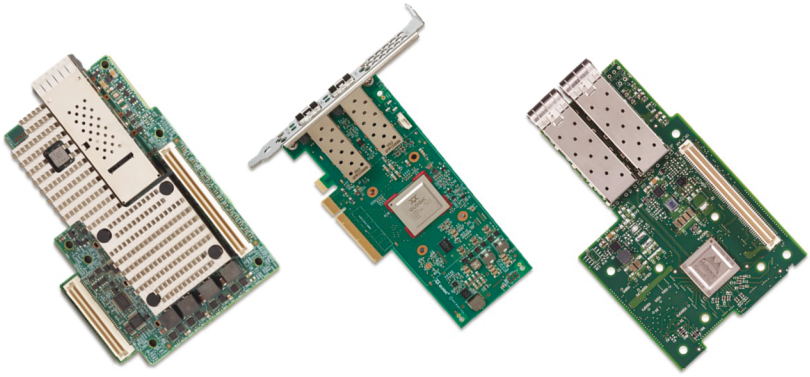
Electronic Design

BILL LEE, Director of Marketing, Mellanox,
ROBERT LUSINSKY, Director of Marketing, Broadcom Ltd.

11 Myths about RDMA over Converged Ethernet (RoCE)

Although RDMA over Converged Ethernet (RoCE) has been well received by the enterprise storage and networking industry, some misinformation about the interconnect technology still remains.



RoCE-capable adapter cards are available today from InfiniBand Trade Association members (left to right) Broadcom, Cavium, and Mellanox Technologies.

Remote direct memory access (RDMA) is a well-known technology at the heart of the world's fastest supercomputers and largest data centers. In short, RDMA is a remote memory-management capability that enables server-to-server data movement directly between application memories without CPU involvement. Offloading data movement from the CPU will result in performance and efficiency gains, while also significantly reducing latency. RDMA first became widely adopted in the High Performance Computing (HPC) industry with InfiniBand, but is now being leveraged by cloud, storage, and enterprise Ethernet networks with RDMA over Converged Ethernet (RoCE).

Given its broad expertise in RDMA technology, the InfiniBand Trade Association (IBTA) developed the RoCE standard and released the first specification in 2010. Although RoCE has been well-received by the enterprise storage and networking industry, especially by those wanting to accelerate application performance without overhauling their existing Ethernet infrastructure, there's still some misinformation that continues on about the technology. Read on as we lay out and address the 11 most common myths surrounding RoCE.

1. ROCE REQUIRES A LOSSLESS NETWORK.

Initial deployments of RoCE required configuring the network to be lossless. However, the most advanced implementations of RoCE are resilient to packet loss and are able to run over ordinary Ethernet networks without the need for priority-based flow control. This resilient RoCE enables cloud, storage, and enterprise customers to deploy RoCE more quickly and easily while accelerating application performance, improving total infrastructure efficiency, and reducing cost.

2. ROCE DOESN'T SCALE.

RoCE is currently deployed within Microsoft Azure Cloud, one of the largest cloud service providers, connecting tens of thousands of their compute and storage nodes.

3. ROCE ISN'T ROUTABLE.

In 2014, the IBTA added routing capability to the specification, and member companies are shipping RoCE adapters supporting IP routing today. The release enabled routing across Layer 3 networks, extending RoCE to provide better traffic isolation and enable hyperscale data-center deployments. In fact, the vast majority of RoCE deployments are routed across Layer 3 networks.

4. ROCE ONLY WORKS OVER SHORT DISTANCES.

While the best latency performance is achieved over short distances, RoCE frames can travel over any wire that's traveled by traditional Ethernet frames. This includes between floors and buildings over LR and LR4 as well as over Metro Ethernet, supporting distances up to 10 km.

5. COMMON TRAFFIC-MANAGEMENT AND -MONITORING TOOLS DON'T WORK WITH ROCE.

RoCE utilizes IPv4 and IPv6 encapsulation on Ethernet, the same as most other Ethernet traffic. This allows RoCE to be monitored and managed with existing tools.

6. ROCE IS NOT AN OPEN STANDARD.

RoCE was developed by the IBTA under the same standardization processes as other parts of the InfiniBand architecture. Multiple IBTA members contributed to defining the open standard, which required all IP to be made available under non-discriminatory licensing terms. The IBTA continues to unite leading industry players with a common goal of advancing the RoCE ecosystem in a vendor-neutral, community-centric manner.

7. ROCE CAN'T HANDLE ADVANCED ETHERNET SIGNALING RATES.

RoCE was defined to run over IEEE 802.3 Ethernet, and as such can run over any speed defined by that organization. As of the publication of this article, there are RoCE adapters supporting 25, 40, 50, and 100 Gb/s.

8. ALL RDMA OVER ETHERNET TECHNOLOGIES OFFER THE SAME EFFICIENCY AND LATENCY BENEFITS.

The Ethernet protocol is immensely popular and considered by many to be the backbone of the modern data center. As data centers began to adopt faster Ethernet networks, IT managers also saw a greater need for the lower latency capabilities of RDMA.

This demand spawned two RDMA over Ethernet protocols—RoCE, which has been widely adopted, and iWARP, which has seen only minimal support. Although both solutions offer RDMA capabilities, benchmark data comparing RoCE versus iWARP at 10 and 40 Gb/s shows that RoCE at each speed delivers lower latency and higher data throughput across all message sizes.

9. ROCE LACKS SUPPORT FROM MULTIPLE VENDORS.

In addition to multiple vendors coming together to define the specification, a growing number of companies are either shipping or have announced RoCE hardware and software solutions as shown in the RoCE Product Directory on the IBTA's RoCE Initiative site.

10. ROCE INTEROPERABILITY BETWEEN DIFFERENT VENDORS IS UNRELIABLE.

Furthering the point above, RoCE solutions undergo rigorous interoperability testing at IBTA Plugfests, which are held twice a year at the University of New Hampshire InterOperability Laboratory. The test results are published in a RoCE Interoperability List that's designed to support data-center managers, CIOs, and other IT decision makers with their planned RoCE deployments in enterprise computing systems. The free-to-download list provides valuable information on cross-vendor interoperability and contains a variety of RoCE devices, including 10, 25, and 40 GbE RNICs, switches, and SFP+, SFP28, and QSFP cables.

11. ROCE IS DIFFICULT AND EXPENSIVE TO DEPLOY.

As with any new capability being integrated across server, adapter and switching technologies, additional technical understanding and guidance may be needed to take full advantage of RoCE. To help IT professionals implement RoCE capabilities in their systems, the IBTA recently teamed up with Demartek, a computer industry analyst organization, to create the RoCE Deployment Guide. The document explains the advantages of RoCE and shows IT managers how to effectively deploy RoCE-enabled solutions quickly and with ease. RoCE adapters are priced competitively to other, general purpose Ethernet adapters. RoCE can be deployed on existing Ethernet switch fabrics and use the same Ethernet cables, so there is no additional costs there either.

BILL LEE is a Director of Marketing at Mellanox Technologies, responsible for marketing operations, market analysis, and partner relationships. Bill has been working in the networking industry for over 20 years in both engineering and marketing roles. Prior to joining Mellanox, Bill worked at various companies, such as National Semiconductor, Galileo Technology, and Marvell Technology Group. He is currently the co-chair for both the InfiniBand Trade Association and OpenFabrics Alliance Marketing Working Groups. Bill received his BSEL from the California Polytechnic State University in San Luis Obispo.

ROBERT LUSINSKY is a Director of Marketing at Broadcom Ltd., responsible for inbound and outbound marketing of Ethernet controllers. Robert has been working in the technology industry for over 20 years in both marketing and engineering roles, holding several patents in networking and system design. Prior to joining Broadcom, Robert worked at various companies such as Phoenix Technologies and Xiran, and is currently the co-chair for the InfiniBand Trade Association Marketing Working Group. Robert received his BSEE and MSEE from California State University, Fullerton, and his MBA from the University of Southern California.