Mellanox Spectrum®— The RoCE-Ready Switch

Automatic RoCE Acceleration With Effortless Configuration

Ethernet is ubiquitous in today's data centers and cloud networks as a unified, scale-out infrastructure fabric that connects compute and storage. With increasingly powerful CPUs and GPUs and faster storage being utilized in scale-out infrastructures, the network fabric must also be sped up in order to transport the intensive data movements between compute/storage servers in the data center/cloud. High performance fabrics must transfer data between servers with low latency, between CPU, GPU, memory, and storage, so that the amount of time it takes to access remote data is almost the same as accessing local data. The resulting transparent data locality has become a mandatory requirement for high-performance and distributed applications, including:

- Machine Learning
- NVMe-oF
- In-memory databases
- Distributed file systems
- VM migration

KEY BENEFITS

- Effortless RoCE deployment
- End-to-end RoCE acceleration
- Real-time RoCE visibility for easy troubleshooting
- Single pane-of-glass RoCE management

RDMA (Remote Direct Memory Access) was developed to addresses this challenge. By eliminating latency-expensive data copying, CPU/GPU interrupts, and context switching, RDMA delivers the lowest latency for data tranfers between servers while offloading CPUs/GPUs for much improved server efficiency. RDMA over Converged Ethernet (RoCE) is the most-adapted RDMA implementation for data center/cloud applications over Ethernet. These applications include Microsoft Storage Spaces Direct, VMware vSphere and vSAN, Spark, Hadoop, Oracle RAC, and AI/ML frameworks. In this context, RoCE has become the eponym of low-latency networking in data centers and clouds today.

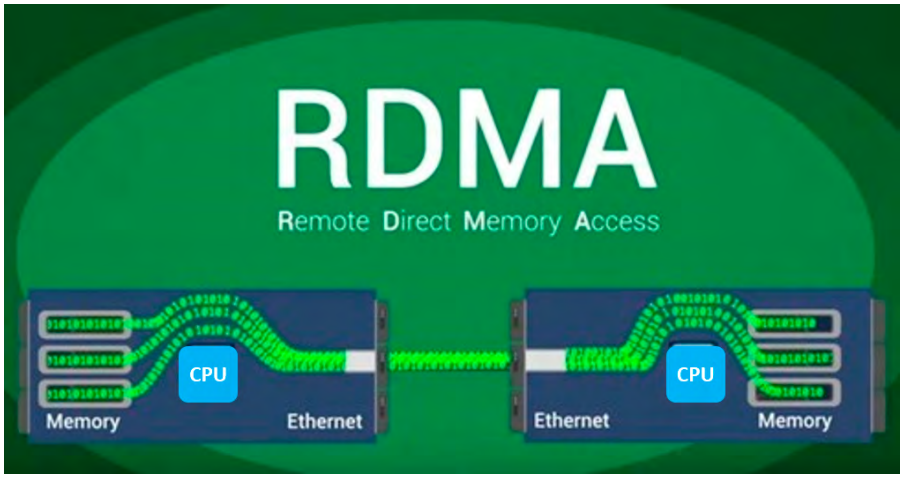

Figure 1. Remote Direct Memory Access (RDMA)

With its heritage in high-performance computing, Mellanox leads RDMA/RoCE technology development and offers the most mature and advanced RoCE solution in the industry. In particular, by being the only vendor that offers a complete end-to-end RoCE solution, Mellanox enables RoCE at its best in any Ethernet network, regardless of speed, topology, and scale.

©2020 Mellanox Technologies. All rights reserved.

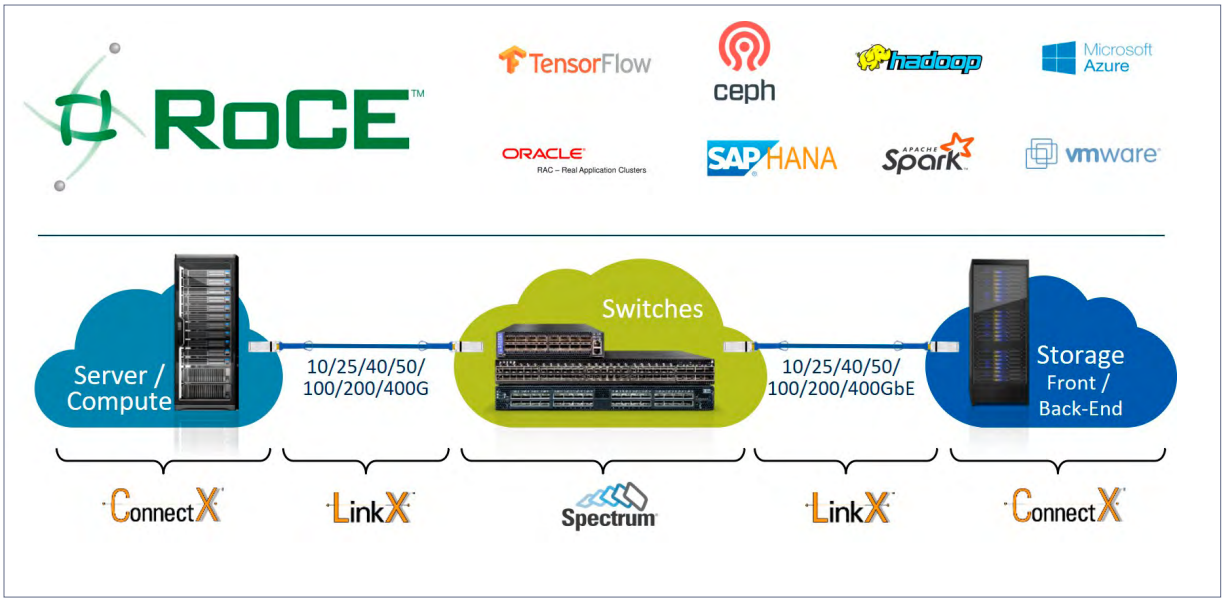

Figure 2. End-to-end Mellanox RoCE

While Mellanox ConnectX® network adapter cards provide zero-touch RoCE-ba hardware-acceleration in the server, Mellanox Spectrum switches deliver RoCE optimization in the network fabric. Mainly, Mellanox Spectrum switches optimize RoCE deployments in the following key areas:

- Effortless RoCE configuration
- Automatic RoCE acceleration
- Real-time RoCE visibility

Effortless RoCE Configuration

Depending on deployments, configuring the network fabric for RoCE involves multiple steps – classifying ingress traffic flows, setting QoS for these flows, and enabling congestion control notification, for example. Manually completing these steps for the switches is not trivial and often error-prone. In contrast, Mellanox Spectrum switches simplify the RoCE configuration with a single "roce" command, which applies a best-practices configuration for optimal performance.

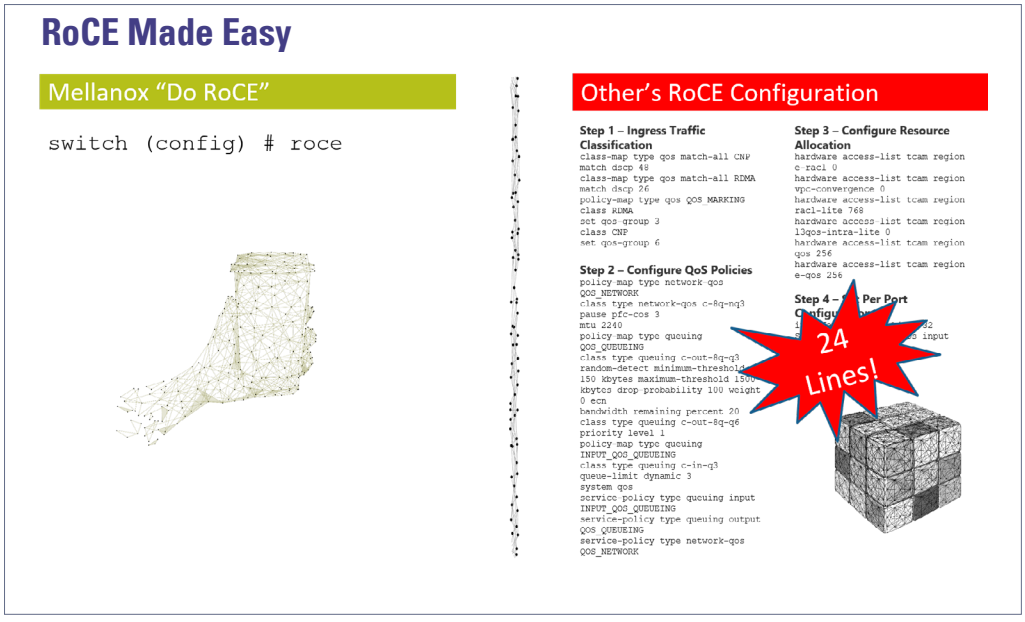

Figure 3. Command Mellanox RoCE Configuration

©2020 Mellanox Technologies. All rights reserved.

Mellanox also offers a GUI for easy RoCE configuration with its network orchestrator, Mellanox NEO®. With one click, NEO automatically configures RoCE on Spectrum Switches as well as ConnectX NICs for fabric-wide end-to-end configuration.

Automatic RoCE Acceleration

Automatic RoCE acceleration first stems from the high-performance and low-latency design of Mellanox Spectrum switches. Mellanox Spectrum switches provide line-rate throughput and ultra-low port-to-port switching latency at all speeds and packet sizes, with zero avoidable packet loss. Employing a shared-buffer architecture enables Mellanox Spectrum switches to deliver high performance and low latency fairly and predictably, which is crucial for software-defined platforms to run RoCE per defined priorities and policies without concerning the underlay switch characteristics.

Automatic RoCE acceleration also comes from innovations in advanced congestion control. Mellanox Spectrum switches support both RoCEv1 and RoCEv2. In supporting per-flow congestion notification using Explicit Congestion Notification (ECN), Mellanox Spectrum switches offer a FAST ECN feature, which allows faster responses to congestion events. Once congestion is detected, instead of marking packets as they enter the queue (at the tail of the queue), Mellanox Spectrum switches mark packets when they leave the queue (at the head of the queue). As a result, the congestion notification is received up to milliseconds sooner. Earlier received alerts, in turn, reduces the chance of congestion occurring, and improves overall application performance.

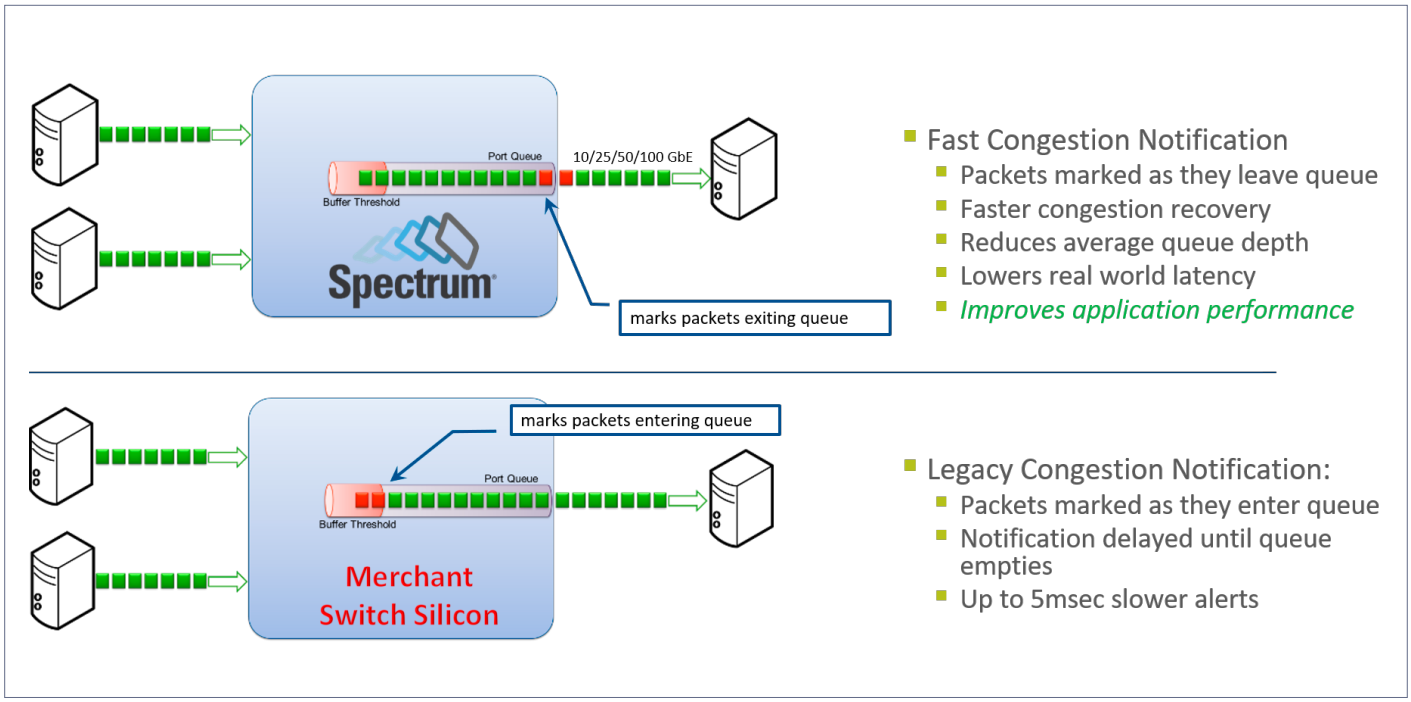

Figure 4. Faster Congestion Notification by Mellanox RoCE

Fully shared buffers in Mellanox Spectrum provide better capabilities at absorbing microbursts while avoiding the need to send congestion notifications. Better per-flow explicit congestion notification (ECN) handling prevents congestion spreading to "victim flows," enabling RoCE to be used on a large scale without complex traffic engineering.

Real-time RoCE Visibility

Real-time network telemetry is critical to manage, orchestrate, and troubleshoot/remediate the network, especially for latency-sensitive RoCE data flows. With a single command, "show roce" Mellanox Spectrum switches provide advanced RoCE telemetry for real-time visibility of the RDMA traffic, including counters of RoCE traffic and non RoCE traffic, congestion counters, as well as current & high-water buffer usage.

©2020 Mellanox Technologies. All rights reserved.

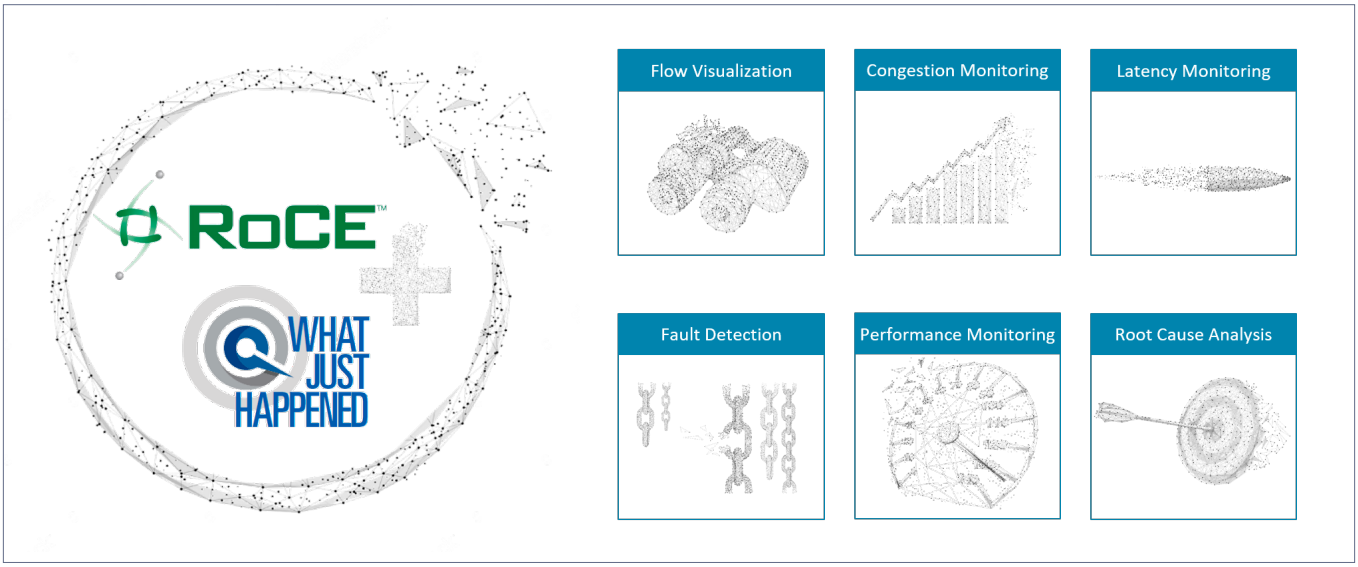

Figure 5. RoCE Monitoring with What Just Happened™

Mellanox's What Just Happened™ (WJH) advanced telemetry technology provides additional hardware accelerated root-cause analysis to RoCE data flows - to continuously validate flow-based configurations, detect traffic patterns and congestion conditions, and instantly inform you of when things go wrong and why they go wrong.

In addition, configuring and monitoring RoCE can be done from a single-pane-of-glass within the NEO platform, which provides flow-based network visibility, one-click network provisioning, automated monitoring and reporting, and quick troubleshooting.

Summary: Mellanox Spectrum is the Only RoCE-Ready Switch

With business applications running on fast NVMe storage and ever-growing extensive data sets, RoCE is the de facto implementation of low-latency networking for these applications. RoCE is designed to support any Ethernet networks; however, RoCE works the best and the easiest with Mellanox end-to-end Ethernet solution. While Mellanox ConnectX NICs provide hardware RoCE acceleration to make RoCE transparent, Mellanox Spectrum switches do the same inside the network fabric. Mellanox Spectrum switches provide effortless RoCE configuration with a single command, real-time RoCE visibility for fast diagnostics and automatic acceleration for all RoCE use cases.

About Mellanox

Mellanox Technologies is a leading supplier of end-to-end Ethernet interconnect solutions and services for enterprise data centers, Web 2.0, cloud, storage, AI and networked edge. More information is available at: www.mellanox.com

350 Oakmead Parkway,
Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400
Fax: 408-970-3403
www.mellanox.com

© Copyright 2020. Mellanox Technologies. All rights reserved.
Mellanox, Mellanox logo, ConnectX and Mellanox Spectrum are registered trademarks of Mellanox Technologies, Ltd.
What Just Happened is a trademark of Mellanox Technologies, Ltd. All other trademarks are property of their respective owners.

60366SB
Rev 1.0