Enabling the Modern Data Center – RDMA for the Enterprise

INFINIBAND℠
TRADE ASSOCIATION

Highly Efficient Fabric Technology for Today's and Tomorrow's Workloads

May 2019

Preface

This book is a companion to Introduction to InfiniBand™ for End Users published by the InfiniBand™ Trade Association (IBTA). Introduction to InfiniBand for End Users focuses on the technology behind the high-performance network used by science and industry. Furthermore, it provides insight into its value in different applications, such as High-performance Computing (HPC), and enterprise and cloud data centers, plus how to approach designing for InfiniBand networks.

While HPC centers have widely adopted InfiniBand for its high performance and low latency, InfiniBand and its related Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) are equally valuable in any data center environment where efficient computing is in demand. Thus, this book expands on the application and value of RDMA in data centers serving the enterprise, the cloud, and the World Wide Web.

Who Should Read This Book

This book is written for data center architects, IT managers, Chief Technology Officers, and Chief Information Officers—key technologists and decision makers for the enterprise. Readers should have the essential knowledge of how a network serves the data center and its users, the standard network software stack model, and other basic concepts of the network infrastructure. Having this foundation will help you apply the value of RDMA to issues faced in today's data center.

A Fresh Look at I/O in the Data Center

The establishment many years ago of distributed processing and load balancing in data centers enabled a significant increase in data center efficiency. The advent of server virtualization increases this efficiency by enabling the allocation of compute, storage and networking resources in the infrastructure. Architectural changes in the data center are also demanding we take a new look at I/O models to address long-standing issues, such as improving resiliency and scalability, providing greater flexibility in allocating server resources, or facilitating cloud computing usage or other deployment models. HPC deployments around the world have already addressed many of these issues and paved the way for using similar solutions in the enterprises'

data centers. A survey over the last ten years of the Top500* list (www.top500.org) of the world's fastest computers illustrates how low-latency, efficient interconnects have largely transformed the way super-computing clusters are built.

So how can we begin taking a fresh look at I/O?

First of all, it helps to acknowledge that IT has accepted and taken advantage of many disruptive technologies over the years—and significantly benefited from them. Virtualization and moving away from the one-server, one-application paradigm has been among them. I/O buses have evolved from the parallel PCI*, to PCI-x*, to the serial PCIe* bus, now in its third generation. We've moved from 10/100 Mbps server network cards to 1 Gbps to 10 Gbps over several years, and now the industry is providing 25, 40, 50, 100 and 200 Gbps server NICs. Each migration requires a significant impact in terms of IT having to replace or upgrade each server with new hardware. But the investments are worth the work.

Secondly, to help take a fresh look at I/O, it helps to be willing to expel all our traditional assumptions about how I/O works, such as there is one dominant I/O interface, the OS owns the network resources, kernel involvement is required, data must be copied between application and kernel space, the underlying media is a disk drive, etc. These have been long-standing ideas we've grown to accept as normal. But are these necessary to an efficient I/O model? Not necessarily.

InfiniBand and RoCE Networking Standards

InfiniBand and RoCE architectures are standards that define Remote Direct Memory Access, or RDMA, over high-speed fabrics. RDMA is a method of directly accessing application memory on distributed compute or storage nodes in a cluster of servers. Like other data networking protocols, RDMA requires both a software stack and hardware layer to exchange data. Unlike the common TCP/IP over Ethernet, RDMA enables applications to directly access remote application memories without the assistance of the CPU processing the data in the kernel.

In this book, "InfiniBand" refers to InfiniBand adapter and switch hardware. "RoCE" (RDMA over Converged Ethernet) refers to Ethernet adapter and switch hardware that support IBTA RDMA. Both technologies utilize a software layer available through all major OS distributors or customized by application programmers. There are other variants of RDMA, and other networking technologies that incorporate similar elements, but these are outside the scope of this book. For more information, visit the IBTA website at http://www.infinibandta.org/.

Contents

Contents continued

Section 1 – The Case for RDMA in the Enterprise

A Look at Today's Enterprise Data Centers

In the last ten years, enterprise data center architects have seen a flurry of innovative new technologies, practices, and usages driven by both software and hardware advancements. These have delivered a multitude of benefits to businesses, while simultaneously introducing new challenges. Consider these examples:

- Virtualization has enabled dramatic consolidation of workloads onto fewer servers, reducing capital and operating costs. But, running several Virtual Machines (VM) on a single server can significantly increase data communications traffic, potentially creating bottlenecks at the host OS and server for I/O access.

- Virtual desktops enable centrally located business resources in the data center, making it easier to manage both hardware and software, compared to scattered desktop systems and applications. Users, though, might experience lags as they all try to access their data at once—especially on the same server.

- The Software-defined Data Center (SDDC) and private clouds have turned data centers and IT shops into valuable service delivery engines with self-service portals, enabling users to quickly and easily provision and access the business resources they need. But, networks and storage must now keep up with much more flexible data center architectures.

- SDDC-based Hyper-convergence solutions enable greater flexibility through software-defined services, placing much more pressure on I/O in the server.

- Flash, SATA/SAS SSD, NVMe, Persistent Memory (PMEM) and NVMe-oF based All Flash Array (AFA) storage solutions deliver much faster performance, exposing or exacerbating bandwidth bottlenecks and high latency sensitivities in traditional network connections.

- The Internet of Things (IoT) Edge computing and the continued growth usage of mobile devices, have opened an entirely new domain where hosted compute and storage resources support remote devices that access them, facing rising congestion at the network and storage layer in the data center.

Productivity benefits from these new usage models and technologies are realizable only with a balanced infrastructure that offers optimal performance in terms of CPU, memory and storage utilization. An unbalanced system can reduce the data center's effectiveness and efficiency as latencies increase and the overall solutions effectiveness decreases.

Leveraging the Network for Efficiency

As mentioned before, servers are not the only data center component that can boost efficiency. In fact, adding compute capability alone can slow down work because of increased network congestion. Achieving higher efficiency heavily depends on having a balanced system where all of the compute, memory and storage resources are fully utilized. The overall network, from protocol stacks down through adapters, switches, and network links, is a critical part of that balance.

When the network is self-sustained and minimizes the use of the CPU to run the data communication task, it maximizes throughput and minimizes latency, enabling services to run more efficiently. With a balanced architecture, more work can get done with fewer resources. Fewer resources mean lower cost and power/cooling demand.

10 Gbps network connections that used to be very common in the core of yesterday's data centers and even distributed across the enterprise have become obsolete, since, a modern 32-core server, which includes NVMe and PMEM often needs at least 25 Gbps to be effective. The networking industry continually works at increasing throughput and reducing latency. The industry now offers 25, 40, 50, 100 and even 200 Gbps Ethernet with RoCE capability; InfiniBand vendors just introduced 200 Gbps connections with application-level server-to-server latency below 600 nanoseconds. These are the types of performance levels that the data center of the near future will require.

To support new and next-generation technologies and usages, data centers need efficient, low-latency networks. Today, RDMA enables the highest efficiency possible in the enterprise data center.

Fully understanding RDMA requires revisiting several traditional concepts about network I/O.

The OS View of I/O

Ethernet and TCP/IP have a long history in the data center. They formed the common view of the network stack until RDMA came along in the late 1990s. To understand the benefits of RDMA-capable networks, it helps to unravel what has long been assumed—I/O is OS-centric.

In the traditional view, within the compute platform, the OS typically owns the network hardware resources and the software stack. Applications create sockets to request use of the network resource from the OS, and the OS copies the message before it sends it to the network adapter to be streamed across the wire. On the receiving end, the OS once again handles the data, investigating the target and eventually copying it into the application's user space (Figure 1). This view is about the platform and the hardware.

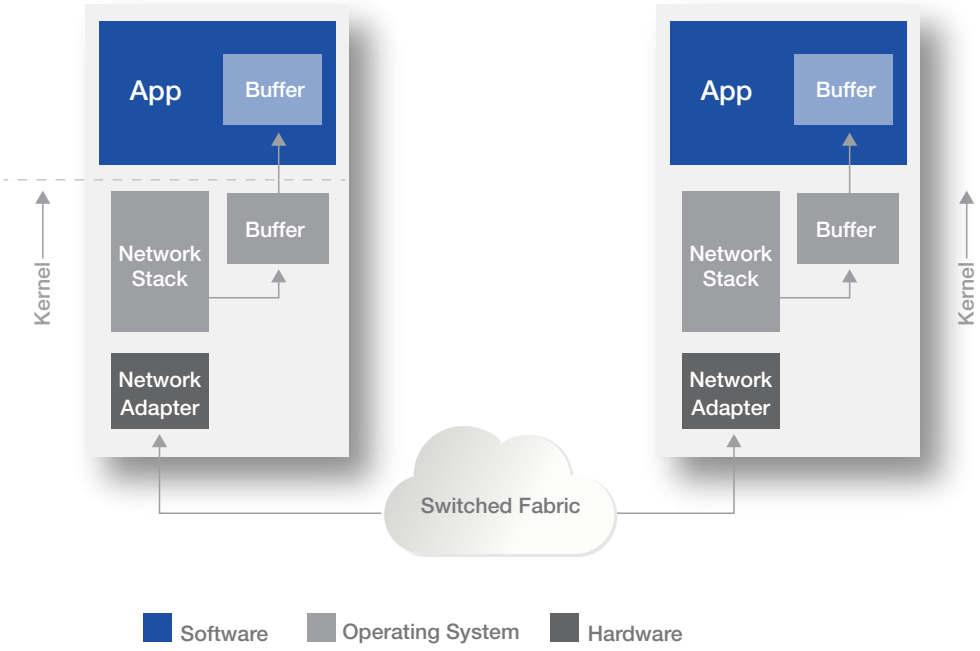

Figure 1 – The OS handles traffic in a traditional top-down networking flow with multiple data copies

The OS-centric view was conceived when PCs and servers (hardware) were occasionally connected islands providing services to a sea of users. It is somewhat analogous to how the community telephone operator of the not too distant past handled—and sometimes listened in on—every phone call. There is, however, another view of I/O.

An Application's View of I/O

An application's view of the network is quite different. Today, data centers (and users) are interconnected consumers, resources, and services in a global archipelago needing instant, and frequent to constant access to each other. They share data, borrow other resources, and split up minor to major tasks within the data center or across the globe. With this view, it's not about the hardware; it's about applications (and users).

From the application's point of view, anything that slows down the transaction or uses CPU cycles is just in the way. Messages and data need to reach another application's memory space, quickly, without delay, wherever the other application is located–in the rack, across the room, or around the campus. Make it happen! Because in today's world of transactions and competitiveness, time is money. Financial services can actually place a dollar-value price tag on a nanosecond of delay in their transactions, and commercial web site operators know a delay of a few seconds can result in the loss of a reader or a sale.

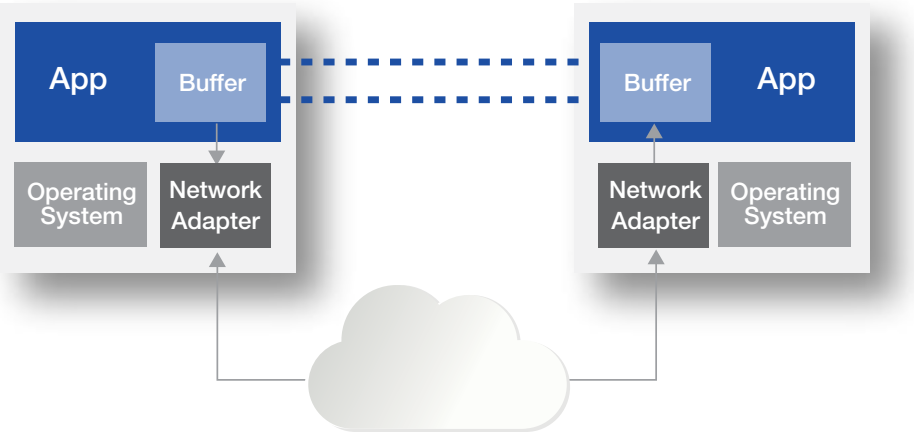

Figure 2 *– The application has direct access to another application's memory in RDMA*

Remote Direct Memory Access (RDMA)

RDMA uses the notion of Channel I/O. Instead of using valuable CPU processing time to handle communications between the application and the network, RDMA directly passes data (files, messages, blocks, etc.) between different application memory spaces, eliminating CPU involvement. (For a detailed discussion on RDMA and InfiniBand Architecture, see Introduction to InfiniBand for End Users, published by the InfiniBand Trade Association.)

Remember our community telephone operator in our earlier example? RDMA is the equivalent of the direct dialing capability that we now have.

In RDMA implementations today, data centers can realize bandwidths up to 100 Gbps with high throughput and very high efficiency (i.e., nearly no CPU overhead required by the transport). Eliminating the OS from data handling adds considerable business benefit in the data center. By eliminating transport CPU overhead, RDMA offload frees the CPU for the other important work needed in the data center—computing. Consider what that means in terms of reducing server count and the overall operations and costs impact of fewer servers.

It's been repeatedly pointed out that the application's view of I/O is as a source of data. RDMA can also be used to connect to storage systems. The idea of eliminating multiple networks, favoring a single resource, is not new to data center managers. RDMA enables convergence into a single, high-speed, low-latency transport for nearly any connectivity, instead of building a separate and expensive network just for storage traffic. Examples of such application are described in ***Introduction to InfiniBand for End Users***.

The Business Case for RDMA

RDMA technology is fascinating in what it can do. But business decisions are not solely based on technology. The beneficial impacts due to fewer servers, improved network efficiency, and long-term investment protection are worth taking a look at.

Reducing Server Count

One of the key differentiators of RDMA is its ability to bypass the CPU in I/O operations. Just as virtualization improves CPU utilization for computing by applying unused CPU cycles to do useful work, RDMA transfers CPU cycles from communication to computing. And, it boosts interconnect performance and efficiency, while reducing latency. Modern applications use more network traffic, at higher speeds than ever before, and without RDMA, this requires increasing amounts of CPU cycles to manage.

Cloud data centers pay for CPU cycles in the form of servers to enable their services. Purchase and operation of servers consumes a large part of the data center budget. Making available more cycles per server enables more VMs, applications, and users per server, which means fewer servers needed for the same workload, or added resources that can support greater workloads. Fewer servers also mean less power and cooling required, less floor space needed, etc.

Improving Network Efficiency

Network fabrics must deliver all the bandwidth required by the applications running in the data center. Overcoming slow links and inefficient topologies means using additional equipment to maintain balance in the system. The most efficient data centers use the most efficient fabrics.

Improving the Topology

Data center server and network topologies are changing; they're looking more like server clusters every day. In today's virtualized data center, traffic increasingly flows "East-West" (i.e., server-to-server or server-to-storage) vs. "North-South" (i.e., user-to-server or server-to-user). That is, there's a higher percentage of sharing among compute nodes than traffic traversing the entire network as seen in traditional client-server (one-to-one) usages. Thus, industry analysts in the data center market segments, and developers and manufacturers in the industry, have identified a growing need to flatten the data center network to make it more efficient and effective. Network equipment manufacturers now market solutions that support this approach. InfiniBand was designed with clusters in mind—applications talking directly to applications. Thus, its architecture has long realized the very network topology that is being called for to make data centers more efficient.

Creating More Efficient Switching

Traditional network topologies are not centrally managed. Network switching often relies on the Spanning Tree Protocol (STP) to choose a best pathway through a network. This algorithm often results in many underutilized ports and unnecessary traffic congestion, simply because STP blocks all but one path between servers.

The InfiniBand Architecture was the first open standard to define a centrally managed fabric. Thus, network managers can optimize RDMA routes for particular applications, data types, and endpoints. Ports are not blocked, unless the network manager does so, reducing or eliminating the need for oversubscription and the associated blocking. At present, this type of control

for RDMA is provided whether running on an InfiniBand or Ethernet transport (using RoCE) using Software-defined Network (SDN) architecture.

RDMA Delivers Performance and Efficiency

InfiniBand and RoCE can achieve a high level of performance partly because, unlike traditional Ethernet, these networks were designed to be lossless from the beginning[1] . It natively provides the flow control and guaranteed delivery required by today's latency-sensitive data streams and high-priority transactions, while directly writing and reading a target application's memory space. Data is successfully delivered the first time and in order, eliminating the retransmissions that are characteristic of traditional TCP over Ethernet. This allows much more traffic to cross the fabric, because retransmissions are not taking up bandwidth or increasing latency.

Leveraging Ethernet with RDMA to Protect Investments

Many data centers are committed to Ethernet. Gigabit Ethernet is common in most traditional data centers, with 10 and higher speed Gigabit Ethernet being deployed in racks. Data Center Bridging (DCB), now a standard feature in most switches, adds losslessness to Ethernet networks. The newer technologies boost performance, but still rely on Ethernet with its CPU involvement and CPU context switches with every packet transmission. Fortunately, Ethernet can leverage RDMA over Converged Ethernet (RoCE) to protect existing investments, gain increased bandwidth, and increase CPU productivity. And, RoCE, utilizing Data Center Bridging's Per Priority Flow Control (PFC) or Explicit Congestion Notification (ECN), can achieve performance levels similar to InfiniBand. Both PFC and ECN are widely available in today's data center.

Advancing Bandwidth Roadmap

RDMA is independent from the speed the data is transferred over the wire, so as RoCE and InfiniBand technologies advances so will RDMA performance. RoCE is widespread at 10 Gbps with 25, 40, 50, 100 and 200 Gbps starting to be deployed. Today HDR 200 Gbps InfiniBand is being widely deployed. Both technologies have strong roadmaps [2,3] .

[1] Lossless and lossy are not qualitative judgments about transports.They are standard terms indicating how, not how well, a transport works. While lossy, Ethernet is still an incredibly reliable and simple transport, which is why it is so ubiquitous.
[2] http://www.ethernetalliance.org/roadmap/
[3] https://www.infinibandta.org/about-infiniband/

Transitions

Disruptive technologies, like virtualization, have brought significant improvements to the data center. The cost of disruption is worth the benefit. RDMA is as disruptive a technology as any other innovative solution. But that doesn't mean its adoption must have a major impact on IT operations and costs. Just like virtualization and the migration of internal server I/O to PCIe, transitioning networks in the data center to RDMA can be done strategically to have a low impact with high benefit over time. There are many paths to RDMA. The route a business takes is driven by its needs and budgets. Similarly, the path any data center takes to RDMA depends on its needs and objectives over the long term and the investment it is willing to make along the way. We will look at these different paths in Section 3. But first, let's look at some key RDMA aspects in Section 2. It is these key aspects that make it possible to take different paths that eventually converge to acheive the benefits of RDMA.

Section 2 – RDMA Illustrated

An RDMA model is very similar to the traditional OSI model; and, in fact, RDMA can be implemented using either a traditional Ethernet network or InfiniBand fabric. RDMA does communicate differently from TCP/IP over Ethernet, but that's what makes it so efficient and fast and beneficial to data centers.

This section will illustrate the key aspects of RDMA, because understanding these points will help managers make informed decisions about how and where to deploy the technology within the data center.

Choices Mean Flexibility

RDMA enables application-to-application communications, so applications can read and write each other's' memory spaces. The InfiniBand Trade Association's Introduction to InfiniBand for End Users describes the technology in more detail. However, a basic understanding of the architecture is enough to see how data centers can intelligently and efficiently implement RDMA to meet their objectives.

RDMA Technical Overview

The RDMA technology described in this book revolves around the industry standard InfiniBand Architecture specification and its different implementations. Fortunately, the creators of the InfiniBand Architecture specification integrated flexibility in its design to allow choices in how to implement it. First, let's take a look at the overall architecture.

The InfiniBand Architecture comprises the following as illustrated in Figure 3:

- An Application Programming Interface (API)—how an application takes advantage of RDMA through the RDMA message service
- An RDMA message service—enveloped in the RDMA software, providing access to the RDMA hardware
- A Host Channel Adapter (HCA) or Converged Network Adapter (CNA) (also simply adapter in this book)—the server-located InfiniBand or Ethernet hardware
- Interconnect—a network of cabling, switches, and routers (InfiniBand or Ethernet

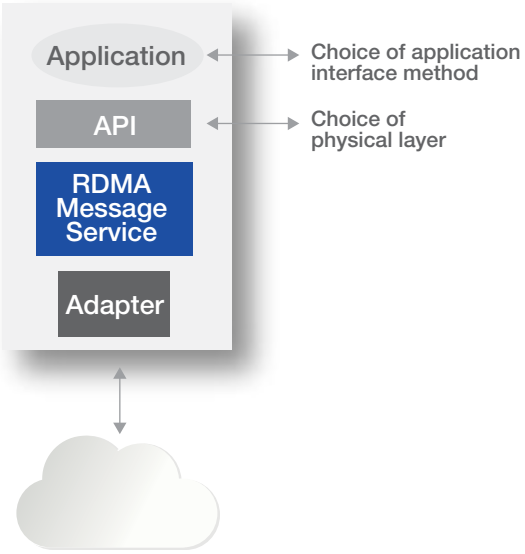

Figure 3 – The 'parts' of RDMA

Note that the specification provides choices in the physical interconnect and the APIs. Choices mean opportunities to intelligently deploy a technology in a way that meets the company's needs. They provide control points for an organization to take advantage of the cost/benefit and technology evolution in ways that makes sense for the business. Which choices to make and when to implement them come down to the company's ultimate objectives, the acceptable tradeoffs, and the investment a business is willing to make.

Hardware Implementations

The InfiniBand specification describes two physical layer implementations for the technology (Figure 4):

- RDMA over InfiniBand (or simply InfiniBand)
- RDMA over Converged Ethernet (RoCE)

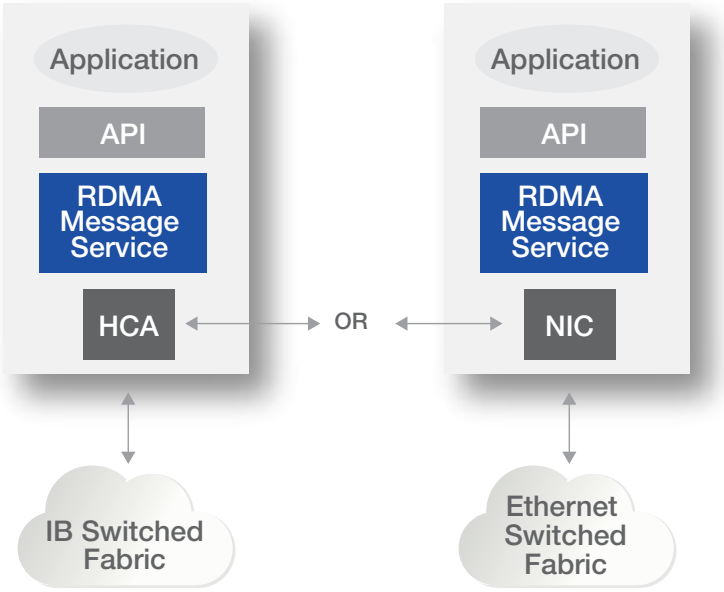

Figure 4 – Two hardware implementations for RDMA

These two choices create different opportunities for the data center, where it can benefit from RDMA. With multiple possibilities, data centers can test, scale, and adapt their deployments as needed.

Members of the InfiniBand Trade Association design and manufacture the hardware components for different physical layer implementations—including native InfiniBand adapters, InfiniBand switches and routers, and Ethernet adapters suitable for RoCE. The InfiniBand Trade Association web site (www.infinibandta.org) lists providers of InfiniBand products and provides contact information for IBTA integrators who can help companies define and implement solutions. The RoCE Initiative web site (www.roceinitiative.org) has similar resources for RoCE products and integrators.

RDMA over InfiniBand – This is the traditional deployment of InfiniBand adapters, InfiniBand switches and routers, and appropriate cabling. A fully native InfiniBand implementation (both hardware and API) offers the lowest latency, highest bandwidth, and highest efficiency of the three implementations (Ethernet, InfiniBand, and RoCE), returning the largest benefit. And it can do this while delivering the network architecture that enables the potential of full virtualization and cloud computing.

RDMA over Converged Ethernet (RoCE) – RDMA over Converged Ethernet (RoCE) essentially places the InfiniBand transport layer onto the Ethernet data link layer, providing the remote DMA capability, kernel bypass, and other benefits not part of traditional TCP over Ethernet. RoCE delivers much of the benefits of InfiniBand Architecture, only over Ethernet. This is good news for data centers committed to Ethernet.

The advancements made in Ethernet over the last few years have enabled it to make many of the same delivery guarantees that native InfiniBand does. With the Data Center Bridging (DCB) extensions that include PFC and ECN, the ubiquitous networking standard of the last 30 years is smarter than ever and offers less lossy characteristics. RoCE benefits from these advances, but it does not require them. If a company wants to improve Ethernet performance with DCB or ECN it will not require hardware investment in buying special switches and routers, since most enterprise and carrier-grade switches already support these features.

In addition, the RDMA software stack is free, making the switch from TCP/IP to RoCE an easy choice for the data centers.

Software Implementations - API Choice

The industry around RDMA has enabled flexibility in RDMA implementation in the data center. The IBTA has defined the hardware standard for RDMA over InfiniBand and Converged Ethernet.

Choices Matter

The choices of hardware and API give a company many opportunities of how to get the most benefit out of the cost/benefit and technology evolution curves. Figure 6 illustrates how these benefits can be achieved.

- Protect investments in Ethernet and run your socketed apps using RDMA
- Protect investments in Ethernet, but gain additional performance with RoCE and Verbs
- Take advantage of the InfiniBand higher performance while maintaining legacy IP-based applications
- Achieve maximum benefits from the features of the IB fabric and RDMA-based applications

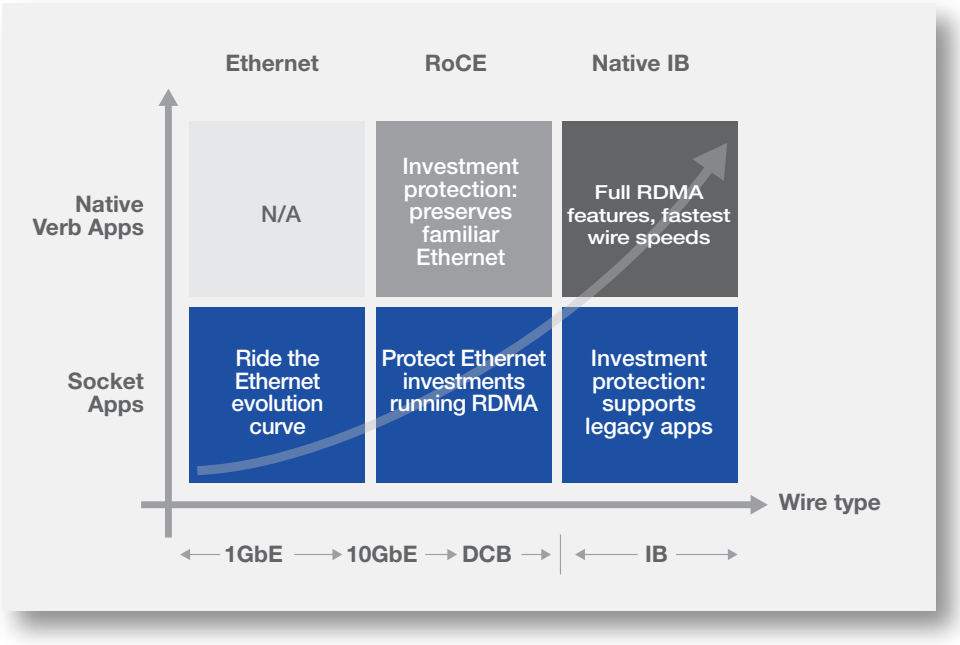

Figure 6 – Flexible choices enable greater control for the organization

Section 3 – RDMA at Work

Enterprises are increasingly engaging virtualization and private cloud. Scaled up analytics and scaled out big data is becoming a fact of life for analytics in the enterprise, the latter requiring IT to build out big data clusters, like what Hadoop uses. Software Defined Networking (SDN), Software Defined Storage (SDS), and the Software Defined Data Center (SDDC) are all emerging as new approaches to how companies build and operate their IT infrastructure.

Efficiently deploying software-defined everything relies on balance among the compute-memory complex, storage, and fabric. New workloads are entering the data center, including workloads for which RDMA has most often been used today—High Performance Computing. These new usages and approaches all can benefit from a faster, more efficient fabric that RDMA enables. Thus, such changes in the data center over the last few years also demand we take a new look at I/O.

RDMA in the Modernized Data Center

As companies look at how to modernize the data center toward more efficient operations, to enable self-services and high-performance data analytics, RDMA should be part of the architecture considerations.

RDMA for Application Servers and Remote Desktops

One of the benefits of RDMA pointed out earlier is hardware offload: freeing CPU cycles to do more computing and less communicating. Considering the size of enterprise and cloud data centers today, freeing up CPU cycles in a platform can mean large savings in scaling out a rack or row. Or it can provide more application processing headroom to an existing configuration, delaying capital expenditures while accommodating a growing user base.

RDMA in Virtualized Environments

Data centers are adopting virtualization for server consolidation and to enable software-defined services. To support RDMA in virtual machines and appliances, leading virtualization software vendors and open source developers have built in support for running RDMA in the virtual machine (VM), including:

- Microsoft Windows* Server— Since the release of Windows Server 2012 (including Windows Server 2016 and Windows Server 2019) RDMA has supported the SMB Direct protocol, which enables higher performance and most reliable data communication[4].

- Microsoft SQL 2016 see HPE, Mellanox, and Micron solution brief
- VMware—vSphere* version 4.0 and above supports various RDMA functionality in its hypervisor technologies, including live migration[5]. RoCE support in vSphere is also now available[6].
- Red Hat KVM—Red Hat has supported RDMA in its OS for some time, and it support live migration using RDMA.
- Oracle Virtual Machine (OVM)—Oracle VM Server supports RDMA (InfiniBand) when deployed within Oracle Exalogic Elastic Cloud [7].

RDMA for the Software Defined Data Center

The Software Defined Data Center (SDDC) is based on a scale-out architecture, whereby nodes are interconnected to provide the various virtualized services. Thus, the network becomes critical to supporting the performance of compute services in virtual machines and movement of data in scalable software defined storage. While traditional Ethernet is advancing in link speed, link speed is only one factor contributing to the overall efficiency and aggregate bandwidth capability of the network. Latency and CPU overhead also have an impact on the SDDC's performance, and thus its ability to support the agility and cost-effectiveness that IT seeks from an SDDC implementation. As we've seen, RDMA offers lower latencies and CPU overhead, and it is being deployed in the most advanced data centers worldwide.

RDMA for Software Defined Networks

The Software Defined Network (SDN) enables the control plane's decision-making functions to be separated from the data plane's data movement activities. It allows greater intelligence to be built into network functions and maximizes the efficiency of the transport. Network Function Virtualization is a big driver in SDN, supporting standardized control functions that can run on a commodity server instead of expensive, specialized hardware.

The InfiniBand Architecture specification defined one of the first software defined networks. InfiniBand's Subnet Manager (SM) software is responsible for discovering, configuring, activating, and managing a network with up 48,000 nodes to optimize traffic flow. Coupling the SM with RDMA enables InfiniBand's network efficiency with high throughput, low latency, low CPU overhead, and extreme scalability.

[4] *https://technet.microsoft.com/en-us/library/jj134210(v=ws.11).aspx*

[5] *https://openfabrics.org/images/eventpresos/workshops2013/2013_Workshop_Wed_0830_Bhavesh-Davda-vRDMA.pdf?7646d4&7646d4*

[6] *https://www.hpcwire.com/off-the-wire/mellanox-announces-software-driver-support-connectx-4-ethernet-roce-vmware-vsphere/*

[7] *http://www.oracle.com/technetwork/server-storage/networking/documentation/o12-020-1653901.pdf*

RDMA for Storage Arrays and Software Defined Storage

Software Defined Storage (SDS) and Hyper-converged systems, like other aspects of the SDDC, require fast data movement across a highly scalable pool of storage media. With RDMA, enterprises can implement fast data movement between storage pools and servers over networking standards like iSER (iSCSI over RDMA), SMB Direct (SMB 3.0 over RDMA), NVMe-oF (NVMe over Fabric) using either native InfiniBand or RoCE, and gain the benefits of RDMA performance for a variety of SDS technologies, such as Ceph*, Lustre*, Microsoft Windows Storage Spaces Direct* (S2D), and Gluster*. Mellanox Technologies, Microsoft, Red Hat, Samsung, SanDisk, Scalable Informatics, Seagate, Supermicro, and Storage Foundry ran benchmarks of RDMA for SDS, showing significant increases in read performance, reduction in write latency, and overall efficiency for large block operations[8]. Both InfiniBand and RoCE can be used in hyper-converged infrastructure (HCI) clusters that combine compute and storage in the same servers.

RDMA for Private Clouds

More and more enterprises are deploying virtualization for server consolidation, greater business agility, scalable computing, and private and public cloud deployments. But virtualization can generate more I/O traffic, since more services and applications in each physical server are using the network. And, servers today are communicating across multiple rows, not just within a single rack. This is an architecture ripe for RDMA.

Additionally, the architecture of emerging data center application-specific appliances and private clouds involve entire systems of shared resources interconnected in a grid in the mid-tier. As a compartmentalized function, say a transaction processing system or a complete private cloud, the interconnect becomes a critical component and a perfect application for an RDMA solution, supporting the throughput demands of high transaction rates.

RDMA and the Internet of Things (IoT)

The Internet of Things is delivering geometrically increasing amounts of data to data centers—both private and public cloud—that will eventually be funneled to big data analytics. Moving this data among storage and nodes falls on the fabric, requiring the network to be both efficient and flexible. As we've seen, these are characteristics for which RDMA has been designed, making it ideal for IoT.

[8] http://www.mellanox.com/related-docs/solutions/RoCE_MSFT_StorageSpaces_SB.PDF

RDMA for Databases and Analytics

While many enterprises rely on scaled-up, monolithic platforms for in-memory database analytics, big data and in-memory cloud platforms are changing how companies will use analytics. Big Data is already driving large scale-out solutions in the enterprise. Data movement is critical to the performance of these distributed systems. Considering Hadoop's shuffle phase alone, RDMA offers a significant performance benefit for high-performance data analytics, plus the ability to reduce critical latencies for real-time analytics on NoSQL databases.

Many companies are utilizing backend databases built on a cluster of servers communicating with a SAN. To increase scalability and performance by reducing latency, for example in lock operations, RDMA's low latency in single-digit microsecond (and less) values will result in the database engines moving more data and waiting less on locks so they can process more transactions.

RDMA for Machine Learning (ML) solutions

As the world continues to march forward with its digital transformation journey, a growing amount of data is becoming available for enterprises to make more accurate business decisions. Organization must now change their traditional thinking to develop "digital thinking" skills for their decision process. Not having such capabilities will put companies at a disadvantage versus competitors. Any DevOps decisions need to be based on a pure analysis of relevant data, this will help remove traditional silos, improve the communication between different teams within organizations and thus boost corporate efficiency.

This trend hasn't been hidden from the Enterprises decision makers that have already started to build and use AI/ML solutions. These solutions not only include CPUs but also other high-performance components, like GPUs, faster All Flash Array (AFA) storage that utilize NVMe or Persistent Memory (PM), which are all connected by high performance networking gear required to handle the massive data communication within the AI/ML clusters. Faster data communication results in faster analytics and a company that is first to make the right decision, gains an edge over its competitors. This is very similar to companies that run High Frequency Trading (HFT) where every nanosecond counts, and clusters are connected by the fastest networking technology -
in these environments the analytic cluster is located very close to the stock exchange data, even within the same building, to minimize the delay that optical cables inject.

In addition to building the AI/ML or HPC clusters, enterprises must also embrace a Hybrid strategy to enable higher scalability and higher efficiency. For example, Oracle just released their services to run over clusters connected by RoCE and the same is with VMware that developed AI/ML solutions over vSphere which runs natively over RoCE.

Deployment Strategies

RDMA is applicable anywhere applications need to communicate fast and efficiently with each other or with storage. The choices for where to deploy RDMA give an enterprise control points along the cost/benefit curve to maximize the benefits the technology offers within their IT budgets.

RDMA is not an all-or-nothing, forklift upgrade. RDMA can be integrated by stages into nearly any data center to optimize the ROI while protecting the investment in existing infrastructure. Or, it can be deployed as a full-on, optimized network solution for maximum performance, the way technical computing clusters and HPC systems are done. From experimenting with the technology to a full deployment rollout, there are many opportunities to take as much advantage of the technology as an organization chooses—incrementally or comprehensively.

The data center has long been modeled in three tiers interconnected by a hierarchical IP network of subnets (Figure 7). But, the data center is evolving: the network is getting flatter; traffic is flowing more East-West to fulfill the needs of a grid-based cloud architecture. This is RDMA territory.

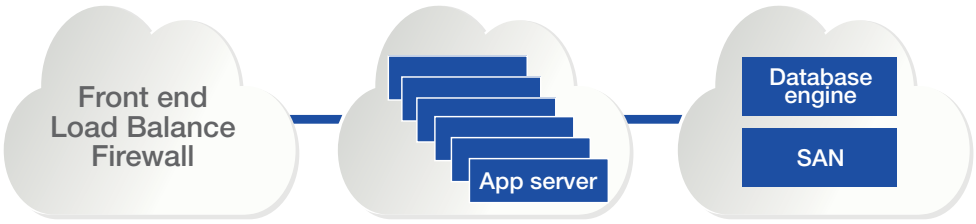

Figure 7 – Traditional DC architecture model

Identifying Good Candidates

RDMA is a good fit where efficient network communication is required, e.g. parallel application execution, storage appliance supporting many servers, live migration of virtual servers; and the more critical the performance, scalability, or latency of the communication, the more appropriate. Look for such configurations as you explore the possibilities. Some examples might include:

- Inside the backend database engine cluster
- Between the database system and the storage network
- In the storage cluster network itself
- A deployment of hyper-converged servers
- A subnet of application servers
- A virtualized or cloud-based data center

Also, consider how soon you might flatten areas of your network, say transitioning from a top-of-rack routing strategy to end-of-row. This is a good time and place to consider deploying RDMA, especially an InfiniBand fabric because of its centrally managed approach, low latency, and bandwidth roadmap.

InfiniBand or Ethernet (RoCE) as the Transport

For the data center, does it make sense to deploy InfiniBand or continue to upgrade the Ethernet infrastructure? As pointed out earlier, either is possible (Figure 8), or they could even be combined. Bridging appliances can connect an InfiniBand cluster to an Ethernet network. While the Ethernet upgrade progresses, deploying RoCE CNAs and DCB-enabled switches will avail the tier with a significant benefit, with least resistance in hardware changes and budget reassessments. A traditional datacenter can easily deploy scale-out storage or database appliances that use InfiniBand internally but talk to the clients using Ethernet.

A full InfiniBand deployment can be done later, and in stages.

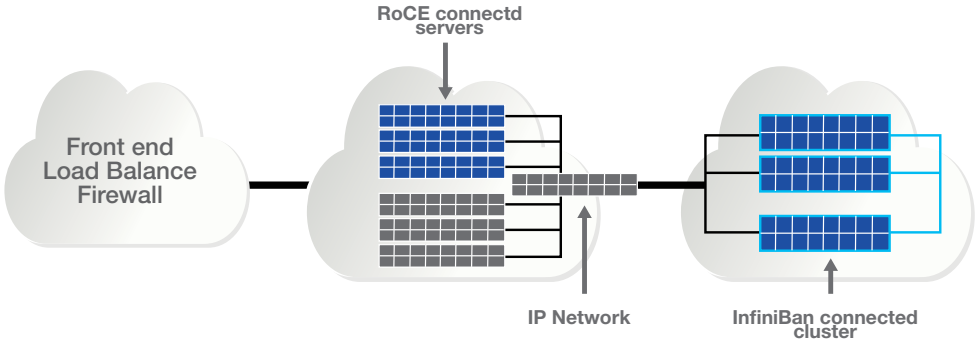

Figure 8 – Upgrading to RoCE in the mid-tier

Appendix A Further Reading/References

Many good references exist in the body of literature on the web, from academic and research institutions, and from industry. A few are provided below.

By Author

Beck, Motti. "Maximize Your Software-Defined Data Center Infrastructure Efficiency with RDMA-Enabled Interconnects."

Beck, Motti. "How to Achieve Higher Efficiency in Software Defined Networks (SDN) Deployments."

Beck, Motti. "Double Your Network File System (NFS) Performance with RDMA-Enabled Networking"

Gaiser, Lee; Brian Kraus and James Wernicke. "Implementation & Comparison Of RDMA Over Ethernet;" Los Alamos National Labs, http://institute.lanl.gov/isti/summer-school/cluster_network/projects-2010/Team_CYAN_Implementation_and_Comparison_of_RDMA_Over_Ethernet_Presentation.pdf

Grun, Paul. Introduction to InfiniBand™ for End Users; InfiniBand Trade Association, 2010; www.infinibandta.org

Grun, Paul. "Storage at a Distance: Using RoCE as a WAN Transport;" System Fabric Works, https://slidex.tips/download/storage-at-a-distance-using-roce-as-a-wan-transport

Skorupa, Joe; Hewitt, Jeffrey; Zeng, Evan. "Use Networking to Differentiate Your Hyperconverged System," Gartner, 2016; https://www.gartner.com/document/3332218

From Organizations

InfiniBand Trade Association. "InfiniBand Roadmap" https://www.infinibandta.org/infiniband-roadmap/

On VMware
http://www.mellanox.com/pdf/whitepapers/RDMA_Performance_in_Virtual_Machines_using_QDR_InfiniBand_on_VMware_vSphere5.pdf

Bhavesh Davda and Josh Simons from VMware present: RDMA on vSphere: Update and Future Directions. Recorded at the Open Fabrics Workshop

http://cto.vmware.com/wp-content/uploads/2012/09/RDMAonvSphere.pdf

On Microsoft Hyper-V

https://downloads.openfabrics.org/Media/Monterey_2012/2012_Workshop_Tues_SMB2_2_RDMA_Talpey.pdf

Appendix B RDMA Consortium

The InfiniBand Trade Association

Founded in 1999, the InfiniBand® Trade Association (IBTA) (www.infinibandta.org) is chartered with maintaining and furthering the InfiniBand™ Architecture specification, defining hardware transport protocols sufficient to support both reliable messaging (send/receive) and memory manipulation semantics (e.g. remote DMA) without software intervention in the data movement path. These transport protocols are defined to run over Ethernet (RoCE) as well as InfiniBand fabrics. The IBTA is led by a distinguished steering committee that includes Broadcom, HPE, IBM, Intel, Mellanox Technologies, Microsoft and QLogic. IBTA members represent leading enterprise IT vendors that are actively contributing to the advancement of the InfiniBand specification.

The IBTA conducts compliance and interoperability testing of commercial InfiniBand and RoCE products. It has successfully added hundreds of products to its Integrators' List Program.

The organization unites the industry through IBTA-sponsored technical events and resources. It actively promotes InfiniBand and RoCE from a vendor-neutral perspective through online communications, marketing and public relations engagements. For more information, visit http://www.roceinitiative.org/