# INFINIBAND'S DATA CENTER MARCH

## JULY 2012


It's becoming clear that the optimal interconnect architecture at every level within the computing system today is a switched fabric. With today's converged adapters and offloaded traffic "intelligence," switched solutions offer better performance and throughput than more democratically accessed bus architectures. Even the Ethernet of the future with added data center bridging and lossless delivery features is acting more and more like a fabric than the old school contentious lossy bus it's been traditionally.

One of the oldest, most successful pioneers in high-speed switched fabric technology is InfiniBand. For the last decade, InfiniBand has made steady progress in the HPC world by stitching large clusters of compute nodes closer together to accelerate system performance. Despite InfiniBand's higher bandwidth, lower latency and overall cost-efficiency, from its inception it has faced an uphill battle for broader adoption. Today, however, significant use cases and technology trends beyond HPC are emerging that are making InfiniBand an attractive, if not inevitable, choice for the core of the enterprise data center.

In this technology profile, we'll take a fresh look at where InfiniBand stands today and its growing role in the next generation data center. We'll examine important technology and solution trends where higher-speed switched fabrics are required for optimal performance in several dimensions – solutions like Big Data, web-scale applications, scale-out storage, and virtual I/O for mission-critical applications. Along the way we'll address some high-level cost and adoption risk concerns.

Data center architects need a high-speed switched fabric at the core of the new data center. Today that choice is clearly InfiniBand.

## INFINIBAND'S GROWING ROLE IN THE DATA CENTER

InfiniBand is a network interconnection protocol delivered via a flat, switched fabric architecture under centralized management. It differs from Ethernet in that it was designed from the start for extremely low latency, high throughput and lossless delivery. Although perhaps first envisioned as a way to interconnect almost everything in the data center from CPU's to edge peripherals to all the servers and storage, practically InfiniBand took root in the HPC world as the primary cluster network with a particularly high adoption in scale-out supercomputing.

InfiniBand currently sits in a broad sweet spot between motherboard-based switching like PCIe and wide area network packet switching, although there are solutions that stretch InfiniBand across the WAN (e.g. Obsidian's Longbow, Bay Microsystem's IBEx). In the data center InfiniBand is often leveraged today for "intimate" remote direct memory access (RDMA) that stitches together multiple compute nodes into HPC powerhouse systems, for front-side I/O to external high performance shared storage, for back-side I/O within high performance scale-out storage solutions, and for providing high performance intranetworking for mission-critical applications.



87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

InfiniBand is an ideal fabric solution for data center consolidation architectures where the goal is to reduce the total amount of physical compute assets deployed and bring the rest into a homogeneous managed environment. Because of its single cable premise that effectively supports multiple higher-level protocols simultaneously while providing the high bandwidth required for dense infrastructures, it becomes a "virtualized" networking solution that in-turn naturally serves as a high-performing and flexible interconnect for any highly virtualized computing environment.

## NETWORK BOTTLENECK IN THE NEXT GEN DATA CENTER

Emerging and evolving IT technologies demand ever greater I/O throughput, effectively shifting where bottlenecks occur in the total data flow through the core data center. For the last decade commodity Ethernet and Fibre Channel have been able to provide sufficient network performance in most non-HPC systems, but that is clearly changing today with technologies like these that can overwhelm a slow network fabric:

- **Remote Direct Memory Access** (RDMA) – Between Big Data clusters, virtualized and cloud server pools, and distributed applications (and of course HPC), RDMA is becoming the standard way to offload inter-process communications from CPU's and provide low latency message passing, enabling large scale-out architectures and simplifying multi-tier application design.
- **PCIe 3.0+** – The next generation of PCIe can provide I/O at more than 100Gb/s bi-directionally off the motherboard to feed those next generation chipsets. Intel's SandyBridge processors for example are spec'd to handle 12GB/s in bi-directional bandwidth.
- **Flash SSD** and increasing server-side storage distribution will soon evolve beyond read-only cache to enable distributed storage array technologies at large scales, require massive low-latency interconnection.

InfiniBand has not only survived but through a decade of challenges has clearly matured into a mainstream enterprise network fabric option. Because it continues to surpass Ethernet in bandwidth, latency, and ultimately total costs, InfiniBand has been gaining adoption slowly and steadily as a data center fabric even throughout the economic downturn. According to TOP500.org, more than half of the top 300 supercomputers in the world leverage InfiniBand, and it has become a great choice not only for HPC, but also (as it was initially envisioned) as a core fabric that runs web, cloud, and other dense scale-out data center implementations. Its original general-purpose vision as a converged data center network may be finally arriving.

## TRADITIONAL USE CASES EVOLVING

Ultimately the data center of the future offers idealized virtual IT in cloud-like service ways but built over highly dense converged physical assets. The key to building and supporting ultra-dense, highly virtualized environments is to ensure that all of the system resources can be effectively utilized at optimum levels. Currently and for the foreseeable future, CPU performance on the motherboard with ever more sockets and cores will outpace the bandwidth of networks, so that making the most of any dense data center will require the fastest interconnect fabric.

As data centers evolve, there will be a growing need to remove network and I/O bottlenecks that could idle those pricey CPUs. Performance optimization is a constant game of removing the slowest bottleneck in the total system spanning CPU, memory, storage and distributed processes. The critical point in many of today's newer data center architectures is the converged network – the fabric that interconnects all the components together (see sidebar – "Network Bottleneck in the Next Gen Data center").

The applications where InfiniBand has prospered the most have traditionally been HPC implementations – either supercomputing or supporting ultra-dense number crunching applications. However, the mainstream data center is fast evolving toward higher density and higher efficiency designs which enable businesses to more effectively apply compute, network, and



87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

storage technologies to their information problems. These dense and fluid architectures increase flexibility and enable a dynamic automation that obviates many manual IT management challenges of the past. As we talk to end users in the market today, it is clear that a few dominant new use cases are rising to the top where the practitioner's attention is more and more often drawn to InfiniBand. These use cases include:

- Big Data and Big Database Solutions
- Virtualized Cloud Infrastructures
- Web Scale Applications
- Scale-out Shared Storage

Let's take a look at each of these use cases in more depth.

## BIG DATA GETS BIG BOOST

Vendor-specific BI and scalable Big Data solutions like Oracle's Exadata have employed InfiniBand internally on the backend for years. Clearly these kinds of applications require high computational density and massive internal data flows. For architectures like Exadata that are scale-out, InfiniBand is an ideal fabric solution.

Those deploying open source Hadoop and Hadoop-like solutions should take note and evaluate how much more powerful their "white box" Big Data cluster could be if it was interconnected with InfiniBand. Since InfiniBand used within a Hadoop cluster can double the throughput or more (see sidebar "Accelerating Hadoop"), the savings from requiring fewer nodes or doing more analysis faster can more than cover the incremental fabric cost over the default traditional networking. Fewer nodes or more work per node even furthers green initiatives.

Regardless of the current hype around Big Data analysis, we are expecting to see more businesses build competitive custom applications in-house that leverage ever larger amounts of data. These new mission-critical business applications will transcend traditionally understood financial or transactional-based applications in both requirements for agility and unstructured data consumption. As long as data continues to grow in both volume and variety, high performance data movement will remain a major data center networking challenge.

## ACCELERATING HADOOP

When working with Big Data, one of the architectural challenges is efficient data access. Apache Hadoop is designed to run its distributed analysis (Map-Reduce) functions close to the data to avoid as much data movement as possible, but for many real-world algorithms there is still a lot of data that needs to be "shuffled" around the cluster during processing. The node interconnect can easily become a critical bottleneck throttling total cluster efficiency. Running IP over the latest high-rate InfiniBand (IPoIB) is one approach to improving large cluster performance.

An even better approach may be to modify Hadoop itself. Leading InfiniBand vendor Mellanox provides an Unstructured Data Accelerator (UDA) plug-in for Hadoop that enables it to take direct advantage of RDMA between nodes in the cluster. Once integrated, Hadoop seamlessly employs RDMA for merge-sort steps.

Mellanox claims huge speedups on larger datasets – double the throughput and half the job execution time per node. While we expect results may vary by analytical algorithm, it's clear that a cluster could be sized much smaller to get the same amount of work done in the same time.

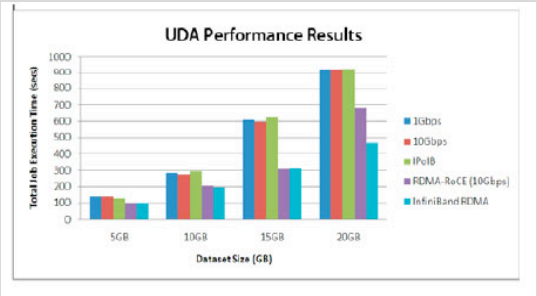


*Fig. 1 From www.mellanox.com/content/pages.php?pg=hadoop.*

## VIRTUAL I/O FOR THE VIRTUALIZED DATA CENTER

As enterprises gain more experience with virtualization, they naturally discover that once they virtualize servers they then need to continue



87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

virtualizing all the rest of the component resources in order to make the whole system more cost-effective, manageable, and dynamic. Virtualizing servers without addressing network and storage can limit the ability to dynamically migrate workloads, support mission-critical applications, and offer elastic on-demand cloud-like services. In particular, availability and performance requirements for supporting mission-critical applications within a virtual environment can drive the adoption of virtualization across the whole I/O data path including shared storage and connecting networks.

We already have virtual NIC ports, virtual HBA's, virtual network switches, and virtual LANs. But the underlying physical connectivity and required dynamic reconfiguration for cloud quality virtualization is difficult when virtual server hosts have to employ and manage multiple physical NICs, HBA's, and other adapters. Taneja Group believes that converged adapters are the future, and that the best of those adapters will connect to a high-performance fabric such as InfiniBand that can support all kinds of virtualized networking and storage protocols simultaneously in one cable.

One key use requirement for high-end virtualization is the need to support VM mobility. For a VM running a mission-critical app to easily move from one host to another, it has to seamlessly and quickly take its entire network and storage "perspective" with it wherever it goes. This means the physical fabric supporting it must be equally connected and capable at every host. If each host has multiple physical adapters of differing kinds and hierarchically (i.e. routed) different physical networking connections then "liquefying" VM movement is quite a difficult challenge. A converged, flat, network fabric like InfiniBand presents one "fat pipe" that can be logically carved out and dynamically re-provisioned as needed, which makes it ideal for dense, highly mobile virtual implementations.

## SCALE_OUT WEB NEEDS TIGHT INTERCONNECTION

Web-based businesses and those with key business applications that run at web scale - quickly becoming most large enterprises - can attack their optimization challenges by leveraging higher speed fabrics. These web-sized apps require infrastructure that not only supports virtualized, cloud-like compute resources, but additional mobility and agility to reconfigure dynamically – and preserve a margin no matter what the current data flow or interconnect requirement is. InfiniBand's flat address space here is a boon to service providers and large web-based businesses alike.

There are a number of network performance considerations, but it's clear that latency in InfiniBand has and will always be significantly smaller than the latency of Ethernet. There are many applications where throughput can be aggregated and delivered across Ethernet just as well as any other network architecture. But there are many more applications where data flows are small and multitudinous (e.g. random storage writes, memcached/RDMA, message queues). Cutting network latency in half can improve application-level performance and throughput by the same factor or more, which can mean significant competitive advantage and seriously reduced requirements for infrastructure leading to large cost savings. And when comparing production available solutions, InfiniBand continues to offer both a latency and a bandwidth edge over Ethernet and is predicted to do so into the future with an upcoming data rate of 100Gb/s expected to reach production by 2014 if not sooner.

## SHARING AND CONVERGENCE DRIVE DENSITY

As part of the move towards cloud-like computing infrastructures both private and public, the underlying compute and storage resources need to offer massive scale-out capability, usually accomplished with some kind of RAIN architecture. For example, check out key storage vendors like EMC, which offers its Isilon grid-like storage connected on the back-end with InfiniBand. It makes sense as data centers become more dense that to squeeze the most juice out of your assets, your front-side interconnect should be able to match the back-side capability. In other words, if



87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

InfiniBand is the choice of high-performance, scale-out storage subsystems internally, it should be considered for the front side connectivity to that storage too.

There is a recent movement towards integrating more large, fast storage directly in each server (e.g. Fusion IO cards). This expanding server-side storage will require an InfiniBand-like interconnect fabric if it is to ever be more than advanced cache, as it will either have to directly share data with other servers or tightly integrate with external shared-storage devices. The net effect of converging storage and servers basically leads also to converging front-side storage I/O with back-side storage I/O.

It's clear that converged fabrics have green benefits in the form of increasing density and computing throughput leading to infrastructure savings, and thusly to power and cooling benefits. InfiniBand offers a greater convergence opportunity than Ethernet even though Ethernet is the more widely adopted, general-purpose networking platform. The real fabric question for the data center should be about how to handle ever-increasing virtualization efforts and the inevitable dense-computing, cloud-building projects. These architectures will be built with more and more building block-like modules containing standardized components, rather than from widely heterogeneous compilations of scattered vendor products that might require trading off performance and cost for lowest common denominator interoperability.

## SURPRISING COST SAVINGS OF SPEED

When comparing the cost of InfiniBand to an alternative solution, you should start by considering the total physical costs of the alternative for all the various connectors, adaptors, cables, switches and racking required for networking and storage. When just the fabric and cables are totaled up, InfiniBand could cost less than half of a comparable high-speed Ethernet solution, depending on extra top-of-rack switching and SAN requirements. Remember that Ethernet is hierarchical in nature and generally requires more switches and cables just to plug everything in. Then account for the reduced power and cooling requirements, and any reduction in hard infrastructure that might be saved due to the increased performance of the network.

With a converged, flat-fabric network, by definition cabling is simpler which can lead to improved air flow. In addition, fewer mistakes can be made (when a spanning tree falls in the data center, everyone hears the screams!) than with hierarchical routed networks requiring crowded and potentially confusing cabling and switch configurations.

## INFINIBAND INDUSTRY HEALTHY AND GROWING

There has been some consolidation in the InfiniBand industry in the last few years with Mellanox emerging as the market-leading supplier of HCA's and switches. Despite some concern about the current dearth of market competition, Intel's recent acquisition of Qlogic's InfiniBand portfolio should alleviate those fears. Mellanox's revenue and production have been steadily increasing even through the economic downturn, and now Intel is back on the InfiniBand bandwagon.

Mellanox now offers virtual port adapters on their HCA's that can be configured to either Ethernet or InfiniBand as a way to help lower total investment costs and ease transition to InfiniBand over time, clearly thinking about how to help data centers more painlessly adopt InfiniBand at the core. Intel is charging back into the InfiniBand HPC market, but could easily leverage its newly acquired InfiniBand IP in several other areas. They have been bringing switching based technology closer to the CPU with every chip generation, and we would not be surprised to see a new generation of IO chipsets post-PCIe that implement a more native type of InfiniBand support.

Microsoft Windows Server 2012 has built-in leverage for RDMA through its SMB Direct feature, and we expect that derives from Microsoft's desire to continue pushing Windows Server higher in the enterprise stack. We also expect to see more native support for InfiniBand emerging in both proprietary and open source server virtualization hypervisor and cloud stacks this year.



87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

The major server vendors like HP (Converged Infrastructure) and Dell (vStart) all build and ship bundled server/InfiniBand solutions in their respective converged infrastructure stacks – and of course InfiniBand has been available in production for over a decade. This isn't new technology that already stressed IT staff have to puzzle out from scratch and pray that it "sticks". Rather, since it's proven, reliable, constantly evolving, and tested at massive scale, InfiniBand is here to stay.

Standards evolve slowly, and the bigger the standard the slower the pace. Ethernet is slowly evolving with some great new features like Data Center Bridging that will enable partitioning bigger pipes into virtual smaller ones, which can then be dedicated to specific applications. Combined with some other emerging features to create more lossless self-paced data flows, Ethernet appears to be learning a lot of the lessons that fabrics like InfiniBand pioneered. Still, the proven InfiniBand roadmap far exceeds expected advances in Ethernet performance into the conceivable future.

## TANEJA GROUP OPINION

Since the game in IT optimization is constantly shifting to address the next "biggest" obstacle, it's a wonder why some data center architects continue to play the safe but slower interoperability card of Ethernet instead of working to lessen the negative impact constricted networks have on their relatively speedy compute resources. If there are I/O bottlenecks that idle pricey CPUs in today's denser, more homogeneous data centers, now is the time to ensure the best interconnect available.

Many of the "extreme" requirements leading to InfiniBand used to only apply to HPC and a few other specialized needs, but moving forward hundreds of mainstream mission-critical applications will be hosted on denser, virtualized clouds of infrastructure with similar interconnect requirements. In a way InfiniBand is being pulled up by its bootstraps by the exigencies of virtualization, cloud and Big Data.

InfiniBand offers lower latency, vast bandwidth, a flat network, better power consumption, and overall lower costs. In virtual server environments, you can effectively run more VM's per host (because each VM will get its job done faster), gain better storage IO performance and management, and ultimately save on total server count and power, potentially enabling a lower total cost of ownership.

Because Ethernet is so entrenched for reasons such as the cost of migration and resistance to adopting new networking technology, some HCAs (adapters) today support both InfiniBand and Ethernet as virtual port configuration options. Companies looking to migrate safely or that have significant legacy infrastructure could look to these dual protocol adapters to lower adoption risk, although we believe the next generation data center core will directly implement high-speed fabrics as a matter of course.

As hardware refresh cycles pick up again with the burgeoning economy, and more physically dense, converged resources become the norm, we expect InfiniBand adoption to continue at a high rate in HPC but really accelerate its march on the data center core. The emerging use cases outlined in this profile -Big Data, web-scale applications, virtualized I/O, and scale-out storage - are collectively going to challenge the data center fabric. Next generation data center architects simply can't afford to let the fabric become a performance bottleneck across this whole new set of demands. The only fabric completely up to this challenge, in fully proven production today and supported by a believable and blisteringly fast future roadmap, is InfiniBand.



87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557