RoCE Accelerates Data Center Performance, Cost Efficiency, and Scalability

January 2017

RoCE

The need to process and transport massive blocks of data for big data analytics and IoT mobile users continues to grow. Along with the increase in data, ever growing database and cloud transactions are forcing data center managers to adopt new approaches to data center design. Fortunately, the acceleration of network and data access speeds combined with the rise of multicore CPUs creates new design options that can make data centers more efficient, less expensive, more flexible, and future proof.

"RoCE (RDMA over Converged Ethernet, pronounced "Rocky") provides a seamless, low overhead, scalable way to solve the TCP/IP I/O bottleneck with minimal extra infrastructure."

The transition from the slow access speeds of spinning HDDs (hard disk drives) to next generation SSDs (solid state disk drives) has led to a major I/O bottleneck that has limited data center performance. Fortunately faster networks at 25, 50, and 100 Gb/s have begun to address this I/O bottleneck. These new high speed networks, together with the emergence of even faster SSDs and promising new NVM (non-volatile memory) technologies are shifting the bottleneck to the CPU. Conventional data transfer over TCP/IP requires transport processing and interrupts that congests the CPU and exacerbates this bottleneck. To overcome precisely these limitations, the InfiniBand Trade Association (IBTA) introduced RDMA (Remote Direct Memory Access) technology for HPC (high performance computing) applications 15 years ago. As its name implies, RDMA manages memory I/O in a way that doesn't consume CPU resources.

RoCE (RDMA over Converged Ethernet, pronounced "Rocky") provides a seamless, low overhead, scalable way to solve the TCP/IP I/O bottleneck with minimal extra infrastructure.

Figure 1 compares storage and memory access speeds from legacy HDD to ubiquitous SSD to lightning fast emerging NVM technologies like 3D XPoint technology coming from Intel and Micron Technologies all the way down to L1 Cache speeds. RoCE is only limited by Ethernet speeds: 10, 25, 40, 50, and 100 Gb/s being installed with 200 and 400 Gb/s on the roadmap. In just a few years NVM will merge with RAM. Since Ethernet speeds are setting the pace, when the boundary between NVM and RAM disappears, RoCE will be essential.

RoCE™

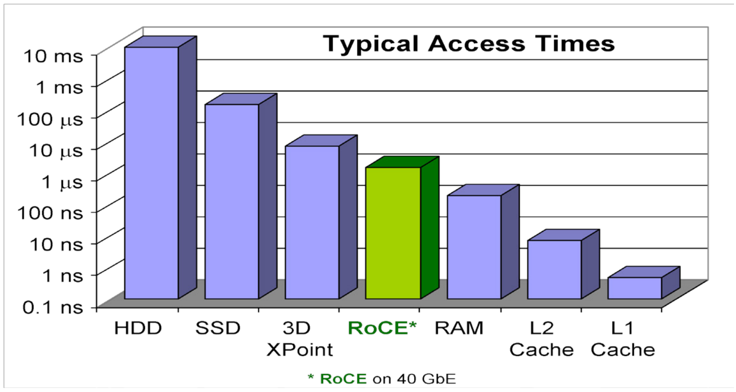

Figure 1: Comparison of typical storage and memory access times.

RoCE Frees the CPU From Performing Storage Access

Conventional access to NVM swamps the CPU with TCP/IP processing requests. RDMA technology opens up the TCP/IP bottleneck and accelerates application performance by bypassing the CPU, Figure 2. RoCE, RDMA over Converged Ethernet, provides the benefits of RDMA by introducing RoCE adapter cards.

"RDMA transfers are performed in a RoCE adapter with no involvement by the OS."

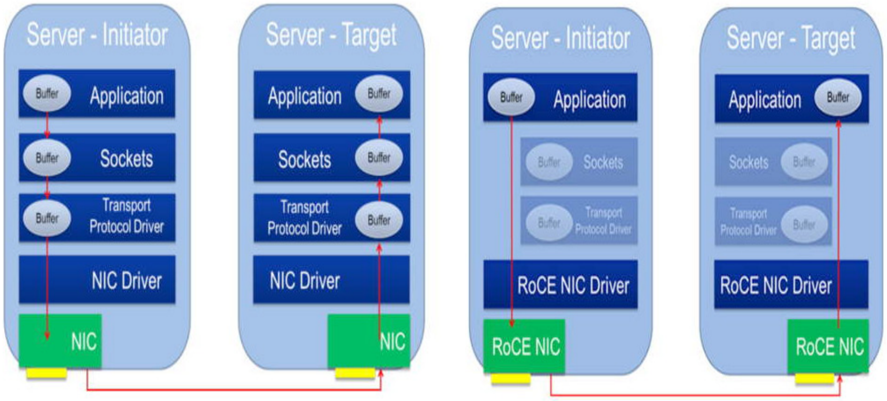

Figure 2: On the left, conventional I/O through TCP/IP requires CPU cycles but on the right RoCE bypasses the CPU.

In a presentation at the 2014 Open Networking Summit conference Albert Greenberg, Director of Development at Azure Networking said, "RoCE enabled at 40 GbE for Windows Azure Storage, achieved 40 Gb/s… Just so we're clear, [that's] 40 Gb/s of I/O with 0% CPU." He went on to say, "Microsoft's Azure allows groups of servers to share their memory with latency that approaches the channel group delay. Think of it as cloud scale, hyper-converged, efficient I/O."[1]

Direct memory access (DMA) has always been a built-in feature of PCs. DMA allows internal peripheral components—disk drive controllers, sound, graphics, and network cards, etc—to read from and write to system memory independently of the CPU. RDMA generalizes DMA to network adapters so that data can be transferred between applications on different servers without passing through the CPU or the main memory path of TCP/IP (transmission control protocol). RDMA completely removes the TCP/IP overhead—from the send request + receipt acknowledgment + permission + memory buffering sequence—bypassing the operating system; instead, the NIC (network interface controller) accesses memory directly.

RDMA transfers are performed in a RoCE adapter with no involvement by the OS. The RoCE adapter uses key features originally defined for InfiniBand but operating on the standard Ethernet physical, link, and IP networking layers.

As the following case studies show, RoCE increases IOPS (input/output operations per second), reduces latency, accelerates message transfer, increases file storage, and reduces cost.

Case Study - RoCE Delivers More IOPS With Less Than Half the CPU of TCP/IP

Mellanox Technologies worked with Dell Computers to improve access to storage without taxing the CPU. They compared the performance of Microsoft Storage Spaces SMB (server message block) running TCP/IP and RoCE-based SMB Direct, both on 10 Gigabit Ethernet (GbE) networks. Figure 3a compares IOPS performance of TCP/IP and RoCE running different read, write, and OLTP (online transaction processing) workloads; RoCE throughput is consistently higher in every case. Figure 3b shows that 10 GbE TCP/IP uses more than twice the CPU of 10 GbE RoCE[2] for both read and write operations.

"The data access latency for TCP/IP traffic increased by a factor of ten, four times more than the latency increase for RoCE traffic."

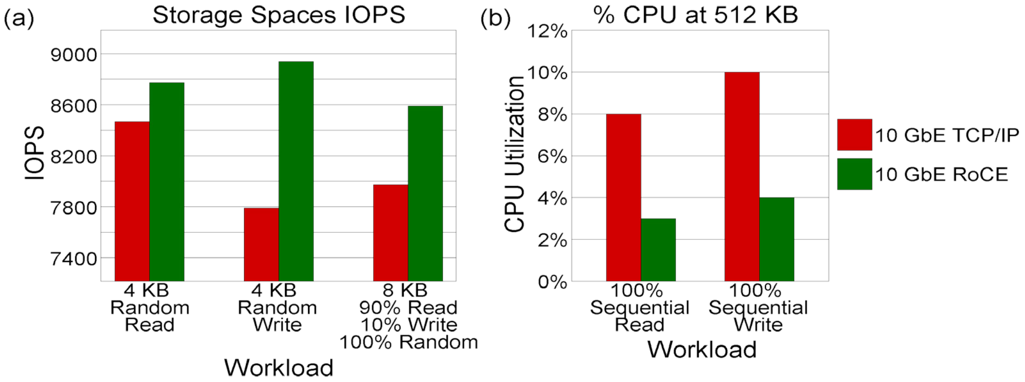

Figure 3: Comparison of 10 GbE TCP/IP and 10 GbE RoCE; (a) IOPS performance and (b) CPU use.

Case Study - RoCE Cuts Data Access Latency by 400%

Many applications access large amounts of data from the Microsoft SMB file system. In so doing, they generate high data rate traffic running over TCP/IP that can lead to congestion and degrade application throughput, transaction rates, and response time. In a test of the most widely deployed RDMA technology on Windows Server 2012, a SMB Direct remote file server was configured to function like local storage over 40 Gb/s RoCE with applications that used Microsoft SQL Server and Microsoft Storage Server. Traffic was maintained at the assigned bandwidth in order to compare congestion between TCP/IP and RoCE based data flows.

The data access latency for TCP/IP traffic increased by a factor of ten, four times more than the latency increase for RoCE traffic[3].

Case Study - RoCE Increases File Storage I/O by 82%

In another example of Windows Server 2012 using RoCE, Lenovo configured a system that blew away its benchmarks. By using Microsoft SMB Direct with a RoCE NIC (R-NIC) on a 10 GbE network, Lenovo[4]:

- Increased file storage I/O by 82%
- Reduced I/O response time by up to 70%
- Improved server power efficiency up to 80%, squeezing 750 more IOPs per watt.

Case Study - RoCE Cuts Cost 50%

Data center managers can dramatically improve data processing speed by using RoCE to implement modern scalable software-defined storage architectures. A major financial institution that struggled to manage its increasing storage infrastructure demands installed a 40 GbE RoCE alternative to its FC (Fibre Channel) based SAN. Running Storage Spaces over SMB Direct they were able to[5]:

- Accelerate storage access by at least 250% and as high as 900%
- Run virtual machines 880% faster with 12.5 times faster storage migration
- Reduce the number of IO ports by 400%
- Lower storage costs by 50%—the cost savings resulted from Microsoft's Storage Spaces compatibility with industry-standard hardware, cutting the cost of proprietary hardware and software required by the traditional FC-based SAN.

> *"Data center managers can dramatically improve data processing speed by using RoCE to implement modern scalable software-defined storage architectures."*

RoCE Alters Design Optimization

SAP applications running on DB2 pureScale[6] keep large amounts of application data in memory. A Linux kernel uses the remaining memory to optimize I/O operations. Access to data in memory is almost 100,000 times faster than HDD access.

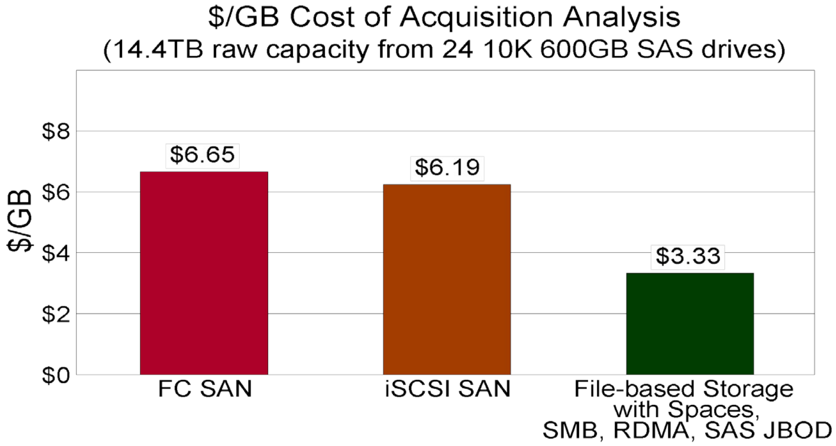

Figure 4: Comparison of cost/GB between Fibre Channel, iSCSi, and RoCE.

IBM's DB2 pureScale, a scalable database cluster based on RDMA over a high-speed interconnect is the ideal RoCE application. Each cluster member has direct memory-based access to the centralized locking and caching services of the pureScale server. A cluster CF (caching facility) server manages the synchronization of data and locks between all DB2 pureScale members to maintain data consistency so that cluster members can continue their statement execution.

Communication to and from the CFs is predicated on the low latency of RDMA network fabric. By using RDMA on either InfiniBand or RoCE, the CF can directly access memory on any member and any member can access memory on the CF without using CPU time.

When an SQL processor issues a page read request, if the CF has the page cached it sends that page to that member where it's read into the LBF (local buffer pool) which eliminates the need to read it from disk. When the CF receives a page write request, the modified page is sent to the CF and cached in the GBP (group buffer pool). At this point, the CF has the most recent version of the page and proceeds to mark invalid any other copies of the page that exist in the LBPs of other members. The ability to mark these pages invalid without interrupting the members enables scalability of write-intensive applications.

"Without RoCE, the many-low-powered design would require I/O transactions across networks that would reduce the availability of the already low powered servers."

Transactional workloads dictate the computing power required by SAP applications. CPU power can be provided by either a few high power database servers or many low power servers.

RoCE alters the usability versus flexibility optimization point of the many-low-powered versus few-high-powered design options. On the one hand, the design with a few high powered servers reduces the number of operating system images and database servers that must be maintained. On the other hand, a design with many lower powered servers offers the ability to quickly and inexpensively scale the total CPU power to fit immediate need. The few-high-powered design can handle most of its I/O at full internal bus data rates. Without RoCE, the many-low-powered design would require I/O transactions across networks that would reduce the availability of the already low powered servers.

Since the CF shares data pages in the global cache across the cluster through RoCE at full GbE data rates without using processor time or I/O cycles, the transaction overhead disappears and the data center design optimization point shifts decidedly in favor of scalable designs with large numbers of less expensive, lower CPU power database servers.

RoCE Future-Proofs the Data Center

We've seen several impressive but conventional RoCE implementations. The RoCE story gets much more interesting when the lines between storage and memory blur.

Systems developed to access spinning disks were designed with storage bottlenecks in mind. Figure 5 shows the dramatic improvement in sustained access bandwidth that can be achieved by replacing SATA HDDs with NVMe SSDs; just one NVMe SSD provides the sustained bandwidth of about 50 HDDs. Impressive as this may be, in the next few years storage performance will catapult ahead and make RDMA in general and RoCE in particular data center necessities.

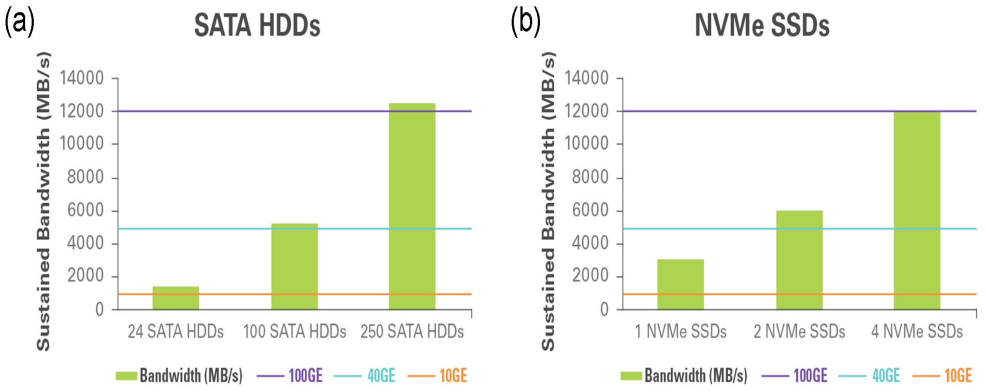

Figure 5: Sustained bandwidth for (a) SATA HDDs and (b) NVMe SSDs

"RoCE gives you the benefits of InfiniBand on Ethernet infrastructure with the same cabling and switching that you already have."

For example, Intel and Micron Technologies are collaborating on the development of a new form of phase-change based solid state NVM that is expected to be 1,000 times faster than flash. It's called 3D XPoint or Optane. The idea is to create random access technology in a three dimensional design with perpendicular wires connecting submicroscopic columns that are 10 times more dense than conventional memory. The XPoint ("cross" point) die has two layers and a crossbar design. Where NAND data is addressed by multi-kilobyte blocks, 3D XPoint NVM can be addressed byte by byte with latency of seven microseconds or less. XPoint chips can be mounted on DIMMs, right on the memory bus, blurring the distinction between storage and memory.

If 3D XPoint and similar new persistent memory technologies can be used to merge RAM with what we tend to think of as disk storage, everything changes. Include RoCE in this vision running on the 400 Gb/s networks that will emerge at the same time, and the distinction between local and cloud computing becomes even hazier.

Introducing such fast NVM technology on full speed RoCE changes the very concept of what we think of as a data center: storage-class persistent memory with microsecond latency SSDs on an NVMe DAS (direct access storage) fabric shared by any number of hot-swappable servers.

Converting Your Data Center to RoCE

RoCE gives you the benefits of InfiniBand on Ethernet infrastructure with the same cabling and switching that you already have. The upgrade path consists primarily of installing RoCE adapter cards and drivers. All Ethernet NICs require RoCE network adapter cards. RoCE drivers are available in Red Hat, SUSE, Microsoft Windows and other common operating systems.

The RoCE conversion can be a wholesale transformation or a staged upgrade. As demand grows, you can simply add servers that are already equipped with RoCE cards. However, it's important to keep in mind that RoCE adapters can only communicate with other RoCE adapters; configurations that attempt to mix adapter types, say RoCE adapters combined with InfiniBand adapters, will probably revert to TCP/IP.

To get all the benefits of RDMA, you have to commit to either RoCE, InfiniBand, or some other RDMA technology at the design stage.

Since SMB Direct is the most widely deployed RoCE technology, let's take a close look at a couple of designs. SMB is an application-layer network protocol with shared access to files and serial ports so that remote file servers work like local storage in applications that use Microsoft SQL Server and Microsoft Storage Server.

"RoCE extends NVMe to NVMe-oF (NVMe over Fabrics) so that it can access remote storage systems at GbE speeds with latency comparable to local attached NVMe."

Deployment of SMB Direct with RoCE adapters requires:

- Two or more computers running Windows Server 2012 R2 or Windows Server 2012.
- One or more RoCE network adapters for each server.
- One or more switches for the appropriate GbE implementation with Priority Flow Control (PFC) capability.
- Cabling, typically enhanced small form-factor pluggable (SFP+) connectors for lower rates like 10 or 25 GbE or QSFP connectors for higher rates like 40, 50, or 100 GbE switches.

Here are two examples:

Two computers using 10 GbE RoCE network adapters: the minimum configuration with a single file server and server running Hyper-V requires two RoCE network adapters and a network cable.

Ten computers using dual 100 GbE RoCE network adapters: A high performance private cloud configured with a two-node file server cluster and an eight-node Hyper-V cluster. Two 100 GbE RoCE network adapters are needed for each system which adds up to 20 RoCE network adapters with a 20-port switch. Since the network adapters are limited by the PCIe (Peripheral Component Interconnect Express) host bus adapter, this system requires 16 lane PCIe 3.0 slots to achieve maximum performance.

Other common RoCE deployments include NVMe-oF, NFS, and iSER:

- NVMe (NVM Express) helps eliminate the data throughput and latency bottlenecks caused by SAS and SATA interfaces that were designed to support spinning HDDs in local storage systems. RoCE extends NVMe to NVMe-oF (NVMe over Fabrics) so that it can access remote storage systems at GbE speeds with latency comparable to local attached NVMe.
- NFS (Network File System) over RoCE improves application performance on NFS servers much the way RoCE improves SMB Direct.
- The iSER (iSCSI extensions for RDMA) network protocol extends iSCSI (internet small computer system interface) to RoCE.

The RoCE specification[7] can be downloaded from the IBTA (InfiniBand Trade Association) website, http://www.InfiniBandta.org. The spec was first released in 2010 and began large scale deployments in 2013. As this paper is produced, the latest draft is Release 1.3, March 2015.

The RoCE Opportunity

The RoCE concept is simple: RDMA access to memory and storage across Ethernet infrastructure without the CPU processing required by TCP/IP or the equipment required by InfiniBand. RoCE is the obvious choice to improve computing performance in a way that offers efficient, inexpensive scalability.

- **For Cloud Computing:** RoCE offers efficient, scalable clustering and higher performance virtualized servers in VMWare, Red Hat KVM, Citrix Xen, Microsoft, Amazon EC2, Google App Engine.
- **For Data Storage**: RoCE delivers higher throughput on systems like Microsoft SMD Direct and Lustre.
- **For Data Warehousing:** RoCE enables significantly higher job operations per second, linear scaling with cluster size, maintains table scan time in the face of exponential growth in database table sizes, and is ideal for Oracle RAC, IBM DB2 PureScale, and Microsoft SQL.
- **Financial Services:** RoCE unleashes scalable CPU performance on low latency applications like Tibco, Wombat/NYSE, IBM WebSphere MQ, Red Hat MRG, and 29West/Informatica.
- **Web 2.0:** RoCE minimizes response time, maximizes jobs per second, and enables highly scalable infrastructure designs. It's ideal for applications like Hadoop, Memcached, Eucalyptus, and Cassandra.

"RoCE is the obvious choice to improve computing performance in a way that offers efficient, inexpensive scalability."

Three emerging trends present data managers a terrific opportunity: First, RoCE is inseparable from Ethernet and 100 GbE is deploying now and 400 GbE is on the horizon. Second, NVM technology is about to blur the distinction between storage and memory. And third, compared to the improvement rate of Ethernet and NVM, CPU core processing power is increasing slowly. By implementing RoCE now, whether through a staged upgrade or a complete transformation, you can open the CPU bottleneck and meet demand as needed with the ideal number of processors.

References

[1] “Windows Azure: Scaling SDN in the Public Cloud,” Albert Greenberg, Keynote Address to Open Networking Summit, 2014.

[2] “Double Your Storage System Efficiency,” Solutions Brief from Mellanox Technologies, 2015.

[3] “Best Practices for Deployments using DCB and RoCE,” White Paper from Emulex and Cisco, 2015.

[4] “Lenovo Accelerates File Storage I/O by up to 82 Percent,” Solutions Brief from Emulex, 2015.

[5] “Major Financial Institution Keeps its Data Center Efficiency One Generation Ahead,” Case Study from Mellanox Technologies, 2015.

[6] “SAP Applications with the DB2 pureScale Feature on SUSE Linux Enterprise Server and IBM System,” T. Ziegler et. al with contributors from IBM, SAP, and Novell, SAP May 2011.

[7] “The InfiniBand Trade Association Architecture Specification, Volume 1, Version 1.3 Annex A 17: RoCEv2,” The InifiniBand Trade Organization, 2015.

About The RoCE Initiative

RoCE™

The RoCE Initiative promotes RDMA over Converged Ethernet (RoCE) awareness, technical education and reference solutions for high performance Ethernet topologies in traditional and cloud-based data centers. Leading RoCE technology providers are contributing to the Initiative through the delivery of case studies and white papers, as well as sponsorship of webinars and other events. For more information, visit www.RoCEInitiative.org.