November 14, 2018

MECHANISM TO DISCOVER END-TO-END LOSSLESS NETWORK CHARACTERISTIC

Mouli Vytla

MECHANISM TO DISCOVER END-TO-END LOSSLESS NETWORK CHARACTERISTIC

AUTHORS:
Mouli Vytla

ABSTRACT

Techniques are described herein for a mechanism that is critical for Remote Direct Memory Access (RDMA) over Converged Ethernet version 2 (RoCEv2) adoption. This mechanism may enable discovering end-to-end lossless network characteristics in RoCEv2.

DETAILED DESCRIPTION

Remote Direct Memory Access (RDMA) over Converged Ethernet version 2 (RoCEv2) requires a lossless network and assumes that lossless characteristics of the network are preserved across Layer 3 routers connecting Layer 2 subnets.

RoCEv2 relies on a combination of various mechanisms for lossless networks. These mechanisms include Priority-based Flow Control (PFC), Enhanced Transmission Selection (ETS), Class of Service (CoS) as defined by Institute of Electrical and Electronics Engineers (IEEE) 802.1p, Differentiated Services Code Point (DSCP) and Data Center Bridging Capability Exchange (DCBX) protocol, Congestion Management, etc.

The underlying assumption is that these mechanisms are supported by all network nodes that carry RoCEv2 traffic. These nodes are typically Network Interface Cards (NICs) sitting in a server, Access or Top of Rack (ToR) switches, End of Rack (EoR) switches, and Layer 3 routers.

In addition to the requirement that all nodes support these mechanisms, proper mapping and resource allocation needs to be in place end-to-end. (e.g., CoS to DSCP mapping, which queue/priority is designated as lossless, how switches/routers allocate/reserve buffers/bandwidth for queues marked for lossless traffic).

Enabling PFC, configuration of priority mapping and resource allocation for lossless traffic class, is a manual operation and hence left to the network administrator's due diligence as well as error prone.

Each vendor may implement PFC and lossless traffic classes in different ways.

Published by Technical Disclosure Commons, 2018

RoCEv2 adoption is being deployed in very few controlled and manually provisioned network environments. RoCEv2 requires the network to offer lossless treatment of RoCEv2 traffic. However, there is no mechanism that exists today to (1) discover/verify whether the entire network is ready to offer lossless treatment of RoCEv2 traffic; (2) check whether treatment of RoCEv2 traffic by each node is consistent; (3) provide network operators with visibility into how each node handles RoCEv2 traffic; and (4) automatically convey and provision intended RoCEv2 lossless network QoS policy on all nodes. There is no mechanism today to discover end-to-end network readiness/compliance for carrying lossless RoCEv2 traffic.

RoCEv2 adoption is significantly impeded by lack of a mechanism to ensure network readiness for RoCEv2 traffic. The lack of a mechanism to discover end-end readiness of lossless network characteristic impedes RoCEv2 adoption and causes deployment and operational issues as well as data loss issues.

Provided herein are techniques to discover end-to-end readiness of lossless characteristics of a network and to configure all nodes with lossless traffic class for RoCEv2 readiness/compliance.

RoCEv2 involves UDP/IP traffic and may be routed through the network.

While a new protocol could be proposed to discover RoCEv2 network readiness, it is preferable to leverage IP / Internet Control Message Protocol (ICMP), which is supported by all nodes. Thus, described is a mechanism that builds on top of IP/ICMP.

Provided is an algorithm and workflow to discover end-end readiness of lossless characteristics of a network.

A new IP option for RoCEv2 Flow Control (“IP_OPT_ROCE_FC”) and a new ICMP type (“ICMP_TYPE_ROCE_FC”) are defined. The IP_OPT_ROCE_FC option can accompany any IP packet. To check for RoCEv2 support in the network, a unicast IP packet (with the IP_OPT_ROCE_FC option) is sent by a RoCEv2 traffic Source to a RoCEv2 Destination.

Each node handles the IP packet (with the IP_OPT_ROCE_FC option) as follows. Upon receiving an IP packet (with the IP_OPT_ROCE_FC option), the node checks itself for RoCEv2 readiness and sends an ICMP packet (with ICMP_TYPE_ROCE_FC) back to the Source. The response indicates support for RoCEv2 readiness as either not supporting

the RoCEv2 lossless characteristic or supporting the RoCEv2 lossless characteristic. ICMP_TYPE_ROCE_FC can specify the exact details of what is supported, how queues are mapped, bandwidth reserved, etc.

Upon receiving the IP packet (with IP_OPT_ROCE_FC option), the node forwards the original packet to the next hop (towards the destination). All intermediate nodes check themselves for RoCEv2 readiness and forward the original packet to the next hop. The Destination node (typically a NIC or server) sends an ICMP packet (with ICMP_TYPE_ROCE_FC) to the Source, indicating its support for RoCEv2 and also indicating that it was the final destination of the original packet.

The Originator (Source) would have received ICMP packets (with ICMP_TYPE_ROCE_FC) from all the nodes along the network path that would be taken for RoCEv2 traffic. By analyzing ICMP_TYPE_ROCE_FC responses from all nodes, it can be determined whether the network is RoCEv2 ready. It is also possible to check whether various mechanisms (e.g., PFC, ETS, CoS/DSCP mapping for lossless traffic class, etc.) are consistent among all nodes and detect mismatched configurations among nodes along the network path.

Also described herein is a protocol to configure all nodes for RoCEv2 readiness/compliance. If a node supports RoCEv2, IP_OPT_ROCE_FC and ICMP_TYPE_ROCE_FC can be extended to carry PFC/ETS/CoS/DSCP information to designate a specific class/priority of traffic as lossless and provision (auto-configure) each node along the network path for RoCEv2 readiness. This ensures consistent treatment of RoCEv2 traffic through the entire network path taken by RoCEv2 packets.

An example packet flow with PING is provided as follows. First, a ICMP Echo Request (PING) packet (with the IP_OPT_ROCE_FC option) is sent from the Source to the Destination. Each intermediate node receives the PING packet (with the IP_OPT_ROCE_FC option) and sends an ICMP packet (with the ICMP_TYPE_ROCE_FC) back to the Source. Each intermediate node forwards the PING packet (with the IP_OPT_ROCE_FC option) to next hop towards the Destination.

Figure 1 below illustrates an example IP Option RoCEv2 Flow Control (IP_OPT_ROCE_FC) type.

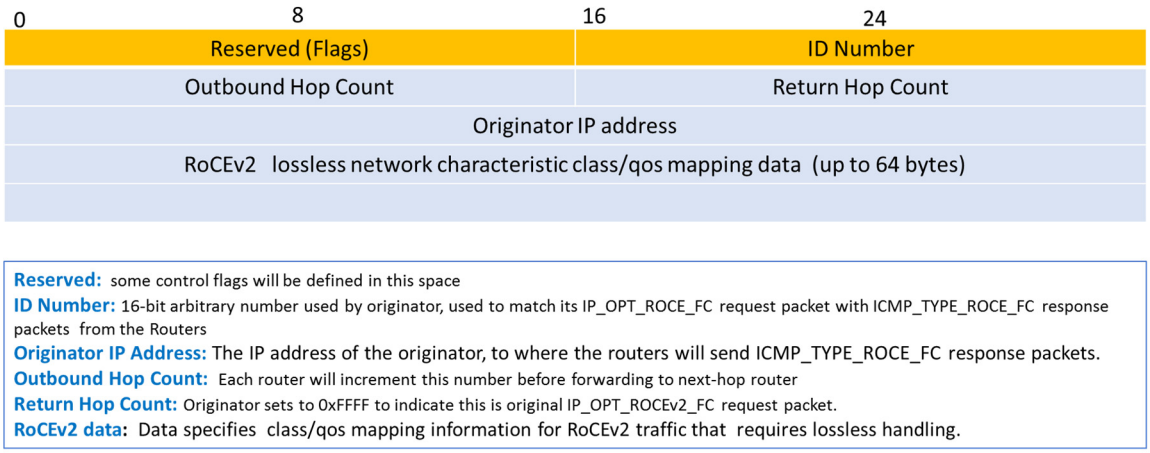

Figure 1

Figure 2 below illustrates an example ICMP Type RoCEv2 Flow Control message (ICMP_TYPE_ROCE_FC).

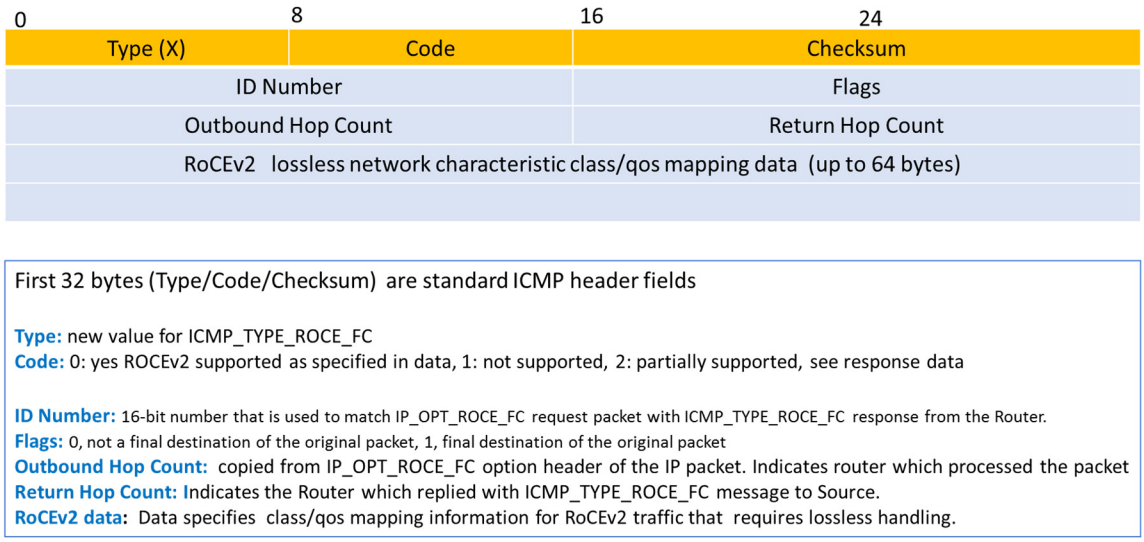

Figure 2

The Destination receives the PING packet (with the IP_OPT_ROCE_FC option) and sends an ICMP packet (with the ICMP_TYPE_ROCE_FC) back to the Source. The Destination also sends an ICMP Echo reply (PING reply) to the Source. The Source analyzes all ICMP_TYPE_ROCE_FC responses received and determines whether the entire network is ready for RoCEv2. The Source also identifies any inconsistencies in the treatment of RoCEv2 traffic by nodes.

An intermediate router handles packets as follows. First, the router increments OutboundHopCount in the IP_OPT_ROCE_FC option of the original packet before

forwarding the packet to the next hop. Next, the router copies the incremented OutboundHopCount from the IP_OPT_ROCE_FC to the newly generated ICMP_TYPE_ROCE_FC message and sets the flags in the ICMP_TYPE_ROCE_FC message to zero, indicating that it is not the final destination of the packet. The router then sends the ICMP_TYPE_ROCE_FC message to the Source/Originator.

The final Destination handles packets as follows. The Destination does not increment the OutboundHopCount in the IP_OPT_ROCE_FC option of the original packet (since it is the final Destination). However, the Destination does copy the OutboundHopCount from the IP_OPT_ROCE_FC to the newly generated ICMP_TYPE_ROCE_FC message and sets the flags in ICMP_TYPE_ROCE_FC message to one (indicating it is the final Destination of the packet). The Destination then sends the ICMP_TYPE_ROCE_FC message to the Source/Originator.

Successful packet forwarding from the Source to the Destination may occur as follows. If the packet was forwarded successfully to the Destination, the Source would have received N response packets (ICMP_TYPE_ROCE_FC), each carrying an OutboundHopCount that indicates the router that processed original packet and generated the ICMP message. The response message informs the Source whether each node along the network supported RoCEv2.

In certain instances, one or more packets may be lost. If the source did not receive all response packets (i.e., it received packets with OutboundHopCount 1...N and N+2...M, missing a packet with OutboundHopCount N+1), this indicates that the network is not reliable or RoCEv2 lossless quality cannot be assumed.

As each node responds with its RoCEv2 lossless offering capability, the Originator knows precisely how each responding node supports RoCEv2 lossless offering capabilities (e.g., PFC, ETS, etc.). Even when the discovery packet is dropped midway, the Originator knows how many nodes and which specific nodes have replied with a response.

This is a software implementation that leverages and extends the IP/ICMP stack which exists across virtually all networking products. Regardless of hardware / Application Specific Integrated Circuit (ASIC), the software can query platform support for RoCEv2, construct a response, and reply to the Originator.

The ICMP is a well-understood and supported protocol. It has room for extensions, and as such the ICMP stack may be extended. The software implementation is debugable and updatable. Troubleshooting tools can be extended to display the RoCE_FC option described herein. Security Access Control Lists (ACLs) may be formulated at routers to allow/drop a specific ICMP code/type. Because extended ACLs already have support for the ICMP code/type, no new requirements are placed on the hardware/platform.

A verified ICMP path is guaranteed to be the path that RoCEv2 packets would take. The packet may be any IP packet (with the IP_ROCE_FC option) that is routed normally. The intent is for this discovery packet to mimic a RoCEv2 packet (IP/UDP). Each node in the forward path generates the ICMP response back to the Originator (as the discovery IP packet is routed by each node). Thus it is guaranteed that RoCEv2 packets will take the same path as the path taken by the discovery packet. The discovery packet is sent with the same RoCEv2 Source (listed as the Originator address in the IP_ROCE_FC option) and to the same RoCEv2 destination.

The RoCEv2 lossless characteristic requires a set of functionality from each network node (e.g., the ability to honor and generate priority flow control xon/xoff, mapping and handling a certain class of traffic as lossless, enhanced traffic scheduling, how RoCEv2 traffic is handled during congestion, no-drop criteria, etc.) that is a combination of ASIC/system architecture/design level support.

It is critical to ensure support for the above from each node. It is also critical to obtain feedback from each node and its level of support for the above. In order to determine end-to-end support and operationally establish a baseline that is supported by each node, a well-defined protocol is required. It is critical to be able to discover RoCEv2 lossless compliance at any time in a production network, as network topologies change, the network nodes get software updates, new hardware gets deployed, etc.

Dynamic and extensible, the RoCEv2 Source can specify what lossless criteria it expects from the network. Each node answers its support and to what extent/variant the support is implemented. For example, a node might only support priority flow control for two levels (high and low) as opposed to supporting eight levels.

In summary, techniques are described herein for a mechanism that is critical for RoCEv2 adoption. This mechanism may enable discovering end-to-end lossless network characteristics in RoCEv2.