# RoCE at a Crossroads

RoCE is a specification for encapsulating InfiniBand frames directly within Ethernet frames, by replacing the IB Local Routing Header with an Ethernet MAC header. The name RoCE (RDMA over Converged Ethernet) emphasizes the need for a lossless fabric to run the InfiniBand protocol, and Converged Ethernet (also called DCB) is required to provide lossless operation.

This paper argues that upon closer examination, the "CE" in the name is revealed to be a misnomer at best, since in a dedicated cluster fabric with a single traffic type, the CE suite of protocols effectively boils down to Ethernet's old PAUSE scheme [6]. In other words, the CE mechanisms relevant to reducing packet loss in a clustered fabric, where mixing different traffic priorities (e.g. storage and networking) isn't required, have been available in "normal" Ethernet networks since 1997, when the IEEE802.3x PAUSE standard was introduced.

The paper therefore concludes that by depending on a Layer-2 scheme of limited usability, RoCE lacks the essential requirements for scalability and reliability. Furthermore, even if the protocol definition is changed to run over IP in order to address routability, it will still lack critical congestion and reliability mechanisms, i.e. the reasons why the IETF standard for RDMA over Ethernet (iWARP) was designed to run over TCP/IP. The following sections discuss the technical aspects in more detail.

## Zero Loss Ethernet

PAUSE works on a link level (back-to-back) basis to temporarily stop the sender from transmitting and overrunning the receiver. However, it is well recognized that enabling PAUSE in a large network is fraught with risk of congestion propagation, and there are numerous online postings of IT staff reporting such occurrences, i.e. a hotspot somewhere in the network results in a generalized state of paralysis in the whole network (see [1], [2], [3] and [4]). These concerns lead most switch vendors to recommend restricting PAUSE to first level switches [2], if enabled at all, effectively limiting its utility in multi-tier topologies. There is no reason to believe that such recommendations no longer apply today.

Given the concerns about deploying PAUSE across a large Layer-2 network, some may look at Quantized Congestion Notification protocol, which is the only remaining potentially relevant component of the CE suite. However, assuming that QCN would provide lossless operation is misconstruing the purpose of the protocol [5]. In fact, QCN is an un-deployed, unverified, and unusual way of effecting congestion control, which involves all switches in the network sending explicit congestion notification frames back to the sources, requesting rate control changes. In a network with tens of thousands of nodes, this traffic is going to be substantial, especially at times where the network is congested, i.e. in trouble. Furthermore, every single one of the sources must be changed to implement TCP like congestion avoidance behavior at the MAC layer. Otherwise, the scheme falls apart. Finally, QCN is limited to a single subnet, i.e. it does not work across IP network boundaries.

It is unwise to assume that such explicit congestion signaling has never been thought of before, or that the reason it hasn't been adopted is unrelated to the problems it introduces. In fact, these considerations have resulted in a split between proponents of RoCE with QCN and those against it.

Independently, similarly to all new protocol development that impact the infrastructure, QCN is expected to take a number of years before the feasibility, interoperability, scalability issues are worked out.

## RoCE Scalability and Routability Limits

Due to the reliance on raw Ethernet mechanisms, RoCE neither can communicate across networks nor can scale past one Layer-2 subnet or worse – the only known RoCE installations are limited to a single switch – notwithstanding that it requires a brand-new, expensive switch infrastructure with large buffers.

Making RoCE routable by encapsulating it within IP headers won't just require finding a new name. In fact, concerns about the feasibility of lossless network communications are multiplied once traffic goes over IP, since the Ethernet loss avoidance mechanisms do not work across an IP hop. More fundamentally, IP does not fit the lossless model at a basic level. It may be useful to recall that a central tenet of the Internet Protocol design and implementation is *best effort delivery*. RFC 791 clearly states in the introduction that "*there are no mechanisms to augment end-to-end data reliability, flow control, sequencing, or other services commonly found in host-to-host protocols*". Therefore, assuming lossless operation across IP boundaries in a multi-tier network is stretching reality to a whole new level.

On the other hand, avoiding IP altogether imposes severe restrictions beyond network scale. In multi-tenant data centers, encapsulation protocols such as NVGRE and VXLAN require building overlays over an IP layer in order to virtualize the network infrastructure. This effectively precludes RoCE and similar protocols requiring absolute zero loss from operating in such environments. It is also interesting to note that the absence of IGMP means multicast is implemented via broadcast in RoCE, where every multicast packet results in a packet flood, and it does not take very many nodes attempting multicast in a data center to completely choke the fabric.

Inevitably, getting out of this impasse either requires re-designing RoCE or dictates a topology with a first tier RoCE connectivity and an IB backbone (which has to also carry non-RDMA IP traffic), not to the dislike of IB vendors (whose interests lie in keeping IB ahead of Ethernet). However, one is left with a dysfunctional virtualization infrastructure, and a management and troubleshooting nightmare.

## iWARP Solution

The majority if not all of data centers today rely on TCP/IP as a transport, which carries all Web traffic, and underlies Hadoop and popular Cloud stacks. iWARP, being indistinguishable on the wire from other TCP/IP applications in both format and behavior, does not introduce any unknowns or new challenges to large datacenter environments like RoCE does. It also does not require expensive switches, complex configuration, special treatment or separate debugging tools.

A key part of TCP, congestion control has been fine tuned over 3 decades to make efficient use of all available capacity in a network, whereas experimental studies at a major OEM showed that a network could not be run at more than 40% of capacity to accommodate RoCE. A TCP Offload Engine that is capable of handling packet loss at silicon speeds is even more adept than a software stack at enabling maximum utilization of a fabric, in addition to increasing host CPU efficiency. Thus, the fact that iWARP is built on offloaded TCP/IP means that it will bring about both of these efficiencies, as it inherits all TCP/IP's attributes and levels of maturity acquired through decades of use in a wide range of the most demanding environments.

A proper packet processing engine architecture can produce a very efficient and high performance TCP offload engine. The mythical "TCP processing overhead" is composed of 2 clocks at 500MHz speed in today's state of the art pipelined implementation, a cost far lower than other basic components of a modern NIC. An iWARP implementation that is built on top of an efficient TOE scales to 40G and 100G with latency comparable to IB. There is no need to rediscover issues long resolved, or to design and iterate for years on a brand new set of mechanisms to prop RoCE up, or to make it routable and scalable. There is no need to replace the entire network infrastructure. iWARP is a safe and proven choice for building a scalable high performance clustering fabric, and it is available today at 40Gb with 1.2usec end-to-end hardware latency [7].

## Summary

iWARP is the no-risk high performance solution for 40Gb Ethernet clustering, leveraging TCP/IP's mature and proven design, with the required congestion control, scalability and routability, preserving existing hardware and requiring no new protocols, interoperability, or long maturity period. RoCE in contrast is an expensive journey of experimentation, rediscovery and frustration that would take years to complete, if it ever does.

## References


[1] Network World Fusion, *Vendors on Flow Control*
[2] Dell Power Connect Team, *Flow Control and Network Performance with PowerConnect*
[3] Virtual Threads Blog, *Beware Ethernet Flow Control*
[4] Mailing Archive *Flow Control and 10G Interfaces*
[5] Cisco Nexus 5000 Switches, *Quantized Congestion Notification and Today's Fibre Channel over Ethernet Networks*
[6] Chelsio Communications, *A Rocky Road for RoCE*
[7] Chelsio Communications, *Ultra Low Latency Report*