CLABBY ANALYTICS

Research Report

How RDMA Is Helping to Drive the Sale of 10/40Gbps Switches

Executive Summary

RDMA (remote direct memory access) is a facility that enables data to be moved from the memory/storage of one system to the memory or storage of another system – *at line speed*.

The types of workloads that benefit from the use of RDMA are network-intensive applications *that suffer from bandwidth/latency-related data retrieval issues*. These include:

- Large scale simulations, rendering, large scale software compilation, streaming analytics and trading decisions – the kinds of applications found most often in massively parallel, high performance computing (HPC);
- Hyper-appliance, hyperconverged and hyperscale environment where large volumes of data needs to be moved between servers and associated storage; and,
- Workloads where network latency slows database performance and interferes with virtual machine (VM) density. For instance, in this report, we show how RDMA improves Microsoft SQL Server database performance and increases VM density.

RDMA accomplishes this fast memory-to-memory/storage-to-storage data transfer by avoiding overhead associated with operating system controls, device driver calls, and by not making copies of the data in transit. Specialized RDMA adapters/switches handle communications transport at the hardware level. With data management and communications transport obstacles out of the way, network latency decreases greatly, while the performance of network intensive applications improves.

Further, by bypassing the operating system, RDMA has no impact on the CPU. By lightening communications processing, RDMA users are able to reclaim processing power (in the report described above, testing shows that 30% of a CPU's processing power can be reclaimed). This benefits the enterprise in two ways: 1) processors can do more work – yielding a better return-on-investment; and, 2) fewer processors are needed to process workloads – which can result in saving BIG MONEY on software licenses (because software is usually priced by the number of CPU cores being used).

To further improve performance, systems and network managers are now building faster networks – moving from 1Gbps (gigabits per second) Ethernet adapters/switches to 10Gbps and 40Gbps Ethernet and/or to InfiniBand adapters/switches. RDMA is helping drive the adoption of faster adapters/switches.

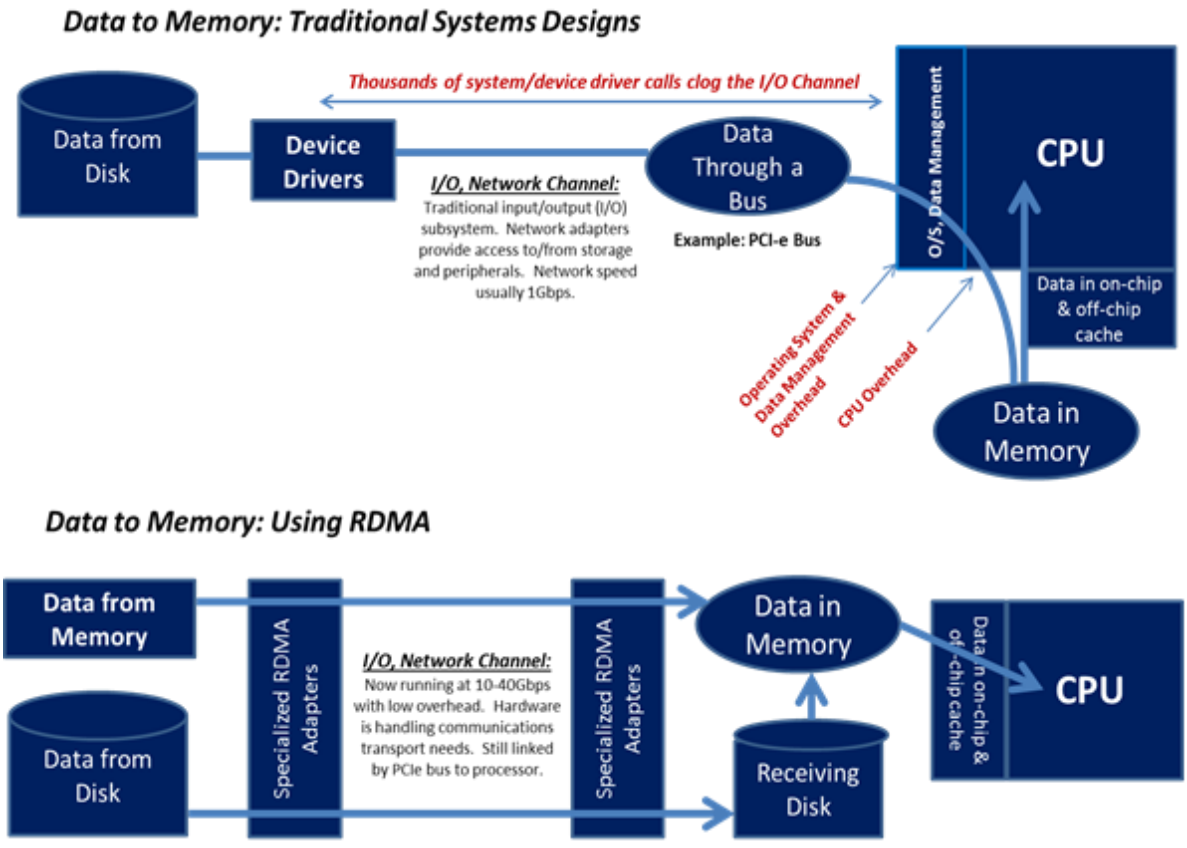
In this *Research Report*, *Clabby Analytics* takes a closer look at several RDMA-related market trends. We are now seeing:

- Rapid adoption of RDMA-enabled 40Gbps Ethernet adapters and switches (see this report that describes "***dramatic uptake in 40GbE data center adoption***");
- ***A shift in system designs that threatens blade deployment*** as information technology (IT) executives move to hyper-appliance, hyperconverged and hyperscale designs;

- Rapid adoption of RDMA in ***high performance computing (HPC) environments***;
- ***Accelerated cloud performance*** due to faster delivery of cloud services such as backups and data replication (which assists in data recovery); and,
- ***New ecosystem development*** activities around RDMA.
    - For instance, Microsoft, Violin Memory and Mellanox have joined together to create an RDMA-enabled server/storage environment (see this report). We believe that Microsoft's end of life for Windows 2003 will encourage users to upgrade to Windows 2012 – an operating environment that includes code that takes advantage of RDMA, Violin Memory Flash storage and Mellanox adapters and switches.

A Closer Look at RDMA – It Is About Eliminating Bottlenecks to Move Data More Quickly
In traditional system designs, the network and I/O subsystem is often seen as a major performance bottleneck. Lack of enough network bandwidth combined with thousands upon thousands of system/device driver calls serve to bog down network performance. When data finally arrives to a system, operating system and data management functions are performed as that data is placed into memory. CPU cycles are also burned in this communications process. Ultimately the data arrives in memory where it can then be placed into cache for processing by the CPU (see Figure 1).

Figure 1: Traditional Disk-to-Memory Versus RDMA Data Passing



Source: Clabby Analytics – February, 2015

Now consider how data is moved from storage or memory using RDMA. In short, data is passed to specialized RDMA adapters (hardware) that handle communications transport. That data is then sent to an address on the receiving system (memory or storage). The operating system need not become involved in managing this data transfer – nor do CPU cycles need to be burned managing communications. ***Figure 1 also makes the point that with faster data delivery to memory or storage, network bandwidth should be changed to exploit RDMA*** – moving from the typical 1Gbps found in many of today's data centers to 10 or 40 (or more) Gbps.

> ***With low overhead and faster networks, RDMA can move data tremendously faster than traditional networked environments. The bigger the pipes (10, 40, 100 Gbps), the faster data can be moved from disk-to-disk or memory-to-memory. Combine this with no copying of data and little CPU involvement in the data transfer to disk or memory, and the result is that RDMA-enabled systems can perform orders of magnitude faster than traditional designs.***

Impact on System Designs: Blades, Hyper-appliance, Hyper Converged, Hyperscale, HPC and the Cloud

As shown in the previous section, RDMA is all about reducing overhead in order to move data faster between systems and storage environments. From our perspective, RDMA – *when combined with bigger pipes* (faster networking such as 40Gbps adapters and switches) – has a major positive performance impact on applications and databases as data is served more rapidly to memory and storage devices.

Here's what we are seeing as a result of the growing adoption of

- Blade chassis will need to be redesigned to handle speeds faster than 10Gbps;
- Hyper-appliances, which now use Mellanox adapters and switches, will start to replace blade architecture;
- Hyperconverged systems, which also use Mellanox adapters and switches, will also replace blade server/storage architectures;
- Hyperscale system designers will continue to adopt high-speed RDMA enabled switches and adapters.
- HPC vendors, such as Dell, are also already making use of Mellanox adapters and switches in their designs; and,
- The use of 10, 25, 40 and 50Gbps adapters/switches will increase in cloud environments in order to facilitate faster delivery of data-centric cloud services (such as back-up and data recovery).

First – Some Definitions

All of the bullet points in the previous section warrant definition:

- ***Blade servers*** are compressed server environments that fit on boards that load into chassis. These chassis usually do not host storage, but do have a 10Gbps backplane (and sometimes higher speeds) for communications within the chassis as well as to external devices. Slots are also available for higher speed network solutions;

- ***Hyper-appliances*** are workload-specific appliances that focus on workloads such as transaction processing, analytics, Hadoop workloads and more. Examples of these appliances include: IBM PureData System for Transactions; IBM PureData System for Analytics; IBM PureData System for Operational Analytics; and HP Converged System for Big Data. It can be argued Oracle's Exalogic Elastic Cloud (a general load-balanced virtualization environment) also fits in this category because it has been specifically optimized for certain workloads. Likewise, we could argue that HP's ConvergedSystem for Client Virtualization (designed specifically for virtual desktop infrastructure deployments) is a hyper-appliance;
- ***Hyperconverged systems*** are those that combine servers, storage and networks into integrated offerings that allow for easy scaling. These designs are more general purpose than workload-specific hyper-appliances, as we argue in this report that compares IBM's hyper appliances to Oracle's Exadata Database Machine (a fine example of a hyperconverged environment). Hewlett Packard also offers its line of HP Converged Systems includings HP ConvergedSystem for Virtualization, HP CloudSystem and HP ConvergedSystem for Collaboration. Further, Lenovo and EMC have partnered to offer converged infrastructure solutions including VSPEX by Flex System for Private Cloud and VSPEX by Flex System for VDI. And, finally, EMC's VMAX and Teradata also offer converged system product families.

Note: From an ecosystem perspective, most of the above-mentioned vendors are using Mellanox InfiniBand switches and adapter as standard in their integrated hyperconverged systems offerings. This includes Oracle, IBM, Lenovo, EMC (VMAX), Teradata and Hewlett-Packard. All of these vendors are well positioned to exploit RDMA transfers – now, and in the future. Further, Red Hat, Oracle and Microsoft already offer RDMA application program interfaces that make it possible for various product offerings to take advantage of RDMA memory and data transfer facilities.

As for storage, Fujitsu, Hyperscale, SaaS scale-out systems have all chosen Mellanox InfiniBand to accelerate storage performance. Further, faster solid state storage (which outperforms mechanical drives) is driving demand for high-speed networking.

The bottom line: enterprises are embracing faster networking for system-to-system and storage-to-storage communications. RDMA will accelerate data transfer on these faster networks.

It is also important to note the way storage is managed in hyperconverged environments. In traditional system designs storage is usually managed in a siloed fashion. Hyperconverged environments are now starting to make heavier use of software defined storage (SDS) to abstract, pool and automate storage functions – essentially breaking-down the storage silo and making storage an integral part of the systems design. SDS is also particularly effective in helping to eliminate single points of failure (because other storage resources are available in the pool if needed).

- ***Hyperscale environments*** allows for systems, storage and networks to be scaled independently. The key difference between hyperscale environments and hyperconverged environments is that hyperconverged environments integrate compute nodes, storage and networks into a singular design – whereas hyperscale

treats each environment separately (so scale can be added to each silo as needed). Typical applications found in this environment include some Big Data analytics, Hadoop and map reduce environments. These environments need to be scaled and quickly provisioned in order to deal with large data sets – and data is constantly on the move across hyperscale systems, storage and networks;

- ***High Performance Computing*** environments are exactly as their name suggests – computer designs that maximize parallel computing functions by providing high performance links between systems and storage. Some of these new workloads include computational analysis, upstream/downstream processing, next-generation genomics, satellite ground stations, video capture and surveillance, 3-D computer modeling, social media analysis, data mining/unstructured information analysis, financial "tick" data analysis, and large-scale real-time customer relationship management environments. Melanox adapters and switches are used heavily in these environments – and the use of RDMA is also increasing dramatically in this market space.
- ***Clouds –*** clouds can be defined by the services that they offer such as Software-as-a-Service, Infrastructure-as-a-Service, Platform-as-a-Service and others. RDMA can be used in cloud environments to move data quickly to various locations (to improve cloud performance) and to serve as a means to quickly backup data as part of a cloud data recovery plan.

What RDMA Means to Blade Server Architecture

As far back as September, 2012, Clabby Analytics has tried to draw attention to shortcomings in blade chassis designs that will limit blade performance over time. More specifically, in this report we pointed out that the blade chassis of most leading blade vendors were limited in terms of communications speeds to 10Gbps. This report went on to cite the research of Daniel Bowers of Ideas International who wrote an excellent research report entitled "Is Your Blade Chassis Obsolete?" in which he described the hardwire connections within various blade designs. As he put it "most of today's blade enclosures have one thing in common: a roughly rectangular circuit board midplane built into the enclosure that connects the individual blade servers to I/O devices through a hard-wired "mesh" of copper connections". Bowers contended that the "style and number of connections in this mesh defines how much I/O bandwidth the blade chassis can handle". We observed at that time that all of leading blade server vendors could handle 10Gb Ethernet — and even converged Ethernet — but only one (IBM's Flex System chassis) was designed to handle higher speeds such as the 40Gb Ethernet and beyond (this has since changed as Oracle, VCE, Hewlett-Packard and Dell all offer converged systems).

Several vendors still make blade chassis with 10Gbps backplane limitations. But since 2012, we have seen several new "converged systems" and hyper-appliance designs come to market that will support speeds of 40Gbps. As mentioned above, IBM's Flex System (now Lenovo's Flex System architecture) was one of the first to do so. But a closer look at Oracle's Exadata and Exalogic system designs shows high-speed Melanox adapters as integral design components. The same holds true for IBM's PureData systems design – and likewise for sever Hewlett-Packard models.

The key point here is that the increasing use of RDMA will cause IT buyers to reexamine blade chassis limitations. For those that have network-intensive applications and are looking for significantly higher performance, a move to converged system architecture should result in exponentially faster processing.

What RDMA Means in Hyper-Appliance, Hyperconverged, Hyperscale and HPC Environments
Very simply, RDMA provides a means for data to be moved rapidly from server memory to server memory or storage device to storage device in each of these designs. To truly maximize performance in each of these environments, RDMA needs to be able to exploit high speed switches.

In a recent conversation with Mellanox Technologies we learned that the company had shipped hundreds of thousands of 40Gbps Ethernet endpoints in Q4, 2014. In fact, a closer look at the company's earning revealed very strong growth in the company's high-speed adapter/-switch businesses – and also progress in broadening its high-speed Ethernet and InfiniBand offerings. Highlights of the company's earnings call include:

- Successful testing of its ConnectX-4 Silicon and STM 25, 40, 50 and 100Gbps Ethernet in Mellanox labs. Mellanox also stressed that it is the first telecommunications vendor to offer full end-to-end 100 gigabit per second InfiniBand interconnect solutions and 25, 50 and 100Gbps Ethernet adapters;
- Mellanox started shipping its 100Gbps InfiniBand solution in Q4, 2014 – with the Minnesota Supercomputing Institute announcing that it will use the company's 100Gbps InfiniBand solution as its high speed network backbone;
- In Q4, 2014, Mellanox also noted that the United States Department of Energy plans to use its 100Gbps InfiniBand product offering to network systems to be installed at Oak Ridge National Laboratory and at Lawrence Livermore National Laboratory;
- At Supercomputing 2014, Mellanox demonstrated its Switch-IB family of 100Gbps InfiniBand switches, setting a world record with port-to-port latency of less than 90 nanosecond. Further, the company's 36-port switch delivers doubled throughput per port while offering half the latency of previous generation InfiniBand switches – while also using less power;
- Mellanox also demonstrated 4, 6 and 8-meter copper cables running at 100Gbps – helping to greatly decrease data center capital expenses using copper instead of fibre; and,
- The company also announced that in the ranking of the most recent TOP500 supercomputers, Mellanox switches were used to connect 45% of the systems.

According to Mellanox, RDMA is a key technology that is being deployed in the largest data centers in the world to improve the efficiency and resource utilization of expensive compute and storage resources. InfiniBand has been widely adopted as the core data fabric within database, analytics, and storage

appliances because of this efficiency of RDMA transport. RDMA is now becoming pervasive with both InfiniBand and RoCE deployments as industry standard networks for cloud and hyperscale enterprise deployments. RDMA enables storage to deliver higher performance at lower cost.

Summary Observations

Well balanced systems should keep processors, memory and I/O very busy ─ but should not overwhelm them. If processing a given workload overwhelms systems resources (processor, memory, I/O), then it is time to consider upgrading the processor/system environment; finding ways to provision additional headroom; and/or find ways to decrease overhead and latency. RDMA attacks this network latency problem.

With RDMA, overhead related to data management, to system calls (operating systems), to copying data, and to CPU communications processing is reduced. RDMA data/memory transfers are not burdened with thousands of device driver/operating system calls nor with data copying – making it possible for data to be transferred between systems at line speed. So, with overhead greatly reduced, the new focal point becomes network bandwidth and related speed. Vendors building hyperconverged systems and storage, and vendors and IT buyers building hyperscale and HPC environments seem well aware of the need for faster data transfer rates – and are building high-speed InfiniBand and Ethernet switching into their product offerings or data center deployments.

We found that many vendors, and several large supercomputing customers are standardizing on Mellanox Technologies high-speed Ethernet and InfiniBand switches and adapters to rapidly move data within and between systems. We are finding these switches and adapters in converged systems such as Oracle's Exalogic and Exalytics systems; in IBM's PureSystems; in EMC/Lenovo joint offerings and in high-end Hewlett-Packard x86 servers. Further, large accounts such as Chevron, Viacom, JPMorgan, Comcast and Airbus are building high-speed hyperscale environments using Mellanox technologies. And the company's own Website contains links to other customers of varying sizes from different geographies – all attesting to major performance gains, energy savings and more resulting from the use of Mellanox products. So, it was only natural for us to approach Mellanox with questions regarding the impact of RDMA on driving higher speed networks into systems and data center designs. And Mellanox confirmed our original assumption: "RDMA is helping to drive the sale of 10 and 40Gbps adapters and switches in hyper-appliance, hyper Converged, hyperscale, HPC and the cloud markets.

Clabby Analytics
http://www.clabbyanalytics.com
Telephone: 001 (207) 846-6662