

# A Rocky Road for RoCE

---

The industry standard clustering protocol over Ethernet is iWARP, published by the IETF in 2007 in RFC 5040 and RFC 5041. iWARP specifies an RDMA Protocol over TCP/IP, which enables running unmodified InfiniBand applications using the same OpenFabrics Enterprise Distribution (OFED) software stack, with similar application performance levels.

Recently, a new Ethernet clustering protocol, named RoCE, has been proposed by an InfiniBand vendor through a trade organization. RoCE is essentially InfiniBand over Ethernet, where the lower virtual circuit layers of the IB stack are replaced with raw Ethernet encapsulation. For an IB vendor, RoCE may be an easy to effect transition to Ethernet, but it comes with serious caveats to its potential users. To start, it is important to highlight a fundamental assumption in running IB over raw Ethernet, namely that the Ethernet network is lossless. Without lossless operation, performance hotspots quickly lead to performance collapse due to the lack of TCP's critical congestion avoidance and management layer. This assumption rides on the tails of a recent push for mixing Fibre Channel traffic with other traditional networking traffic over "Converged Ethernet," where a similar guarantee is necessary, leading to popularizing the myth of "Lossless Ethernet." Let us examine this assumption in more detail.

## *RoCE is IB over Converged Ethernet*

1. The missing IB layers provide a reliable, flow controlled fabric. Ethernet, on the other hand, has no such guarantees, and this simplicity has in fact been one of its main strengths.
2. Converged Ethernet (also called Data Center Bridging or DCB) refers to a collection of proposed enhancements to traditional Ethernet:
  - a. Per-Priority Pause (PPP).
  - b. Enhanced Transmission Selection, used for provisioning bandwidth when multiple traffic types are in use. ETS is not relevant in a network carrying a single traffic type, and in particular, high performance computing environments, where low latency of operation is the critical metric.
  - c. Congestion notification, which is an end-to-end explicit notification scheme. Deploying this protocol requires replacing all network interface cards and switches. More importantly, it is not widely implemented or supported in the switching industry, mainly because of the doubts on its scalability, effectiveness and congestion messaging impact in large networks. In fact, while explicit congestion notification has been proposed many times in the past, it never gained acceptance beyond hop-by-hop uses for these reasons.
3. The "Converged" part therefore boils down to PPP, which provides isolation, and therefore is only relevant when the same wire is carrying different traffic types (priorities). When a single traffic type is in use, there is no difference between CE and regular Ethernet with PAUSE. There is no epiphany.

*The assumption that CE provides lossless operation is therefore based on a false premise*

1. The same concerns that have prevented widespread use of PAUSE in large networks apply: the network runs at the speed of its slowest link, congestion propagates quickly and results in gridlock, i.e. generalized poor performance.
2. Therefore, RoCE is really IB over good old Ethernet, and will have to fall back on inefficient error recovery as the network scales and packet loss increases.
3. Error recovery without congestion control exposes the network to congestion collapse.

In addition to the fundamental issues with running IB over raw Ethernet, RoCE suffers from a number of other serious limitations. First, since it is not routable, it requires flat L2 networks and as a result the issues above are amplified. Secondly, RoCE is not amenable to load balancing and link aggregation, therefore the flat L2 networks needed will have limitations at the core.

### Reality

In conclusion, by throwing overboard critical pieces of the stack which provide stability and scalability, RoCE shines at simple micro-benchmarks in back-to-back or similarly limited deployment scenarios. However, it stands to fail in clustered application performance, where all its limitations would be exposed.

Essentially, RoCE represents an attempt by InfiniBand vendors at enticing the customer with a good Ethernet clustering benchmark story, but switching to selling InfiniBand gear in the end.

With a solid and proven networking stack in place, hardened by years of deployment in increasingly larger clusters, iWARP is now a robust clustering protocol with no application level compromise to make in performance compared to IB. As it stands today at the beginning of the journey, RoCE's road promises to be paved with anything but marshmallows.

### Related Links

[IBM Research Report on IB and 10GbE Performance for HPC Applications](#)

[IBM/Blade Networks Presentation](#)

[Purdue University 10GbE Coates Cluster Whitepaper](#)

### References

[RFC 5040] Recio, R., Culley, P., Garcia, D., and J. Hilland, "A Remote Direct Memory Access Protocol Specification", RFC 5040, October 2007.

[RFC 5041] Shah, H., Pinkerton, J., Recio, R., and P. Culley, "Direct Data Placement over Reliable Transports", RFC 5041, October 2007.