parallel Processing (MPP), such systems are architecturally just a specialized implementation of a cluster.

Today's Blades

Blades are essentially a Lego approach to designing servers in which server modules snap into a chassis rather than sliding into a rack. They were first heavily promoted circa 2000 during the dot-com boom as a way to get more servers into a given square foot of floor space while simultaneously reducing the number of cables needed to connect all those servers together.² In keeping with the priorities of the time, their designers viewed the role of blades through the lens of the Web servers and associated Internet infrastructure. In the years since, blades moved forward, albeit in a bit of a hit and miss way as both blade builders and blade buyers struggled to figure out where unique value relative to rackmount servers truly lay.

Now, a bit under a decade later, blades' place in computing has clarified, at least for the nonce. Contrary to initial expectations, they haven't really become the standard for Internet infrastructure—at least not yet—although they certainly see use there. But megascale Internet service providers remain largely the province of stripped-down rackmount servers of which Google's home-grown creations based on a custom Intel motherboard are only the most extreme example.

Instead we see blades primarily in two distinct roles. The first largely surprised everyone. Blades were supposed to be your quintessential enterprise play. As it turns out, there's been a lot of interest in blades from the midmarket and distributed/-replicated sites where they act as a sort of "datacenter in a box" into which multiple servers, operating systems, switches, storage, and even processor architectures can be packaged together—ready to plug into a power outlet and a network connection without involving IT experts at the site—of which there often aren't any in these kinds of environments. In short, blades here can be best thought of as an integration play.

However, blades also have a major scale play. But not just any sort of scale; large numbers of servers aren't the only criterion. Blades tend to fit best when there are particular optimizations evolved—things like maximum density, extreme scale, and performant interconnects. Blades and their associated software—which often takes the form of add-on functions or modules for a vendor's server management suite (xVM Ops Center in the case of Sun)³—also provide a sort of scale-out server consolidation. Reducing the number of independent physical components in a datacenter and grouping those components into managed entities brings a level of both physical and logical consolidation. There are no hard and fast rules about when to use blades rather than rackmount servers but HPC (whether traditional academic/government, engineering departments within enterprises, or commercial workloads with HPC-like characteristics) is clearly one of the segments that is adopting blades in great numbers.

Now, these two styles of blades often reflect different faces of essentially the same product line. HP and IBM, in particular, have pushed the integration play hard—even to the point of introducing scaled-down chassis tailored for medium businesses and replicated sites. But they've also used the same product families to go after HPC opportunities—albeit typically with different hardware and software configurations.⁴

Sun Microsystems, on the other hand, has placed many of its chips on leveraging blades for massive scale. This approach is much more akin to the first generation of blades—albeit with "fatter" and faster blades—in which compute (and associated memory) is disaggregated from storage and the network. Imperatives include maximizing compute density and interconnect speed. Sun's distinctive focus is on "Red Shift" applications—those growing at faster than Moore's Law rates—which include both traditional high performance computing and newer forms of simulation and analytics. Sun blades aren't limited to

² Various modular designs dated back even further but RLX Technologies first evangelized the products that we'd recognize as blades today. See our [RLX's Slice of the Blade Server Pie](#).

³ See our [Sun Revamps its Management with xVM Ops Center](#).

⁴ See our [Get Shorty](#).

Server module	Sun Blade X6270	Sun Blade X6275
Blades per rack	48 per Sun Blade 6048 Modular System (unibody rack/chassis)	Two nodes per module; 48 modules/96 nodes per Sun Blade 6048 Modular System
Specifications	Per server module	Per node
Processor	Dual-socket Intel Nehalem EP	Dual-socket Intel Nehalem EP
Max speed grade	2.93 GHz (Xeon X5570)	2.93 GHz (Xeon X5570)
Cores per processor	4	4
DIMM slots	18 x DDR3	12 x DDR3
PCI Express interfaces	4 x PCI Express 2.0 x8	1 x PCI Express 2.0 x8
On-board networking	2 x GbE	1 x GbE QDR InfiniBand
Storage	4 x hot-swappable 2.5" HDD or SSD 16 GB Compact Flash	24 GB flash module (SATA)
Management	ILOM Service Processor	ILOM Service Processor

supercomputing. That's their primary design center, however, as the InfiniBand component of this latest announcement only further emphasizes.⁵

Sun's New Blades

Sun's latest announcement includes two new blades: the X6270 and the X6275.

The first, the Sun Blade X6270, is the more conventional. It upgrades the existing X6250 with Intel's new Xeon 5500 ("Nehalem EP") processors, adds a couple more DIMM slots (and switches from FB-DIMM to DDR3), and bumps up I/O bandwidth using the latest generation of PCI Express. Nice improvements all, but more or less what one would

expect 18 months or so down the road from the X6250's introduction.⁶

In the specific context of HPC, the X6270 will likely function primarily as a database or a storage server using either its internal drives or its quad-port SAS interface to external storage. It's not that the X6270 doesn't support high-performance processors—it does—but HPC compute nodes tend to be stripped-down configurations that major in computation and feeding that computation.

Which describes the Sun Blade X6275. The X6275 server module is a dual-node blade. In other words, two complete servers share a single server module enclosure—effectively doubling the compute density relative to an X6270. With the X6275, Sun can configure a total of 96 nodes per Sun Blade 6048 unibody chassis/rack for a total of 192 processors and 768 cores (and up to 9.2 TB of memory using 8 GB DIMMs). A number of

⁵ This shouldn't be taken to suggest that Sun blades are solely about HPC or Web 2.0. Plenty of Sun servers and blades (such as the new Sun Blade X6270) continue to target transactional business applications in which they're paired with virtualization and other platform software associated with more traditional enterprise apps. See our [Virtualization Strategies: Sun Microsystems](#).

⁶ See our [Sun's Constellation of Blades](#).

vendors offer dual-node rackmount servers, but the more common approach in blade servers has been to design some form of “half-height” blade and then plug two of them into a chassis—one on top of the other.⁷ The X6275 doesn’t give Sun a unique position in blade compute density, but it puts Sun clearly in the top tier.

Continuing on the scale theme, Sun also emphasizes its 6048 Modular System as the physical infrastructure of choice to house these blades. Typically, somewhere between about 10 to 20 blades slide vertically into a chassis, which provides shared power, cooling, and management. The chassis then slides into a standard rack. The Sun Blade 6000 chassis is in this vein.

However, Sun’s design center for its new blades is bigger than a chassis. The 6048 is a unibody design that builds the chassis directly into the rack. By eliminating rails and other mechanical components associated with a conventional rack/chassis combination, Sun estimates that it shaves about 500 pounds off the weight of a loaded chassis. With the 6048, the basic blade unit is the rack, rather than the chassis—a step function larger.⁸

This sort of density is starting to push the limits of what can be practically powered and cooled. A rack that is fully loaded with X6275 modules, memory, and I/O will be getting close to drawing (and needing to dissipate) 35 kW.⁹ By way of comparison, a typical enterprise datacenter of a few years back typically kept the power draw of a rack under 10 kW or so—and for many, more than 5 kW was considered “extreme.”

⁷ HP is one exception but its dual-node blade currently only supports low power Intel processors.

⁸ Sun first publicly discussed (on CEO Jonathan Schwartz’ blog) the Sun Blade 6048 as “C48,” initially developed for the Texas Advanced Computing Center in Austin. Although physically integrated, the “shelves” in the 6048 have independent power, cooling, and management electronics just as with a standard chassis design.

⁹ This is, to be sure, a worst case number for fully loaded configurations. In practice, the power draw and cooling load is almost certainly lower. But it’s still substantial relative to recent historical norms.

The Cool Factor

At these power densities, some form of cooling system optimization is required. Although such optimizations are often framed as an “air cooling vs. water cooling” debate, it really comes down to a discussion of where and how water carries away heat. After all, how does heat get carried away from a “normal,” air-cooled datacenter? Well, by the water flowing into the CRAC (Computer Room Air Conditioner) units at the room’s periphery. Thus, the issue is whether to bring water further into the datacenter—not whether to use it at all.

There are a wide range of possible approaches. For example, one of the premier customers for the prior generation of the Sun Constellation System, the Texas Advanced Computing Center (TACC) outside of Austin, uses in-row chillers from APC in enclosed hot aisles. Such a design allows cooling to be focused on where it is really needed rather than covering the whole facility—and, thereby, over-cooling a lot of the volume.

With this latest announcement, Sun is introducing an even more local cooling option: a Sun Cooling Door that attaches to the rear of a rack. We’ve seen a variety of such announcements over the past few years.¹⁰ This one is a bit unusual in that it comes in two variants to support two different kinds of cooling.

One flavor, the Sun Cooling Door 5600, uses R134A refrigerant and is compatible with Liebert XD pumping and chiller units. In this case, water is still only piped to the datacenter chillers; refrigerant lines then connect to the rack. Sun says that this option has the highest energy efficiency and the smallest footprint. (And it also keeps water out of the datacenter proper, which many datacenter managers prefer.)

The Sun Cooling Door 5200 is a straight chilled-water version. It connects to water sources in either the floor or the ceiling, and is intended primarily for datacenters that can simply and cost-effectively

¹⁰ See our [Water Cooling Lives! \(Sortof\)](#).

extend an existing chilled-water system—in which case this will be the least expensive approach.

Currently, the Sun Cooling Door is specifically for the 6048 unibody rack/chassis design, but Sun intends to make it available for its standard racks in the future.

InfiniBand and Latency

Now let's talk about connecting things together.

InfiniBand resulted from the 1999 merger of two competing designs: Future I/O (developed by Compaq, IBM, and HP) and Next Generation I/O or NGIO (developed by Intel, Microsoft, and Sun). From the Compaq side, the roots were derived from Tandem's ServerNet, still used within HP's Integrity NonStop servers. The collective goal was to create an industry-standard "System Area Network"—a high-performance connective fabric for datacenters. A whole mini-industry of silicon, software, host bus adapter, and switch vendors supported InfiniBand.

InfiniBand was invented primarily as an alternative to Ethernet and its associated TCP/IP software stack. TCP/IP evolved as a network and transport layer protocol that provides for reliable transmission of packets across inherently unreliable wide area data links with indeterminate transmission times—a role that it performs well. But, in the context of a datacenter fabric spanning much shorter distances and capable of a significant amount of error correction at the hardware level, Ethernet and TCP/IP just weren't (and aren't) the most efficient approach.¹¹

InfiniBand, on the other hand, was explicitly designed as a datacenter switched-fabric interconnect. InfiniBand provides for communication between nodes in a way that allows for protected direct memory access (DMA) between the communicating devices. This offers two primary advantages: low latency and low messaging overhead. Because InfiniBand can

ensure data integrity in hardware at the link layer, the higher stack levels do not need to incur the overhead associated with error handling.¹²

InfiniBand therefore doesn't take long to pass a message between the nodes in a cluster. (Messages are the fundamental way that nodes interact in a distributed memory architecture such as a cluster.) The exact length of time depends on the software stack in use—in HPC, the message passing software is often a variant of MPI—and the topology of the links and how many hops are needed to get from one end to the other. However, typical InfiniBand latencies are in the single digit microseconds; TCP/IP latencies on Gigabit Ethernet can be at least 10 times longer. The CPU overhead associated with InfiniBand's sending a message and the bandwidth it has available to transfer data also improves on basic GbE.

Now, this delta doesn't always matter. Even many of the mega-supercomputers on the TOP500 list interconnected with vanilla GbE. Workloads that involve largely autonomous processing of relatively modest datasets—often referred to as "embarrassingly parallel" workloads—don't especially benefit from faster interconnects. With tasks such as rendering a movie, individual frames can typically be processed largely in isolation.¹³

However, workloads that have to do a lot of synchronization of parallel tasks or that have to ship a lot of data around will start bottlenecking on I/O rather than computing—and that's where an optimized interconnect like InfiniBand can help. For many types of numerical simulations and modeling, the computational results associated with any point in space affect the results of other points in space so the overall calculation needs to be treated in a more integrated way—which requires more and faster messages.¹⁴

¹² "Data Center Ethernet," a set of standards initiatives that build on top of 10 GbE aims to replicate some of the technical approaches built into InfiniBand but is still in early days. See our [InfiniBand Eight Years Later](#).

¹³ The LINPACK benchmark used as the performance metric to rank the TOP500 is an example of an embarrassingly parallel workload.

¹⁴ See our [Latency Matters!](#) for a much more detailed

¹¹ This description glosses over a great many tangential threads such as offload engines and direct data placement protocols such as VI developed (but never widely used) for Ethernet and TCP/IP networks.

Latency does matter for these types of workloads.¹⁵ And that's why InfiniBand is now in 28 percent of the TOP500 list of the world's largest (publicly acknowledged) supercomputers. That's particularly striking given that the benchmark used to do the ranking, LINPACK, measures system horsepower in the very narrow context of an "embarrassingly parallel" workload that doesn't particularly reward a system for having a design whose components are more tightly integrated or communicate with each other particularly efficiently.

Majoring on InfiniBand

InfiniBand is what's unique about the Sun Constellation System. Many vendors offer InfiniBand connectivity as an option for their blades. However, Sun has truly majored on InfiniBand—even to the point of designing its own high-density switch. This announcement carries on in that vein by moving up to the new InfiniBand QDR speed grade and integrating InfiniBand right onto its X6275 nodes.¹⁶

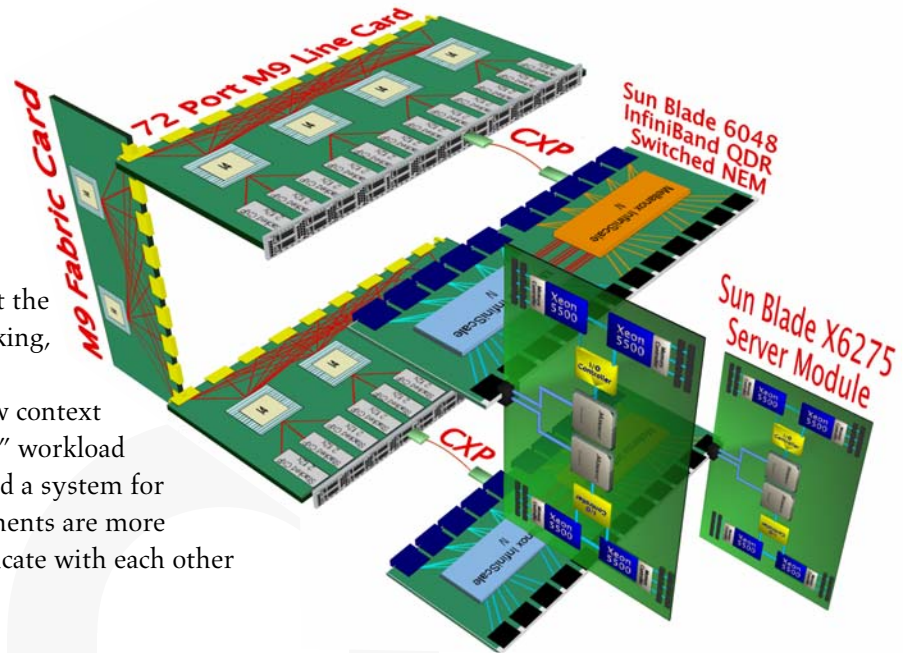
A Mellanox ConnectX QDR InfiniBand 4X HCA sits on each node of the X6275 server module; it connects to the Intel "Tylersburg" I/O hub via a PCI Express 2.0 8X connection.¹⁷ Both the x4 InfiniBand and the x8 PCI Express links have a

treatment of this topic. Many of the specific technologies covered in that note have been supplanted by newer or alternative products, but the basic discussion still applies.

¹⁵ As well as many commercial workloads, perhaps most notably database clusters, although InfiniBand has never really taken off there in a widespread way.

¹⁶ The X6270 can also connect to InfiniBand, including Sun's new QDR switch, at DDR rates using a mezzanine card (InfiniBand DDR Fabric Expansion Module). On-board InfiniBand is an option for the X6275 but Sun says that all its initial shipments will include it.

¹⁷ The 4X and the 8X refers to the number of "lanes" (i.e. serial links) in the InfiniBand and PCI Express connections respectively.



nominal bi-directional data rate (before overhead) of 64 Gbits/sec. Thus internal and external bandwidth are well-matched with each other.

From each X6275 node, the InfiniBand link travels over the midplane to a Sun Blade 6048 InfiniBand QDR Switched Network Express Module (NEM)¹⁸ in the rear of the shelf. This is effectively an embedded switch using two on-board 36-port Mellanox InfiniScale IV QDR switching chips. Each node connects at full bandwidth to a port on one of those two switch chips. Up to four NEMs can be configured per Sun Blade 6048 chassis.

The switch chips provide a one-to-one connection (with an X4 InfiniBand link) among all 24 nodes in a shelf across the midplane without using any cables. In addition, it provides eight 12X links from the NEM to the external InfiniBand "Project M9" switch.¹⁹ The links will typically be optical—copper is limited to a 7 meter distance—and use CXP connectors that were adopted as an X12 InfiniBand standard in 2008.

¹⁸ We'll just call it a NEM, though industry acronym conventions might fancy it a SB6048IBQDRSNEM.

¹⁹ To not further complicate the architectural discussion, we limit this description to the more common fabric architecture that will be used: a multi-stage non-blocking "Clos network." A 3D torus network is also possible, using the NEM's two additional X12 ports.

The Project M9 switch uses nine line cards, each with 24 physical X12 ports (to connect to the NEMs) that split into 72 X4 InfiniBand links internally. To round out the picture, nine fabric cards connect the line cards. Line cards use four InfiniScale IV cards each and the fabric cards two. Thus, each Project M9 switch can support a total of 648 nodes (9 * 72); that's almost seven full racks of X6275 server modules.

At maximum scale, up to eight of the 11U Project M9 switches can be used as part of the same network to connect a whopping 5,184 compute nodes with a claimed server-to-server latency of less than two microseconds.²⁰ Such a configuration would require 216 NEMs, each with only eight cables for fabric interconnect. If that sounds like a lot, consider how large a configuration this is—something like 70 or so racks of gear by the time you add storage, network switches, and so forth. Such a large cluster built with conventional servers connected by InfiniBand would involve thousands of cables. The benefit may not even be so much in the direct cost and labor savings of reducing cables by almost an order of magnitude, but in the greater reliability and reduced problems likely to accrue.

Conclusion

Sun's fixation on high scale design points partly reflects its worldview and biases as a company—its

²⁰ Switch latency is 300 ns. End-to-end latency: ~1.2 us

view of where growth opportunities most lie. Sun CEO Jonathan Schwartz, after all, was an early-on proponent of what today usually goes by cloud computing. And cloud computing, regardless of whatever particular definition of it you favor, inevitably suggests computing that is more consolidated and more centralized—and hence larger-scale.

With respect to blades specifically, Sun's strategy also reflects a need to differentiate itself from others who did blades earlier and more aggressively. As a result, rather than integrating third-party switches or designing SMB-friendly racks and chassis, Sun has decided to take aim squarely at Internet scale and performance computing—whether that computing happens in the “cloud” or within enterprises for whom this class of computing is a strategic asset. And in doing so, largely eschew the more mainstream path being taken by the likes of Dell, HP, and IBM.

Sun's latest blades update reflects all this and takes it even further. The primacy of InfiniBand and the emphasis placed on blade and blade cooling optimization at the rack level in this latest update makes it clear that, probably more than any other vendor, Sun looks on blades as a means to push the envelope of what can be achieved with more general-purpose designs. When Sun talks about blades, it's talking about a lot of blades.