Resilient RoCE: Misconceptions vs. Reality

The Idea

Mellanox recently announced there is an alternative mechanism for RoCE to avoid packet loss, which leverages Explicit Congestion Notification (ECN). With this new software from Mellanox, RoCE can presumably be deployed either with or without PFC, depending upon customer network requirements, infrastructure, and preference. This would supposedly make RoCE easier to deploy for more customers and accelerate adoption of RDMA. The announcement claims RoCE has built-in error recovery mechanisms and that a lossless network has never been a strict requirement and that customers typically configure their networks to prevent packet loss. The limited success of RoCE in a single hyper scale install in a controlled environment is presumed to mean having a wide scale deployment capability similar to traditional Ethernet. This new protocol, dubbed Resilient RoCE (and really the 4th incarnation of RoCE) is promoted as being an easier, more deployable version of RoCE.

The Devil is in the Details

While this announcement implicitly acknowledged that the existing RoCEv2 is non-practical (or this new Resilient RoCE would have been unnecessary), it continues the science experiment, and is misleading. RoCE with ECN requires switches/routers to be ECN capable, so it still will only work on specific switches/routers and is only for greenfield deployments rather than brownfield deployments. In fact, it requires the customer to use the latest Mellanox switches to setup traffic classes and priorities for RoCE, TCP/IP, CNP and other traffic as well as the latest Mellanox ConnectX-4 and ConnectX-4 Lx adapters as it will not work on Mellanox ConnectX-3 adapters. In fact, for each new version/feature of RoCE, new hardware is needed: ConnectX-3 supports RoCEv1 only, ConnectX-3 Pro supports RoCEv2, only ConnectX-4 can support Resilient RoCE, and so on (we have not seen the last of this). New adapters and switches are needed to accommodate each new flavor of RoCE and different variants of RoCE don't interoperate. This software shim layer in the server and the new firmware in the switch which provide the resilience/retransmission run at software speeds. Therefore, right when the network is congested and experiences loss and needs to recover quickly, is when one will not have the performance to do so and will burn out the performance and CPU cycles and hence the value proposition of RDMA in the first place. Therefore, resilience and performance are mutually exclusive when it comes to RoCE. iWARP has not changed since 2007, with current vendors' cards interoperating with each other, and future versions from Intel with the x722 and the Lewisburg chipset on the Purley platform interoperating with existing cards. Using DCB/PFC or ECN, still requires lots of manual setup of hosts and switches to try to minimize packet loss.

The RoCE Reality

RoCE is an ongoing exercise in recreating TCP (a protocol that has taken 35 years to refine) in other layers, and is most likely to go the way of other protocols that attempted to do this (XNS, etc.). RoCE is an experiment that is trying to change the fundamental tenets of Ethernet. Ethernet allows for packet loss, RoCE does not, at least not if you want any kind of performance.

Ethernet is easy to use and deploy, RoCE is not, with some 50 configuration steps required per node (50K steps for 1K nodes, etc. – it adds up). Ethernet is backward compatible, RoCE needs specific switches and can only be used in greenfield deployments and not incremental brownfield installs. Resilient RoCE requires Mellanox switches and only Mellanox cards can be used. This is because this new resilient protocol has a component in the adapter and a component in the switch. iWARP is dual sourced (Chelsio and Intel), Resilient RoCE is sole sourced. With Intel rolling out iWARP on the x722 and Purley platform, iWARP is the default Ethernet RDMA, RoCE is the add-on. Standard iWARP Ethernet allows for the same performance as RoCE, without the hassle. iWARP Ethernet is lower cost, due to economies of scale without the need for special/specific switches and cards. iWARP Ethernet is easy to install/support as it is plug-n-play, RoCE requires 1000's of extra steps depending on cluster size, and switches of the same kind, etc.

Congestion Happens

Congestion in networks happen, whether it is from multiple senders to 1 receiver, a slow receiver, dynamic application/storage network bandwidth changes, addition of more servers/storage, etc. Good network design practices are needed to design a network so that congestion is minimized by having a non-blocking architecture, putting storage close to the users of that storage, having high volume traffic servers that have fatter pipes and other such practices. Requiring a lossless Ethernet network is a bad idea that requires long install time, large support teams, constant monitoring of hotspots and their dynamic changes and retesting of applications on lossless networks to make sure when head of line blocking occurs -and it WILL occur-, application timeouts and crashes are avoided. The best way to deal with congestion is to use TCP and to not reinvent the wheel.

Below table summarizes the differences of iWARP & two recent variants of RoCE.

Metric	iWARP	RoCEv2 with DCB	RoCEv2 with ECN
Allows Packet Loss	Yes, performance is maintained by hardware retransmission, microsecond timers and TCP tunables to disable slow start and enable fast retransmit	No for Performance Yes for Poor Performance	No for Performance Yes for Poor Performance
Ease of Use	Easy 0 extra steps per node	Complicated DCB ~50 extra steps per node	Complicated ECN ~50 extra steps per node
Backward Compatibility	Use on Brownfield and Greenfield with any switch or wireless link	Only supported on DCB switches. ConnectX-3 cards don't support RoCEv2. Only ConnectX-3 Pro does.	Only supported on Mellanox adapters and switches to separate TCP, RDMA, CNP and other traffic. ConnectX-3 Pro cards do not support RoCEv2 with ECN (i.e. Resilient RoCE). Only

			ConnectX-4 does.
Performance	100Gb/s sub 1µs latency	100Gb/s sub 1µs latency	100Gb/s sub 1µs latency
Cost	Lower due to economies of scale. Works with legacy installs (decouples server/switch upgrade cycles). Does not need gateways or routers.	Higher due to needing specific switches for DCB and gateways to talk to the outside TCP/IP world. Switches must be the same brand (if want the same set of ~50 configuration steps)	Higher due to needing specific switches and gateways to talk to the outside TCP/IP world. Switches must be the same brand (if want the same set of ~50 configuration steps)
Support	No increase in support team size as everything works on plug-n-play TCP/IP/Ethernet	Large support team. More hours to get things up, running and to stay running. Modifications needed as cluster grows or traffic pattern changes	Large support team. More hours to get things up, running and to stay running. Modifications needed as cluster grows or traffic pattern changes

Table-1 iWARP/RoCE comparison

Reasons to choose Chelsio iWARP and not go down the RoCE path

- Decouple the server and switch sales cycles and make the sales easier and incremental which for example is a requirement of windows distributed datacenter oriented sales. Leverage the existing switch infrastructure.
- Have an easier time of install/deployment/support and hence accelerate market penetration relative to the competitors who are fumbling with RoCE.
- For switch vendors: not lose market share to Mellanox switches when deploying Resilient RoCE. Resilient RoCE only works with Mellanox switches. This is because the new protocol is partially in the switch and partially in the server, neither of which is multi-sourced at this time, and may never be.
- Intel is releasing iWARP support on Lewisburg chipset and x722. iWARP will be the default on x86 platforms and in the industry, and RoCE will be the add-on. By selecting iWARP, you can avoid a re-investment and disruption in 2017.
- So far, every generation of RoCE has required new hardware. ConnectX-3 only runs RoCEv1. RoCEv2 only works on ConnectX-3 Pro. Resilient RoCE only works on ConnectX-4, and none of these protocols are interoperable. RoCE has not completed its evolution cycle. Any RoCE solution now is guaranteed to need to change. By selecting iWARP, you can preserve your investment.
- Chelsio iWARP is inbox'ed with Client RDMA for Windows 10 – RoCE isn't. This will enable a very high performance Client-Server operation on Windows platforms for such things as clusters or Video Post-Production.
- Not have to buy or sell specialized routers (i.e. Metro-X), hence leaving more budget to instead add servers and compute power.
- Currently RoCE does not scale to the full 128-node configurations very easily (has a hard time leaving the rack).

- If you select Chelsio, you also get the benefit of 5M iSCSI IOPs solution (immediate path to revenue for SANs and access to a $3.6B market), as well as a solid NVMe-Fabrics solution (inbox'ed and QA'ed and tuned with Linux 4.8).
- Chelsio iWARP has been in production for many years with several tier-1 OEMs across many generations of silicon. The T5 silicon itself has been in production for 3.5 years. There is no risk.

RoCE Related Terms

ECN (Explicit Congestion Notification): RoCEv2 makes use of the ECN field in the IPv4 header for signaling of congestion. ECN allows switches to notify hosts when congestion is likely to happen, and the end nodes adjust their data transmission speeds to prevent congestion before it occurs.

RCM (RoCEv2 Congestion Management): Provides the capability to try to avoid congestion hot spots and try to optimize the throughput of the fabric.

Resilient RoCE with ECN: RoCE is InfiniBand over UDP/IP, but ECN works with TCP/IP with the switch/router setting ECN bits and the receiving station sending a TCP/IP congestion message back to the sender, then the sender acknowledging that packet to the receiver saying that traffic was reduced. RoCE ECN or RoCE Congestion Management, has the switch/router set the ECN bits and then the receiving station sending back a UDP/IP message to the sender, with no feedback from the sender to the receiver that it reduced its traffic. It is unclear what happens when UDP/IP Congestion Notification Packet gets lost during congestion.

DSCP (Differentiated Services Code Point): For RoCEv2 packets with IPv4, the DSCP field shall be set to the value in the Traffic Class component of the RDMA Address Vector associated with the packet.

Traffic Separation: The intended use of a distinct set of priorities for RoCEv2, TCP/IP and the other traffic, each set of priorities having a bandwidth allocation. This requires separate queues and buffer resources to traffic on distinct priorities.

Congestion Flow: Sending station sends RoCEv2 packet with ECN bits '10', switch/router marks congestion with ECN bits '11', receiving station sends RoCEv2 CNP back with ECN bits '01', sending station reduces traffic.

CNP (Congestion Notification Packet): The notification message an NP sends to the RP when it receives EC marked packets.

RP (Reaction Point): The end node that performs rate limitation to prevent congestion.

NP (Notification Point): The end node that receives the packets from the injector and sends back notifications to the injector for indications regarding the congestion situation. One CNP is sent to the injector once in X microseconds.

CP (Congestion Point): The switch queue in which congestion happens.

ECT (ECN-Capable Transport): End stations set to 00-not ECT, 01 ECT (1), 10 ECT (0).

EC (Explicit Congestion): Switch/router sets this to 11 upon congestion.

Traffic Patterns: Ethernet-based datacenters are usually more diverse compared to the HPC traffic, which is common to be used in the InfiniBand networks.

Unbiased QCN (Quantized Congestion Notification): Standard QCN puts congestion points in the output buffers of switches which leads to intrinsic unfairness of QCN under typical fan-in scenarios. This is alleviated by installing congestion points at the input buffers of switches. QCN at input buffers cannot always discriminate between culprit and victim flows. To overcome this limitation, a marking scheme of occupancy sampling was proposed. Unbiased QCN puts congestion points at input buffers of switches with a marking scheme of occupancy sampling.

DCTCP Algorithm: Contrary to DCPCP algorithm, with RoCEv2 it is recommended to configure two-threshold ECN marking slope, such that the marking becomes probabilistic. In that way, larger congestion causes longer queues, hence increases the marking probability. Higher marking probability means higher probability to receive CNP for a flow in a time period.

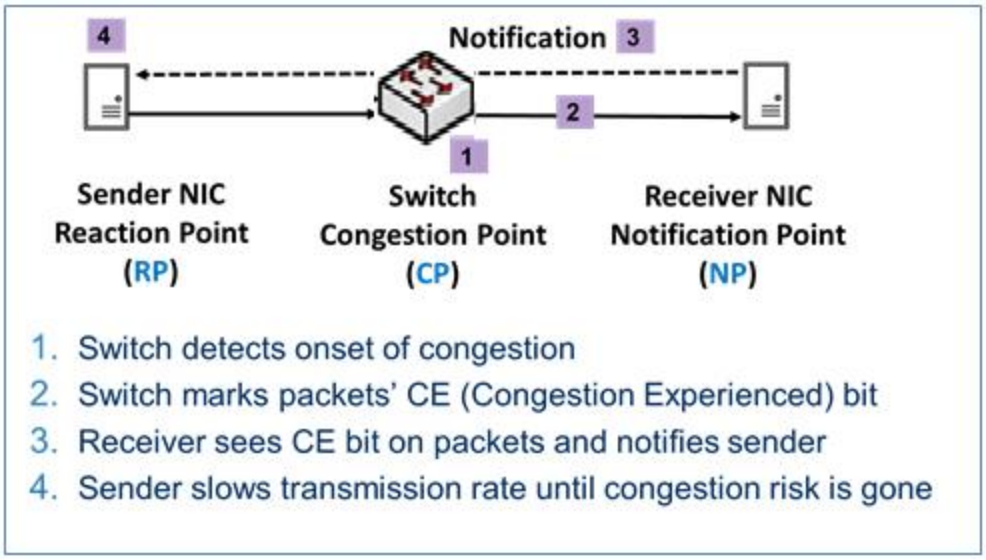

Figure 1 - Resilient RoCE

iWARP Related Terms

None. iWARP and iSCSI packets are indistinguishable from TCP/IP packets and need no special treatment.

Summary

RoCE is work in progress. RoCE is trying to recreate the 35-year evolution of TCP, via a complicated set of protocols that span the switch and server and which have not stabilized yet. RoCE is a protocol which brings no additional value relative to pre-existing iWARP in terms of realizable performance, ease of use, cost, or ubiquity. It is an attempt at reinventing the wheel which does not allow leveraging one's existing investment. It is a protocol that requires lossless capability based on DCBX protocol in the switches, similar to FCoE protocol before it, and hence suffers from the same shortcomings as a result, which probably explains its adoption difficulties. Ultimately, the constant churn of RoCE variants will result in a sole sourced solution from a single vendor, much akin to InfiniBand. With the release of Lewisburg platform from Intel, iWARP will be the default RDMA vehicle in the industry. Selecting iWARP now will allow the benefits of RDMA without having to take on coupling of the server and switch refresh cycle -enabling incremental installs-, purchasing gateways, worrying about data-center topologies, etc. All OS's support iWARP and using iWARP as the Ethernet RDMA vehicle will do away with a whole swath of issues that really don't need to be taken on.

Related Links

Competitive Analysis
Resilient RoCEv4: The Experiment Continues
iWARP RDMA for Microsoft Storage Spaces Direct
iWARP Targets Data Center and Cloud Applications
Adventures In RDMA
RoCE Exposed
RoCE Fails to Scale
RoCE – Plug and Debug
The Case Against iWARP
RoCE is Dead, Long Live RoIP?
RoCE at a Crossroads
RoCE: The Missing Fine Print
RoCE: Autopsy of an Experiment
A Rocky Road for RoCE
RoCE vs iWARP