RoCE: Frequently Asked Questions

The truth about the capabilities and limitations of RoCE hasn't exactly been forthcoming, leaving customers interested in RDMA over Ethernet with many unanswered questions. This paper compiles a list of answers to those frequently asked questions.

Is RoCE the standard RDMA over Ethernet protocol?

NO – The IETF standard for RDMA is iWARP. It provides the same host interface as InfiniBand and is available in the same OpenFabrics Enterprise Distribution (OFED).

Does RoCE lower CPU utilization?

YES – However, applications get lower CPU utilization because of RDMA, not RoCE. Using iWARP provides the same benefits.

Does RoCE reduce memory copies?

YES – Again, zero copy is a benefit of RDMA, also provided by iWARP.

Does RoCE allow user-space I/O?

YES – Similarly to zero copy, user-space I/O is a benefit of RDMA, also provided by iWARP.

Is RoCE an alternative to InfiniBand?

NO – Although RoCE gives good micro-benchmarks results, it lacks critical pieces of the IB stack and is neither scalable nor competitive as an Ethernet solution.

Is RoCE more efficient than iWARP?

NO – Both protocols have similar header sizes and hardware TCP/IP implementations provide similar performance to InfiniBand, without all the limitations.

Is RoCE easy to deploy and use?

NO – Unlike iWARP, RoCE requires a complicated layer-2 configuration for lossless operation, and has been found to be very difficult to deploy, even by experienced IT staff.

Does RoCE take advantage of Ethernet economies of scale?

NO – Unlike iWARP, RoCE does not operate with standard switches, and requires the more expensive, DCB capable types.

Does RoCE inter-operate with switches from different vendors?

NO – RoCE does not work with non-DCB switches, and depends on configuring Priority Flow Control consistently throughout the network, which adds many inter-operability challenges.

Can RoCE share a channel with other traffic?

NO – RoCE is very sensitive to packet drop and requires a dedicated priority channel for its traffic.

Does making QoS configuration changes to a switch affect RoCE's operation?

YES – If a switch is configured to treat QoS classes differently than expected, it may easily result in the collapse of RoCE performance.

Does RoCE restrict QoS traffic marking configuration?

YES – Any switch configured to re-mark traffic priority can break down the uniformity of RoCE frame treatment in the network, and result in dismal performance.

Does RoCE scale?

NO – A RoCE network must have PAUSE enabled in all switches and end-stations, which limits the deployment scale of RoCE to single hop at best.

Does RoCE real application performance match micro-benchmarks?

NO – Although RoCE may perform well in simple single hop scenarios, real application performance can fall short of expectations, particularly when network hotspots are involved.

Does RoCE operate over long distance links?

NO – PFC limits RoCE operation to a few hundred meters.

Does RoCE operate over WAN links or cross subnet boundaries?

NO – PFC does not operate beyond a subnet.

Is RoCE routable?

NO – Although RoCE may use IPv6-like addresses, RoCE does not use a standard IP header and cannot be routed by standard IP routers.

Can I use standard traffic management and monitoring tools for RoCE?

NO – Most traffic management and monitoring tools have been developed for IP applications. RoCE does not use IP and therefore is unrecognized by existing tools.

Can I configure RoCE congestion management to suit my environment?

NO – The congestion management layer for RoCE is non-existent, RoCE being completely dependent on PAUSE for operation.

Summary

RoCE raises many questions when practical deployment issues and limitations are encountered, and the answers are almost always cause of concern to potential users.

The standard for RDMA over Ethernet is iWARP, which uses the familiar TCP/IP stack as foundation. This allows it to use existing hardware, live alongside existing applications and use existing management and monitoring tools. High performance iWARP implementations are available, and compete directly with InfiniBand in real application benchmarks.

iWARP is supported in the same OpenFabrics Enterprise Distribution as IB for Linux, and is similarly available on Windows and BSD systems for a drop-in Ethernet replacement of IB.

Related Links

IBM Research Report on IB and 10GbE Performance for HPC Applications
IBM/Blade Networks Presentation
Cisco 10G for ECLIPSE Reservoir Simulation
Open Fabrics Enterprise Alliance

References

[1] *RFC 5040 - A Remote Direct Memory Access Protocol Specification*
[2] *RFC5041 - Direct Data Placement over Reliable Transports*
[3] *OpenFabrics Enterprise Alliance*
[4] *Priority Flow Control - Building a Reliable Solution, Cisco Systems*
[5] *LAMMPS, LS-DYNA and LINPACK on RoCE vs. iWARP*
[6] *RDMA over Converged Ethernet, A Personal Obsession*