RoCE: The Missing Fine Print

The promise of RoCE, or InfiniBand over Ethernet, is to bring RDMA's benefits to Ethernet. In the interest of truth in advertising, here is the missing fine print.

RoCE is difficult and expensive to deploy

RoCE depends on creating a lossless Ethernet network. Deploying a RoCE solution requires enabling support for a new IEEE standard for Priority Flow Control, part of Converged Ethernet (CE) or Data Center Bridging (DCB). Using PFC implies consistent priority marking and treatment of RoCE frames throughout the network. This translates into strict requirements for inter-operability among switches, and between switches and end-stations. This task has proven challenging even for experience IT staff and network architects, requiring considerable staff resources and time. Furthermore, RoCE requires DCB switches which are far more expensive and complex than standard switches.

In contrast, iWARP does not require any special treatment in the network. iWARP is the easiest to deploy RDMA solution [3].

RoCE may not interoperate with switches from different vendors

PAUSE capabilities and implementations differ widely between different vendors and sometimes even between different product lines of a single vendor, therefore imposing serious deployment restrictions on ROCE. These restrictions are worsened when looking at DCB implementations, small changes in switch configuration can break support for RoCE [3].

In contrast, iWARP uses TCP/IP as transport and is therefore identical to other protocols running over TCP/IP, such as FTP, NFS or HTTP. These protocols run perfectly over all Ethernet networks without requiring any special configuration.

Making QoS configuration changes to a switch may affect RoCE's operation

RoCE depends on priority PAUSE, and this forces dependencies and limitations on the QoS configuration of every switch in the network. If a switch is configured to treats QoS classes differently than expected, it may easily result in the collapse of RoCE performance.

In contrast, iWARP does not require specialized QoS, and although it can make use of network QoS, it gets high performance regardless of the QoS configuration of the switches and end-stations.

RoCE restricts QoS traffic marking and configuration freedom

For the same reasons as above, any switch configured to re-mark traffic priority based on VLAN or other setting, can also result in breaking down the uniformity of RoCE frame treatment in the network, and therefore dismal performance.

In contrast, iWARP can be treated exactly the same as any generic Ethernet traffic, marked and remarked at will.

A RoCE network may not scale as expected

A RoCE network must have PAUSE enabled in all switches and end-stations. The basic PAUSE mechanism, link-level back-pressure, has been available in Ethernet for more than 5 years. It has known limitations which can result in congestion propagating from a hotspot to the whole network, grinding it to a halt. This has limited the deployment of PAUSE to first tier switches, if at all, which directly limits the deployment scale of RoCE.

Furthermore, RoCE uses a new EtherType which is unknown to most switches and bypasses the value added features typically available in today's L2/L3 switches. In particular, they do not benefit from critical features such as load balancing or multi-pathing.

In contrast, iWARP uses proven TCP/IP mechanisms for congestion avoidance and control and therefore does not require enabling PAUSE throughout the network. In addition, iWARP frames being identical to other TCP/IP/Ethernet frames means that they automatically benefit from the decades of performance optimizations in switches, as well as monitoring and debugging tool development.

RoCE real application performance may not match micro-benchmarks

The same problems limiting the scalability of RoCE deployment can directly impact application level performance. Although RoCE may perform well in simple single hop scenarios, real application performance can fall short of expectations, particularly when hotspots are involved.

In contrast, iWARP does not require enabling link layer flow control throughout the network, and is able to handle real application traffic patterns.

RoCE cannot operate over long distance links

A direct consequence of needing PAUSE is a limit of the distance between two peers in a RoCE network. This precludes RoCE from operation over longer than a few hundred meters [1].

In contrast, iWARP can run over long distance links, like any TCP/IP traffic, and is limited only by optical transceiver capabilities.

RoCE cannot operate over WAN links or cross subnet boundaries

There are two critical limitations which prevent RoCE traffic from going across WAN links. First, the lossless network requirement has no chance of getting satisfied once traffic crosses a subnet boundary since PAUSE does not operate beyond that limit. Second, RoCE does not use

IP and therefore is not recognized by IP routers. It therefore cannot be delivered beyond a single subnet.

In contrast, iWARP uses IPv4 or IPv6 and can be routed and delivered without limitations.

RoCE routing is not what is expected

The claim that RoCE is routable does not adhere to the conventional definition of routability, which is generic delivery of packets across potentially heterogeneous IP subnets. RoCE does not use standard IP and cannot be routed by standard IP routers. In reality, it is InfiniBand (IB) subnet connectivity and therefore requires specialized IB gateways, which are single sourced.

In contrast, iWARP is recognized by IP equipment and routed like any other IP traffic.

RoCE requires new traffic management and monitoring tools

Most traffic management and monitoring tools have been developed for IP applications. RoCE does not use IP and therefore is unrecognized by existing tools. Deploying RoCE necessitates a potentially costly upgrade to all network management and monitoring tools.

In contrast, iWARP over TCP/IP leverages all existing tools and investments, resulting in significantly lower TCO.

RoCE congestion management is not configurable

The congestion management layer for RoCE is non-existent, RoCE being completely dependent on PAUSE for operation. This means that a network operator has no control on RoCE's behavior in an infrastructure. This lack of control can severely limit the deployability of RoCE outside small, homogenous clusters.

In contrast, iWARP benefits from decades of experience with running TCP over a wide range of network technologies and in extreme network conditions. Good iWARP implementations offer similar control on TCP's behavior to software stacks, allowing it to be tuned for any network characteristics.

Summary

The potential benefits of RoCE come not from its design, but from RDMA, which provides zero copy and kernel bypass (user-space I/O). iWARP is the Ethernet standard for RDMA, and provides these benefits without compromises or a long list of fine print, while matching native IB in application level benchmarks [2].

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] *Priority Flow Control - Building a Reliable Solution, Cisco Systems*
[2] *LAMMPS, LS-DYNA and LINPACK on RoCE vs. iWARP*
[3] *RDMA over Converged Ethernet, A Personal Obsession*