

# Network Performance Considerations in Virtualized Data Centers

*Motti Beck, Mellanox Technology Inc. and Michael Kagan, Mellanox Technologies Inc.*

## Introduction

Data center virtualization is touted as the best solution that addresses capital (CapEx) and operations (OpEx) expenses, and it is rapidly being adopted as a standard method of operations. By running multiple applications on multiple operating systems on a single server and permitting management of all servers and storage as single pools, virtualization allows data center managers to optimize the use of each server's processing power, minimize the space used, and cut overall maintenance costs. With virtualization, IT managers gain provisioning flexibility and reduce the data center Total Cost of Ownership (TCO) by avoiding massive amounts of physical cabling or manual reconfiguration, by gaining more compute power per Kilowatt hour, by reducing management overhead, and by significantly reducing the number of servers needed.

More powerful servers based on multi-core, multi-processor architecture are the core technology that enables virtualization. Whether in blade server or pizza box configurations, the average server will soon have four processors, and with each processor having eight cores, this will add up to 32 cores running in parallel.

While rising server performance and declining cost/gigabyte for storage systems are helping to drive data center performance, the overall efficiency gain of the data center is heavily dependent on its connectivity capabilities. Extending Amdahl's law to the data center means that under ideal conditions, the performance gain should be proportional to the number of cores; in other words, the computing performance is only as good as the weakest link in the system. Therefore, the goal should be to create not just powerful servers or massive storage pools, but a balanced level of performance that fully utilizes servers, storage, and the network connecting them.

With a balanced data center, performance and productivity are maximized and costs are kept to a minimum. In this article, we'll look at how to accomplish this.

## A Big Engine, Big Gas Tank, Small Fuel Line

Today's virtualized data center has a lot of raw processing performance and a lot of storage, but poor connectivity among systems limits its effective efficiency. In short, if performance is not balanced across the entire data center, it affects the efficiency and results in higher total cost of ownership. There are several factors that contribute to the performance bottleneck:

**Processing is outpacing connectivity** – The more processor cores there are in each server, the more I/O bandwidth that server needs. In the near future the typical number of cores per server will be 32. In typical database applications, each core requires between 300MB/s to 500MB/s IO bandwidth. However even if we'll assume that each core will need only 1 Gb/s of sustainable I/O bandwidth, that means that in the very near future not even 10 Gb/s of I/O will be sufficient to support true concurrent multi core operation in real time. Rather, the I/O will be a bottleneck, causing the server to wait unnecessarily, and thereby wasting energy and computing cycles. This trend has been already recognized and addressed in the PCIe 3.0 standard, and it is expected that the networking technology will “match” it (Figure 1).

- *Motti Beck is a Director of marketing at Mellanox Technologies. Motti can be reached at [Motti@mellanox.com](mailto:Motti@mellanox.com).*
- *Michael Kagan is the Chief Technology Officer at Mellanox Technologies. Michael can be reached at [Michael@Mellanox.com](mailto:Michael@Mellanox.com)*

| PCIe Type | Raw Bit Rate | Interconnect Bandwidth | Bandwidth Lane Direction | Total Bandwidth for x16 Link |
|-----------|--------------|------------------------|--------------------------|------------------------------|
| PCIe 1.x  | 2.5GT/s      | 2Gb/s                  | ~250MB/s                 | ~8GB/s                       |
| PCIe 2.0  | 5.0GT/s      | 4Gb/s                  | ~500MB/s                 | ~16GB/s                      |
| PCIe 3.0  | 8.0GT/s      | 8Gb/s                  | ~1GB/s                   | ~32GB/s                      |

Figure 1: PCIe evolution.

**The need for network convergence** – while server virtualization means fewer physical servers to maintain, the use of multiple networks to carry different types of traffic creates the need for a higher level of network consolidation. Typically, this includes having separate networks for low-latency server clustering, client-server connectivity, storage connectivity and data center management. The clustering network is typically based on InfiniBand, while the client-server management is based on the Ethernet infrastructure. The storage network is based on Fiber Channel or iSCSI (Figure 2). This model increases the overall cost of network operations because operators must maintain different types of network adaptors, increasing number of switch ports and different cables. It also results in lower flexibility and more complex management in the data center. Now days, when data center performance is vital to the company business (and in many cases, it's the business itself), there is a clear need for networking consolidation solutions that complement the server consolidation enabled by server virtualization technology.

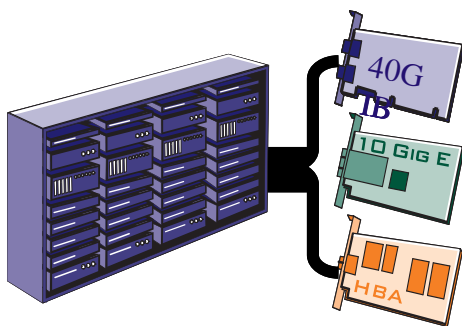


Figure 2: Typical data center network solution

Supporting convergence places a lot of requirements on the data center fabric. First, the fabric must support the different data traffic types, which complicates the overall implementation. Storage traffic usually consists of moving large packets, while server-to-server communication requires both large and small size packets, but the smaller ones are typically latency sensitive. Small size packets are typically being used for real-time server synchronization and the per-packet protocol processing costs associated with bulk data transfer are high when using small packets. So, the converged network must include the right mechanism to support those needs. Also, since the converged network must carry storage traffic, the fabric must be lossless. These needs are being added to the Ethernet standard today. InfiniBand already supports these capabilities, as they were part of its original standard specifications. Fiber Channel over Ethernet (FCoE) or Fiber Channel over InfiniBand (FCoIB) can be used for data center connectivity convergence.

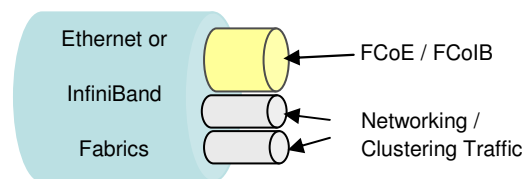


Figure 3: With FCoE/FCoIB, the same fabric carries Fiber Channel along with networking and clustering traffic

In addition of supporting the functional convergence needs, the unified fabric also needs to support higher bandwidth, since it should accommodate not just the server-to-server communication but also the server-to-storage transactions that are typically very bandwidth-intensive. In the case of multi cores server virtualization, not having enough bandwidth to support the server-to-storage traffic makes the data center less efficient; power is wasted because the servers must wait longer than necessary to the read/write the data to storage.

**Virtualization technology requires higher fabric bandwidth** – There is also additional overhead from the virtualization framework itself. To begin with, the hypervisor needs significant network bandwidth to operate and to effectively manage dozens or hundreds of data center servers. For example, the VMware vMotion utility that moves a job from an overloaded server to a less busy server uses network bandwidth in doing this. In addition, Hypervisors now offer new capabilities like Storage vMotion that enables storage virtualization, or High Availability (HA), which include Fault Tolerance mechanisms that create a “shadow” task to each “master” task with automatic failover to the shadow task if the master fails. Also regular maintenance task require a shutdown of specific system, which means massive data transfer over the converged network. All of those tasks put a heavy load on networking resources by increasing the amount of data being passed among servers and storage.

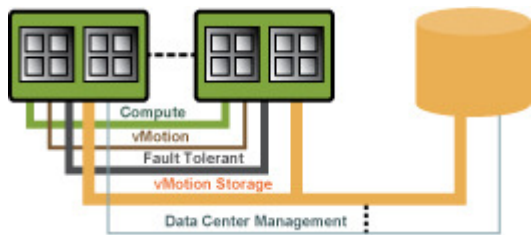


Figure 4: Hypervisor-based data center

**“Blocking” switch architectures limit compute performance** – In today’s data centers, there are also limitations to Ethernet switching efficiency. In order to prevent loops, Ethernet switches are using the Spanning Tree protocol which may cause “contentions” and packets may be lost and retransmit. Energy-efficient data centers can use non-blocking “Fat Tree” technology to minimize unnecessary processing delays. In addition, Spanning Tree is very hard to scale and it will be very difficult to be used in the large data centers that will be needed for providing cloud computing services. This is why, Fat Tree capabilities are being addressed in a new Ethernet standard, but as it is for today’s InfiniBand is (and has been for years) the only fabric technology that enjoys massive deployment of Fat Tree switching.

So we see that I/O performance lies at the heart of an efficient data center, and not having an adequate fabric bandwidth to support the traffic demand, results in higher TCO.

## Letting the Numbers Choose

With the inherent challenges of multi-core servers and virtualization technology, IT architects must be careful in choosing I/O and networking options to prevent the occurrence of bottlenecks. This will undoubtedly require moving beyond Gigabit Ethernet speeds, but there are many options. The following example shows how the choice of networking technology impacts total cost of ownership in data center investments (Figure 5). This example assumes 500 servers that have 16 cores each and a sustainable IO need of 400KB/s per core.

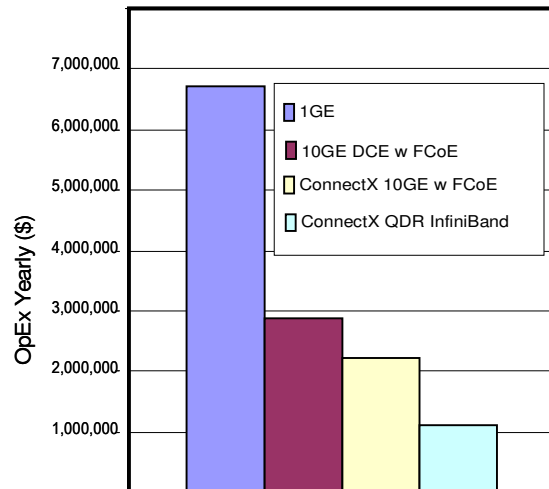


Figure 5: Data Center delta in OpEx for different Interconnects using total of 500 of 16 cores Servers, assuming sustainable IO need of 400MB per core

As we can see, InfiniBand 40 Gb/s (quad data rate, or QDR) networks offer the lowest year-to-year total cost of ownership. InfiniBand is less expensive than any Ethernet based fabric, including DCE with Fiber Channel over Ethernet (FCoE). Good example of taking a full advantage of the InfiniBand advantages is Oracle’s Exadata Data Base appliance. Using the InfiniBand enabled Oracle to come with a very unique “Data Center in the Box” architecture that solved all the bandwidth issues that such a data intensive application needs, and still to maintain the lowest Total Cost of Ownership. However, in less demanding applications that the total IO needs on a

specific server is less than 10Gb/s and there is no sensitivity to latency or to higher CPU utilization, other fabrics, like the 10GE position to be more efficient.

### Why InfiniBand?

InfiniBand's technological advantages as the converged lossless data center fabric derive from several different factors.

**Throughput** – Running at a speed of 40Gb/s, InfiniBand is much faster than any other standard fabric that is available today. Packets are moving among servers or from servers to storage arrive much more quickly, reducing or even eliminating the need for servers to wait for transmissions to complete and thus using less Watts to do the job.

**Latency** – Using its advanced Channel IO Virtualization (CIOV) mechanism and its Remote Direct memory Access (RDMA) capabilities, InfiniBand currently has application latency of less than one microsecond, which is up to six times less than offered by any 10 Gigabit Ethernet controllers available today in the market (Figure 6).

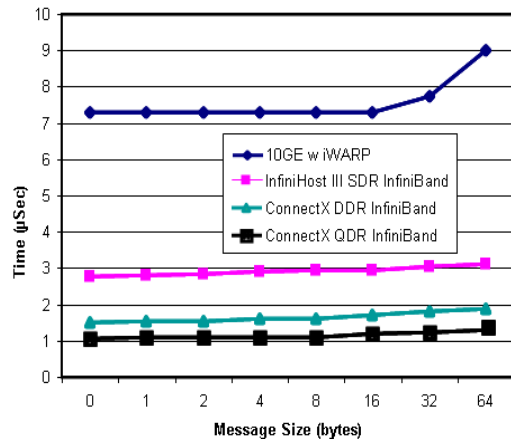


Figure 6: Latency for different interconnects. Ohio State University MPI-level latency test presented in Sonoma 2008 ([http://mvapich.cse.ohio-state.edu/publications/openfabrics\\_pres.shtml](http://mvapich.cse.ohio-state.edu/publications/openfabrics_pres.shtml))

**Power consumption** – With such a large advantage in available bandwidth, InfiniBand is positioned to consume less power compare to any other fabric. Compared to 10GE, InfiniBand QDR consumes three times less power per Gbps (same power but 3x throughput). In addition, supporting 32Gb/s

sustainable connectivity requires only one QDR InfiniBand port, compared to more than 3 ports of 10GE. This reduces the number of adapters, cables, and switches across the data center, thereby reducing power and management costs (Figure 7).

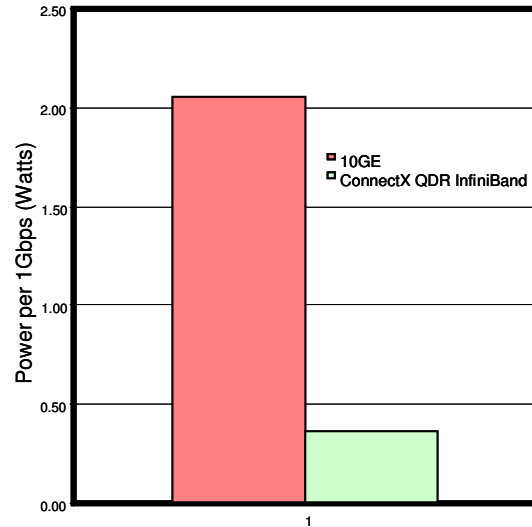


Figure 7: Typical power consumption per 1Gbps traffic including connectivity to SAN

**Management** – With InfiniBand's converged data center network technology, IT managers spend very little time managing the connections among servers, switches, and storage. Rather than having to manage two or three different protocols as traffic moves across the data center, IT personnel can rely solely on InfiniBand.

### Reaching the Goal

Organizations can gain significant improvements in data center operational efficiencies by deploying server virtualization technologies along with a converged network. Using virtualization along with fabric convergence optimizes performance and flexibility while minimizing costs. As we have seen, this requires balanced performance across the data center infrastructure with the reduction or elimination of processing bottlenecks.

As the market moves toward multi-core, multi-processor servers, converged networks and ever-increasing storage pools, InfiniBand is the only proven fabric technology that eliminates processing bottlenecks, and reduces power and management costs in the bargain.