Mellanox TECHNOLOGIES

Zero Touch RoCE

IBM

The Easiest Way to Deploy RDMA on Windows Storage Spaces Direct

Microsoft took RDMA mainstream in Windows Server 2012 by introducing SMB Direct to improve file server performance. Within Windows Server 2016, Storage Spaces Direct (Windows S2D) leverages Microsoft SMB Direct to accelerate east-west traffic and storage access. However, some RDMA technologies rely heavily on Data Center Bridging (DCB) configurations on the switch, requiring a network administrator who is familiar with DCB and its components, ECN and PFC.

Mellanox, the leader in high-performance Ethernet, brings extensive expertise in the RDMA domain and is advancing the RDMA landscape with the introduction of Zero Touch RoCE. Typically, RoCE is deployed over "lossless" Ethernet which requires familiarity with ECN and PFC as well as a switch that is capable of supporting these features. Mellanox RoCE is fully compliant with the RoCE industry standard and utilizes enhanced congestion mechanisms that do not require any special configuration when using Mellanox ConnectX-4 or ConnectX-5 network adapters – a true zero configuration network for RDMA over Windows S2D environments! Install the Mellanox adapters just like you would any NIC, then let Mellanox do the rest to give you, fast and reliable data access. Perfect for small cluster configurations, hyperconverged deployments, to increase efficiency within scale-out based storage architectures and anyone looking to improve storage performance.

Increases Performance, Scalability and Efficiency with Zero Configuration

6X The Throughput

Compared to iWARP

<1usec Latency

VM to VM Communication

<2% CPU Utilization

Delivering I/O at 25Gbps

Performance

Higher Throughput & IOPS

Scalability

Reduce Overhead

Efficiency

Lower CPU Utilization

Mellanox RoCE Delivers Lower Latencies and Higher Performance

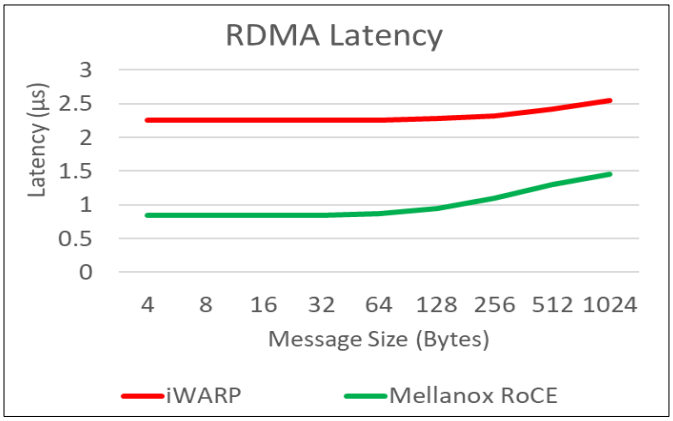
Zero Configuration Required

Accelerate Your Server & Network Infrastructure With Mellanox Zero Touch RoCE to Boost Application Performance and Scalability!

RoCE Advantages over iWARP

An alternative RDMA technology is iWARP, which is unable to achieve the same level of performance as RoCE-based solutions. iWARP uses a complex mix of layers over TCP/IP and a separate RDMA protocol (RDMAP) to deliver RDMA services over TCP/IP. This convoluted architecture is an ill-conceived attempt to fit RDMA into existing software transport framework with respect to congestion handling, scaling, and error handling, causing inefficiency in hardware offload of the associated transport operations. Unfortunately this compromise causes iWARP to fail to deliver on precisely the three key benefits that RoCE is able to achieve: high throughput, low-latency, and low CPU utilization.

RoCE, on the other hand, is a purpose-built RDMA transport protocol for Ethernet, not a patch to be used on top of existing TCP/IP protocols. Because of this, iWARP faces challenges that limit the cost-effectiveness and deployment of iWARP products in comparison to RoCE.



4 Node Microsoft S2D Cluster over Lossy Network

5 Myths about RoCE

Although RDMA has been well received by the storage and networking industry, some misinformation about the interconnect technology still remains.

1. RoCE requires a lossless network.
Initial deployments of RDMA required configuring the network to be lossless (RoCE). However, Mellanox RDMA is resilient to packet loss and is able to run over ordinary Ethernet networks without the need for configuration of priority-based flow control. This enables cloud, storage, and enterprise customers to deploy RDMA quickly and easily with zero configuration.

2. RoCE doesn't scale.
RDMA is currently deployed within Microsoft's Azure Cloud, one of the largest cloud service providers, connecting tens of thousands of compute and storage nodes.

3. RoCE only works over short distances.
While the best latency performance is achieved over short distances, RoCE frames can travel anywhere traditional Ethernet frames can including over Metro Ethernet, supporting distances up to 10 km.

4. RoCE can't handle advanced Ethernet rates.
RDMA was defined to run over any speed defined by the IEEE 802.3 Ethernet standard including 25, 40, 50, and 100 Gb/s.

5. All RDMA over Ethernet technologies offer the same efficiency and latency benefits.
RoCE, has been widely adopted, while iWARP, has seen only minimal support. Although both solutions offer RDMA capabilities over Ethernet, benchmark data comparing RoCE versus iWARP at 10 and 40 Gb/s shows that RoCE beats iWARP at each speed delivers lower latency and higher data throughput across all message sizes.

Mellanox Connect-X Adapters

Mellanox provides a variety of adapter form factors and speeds:

Description	Speed	Ports	Mellanox PN	FC#
ConnectX-4 Lx LP PCIe 3 x8 SFP28	10/25	2	MCX4121A-ACAT	EKAU
ConnectX-4 Lx LP PCIe 3 x8 SFP28	10/25	2	MCX4121A-ACAT	EC2T
ConnectX-4 HP PCIe 3 x16 QSFP28	100	2	MCX456A-ECAT	EKAL
ConnectX-4 LP PCIe 3 x16 QSFP28	100	2	MCX416A-CCAT	EC3L
ConnectX-4 LP PCIe 4 x16 QSFP28	100	2	MCX556N-EDAT	EC67
ConnectX-4 LP PCIe 4 x16 QSFP28	100	2	MCX556A-EDAT	EKAY

Available with ConnectX-4 and greater adapters and requires MLNX_OFED 4.6 or WinOF-2 2.20 and its associated firmware. Feature availability dependent on OEM