Benefits of Remote Direct Memory Access Over Routed Fabrics

Introduction

An enormous impact on data center design and operations is happening because of the rapid evolution of enterprise IT. Organizations are required to get more creative and resourceful about leveraging new technologies while taking advantage of existing infrastructure and limiting overall costs; for example, leveraging an existing storage and compute infrastructure for more efficient but less expensive operations by transitioning toward converged infrastructure. The growing demand for storage shows no sign of slowing down along with the need for ever more compute to analyze the expanding data. Data center managers must manage all of these demands and meet their aggressive business goals while also remaining efficient and keeping costs under control.

Additionally, in virtualized environments there is an increased need for more efficient use of the existing infrastructure while delivering high performance. Technologies that allow a higher density of virtual machines per host and the migration of a virtual machine while it is running—live virtual machine migration—improve the flexibility and efficiency of server operations. However, this means that we need high-speed, low-latency networks and efficient handling of CPU/memory resources on the hypervisors. For example, today's virtual machines can have up to 1 TB of memory and 64 TB of storage each. To copy such a virtual machine in real time from one physical server to another can impose a significant hit on the CPU and overall performance of the source machine. It can also take significant time to copy all of this data to the destination machine, creating demand for more network resources. RDMA over converged Ethernet is the ideal technology to address requirements of high performance, low latency, and low cost.

In this document we describe the advantages of running Microsoft Windows SMB Direct protocol on a Remote Direct Memory Access (RDMA) Layer 3 network, via RDMA over Converged Ethernet (RoCE[1]) transport, using Cisco Nexus® Switches and Mellanox ConnectX adapters. With this architecture the challenges of today's modern data centers can be solved and overall efficiency improved.

Congestion avoidance and bandwidth management are not part of this document. These technologies were enabled during our test but not explicitly tested. See the Appendix for configuration settings.

[1] The v2 version of RoCE is required for Layer 3 environments and has largely replaced the earlier Layer 2 version. The v2 version was tested here, and all references to RoCE refer to the routable implementation.

Audience

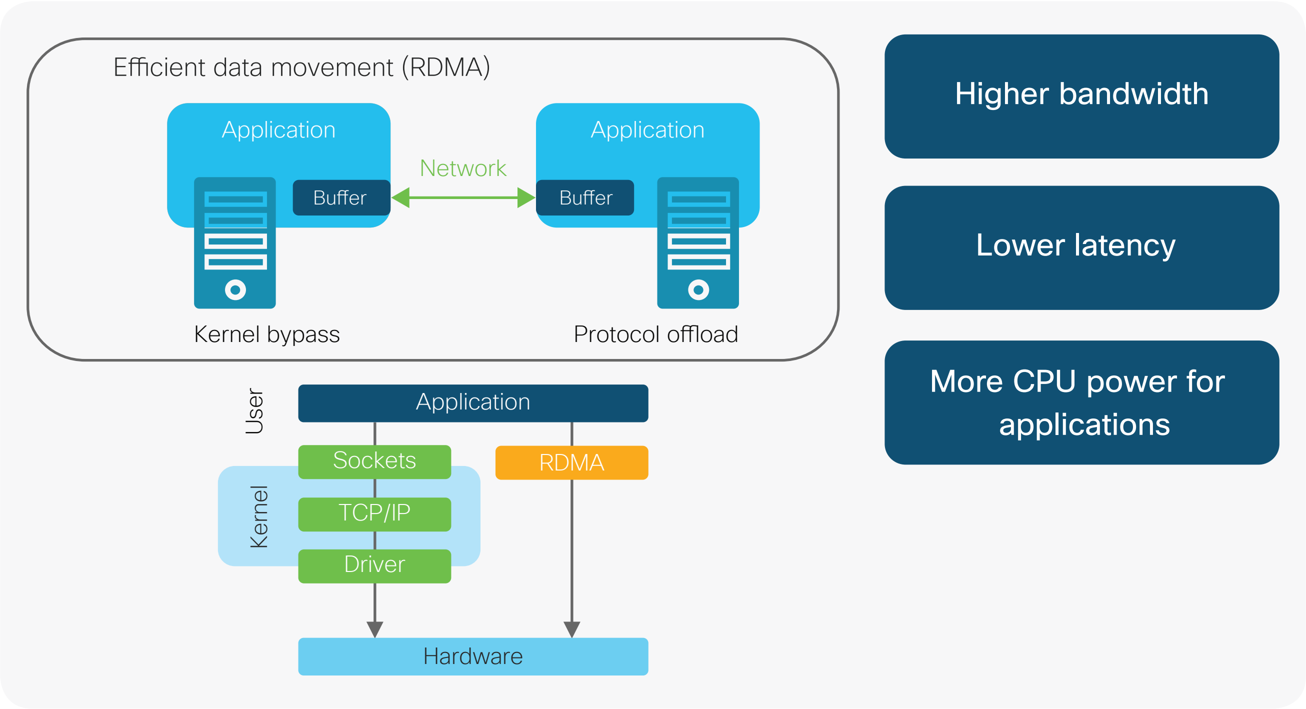
This paper is targeted for IT managers and architects. It showcases how to utilize your network efficiently and gain higher performance using RoCE as an RDMA transport with Mellanox adapters and Cisco® switches. This paper focuses on Windows traffic with the SMB Direct protocol, but the same benefits and technologies also apply to other RDMA-capable storage and server traffic, such as iSCSI Extensions for RDMA (iSER) and NVMe over Fabrics (NVMe-oF).

Technology overview

Remote Direct Memory Access (RDMA)

Remote Directory Memory Access (RDMA) is a technology originally used in High Performance Computing (HPC) environments. The advantages of RDMA is the low latency transfer of information between compute nodes at the memory-to-memory level, without burdening the CPU. This transfer function is offloaded to the network adapter hardware in order to bypass the operating system software network stack. With this technology, network adapters can work directly with the memory of applications, allowing network data transfers without the need to involve the CPU; thus providing a more efficient and faster way to move data between networked computers at lower latency and CPU utilization, freeing up CPU cycles to perform other functions. See Figure 1, below.

Figure 1. Remove Data Memory Access (RDMA)



RDMA over Converged Ethernet (RoCE)

RDMA over Converged Ethernet (RoCE) is the most popular way to run RDMA over Ethernet. It allows direct memory access over an Ethernet network and delivers superior performance compared to traditional network socket implementations because of lower latency, lower CPU utilization, and higher utilization of network bandwidth. For example, Windows Server 2012 and later versions can use RDMA for accelerating and improving the performance of SMB file-sharing traffic and live migration.

The benefits of introducing RoCE in data center infrastructure are:

- Lower cost of ownership, because no separate storage networking infrastructure is needed
- Higher ROI across traditional and modern agile infrastructures
- Improved overall CPU utilization for running applications
- Efficient host memory usage
- Higher throughput and lower latency for compute and storage traffic

RoCE use cases

Private cloud

RoCE is implemented in many private cloud environments because it reduces the networking overhead on the CPU, making more CPU resources available for virtual machines and applications. This translates to greater business opportunity for the cloud data center by monetizing these saved CPU cycles. Traditional computing infrastructure often required separate storage networks, which increased the cost of the infrastructure. Because RoCE runs on Ethernet, it provides a converged network infrastructure with a lower cost of ownership.

Storage

Data storage always struggles to deliver more IOPs, higher throughput and lower latency. Traditional SAN networks are unable to keep up with the demand of today's SSD flash-based storage and real-time applications. RoCE provides a positive impact when faster storage requires faster networks. Higher throughput and lower latency are introduced when using data storage protocols over RoCE. Many high-performance data storage appliances and software-defined storage solutions benefit from implementing RoCE. A leading example is Microsoft SMB Direct, illustrated in this white paper. Other storage solutions support iSCSI Extensions for RDMA (iSER) and NVMe over Fabrics (NVMe-oF), both of which run over RoCE.

Data warehousing and big data

Data warehousing applications, such as Oracle RAC and Microsoft SQL Server, require high levels of availability and performance. The same is true of big data analytics tools such as Apache Spark. With traditional Ethernet as a best effort medium, this was unachievable. Because RoCE delivers deterministic latencies and lossless or near-lossless transport over Ethernet, database, data warehousing, and big data analytics workloads all benefit from RoCE. By using RoCE in these environments customers achieve more job operations, more transactions per minute, more I/Os per second, and predictable scaling of clusters.

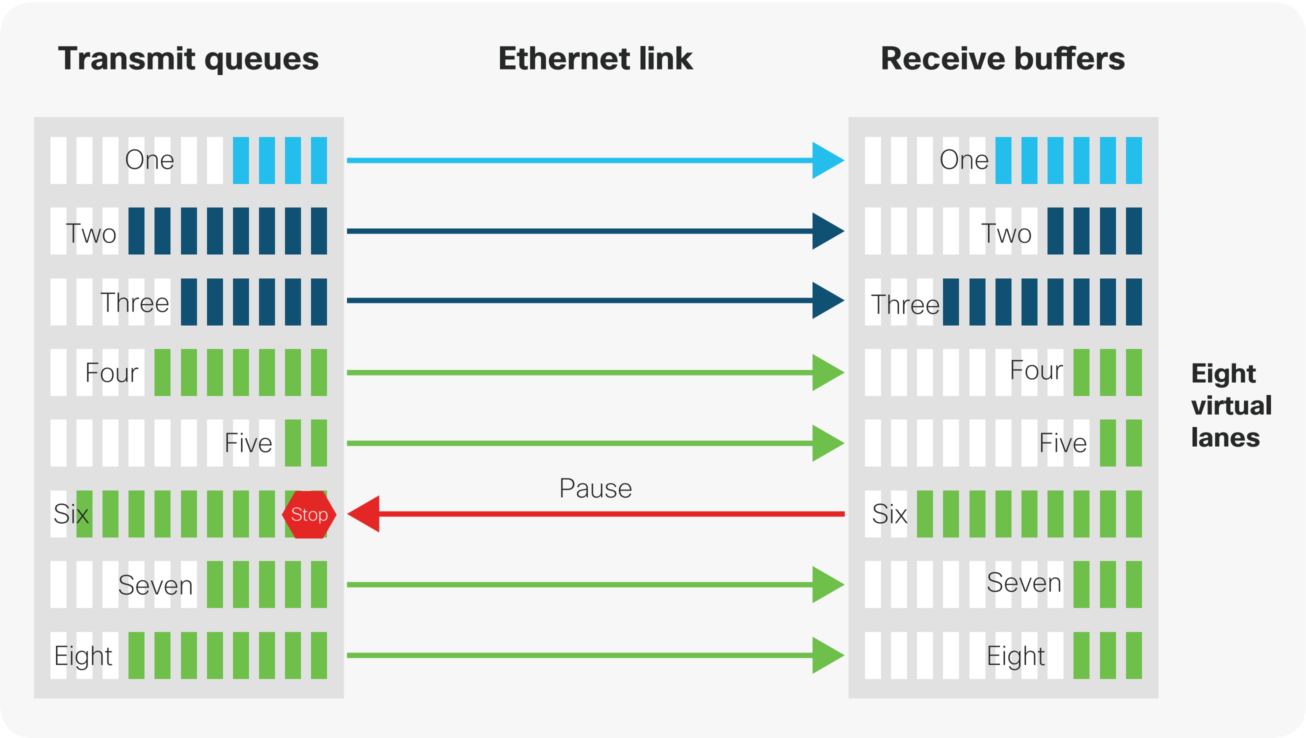
Financial services

Financial institutions looking to reduce transaction time, have been leveraging dedicated networks to provide lower latency. RoCE allows them to achieve the same benefits of reduced transaction times on a converged network infrastructure that still provides predictably low latency, with the accompanying higher throughput, quicker transactions, and higher IOPs performance for storage.

Converged Ethernet

Convergence Ethernet or Converged Enhanced Ethernet (CEE) was defined in 2008 and 2009 by an ad-hoc industry group. The group's goal was to create proposals for enhancements to Ethernet that enable networking protocol convergence over Ethernet. Data Center Bridging (DCB) is the enhanced version of Ethernet that resulted from the work of the Convergence Ethernet group. To meet the group's goals new standards were developed by two separate standards bodies: the Internet Engineering Taskforce (IETF) and the Institute of Electrical and Electronic Engineers (IEEE). One feature added is Priority-based Flow Control (PFC), which is required by some RoCE deployments. IEEE-standard PFC provides a link level flow control mechanism that can be controlled independently for each frame priority. See Figure 2.

Figure 2. Priority-based Flow Control (PFC) lanes

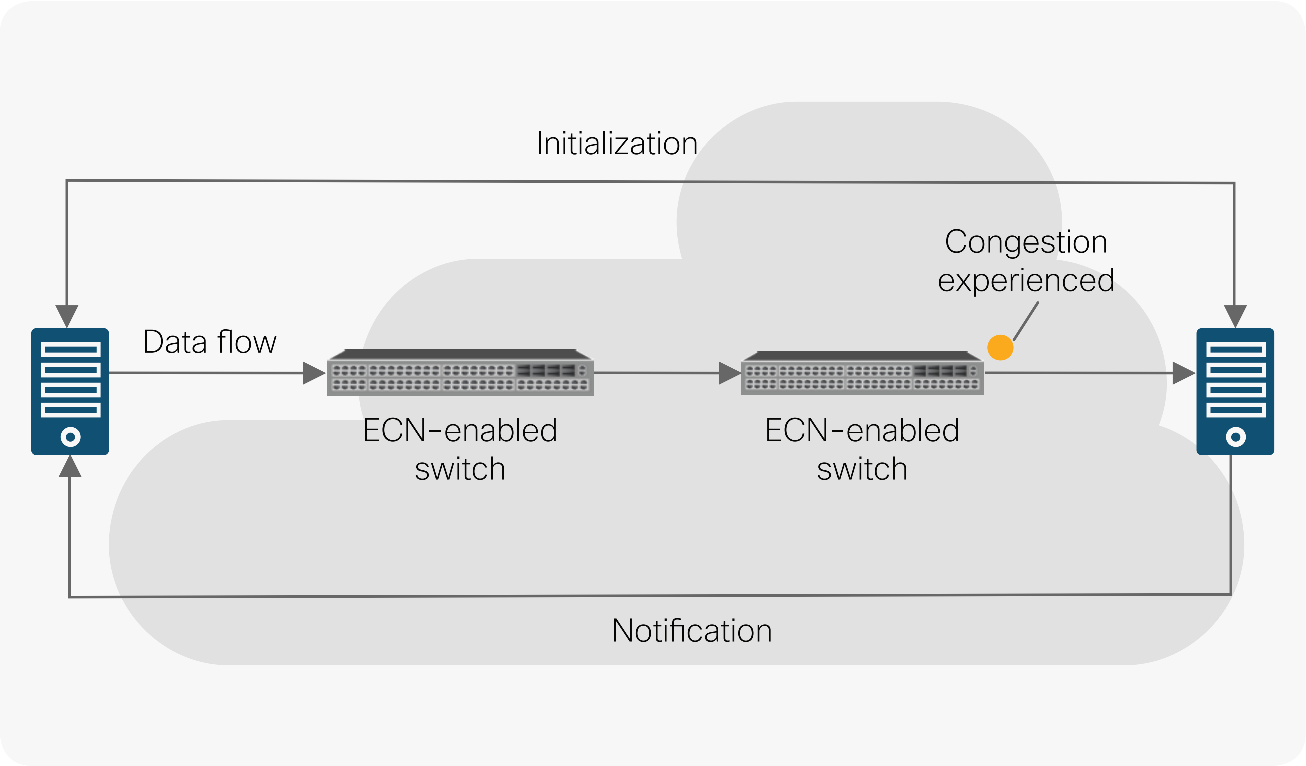


The goal of PFC is to maintain RDMA performance expectations by ensuring a lossless network during congestion situations. Minimizing packet loss allows for the best possible performance for RDMA traffic. To make Ethernet lossless, traffic that belongs to the class needs to be marked with Class of Service (CoS) value that network devices recognize as nondrop class and queue it in a nondrop queue. As a best practice for routed networks all traffic needs to be marked with IP Differentiated Services Code Point (DSCP) in addition to CoS value to be recognized as nondrop class. When a nondrop queue fills up to a maximum threshold, a port will start sending pause frames toward the source of the traffic for the lossless class. This should slow down the sender and mitigate congestion. After buffer usage goes below a minimal threshold, the port will stop sending pause frames and traffic will continue to flow at maximal capacity. This mechanism prevents drops in Ethernet transport, which is best effort by default. PFC performs by per-hop behavior; if the topology is made of multiple network devices, all of the devices need to be PFC-enabled in order to have safe transport for lossless traffic.

Explicit Congestion Notification

Another method of network congestion control is Explicit Congestion Notification (ECN). Many data centers today already use or can easily enable IETF-standard ECN. PFC performs by per-hop behavior, which at times might take more time than others to propagate congestion information to the sender. With ECN, congestion on any network device is propagated immediately to the endpoints and traffic is marked on any network device that experienced congestion (see Figure 4). Because of this, ECN is used more commonly than PFC because it is sometimes considered less disruptive to configure on networks.

Figure 3. Explicit Congestion Notification (ECN)



Not all RoCE Ethernet adapters support ECN for congestion control, but all Mellanox adapters from ConnectX-4 (and ConnectX-4 Lx) onward do so.

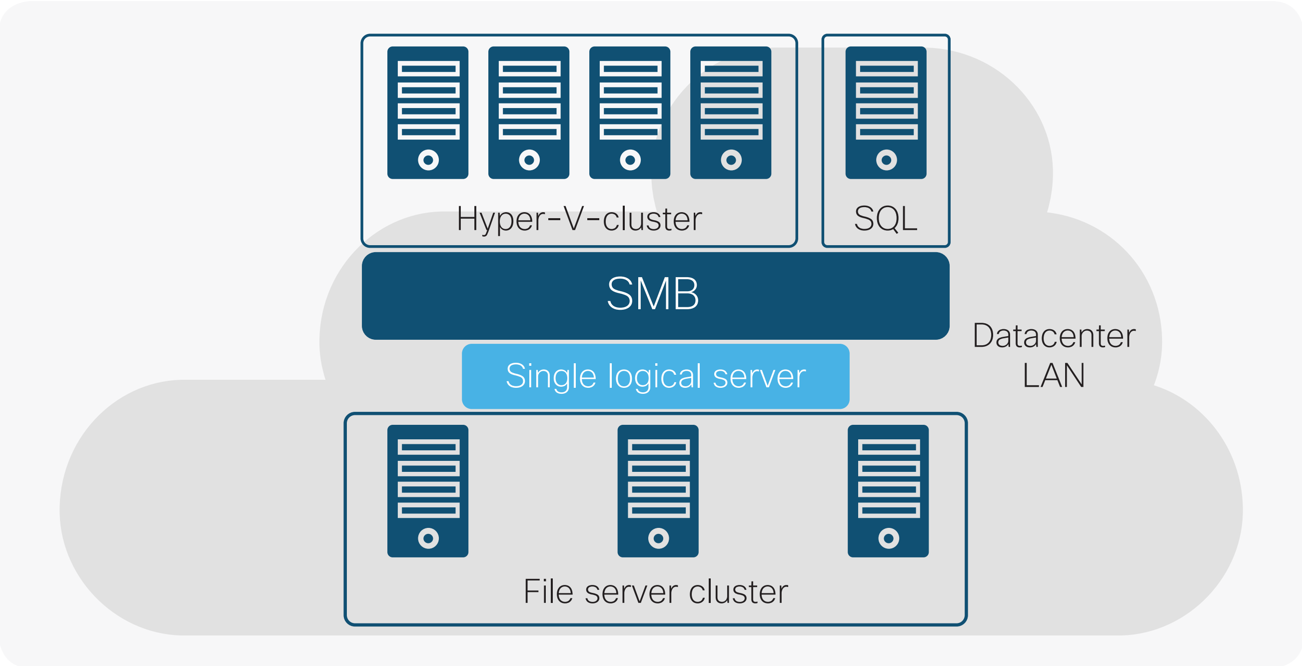
For the optimal RDMA performance in changing and dynamic network environments, both PFC and ECN can be used together. In that case, congestion caused by traffic patterns such as incast can be easily mitigated with ECN, because capabilities that exist anywhere in the data path congestion are signalized to the endpoints. However, if congestion is experienced close to the endpoints, and caused by a burst application on the sender, PFC efficiently mitigates and manages the congestion by slowing down the sender.

RoCE can also be deployed using just PFC or—with advanced RoCE adapters, such as those from Mellanox—with just ECN. Using only PFC or ECN, performance for ordinary applications can be satisfied, but highest performance is usually achieved only by coupling PFC and ECN.

Microsoft SMB 3 and SMB Direct storage protocols

The Server Message Block (SMB), a network file sharing protocol, allows applications on a computer to read and write to files and to request services from server programs in a computer network. Originally designed for Windows, SMB is now also supported by Linux and other operating systems. The SMB storage protocol can be used over TCP/IP or other network transport protocols like RDMA. Using the SMB protocol, an application (or the user of an application) can access files or other resources on a remote server. This allows applications to read, create, and update files on the remote server. It can also communicate with any server program that is set up to receive an SMB client request. (See Figure 5.)

Figure 4. Server Message Block (SMB) 3 overview.



Microsoft supports RDMA with SMB 3.0, called SMB Direct, as a way to speed up storage networking performance and virtual machine migration with Hyper-V. This adds value to Hyper-V deployments using Scale-Out File Server, Storage Spaces, and Storage Spaces Direct.

Overview: Hardware

This section describes the hardware used in this white paper.

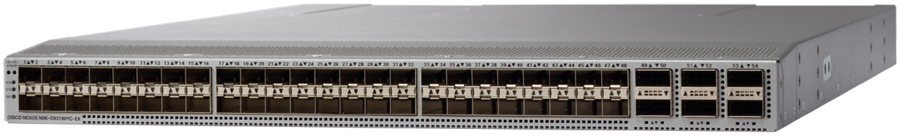
Cisco Nexus 9000 Series

Based on Cisco Cloud Scale technology, the Cisco Nexus 9300-EX and 9300-FX/FX2 platforms are the next generation of fixed Cisco Nexus 9000 Series Switches. The new platforms support cost-effective cloud-scale deployments, an increased number of endpoints, and cloud services with wire-rate security and telemetry. The platforms are built on modern system architecture designed to provide high performance and meet the evolving needs of highly scalable data centers and growing enterprises.

The Cisco Nexus 9300-EX and 9300-FX/FX2 platform switches offer a variety of interface options to transparently migrate existing data centers from 100-Mbps, 1-Gbps, and 10-Gbps speeds to 25 Gbps at the server, and from 10- and 40-Gbps speeds to 50- and 100- Gbps at the aggregation layer. The platforms provide investment protection for customers, delivering large buffers, highly flexible Layer 2 and Layer 3 scalability, and performance to meet the changing needs of virtualized data centers and automated cloud environments.

Switch used for the testing is the Cisco Nexus 93180YC-EX Switch, which provides a flexible line-rate Layer 2 and Layer 3 feature set in a compact form factor. Designed with Cisco Cloud Scale technology, it supports highly scalable cloud architectures. With the option to operate in Cisco NX-OS or Application Centric Infrastructure (ACI) mode, it can be deployed across enterprise, service provider, and Web 2.0 data centers.

Figure 5. Cisco Nexus 93180YC-EX

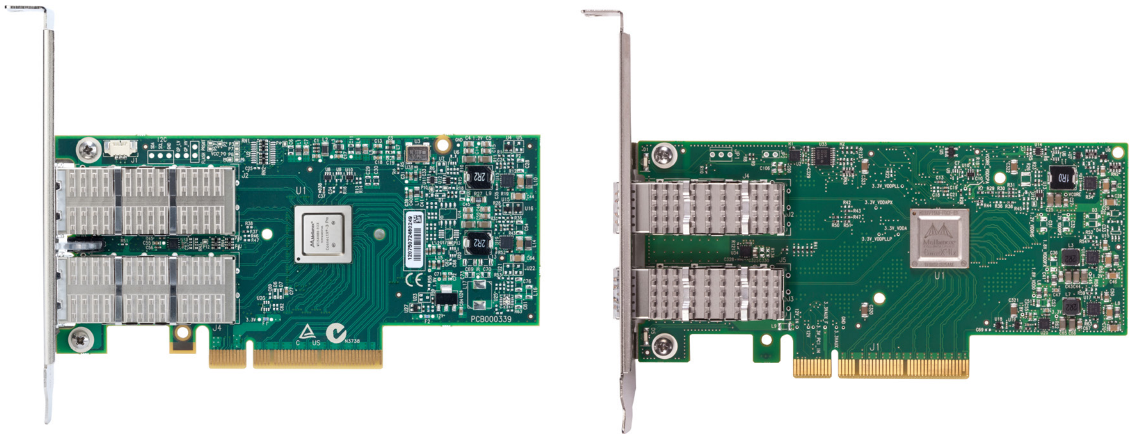


Mellanox ConnectX single- and dual-port adapters

In testing for this white paper, Mellanox ConnectX-3 Pro EN 10/40GbE and ConnectX-4 Lx EN 25GbE adapter cards were used. Like all Mellanox Ethernet adapters, they have hardware offload engines for RDMA using RoCE, which provides the highest performance for public and private clouds, enterprise data centers, storage, and high performance computing environments.

Mellanox ConnectX-3, ConnectX-4, and ConnectX-5 adapters are also compatible with Cisco UCS® C-Series Servers, such as the Cisco C220 and Cisco C240.

Using these cards, clustered databases, parallel processing, transactional services, and high-performance embedded I/O applications will achieve significant performance improvements over commodity NICs, resulting in reduced completion, lower cost per operation, and higher data center efficiency.



Mellanox ConnectX-3 Pro in dual-port (single-port also available) 40GbE (QSFP) and 10GbE (SFP+) versions. ConnectX-4 Lx adapters support 10, 25, 40, and 50GbE speeds.

ConnectX adapters utilizing IBTA-standards-compliant RoCE technology deliver low-latency and high performance over Ethernet networks. Leveraging Data Center Bridging (DCB) capabilities, RoCE provides efficient low latency RDMA services over Layer 3 Ethernet.

ConnectX adapters also effectively address the increasing demand for an overlay network, enabling superior performance by introducing advanced NVGRE, VXLAN, and hardware offload engines that enable traditional offloads to be performed on encapsulated traffic.

ConnectX SR-IOV technology provides dedicated adapter resources and guaranteed isolation and protection for virtual machines within the server. This I/O virtualization provides data center managers better server utilization while reducing cost, power, and cabling complexity.

Applications utilizing ordinary TCP/UDP/IP transport can achieve industry-leading throughput over 10/25/40/50GbE networks. The hardware-based stateless offload engines reduce the CPU overhead of IP packet transport. Socket acceleration software further increases performance for latency-sensitive applications.

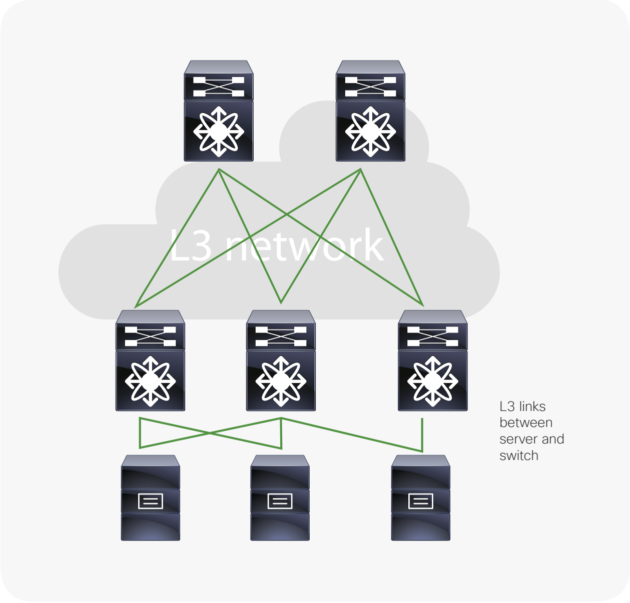
Storage Acceleration with ConnectX adapters and a consolidated compute and storage network achieves significant cost-performance advantages over multifabric networks. Standard block and file access protocols can leverage regular Ethernet (TCP/IP) or RDMA for high-performance storage access.

All Mellanox adapter cards are supported by Windows, Linux distributions, VMware, and FreeBSD, along with top hypervisors, including Windows Hyper-V, Xen, KVM, and VMware. They support Open Fabrics-based RDMA protocols and software and are compatible with configuration and management tools from OEMs and operating system vendors.

Test topology

The testing setup consists of three Cisco UCS servers equipped with Mellanox ConnectX adapters for 40 G or 25 G and running Windows Server 2012 R2, connected via five Cisco Nexus 9300 switches set up in a spine-leaf configuration shown below. All switch-to-switch connections, spine to leaf and leaf to server, in this topology are L3 routed links. Cisco UCS C220M3 Servers are used for Hyper-V nodes 1 and 2. A Cisco UCS C240M3 Server is used as the Windows Storage Spaces FileServer. To showcase the difference benefits of RoCE over TCP, both variations of RoCE and TCP base transfer were tested here. The test was performed with 40G and 25G link speeds.

Figure 6. Test topology



Deployment scenario overview

The Windows Server 2012 R2 offers a new way to connect storage to Hyper-V virtual machines on shared storage. Instead of offering only block-based shared-storage like Fibre Channel or iSCSI, Hyper-V now allows you also to use file-based storage through the SMB 3.0 protocol. This allows you to store virtual machines on an SMB file share.

With block storage the Hyper-V host has to handle the storage connection. This means that, for iSCSI or Fibre Channel, all configuration for the storage connection is done on the Hyper-V host. Examples of this are multipath, iSCSI initiator, and DSM software. With the use of SMB, advanced features like multipathing are enabled by default. SMB 3.0 is built in to the Windows Server 2012 R2.

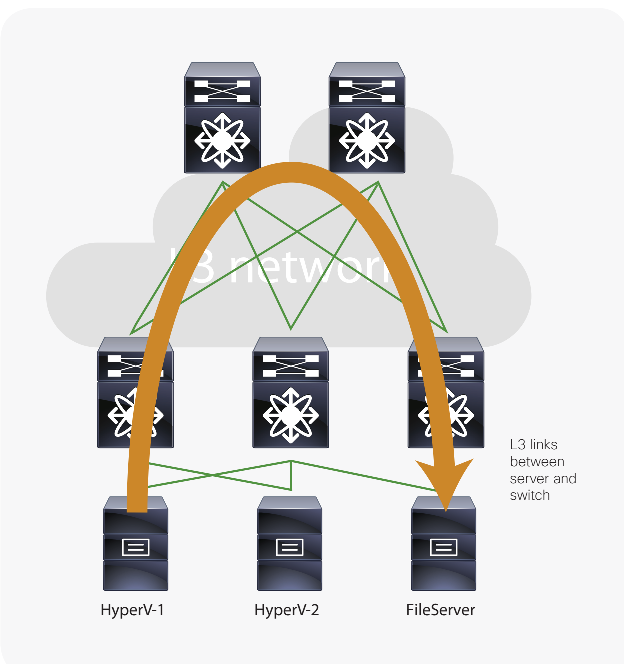
With the use of SMB Direct, which is SMB 3.0 over RoCE, both Hyper-V shared storage and general file-based storage performance are improved. This is because of the added flexibility, compute efficiency, and performance of RDMA.

In the two main use cases, tested for this white paper, we use SMB Direct (RoCE) as the protocol for network storage and Hyper-V live migration. We compare the performance with RDMA enabled and disabled to demonstrate the advantages of RoCE RDMA vs. TCP/IP.

The whole setup was configured to support RoCE. The detailed switches and adapter configurations are described in the Appendix.

Use case #1: SMB Direct for network storage

Figure 7. SMB Direct test case



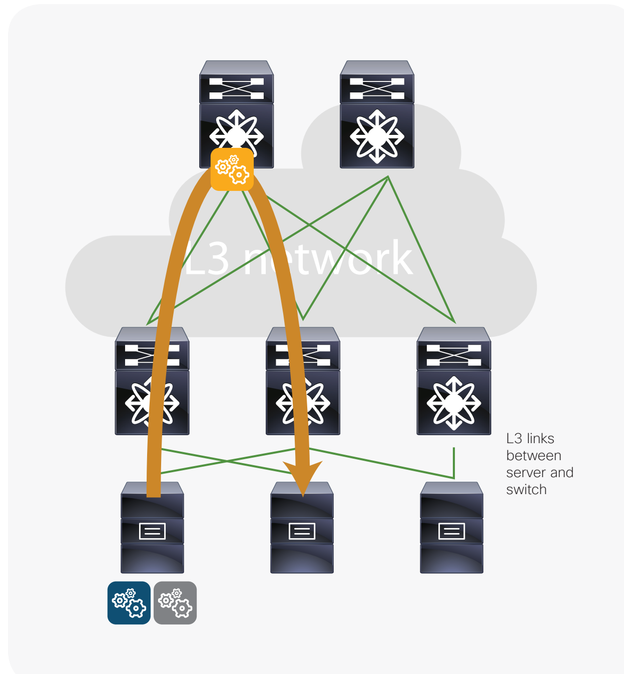
This use case describes the performance testing of SMB 3.0 with RoCE. Key configurations include:

- On the fileserver, a network share is created and mounted on the Hyper-V1.
- Microsoft Diskspd is used to measure the performance of a network copy and the impact on the CPU resources on the source system.

To show the benefits of RDMA, the test uses SMB Direct enabled (RoCE) and disabled (TCP/IP).

Use case #2: SMB Direct for Hyper-V live migration

Figure 8. SMB Direct for live migration test case



This use case describes the usage of SMB 3.0 with RoCE for live migration of Hyper-V virtual machines between Hyper-V servers. It includes 4 virtual machines (50GB of RAM), which were migrated simultaneously form one host to the other. To show the benefit of RDMA, the test was done with SMB Direct enabled (RoCE) and disabled (TCP/IP). In a hypervisor environment, you expect that the CPU load on the receiving hosts is higher than on the sending host.

Results for use case #1 performance tests: SMB Direct for storage – file transfer

Data transfers – RoCE vs. TCP:

Using Microsoft Diskspd, we see the benefit of running SMB Direct based on RoCE as a network storage protocol. We see improved metrics for RoCE in both Read (figure 9-11) and Write (figure 12-14) operations. The metrics were IOPS, bandwidth, and CPU utilization. In this test case, for the sake of simplicity, we conducted the tests with IOs sized 4KB (a typical cluster size) for writes and reads.

Figure 9. Read IOPs RoCE vs. TCP

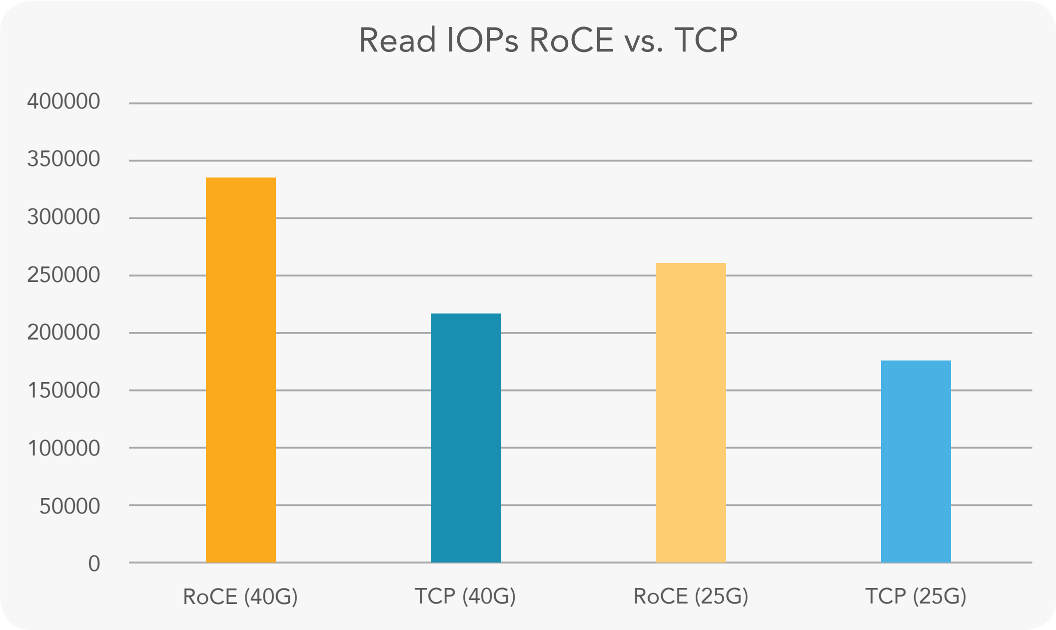


Figure 10. Read bandwidth RoCE vs. TCP in MB/s

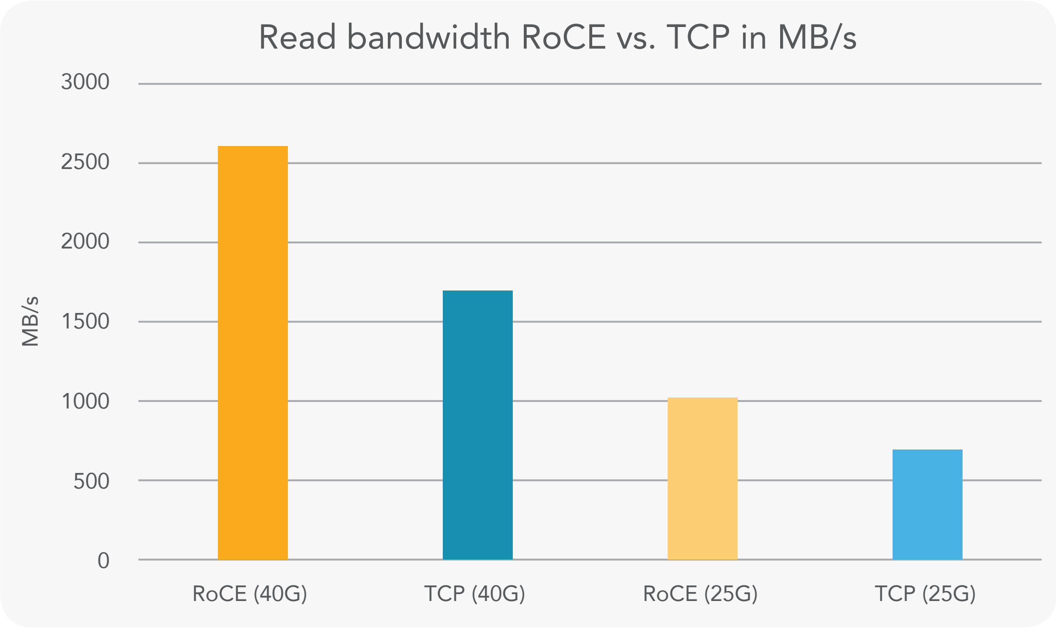


Figure 11. Read CPU utilization RoCE vs. TCP

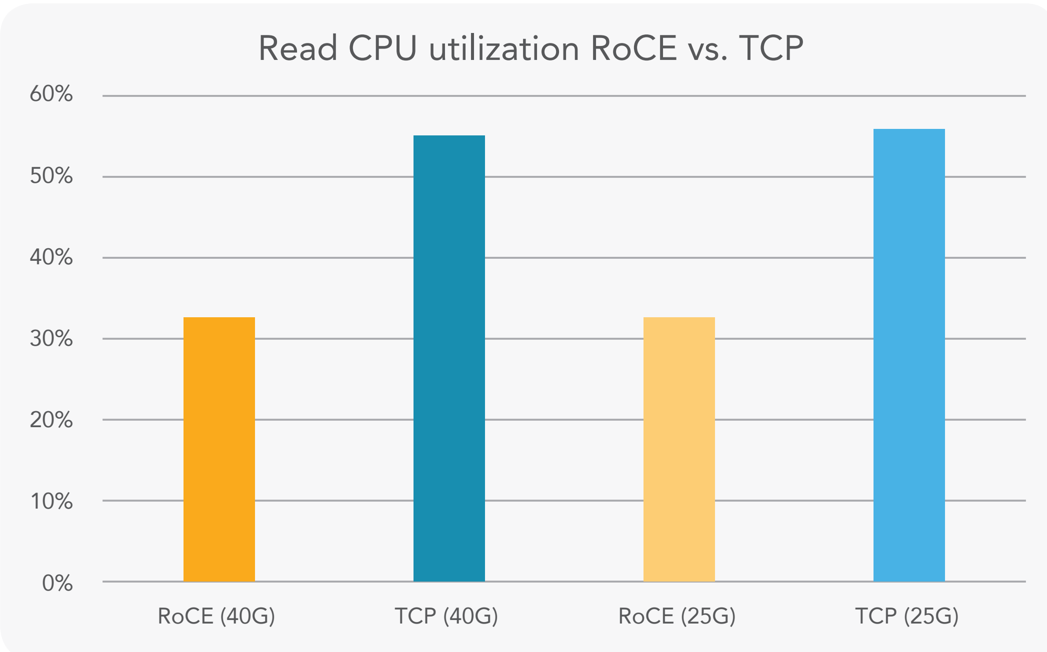


Figure 12. Write IOPs RoCE vs. TCP

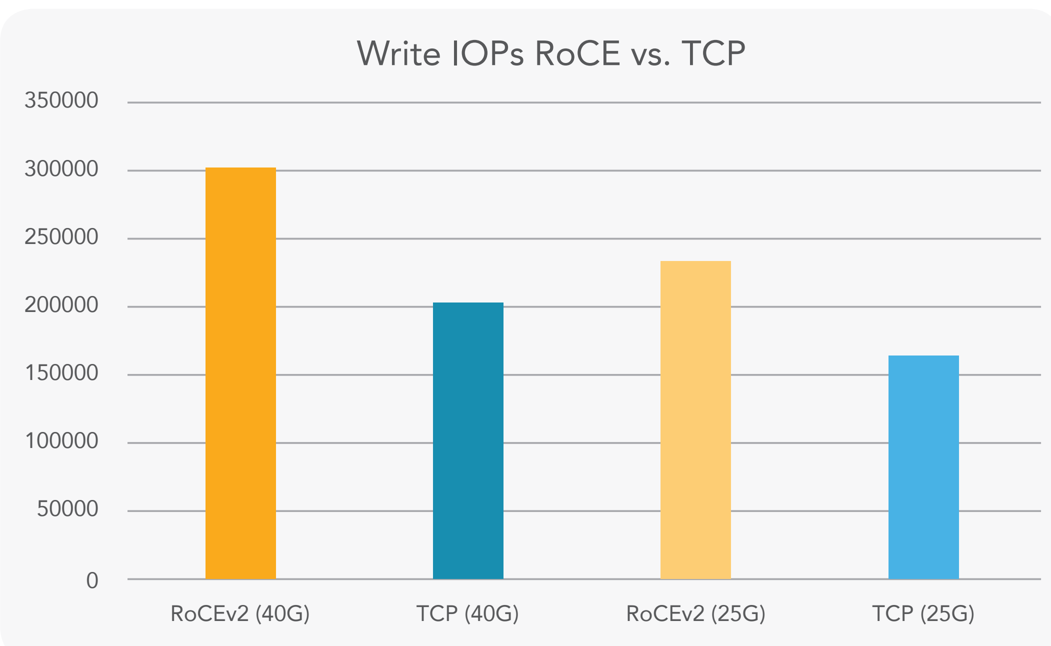


Figure 13. Write bandwidth RoCE vs. TCP in MB/s

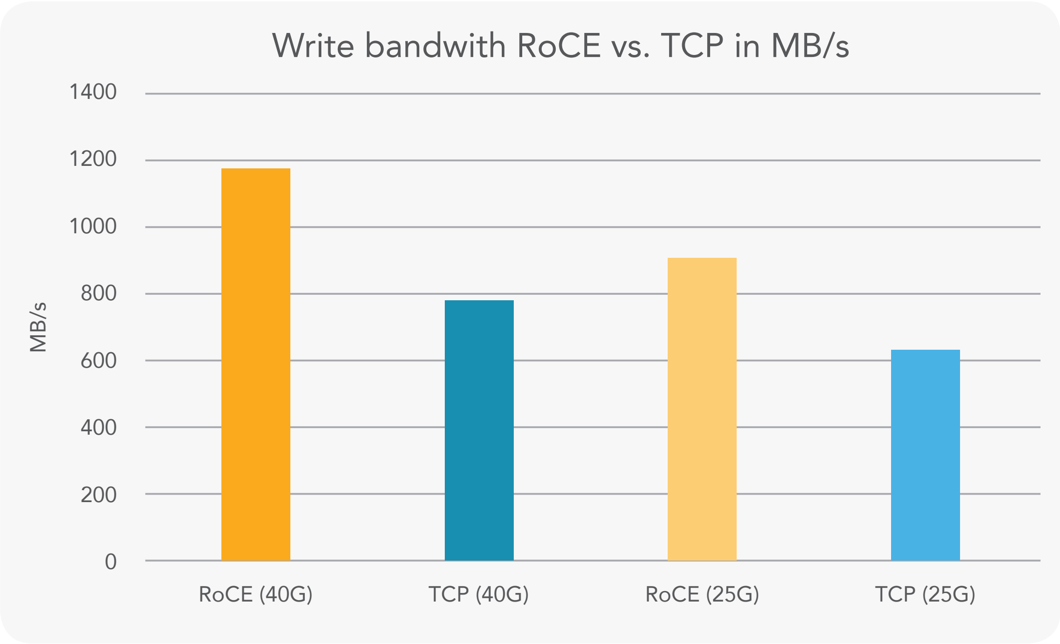
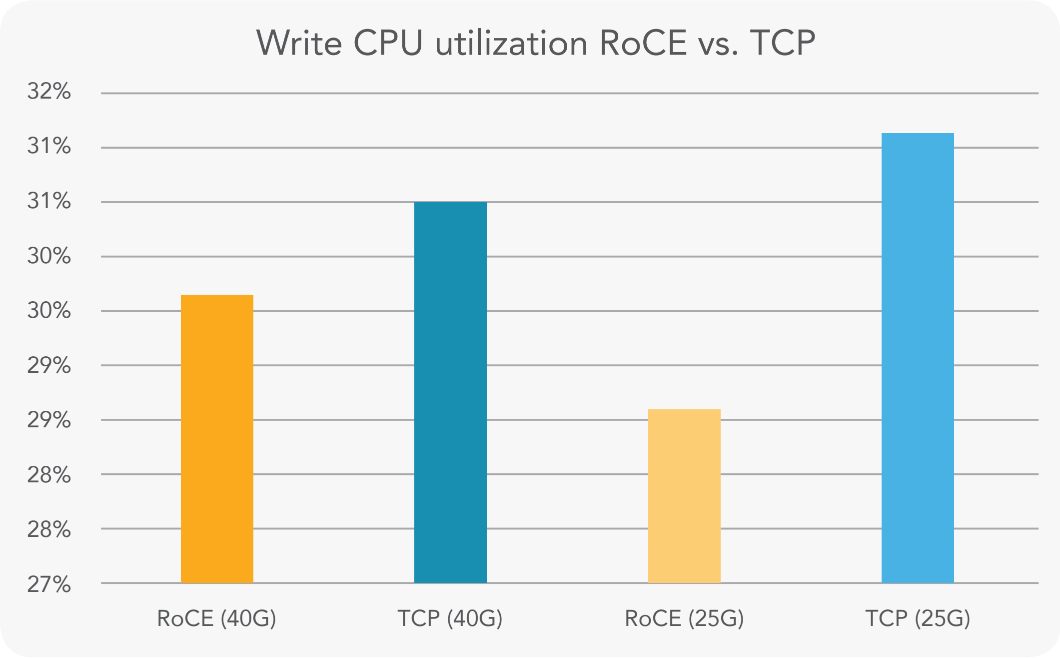


Figure 14. Write CPU utilization RoCE vs TCP



For both read and write file transfers (see the figures above), RoCE performs better than TCP/IP at the same wire speed. We also see less CPU usage even though the bandwidth is higher.

Results for use case #2: Performance test results: SMB Direct for virtual machine live migration

Using PowerShell to test the performance of a live migration between both Hyper-V hosts, we were able to show a similar benefit when running SMB Direct (RoCE) vs. SMB (TCP/IP). Again we see less CPU usage on Hyper-V hosts when using RoCE vs. TCP/IP even though the network wire speed was identical.

Figure 15. Sending host CPU utilization: maximum, average, and minimum, compared between RoCE and TCP

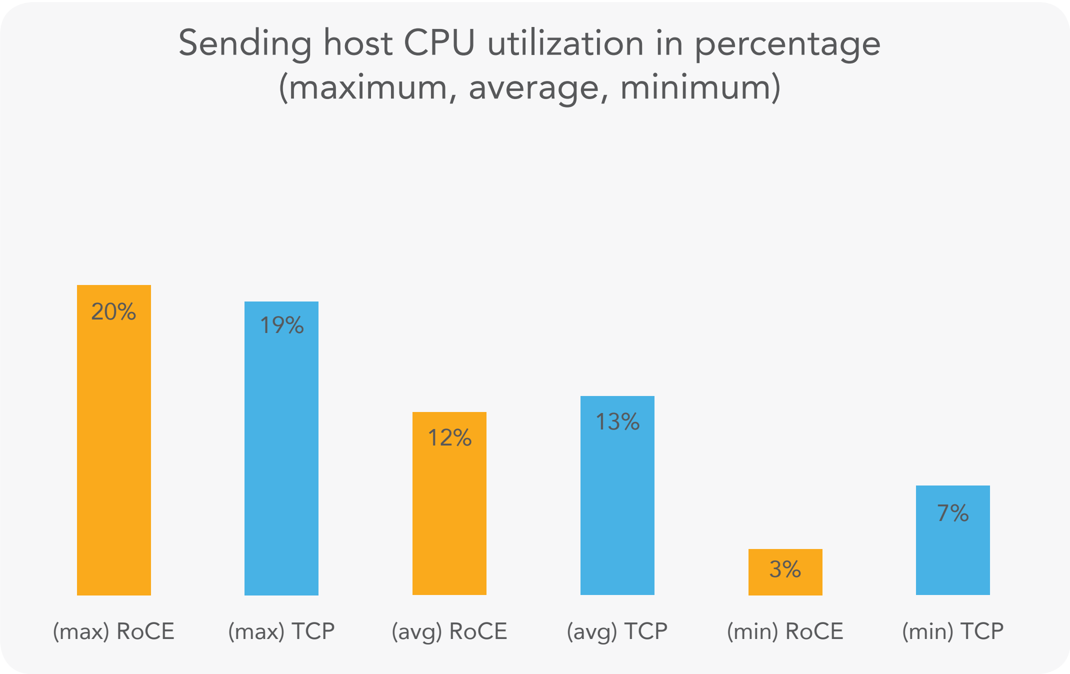
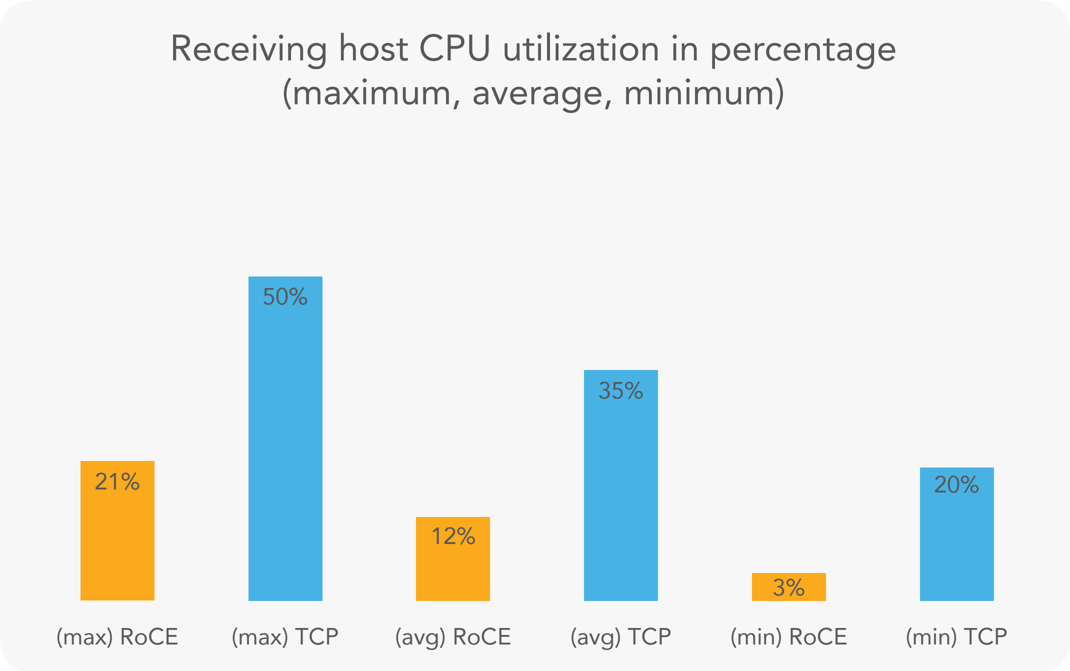


Figure 16. Receiving host CPU utilization: maximum, average, and minimum, compared between RoCE and TCP



As expected, CPU utilization is dramatically less with RoCE.

Conclusions

RDMA is available for Ethernet networks, and improves storage and virtual machine migration performance for Windows server SMB environments. Ethernet networks based on Cisco switches and Mellanox network adapters can easily be configured to take advantage of RDMA technology. In Windows virtualization and bare metal environments, using RDMA frees up CPU resources for application and storage workloads instead of networking. The examples above show that RDMA RoCE protocol performs significantly better than TCP/IP for SMB traffic, not only for pure storage access, but also for application traffic, such as live migration. We see less CPU usage with RoCE even though the network bandwidth produced is higher.

Using Cisco Nexus Switches with Mellanox network adapters can help you to unlock the performance of RoCE in SMB Direct. Based on the technologies and use cases described in this white paper, you can set up a cost-effective storage and application network based on RoCE in you existing data center fabric.

Where to go next

Mellanox community:

- How to configure RoCE in a Windows environment using Global Pause https://community.mellanox.com/docs/DOC-1844
- How to configure SMB Direct over RoCE while using PFC on Windows Server 2012 https://community.mellanox.com/docs/DOC-1853
- How to run RoCE over L2 enabled with PFC (for Linux): https://community.mellanox.com/docs/DOC-1414
- How to configure ECN parameters for Windows: https://community.mellanox.com/docs/DOC-2917
- Link to download Mellanox Windows drivers: https://www.mellanox.com/page/products_dyn?product_family=32&mtag=windows_sw_drivers

Cisco web resources:

- Cisco.com/go/nexus9000
- Cisco.com/go/ucs

Appendix: Configuration overview

The Cisco Nexus configuration used in the test cases for this document. All parameters for buffer tuning, thresholds, etc., are example values. They need to be adjusted to reflect the deployment scenario.

Global Quality of Service (QoS) configuration:

QoS class map to match Cost of Service (CoS) 4/Differentiated Services Code Point(DSCP 34) tagged RoCE traffic:

```
class-map type qos match-all RoCE_qos_class
 match cos 4
 match dscp 34
```

QoS policy map to set qos-group:

```
policy-map type qos RoCE_qos_policy
 class RoCE_qos_class
 set qos-group 4
```

QoS queuing policy-map to ensure bandwidth to RoCE traffic (RoCE traffic is set to qos-group 4, which maps to c-out-8q-q4):

```
policy-map type queuing ROCE_queuing8q-out
 class type queuing c-out-8q-q4
 bandwidth remaining percent 50
 random-detect minimum-threshold 100 kbytes maximum-threshold 1500 kbytes drop-probability 40 weight 0 ecn
 class type queuing c-out-8q-q-default
 bandwidth remaining percent 50
policy-map type queuing ROCE_INGRESS_queuing
 class type queuing c-in-q4
 pause buffer-size 62400 pause-threshold 62400 resume-threshold 60736
```

QoS network class to identify RoCE traffic for no-drop:

```
class-map type network-qos RoCE_network_class
  match qos-group 4
```

QoS network policy to set JUMBO MTU and no-drop to RoCE traffic:

```
policy-map type network-qos RoCE_network_policy
  class type network-qos RoCE_network_class
    pause pfc-cos 4
    mtu 9216
  class type network-qos c-8q-nq-default
    mtu 9216
```

Apply the created policies for no-drop and queuing (ingress and egress) to the system:

```
system qos
  service-policy type network-qos RoCE_network_policy
  service-policy type queuing output ROCE_queuing8q-out
  service-policy type queuing ROCE_INGRESS_queuing
```

Apply classification policy to interface ingress:

```
interface Ethernet1/51/1
  priority-flow-control mode on
  service-policy type qos input RoCE_qos_policy
```

Mellanox adapter configuration used in this test

Adapter/host setting	Recommended value	Tool/GUI/command
Install Mellanox Windows OFED.	WinOF 5.25 (for ConnectX-3 NICs)	Download link.
Set adapter IP addresses.	Appropriate for local network	Network connection GUI; Mellanox local area connection. Click properties. Fill in proper IP addresses.
Verify IP connectivity.	Ping shows a live connection	**ping <IP Address>**
Configure flow control (aka global pause) or PFC.	Flow control = "Rx & Tx Enabled"; or configure PFC	Device manager, adapters, port configuration, advanced tab, set flow control value.
Install DCB if using QoS or PFC.	"Data Center Bridging" installed (Install using GUI or PowerShell.)	Server manager, add roles and features, select server, select features, enable "Data Center Bridging." Or: Windows command line: Install-WindowsFeature Data-Center-Briding
Disable DCBx.	Set to 0.	Set-NetQosDcbxSetting –Willing 0
Check adapter RoCE mode.	Set to 2.	Get-MlnxDriverCoreSetting
Set adapter RoCE mode; "0" = no RoCE; "1" = RoCEv1.	"2" (for RoCEv2)	Set-MlnxDriverCoreSetting – RoceMode 2
Check PFC configuration	**Enabled only for RoCE priority**	**Get-NetQosFlowControl**
Set up VLAN (for PFC).	VLAN ID = <VLAN value> Possible VLAN values 1-4094	Device manager, adapters, port configuration, Advanced tab; enter "VLAN ID" value.

Adapter/host setting	Recommended value	Tool/GUI/command
Enable PFC (if not using global pause) on specific priority; disable it for other priorities.	Enable PFC for priority 4 (or whatever priority you have chosen for SMB Direct traffic).	Enable-NetQosFlowControl –Priority 3 Disabl-NEtQosFlowControl 0,1,2,4,5,6,7
Enable QoS on the interface.	Interface name (alias) being used for SMB Direct	**Enable-NetAdapterQos –InterfaceAlias "<aliasname>"**
Limit bandwidth being used by SMB traffic (optional).	Bandwidth % limit for SMB traffic, i.e. "60" = 60%	**NewNetQoSTrafficClass "SMB" –Priority 3 –Bandwidth 60 –Algorithm ETS**
Verify or change IP addresses.		
Verify RDMA running for SMB.	PowerShell cmdlets	**Get-NetOffloadGlobalSetting | Select NetworkDirect** **Get-NetAdapterRDMA** **Get-NetAdapterHardwareInfo**
Verify SMB RDMA on client.		**Get-SmbClientConfiguration | Select EnableMultichannel** **Get-SmbClientNetworkInterface**
Verify SMB RDMA on server.		**Get-SmbServerConfiguration | Select EnableMultichannel** **Get-SmbServerNetworkInterface** **netstat.exe -xan | ? {$_ -match "445"}**
Verify SMB Connect	Start copy from client, then run PowerShell cmdlets while copy is running.	**Get-SmbConnection** **Get-SmbMultichannel Connection** **netstat.exe -xan | ? {$_ -match "445"}**

C11-741091-00 10/18