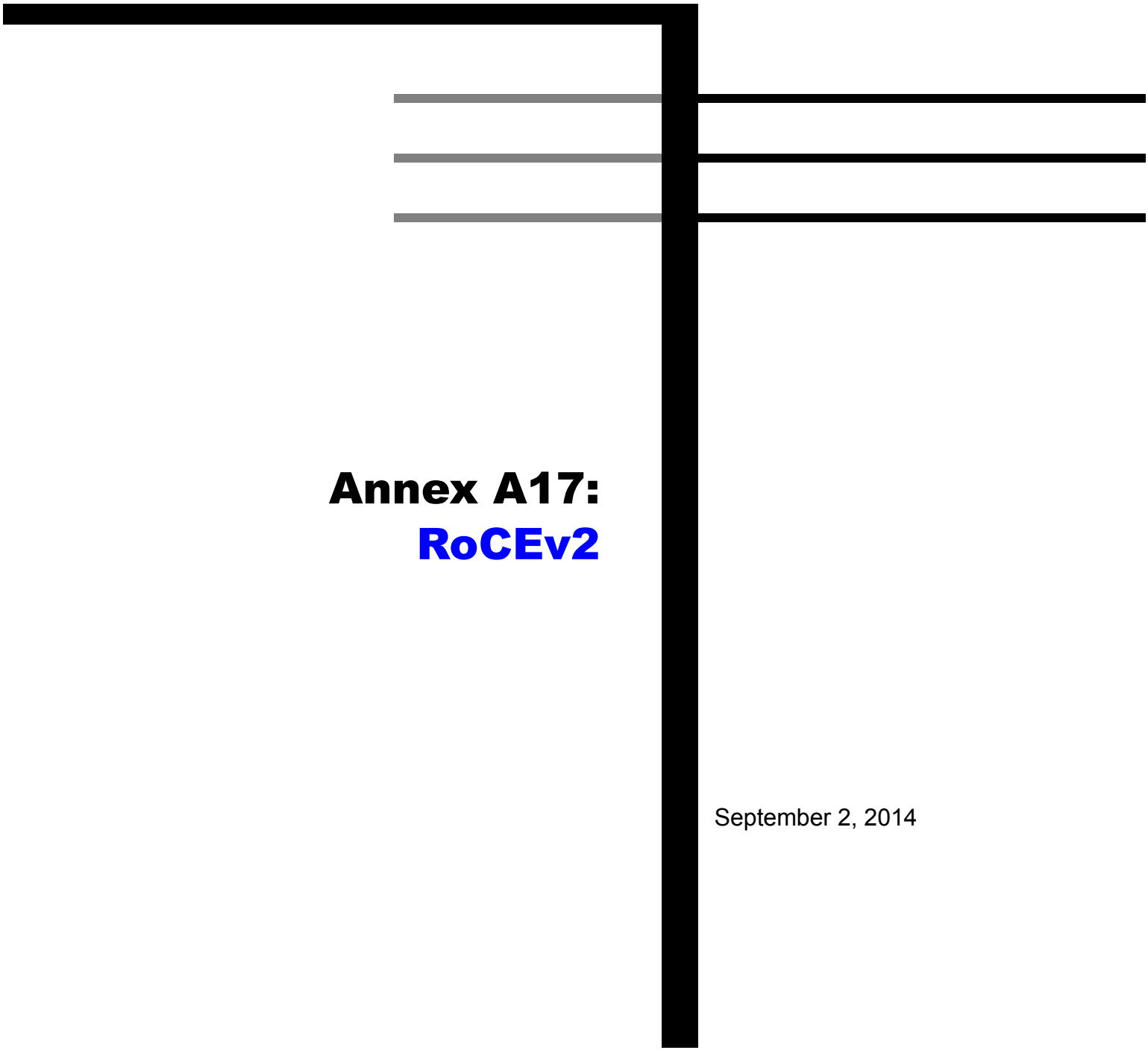
Supplement to InfiniBand™ Architecture Specification Volume 1 Release 1.2.1



Annex A17:
RoCEv2

September 2, 2014

Table 0 Revision History

Revision	Date	
1.0	Sept. 2, 2014	General Release

LEGAL DISCLAIMER

This specification provided "AS IS" and without any warranty of any kind, including, without limitation, any express or implied warranty of non-infringement, merchantability or fitness for a particular purpose.

In no event shall IBTA or any member of IBTA be liable for any direct, indirect, special, exemplary, punitive, or consequential damages, including, without limitation, lost profits, even if advised of the possibility of such damages.

ANNEX A17: ROCEV2 (IP ROUTABLE ROCE)

A17.1 INTRODUCTION

This document is an annex to Volume 1 release 1.2.1 of the InfiniBand Architecture, herein referred to as the base specification. This annex is Optional Normative, meaning that implementation of the feature described by this annex is Optional, but if present, the implementation must comply with the compliance statements contained within this annex. This specification follows the spirit of the RoCE Annex (Annex A16 to the base specification) in defining a new InfiniBand protocol variant that uses an IP network layer (with an IP header instead of InfiniBand's GRH) thus allowing IP routing of its packets.

A17.2 OVERVIEW

A17.2.1 THE INFINIBAND ARCHITECTURE

The InfiniBand Architecture offers a rich set of I/O services based on an RDMA access method and message passing semantics. Included are a variety of transport services, reliable and unreliable, connected and unconnected, support for atomic operations, multicast and others.

InfiniBand defines a layered architecture that specifies the first four layers of the OSI reference stack including the physical, link, network and transport layers as well as an accompanying management framework. In addition, the IB specification defines a software interface and its accompanying verbs which are designed to allow smooth access to the services provided by the InfiniBand Architecture.

A17.2.2 RDMA OVER CONVERGED ETHERNET (ROCE)

RDMA over Converged Ethernet (RoCE) is an InfiniBand Trade Association Standard designed to provide InfiniBand Transport Services on Ethernet Networks[4]. RoCE preserves the InfiniBand Verbs Semantics together with its Transport and Network Protocols and replaces the InfiniBand Link and Physical Layers with those of Ethernet. The network management infrastructure for RoCE is also that of Ethernet.

4. http://www.infinibandta.org/content/pages.php?pg=about_us_RoCE

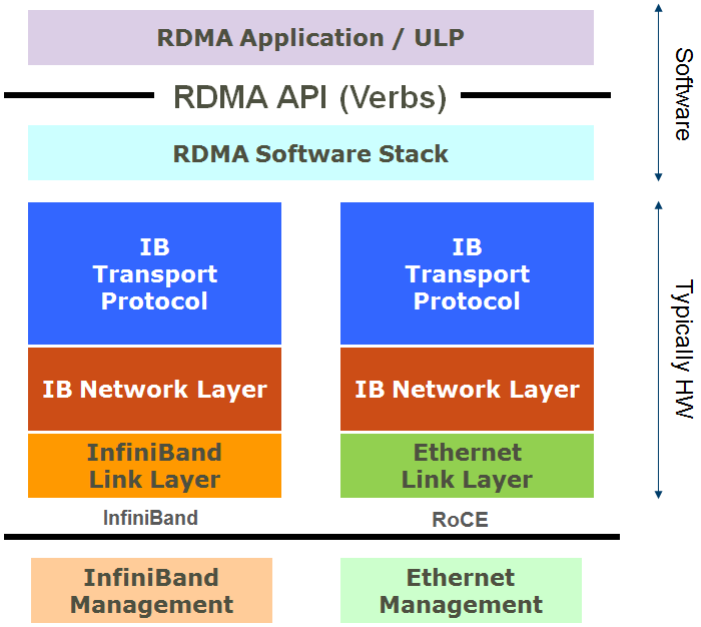


Figure 1 InfiniBand and RoCE Protocol Stacks

A17.2.3 THE NEED FOR (IP) ROUTABLE RDMA

RoCE packets are regular Ethernet frames[5] that carry an Ethertype value[6] allocated by IEEE which indicates that the next header is a RoCE GRH.

Figure 2 RoCE Packet Format

Since RoCE traffic doesn't carry an IP header, it can't be routed across the boundaries of Ethernet L2 Subnets using regular IP routers. Under this scheme, RoCE provides RDMA services for communication within an Ethernet L2 domain.

5. Including VLANs and all other Ethernet header variations as defined by IEEE 802
6. 0x8915

A17.2.4 RoCEv2 (IP Routable RoCE)

RoCEv2 is a straightforward extension of the RoCE protocol that involves a simple modification of the RoCE packet format. Instead of the GRH, RoCEv2 packets carry an IP header which allows traversal of IP L3 Routers and a UDP header that serves as a stateless encapsulation layer for the RDMA Transport Protocol Packets over IP.

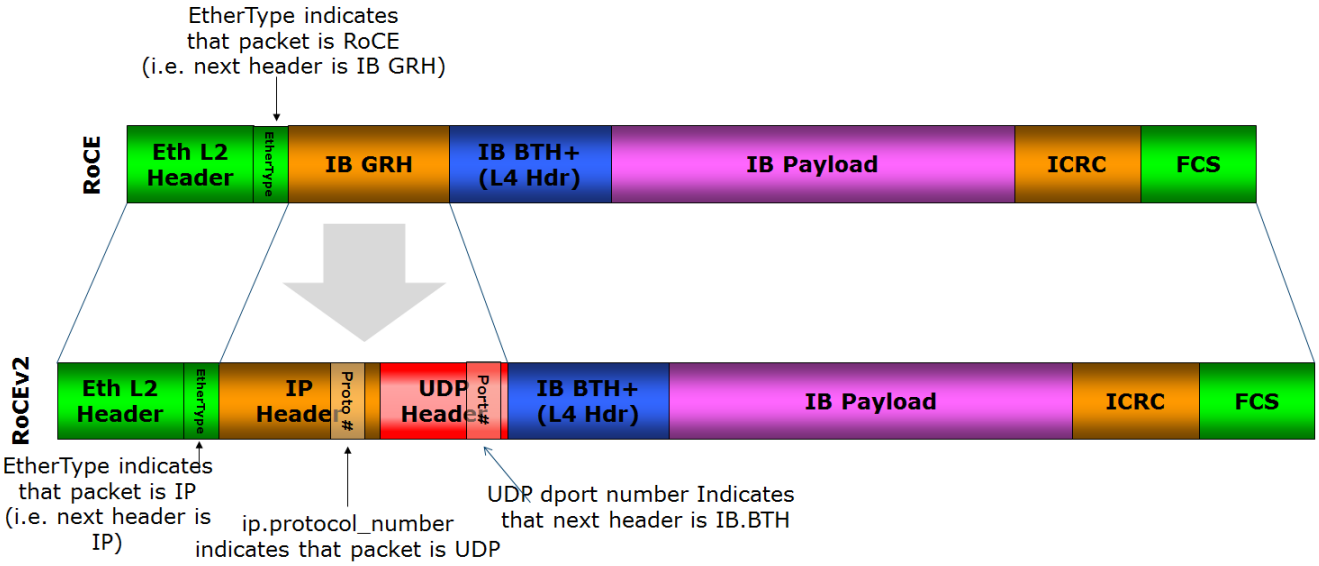


Figure 3 RoCEv2 and RoCE Frame Format Differences

RoCEv2 packets use a well-known UDP Destination Port (dport) value that unambiguously distinguishes them in a stateless manner.

As an additional benefit, following common practices in UDP encapsulated protocols, the UDP Source Port (sport) field of RoCEv2 packets serves as an opaque flow identifier that can be used by the networking infrastructure for packet forwarding optimizations - see Section 17.9.4, "ECMP for RoCEv2," on page 21.

Since this approach exclusively affects the packet format on the wire, and due to the fact that with RDMA semantics packets are generated and consumed below the API, applications can operate over any form of RDMA service (including RoCEv2) in a completely transparent way[7] (see Figure 4).

7. Widespread RDMA APIs are IP based for all existing RDMA technologies

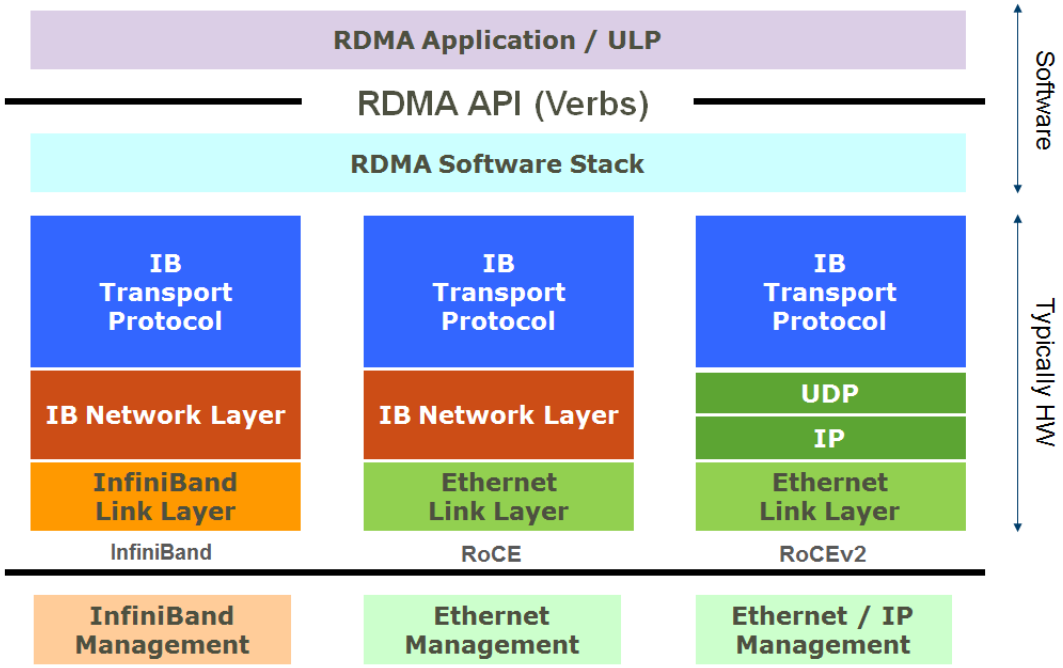


Figure 4 RoCEv2 Protocol Stack

A17.3 RoCEv2 Packet Format

The RoCEv2 Packet format is shown in Figure 5.

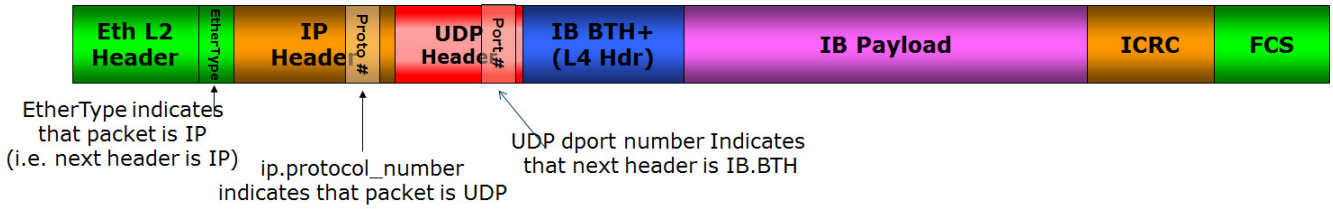


Figure 5 RoCEv2 Packet Format

A17.3.1 EtherTypes and IP Header Fields

RoCEv2 supports both IPv4 and IPv6. The corresponding Ethertype values as well as IPv4 and IPv6 header fields for RoCEv2 packets are described in Section 17.3.1.1, “RoCEv2 with IPv4,” on page 5 and Section 17.3.1.2, “RoCEv2 with IPv6,” on page 6 respectively.

CA17-1: RoCEv2 Ports shall support both RoCEv2 with IPv4 and RoCEv2 with IPv6 packet formats.

CA17-2: RoCEv2 Packets shall conform to the format depicted in Figure 5 with individual fields set as mandated by either Section 17.3.1.1, "RoCEv2 with IPv4," on page 5 or Section 17.3.1.2, "RoCEv2 with IPv6," on page 6.

A17.3.1.1 RoCEv2 with IPv4

The Ethertype value for IPv4 as assigned by IEEE is 0x0800.

The format of the IPv4 header and its fields are specified by the IETF in RFC791, RFC2474 and RFC3168. The sub-sections below define the values for relevant fields in the IPv4 header of RoCEv2 packets.

A17.3.1.1.1 Internet Header Length (IHL)

CA17-3: For RoCEv2 packets with IPv4, the IHL field shall be set to 5.

A17.3.1.1.2 Differentiated Services Codepoint (DSCP)

CA17-4: 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.

A17.3.1.1.3 Explicit Congestion Notification (ECN)

RoCEv2 makes use of the ECN field in the IPv4 header for signaling of congestion as defined by the IETF in RFC3168. See Section 17.9.3, "RoCEv2 Congestion Management," on page 20.

For HCAs that support RoCEv2 Congestion Management, the ECN field in the IPv4 header of a RoCEv2 packet may be set to '01' or '10' to indicate that the packet is subject to marking in the network to indicate congestion.

CA17-5: For HCAs that don't support RoCEv2 Congestion Management, the ECN field in the IPv4 header of a RoCEv2 packet shall be set to '00'.

A17.3.1.1.4 Total Length

CA17-6: For RoCEv2 packets with IPv4, the Total Length field shall be set to the length of the IPv4 packet in bytes including the IPv4 header and up to and including the ICRC.

A17.3.1.1.5 Flags

CA17-7: For RoCEv2 packets with IPv4 the Flags field shall be set to '010' (don't fragment bit is set).

A17.3.1.1.6 FRAGMENT OFFSET

CA17-8: For RoCEv2 packets with IPv4 the Fragment Offset field shall be set to 0.

A17.3.1.1.7 TIME TO LIVE

CA17-9: For RoCEv2 packets with IPv4 the Time to Live field shall be set to the value in the Hop Limit component of the RDMA Address Vector associated with the packet.

A17.3.1.1.8 PROTOCOL

CA17-10: For RoCEv2 packets with IPv4 the Protocol field shall be set to 0x11 (UDP).

A17.3.1.1.9 SOURCE AND DESTINATION IP ADDRESSES

CA17-11: The Source IP Address of RoCEv2 packets with IPv4 shall be set to the IPv4 address encoded in the Port GID entry referenced by the "port" and "SGID index" components of the Address Vector associated with the packet.

CA17-12: The Destination IP Address of RoCEv2 packets with IPv4 shall be set to the IPv4 address encoded in the DGID component of the Address Vector associated with the packet.

A17.3.1.2 ROCEV2 WITH IPV6

The Ethertype value for IPv6 as assigned by IEEE is 0x86DD.

The format of the IPv6 header and its fields are specified by the IETF in RFC2460, RFC2474 and RFC3168. The sub-sections below define the values for relevant fields in the IPv6 header of RoCEv2 packets.

A17.3.1.2.1 DIFFERENTIATED SERVICES CODEPOINT (DSCP)

CA17-13: For RoCEv2 packets with IPv6, the DSCP field shall be set to the value in the Traffic Class component of the Address Vector associated with the packet.

A17.3.1.2.2 EXPLICIT CONGESTION NOTIFICATION (ECN)

RoCEv2 makes use of the ECN field in the IPv6 header for signaling of congestion as defined by the IETF in RFC3168. See Section 17.9.3, "RoCEv2 Congestion Management," on page 20.

For HCAs that support RoCEv2 Congestion Management, the ECN field in the IPv6 header of a RoCEv2 packet may be set to '01' or '10' to indicate that the packet is subject to marking in the network to indicate congestion.

CA17-14: For HCAs that don't support RoCEv2 Congestion Management, the ECN field in the IPv6 header of a RoCEv2 packet shall be set to '00'.

A17.3.1.2.3 PAYLOAD LENGTH

CA17-15: For RoCEv2 packets with IPv6, the Payload Length field shall be set to the length of the IPv6 packet payload starting from the first byte after the IPv6 header (i.e. the BTH) up to and including the 4 bytes of the ICRC

A17.3.1.2.4 NEXT HEADER

CA17-16: For RoCEv2 packets with IPv6 the Next Header field shall be set to 0x11 (UDP).

A17.3.1.2.5 HOP LIMIT

CA17-17: For RoCEv2 packets with IPv6 the Hop Limit field shall be set to the value in the Hop Limit component of the Address Vector associated with the packet.

A17.3.1.2.6 SOURCE AND DESTINATION IP ADDRESSES

CA17-18: The Source IP Address of RoCEv2 packets with IPv6 shall be set to the IPv6 address in the Port GID entry referenced by the "port" and "SGID index" components of the Address Vector associated with the packet.

CA17-19: The Destination IP Address of RoCEv2 packets with IPv6 shall be set to the IPv6 address in the DGID component of the Address Vector associated with the packet.

A17.3.2 UDP HEADER FIELDS

The UDP header format is defined by IETF in RFC768. The sub-sections below define the values for relevant fields in the UDP header of RoCEv2 packets.

A17.3.2.1 SOURCE PORT

The Source Port field in the UDP header of a RoCEv2 packet may be used by network devices as a component in the selection of a route among multiple possible alternative routes - see Section 17.9.4, "ECMP for RoCEv2," on page 21. For that reason, HCAs should use a fixed value across packets where ordering matters between them (e.g. packets of a connected QP).

A17.3.2.2 DESTINATION PORT

CA17-20: The Destination Port field in the UDP header of RoCEv2 packets shall be set to the value allocated by IANA[8] for use with RoCEv2.

A17.3.2.3 LENGTH

CA17-21: The Length field in the UDP header of RoCEv2 packets shall be set to the number of bytes counting from the beginning of the UDP header up to and including the 4 bytes of the ICRC.

A17.3.2.4 CHECKSUM

The Checksum field in the UDP header of RoCEv2 packets should be set to 0.

A17.3.3 ICRC FOR ROCEV2 PACKETS

RoCEv2 implements a 32b end-to-end CRC (denoted ICRC) that covers all invariant fields of the packet and offers protection beyond the coverage of the Ethernet Frame Checksum (FCS) that is usually updated hop-by-hop in the fabric.

CA17-22: The rules for generation/checking of the ICRC of RoCEv2 packets follow the ICRC calculation in RoCE and InfiniBand as defined in Volume 1 of the InfiniBand Specification Section 7.8.1 subject to:

(a) The ICRC calculation starts with 64 bits of '1'.[9]

(b) The ICRC calculation continues with the entire IP datagram starting with the first byte of the IP header up until and including the last IB Payload byte right before the ICRC field itself.

(c) The variant fields in the IP header are replaced with '1s for the purpose of the ICRC calculation/check so that changes to these fields along the way don't affect the calculated ICRC value.

For RoCEv2 over IPv4 the fields replaced with '1s for the purpose of ICRC calculation are:

- Time to Live
- Header Checksum
- Type of Service (DSCP and ECN).

For RoCEv2 over IPv6 the fields replaced with '1s for the purpose of ICRC calculation are:

- Traffic Class (DSCP and ECN)
- Flow Label

8. Once allocated by IANA will be updated to include the actual value

9. This is to make it equivalent to the RoCE(v1) ICRC that runs 64 bits of '1' (dummy LRH) prior to the GRH through the ICRC machine following the spirit of the IB ICRC calculation (IB Spec Vol.1 Section 7.8.1)

- Hop Limit.

(d) UDP Checksum field is replaced with ‘1s for the purpose of the ICRC calculation/check.

A17.3.4 RoCEv2 Inbound Packet Validation

CA17-23: RoCEv2 packets shall undergo validation as mandated by the Base Specification subject to the explicit modifications defined in Section 17.4, “InfiniBand Transport Protocol Spec Considerations,” on page 9.

In addition,

CA17-24: Received RoCEv2 packets, that don’t conform to the rules set in Section 17.3.1, “Ethertypes and IP Header Fields,” on page 4, Section 17.3.2, “UDP Header Fields,” on page 7 and Section 17.3.3, “ICRC for RoCEv2 Packets,” on page 8 shall be silently dropped.

A17.4 InfiniBand Transport Protocol Spec Considerations

This section describes adaptations to elements of normative behavior defined in the InfiniBand Transport Protocol Specification as they apply to RoCEv2.

CA17-25: An HCA containing a RoCEv2 port which claims compliance to this annex shall be compliant with the InfiniBand transport as defined in Chapter 9 of the base specification, subject to the adaptations and exceptions explicitly called out in this section.

A17.4.1 RoCEv2 Addressing

A17.4.1.1 L3 Addresses

For simplicity in the interpretation of the IB Base Specification text, RoCEv2 L3 Addresses are interchangeably referred to as GIDs. As GIDs have the same format as IPv6 addresses, for RoCEv2 with IPv6, the corresponding IPv6 Source IP (SIP) and Destination IP (DIP) Addresses are simply referred to as SGID and DGID. For RoCEv2 with IPv4, the corresponding IPv4 Source IP (SIP) and Destination IP (DIP) Addresses are encoded into the SGID and DGID respectively following common rules for IPv4-mapped IPv6 addresses namely: GID =::ffff:<IPv4 Address>.

A17.4.1.2 L2 Addresses

All references in the Base Specification to the LRH and its fields are Not Applicable to RoCEv2 ports.

A17.4.2 ADDRESS VECTOR

The InfiniBand Specification defines an Address Vector that is used to denote a remote destination and the path parameters selected to communicate with it. (InfiniBand Specification Vol.1 Rev 1.2.1 Section 11.2.2 and Section 11.2.4.2).

The Address Vector fields are manipulated via the “Software Transport Interface” and passed down to the Transport Layer for the generation of outgoing and validation of incoming RoCEv2 packets.

RoCEv2 Address Vectors include L2 and L3 address information as well as other path components (e.g. QoS, Hop Limit, etc). The Address Vector components retain their specified behavior with the exceptions explicitly called out in this section:

As with RoCE, the “Send Global Routing Header Flag” in the Address Vector is always set to TRUE for RoCEv2.

Source L3 Addresses (SGIDs) are not explicitly stored in the Address Vector but obtained by reference. The Address Vector includes a SGID Index that points to an entry in a Source GID Table that is maintained per port as described in Section 17.4.3, “Port GID Table,” on page 11. The GID type as obtained from the selected SGID entry determines the protocol for outbound packet generation - see Section 17.8, “Interoperability with RoCE Endnodes,” on page 18.

As with RoCE, the mechanism for RoCEv2 L3 to L2 Address Resolution is outside of the scope of this specification. It is assumed that implementations follow common practices and use existing OS services to perform this task (e.g. use ARP and routing infrastructure in the host).

For RoCEv2, the DLID component in the Address Vector is generalized to become the “Destination L2 Address“ and may be used to carry an already resolved DMAC across the Software Transport Interface (Create Address Handle verb). In this case, the implementation may use the provided “Destination L2 Address” as-is for the generation of packets associated with the Address Vector in question[10]. Alternatively, by setting the “Destination L2 Address” component of the Address Vector to the value 0xFFFFFFFFFFFF, the L3 to L2 Address Resolution is effectively carried out by the implementation of the corresponding verb (“Create Address Handle” and “Modify QP”).

The “Source Path Bits” component of the Address Vector is not applicable for RoCEv2 ports. The Source L2 Address for RoCEv2 packets is implied

10. This allows L3 to L2 Address Resolution to happen before the generation of the Address Vector as is the case with the current RDMA CM implementation in Linux

by the selected SGID of the Address Vector and obtained by the implementation using common services of the underlying OS infrastructure.

The SL component in the Address Vector is used to determine the Ethernet Priority of generated RoCEv2 packets. SL 0-7 are mapped directly to Priorities 0-7, respectively. SL 8-15 are reserved.

A17.4.3 PORT GID TABLE

Every RoCEv2 port maintains a port GID table that contains all L3 Addresses that have been configured to the port as described in section 10.2.2.1 of the InfiniBand Specification.

Addresses in the RoCEv2 Port GID Table can be of type “IPv4”, “IPv6” or “IB GID[11]”. A new “GID type” attribute is added to the Port GID Table Entries of RoCEv2 ports to denote the L3 Address type.

CA17-26: RoCEv2 Port GID Table entries shall have a “GID type” attribute that denotes the L3 Address type among “IPv4”, “IPv6” and “IB GID”.

Protocol selection for outbound packet generation is based on the GID type of the selected GID Table entry as described in Section 17.8, “Interoperability with RoCE Endnodes,” on page 18.

The software stack is responsible for maintaining the GID table following creation/removal of L3 addresses to the port. This is typically achieved through interaction (e.g. subscription to callback/event services) with the OS and its host administrative interfaces.

A17.4.4 GRH CHECKS

The Base Specification (InfiniBand Specification Vol.1 Rev 1.2.1 Section 9.6.1.2) defines the rules for checking of the GRH (L3 header) of received InfiniBand packets. As RoCEv2 packets carry an IP header instead of the GRH the following rules replace those mandated by the base specification.

All RoCEv2 packets carry an IP header and hence C9-43.1.1 and C9-43.1.2 in the Base Specification are not applicable for RoCEv2.

RoCEv2 packets have a Next Header / Protocol field set to 0x11 (UDP) and hence C9-44 of the Base Specification is not applicable for RoCEv2.

A17.4.4.1 IP VERSION

Compliance statement C9-45 of the Base Specification is replaced by:

11. For interoperability with RoCE as described in Section 17.8, “Interoperability with RoCE Endnodes,” on page 18

CA17-27: For RoCEv2 with IPv6, if the version number is anything other than 6, the packet shall be silently dropped. For RoCEv2 with IPv4, if the version number is anything other than 4, the packet shall be silently dropped.

A17.4.4.2 ADDRESS VALIDATION RULES

The Base Specification mandates L3 Address validation rules for inbound packets (InfiniBand Specification Vol. 1 Rev 1.2.1 Section 9.6.1.2.3). These rules apply to RoCEv2 packets. For the purpose of these checks, the Source and Destination GIDs of RoCEv2 packets with IPv6 are the IPv6 SIP and DIP addresses respectively. For RoCEv2 with IPv4, the SGID and DGID are respectively obtained from the IPv4 SIP and DIP addresses following the common practice used to map an IPv4 address into an IPv6 one namely: GID =::ffff:<IPv4>.

In addition, the DGID check is amended to include verification of protocol type as detailed in Section 17.8, “Interoperability with RoCE Endnodes,” on page 18

A17.4.5 UNRELIABLE DATAGRAM (UD)

A17.4.5.1 UD COMPLETION QUEUE ENTRIES (CQES)

For UD, the Completion Queue Entry (CQE) includes remote address information (InfiniBand Specification Vol. 1 Rev 1.2.1 Section 11.4.2.1). For RoCEv2, the remote address information comprises the source L2 Address and a flag that indicates if the received frame is an IPv4, IPv6 or RoCE packet.

A17.4.5.2 SCATTERING OF THE L3 HEADER IN UD

The first 40 bytes of user posted UD Receive Buffers are reserved for the L3 header of the incoming packet (as per the InfiniBand Spec Section 11.4.1.2). In RoCEv2, this area is filled up with the IP header. IPv6 header uses the entire 40 bytes. IPv4 headers use the 20 bytes in the second half of the reserved 40 bytes area (i.e. offset 20 from the beginning of the receive buffer). In this case, the content of the first 20 bytes is undefined.

A17.4.6 IB RAW DATAGRAMS

The InfiniBand Architecture defines a Raw service which does not use the InfiniBand transport (InfiniBand Specification Vol.1 Rev 1.2.1 Section 9.8.4). The Raw services as defined in the base specification are provided by the InfiniBand link layer. Similarly to RoCE, since RoCEv2 does not use the InfiniBand link layer, IB RAW datagrams, namely Raw Ethertype and Raw IPv6, are not applicable for RoCEv2.

CA17-28: An implementation of an HCA claiming conformance to this annex shall not support the concept of IB Raw Datagrams on a RoCEv2 port.

As follows from the above, all references to Raw Packets in the Base Specification are not applicable to RoCEv2 ports.

A17.4.7 INFINIBAND PARTITIONING

Methods to populate the P_Key table associated with a RoCEv2 port are outside the scope of this annex. Note that this annex relies on the partition table being initialized at power on time with at least the default P_Key as described in Chapter 10 (Software Transport Interface) of the base specification. The P_Key contained in the BTH is validated for inbound packets as required by the packet header validation protocols defined in Chapter 9 of the base specification.

A17.4.8 INFINIBAND CONGESTION CONTROL

Congestion Management for RoCEv2 is specified in Section 17.9.3, "RoCEv2 Congestion Management," on page 20. InfiniBand Congestion Control as defined in Annex A10 of the base specification is not applicable to RoCEv2 ports. Thus, a CA claiming compliance to Annex A10 for its InfiniBand ports is not required to support any of the port attributes, counters or controls required by Annex A10 for its RoCEv2 ports.

CA17-29: The B (BECN) and F (FECN) bits in the BTH devoted to congestion control as defined in Annex A10 of the base specification are unused and shall be ignored by a RoCEv2 port.

A17.4.9 INFINIBAND QOS

QoS Management as defined in Annex A13 of the base specification is based on InfiniBand Link Layer capabilities that are not applicable to RoCEv2 ports. Thus, a CA claiming compliance to Annex A13 for its InfiniBand ports is not required to support any of the port attributes counters or controls associated with its RoCEv2 ports.

A17.5 INFINIBAND VERBS CONSIDERATIONS

The following sections specify modifications to InfiniBand verbs required for RoCEv2 ports.

CA17-30: RoCEv2 HCAs shall adopt the modifications to verbs described in this section.

A17.5.1 QUERY HCA

The Port Attribute List Output Modifier for this verb when associated with a RoCEv2 port is changed as follows:

- The Base LID & LMC fields are unused and shall be ignored.

- The maximum number of virtual lanes supported by this HCA is unused and shall be ignored.
- The Optional InitTypeReply value is unused and shall be ignored.
- The Subnet Manager address information for each RoCE port of this HCA is unused and shall be ignored.
- The CapabilityMask bits IsSM, IsSMDisabled, IsSNMPTunnelingSupported, IsClientReregistrationSupported are unused and shall be ignored.
- A new value ("RoCEv2") shall be added for the port-type attribute of the Port Attributes list to denote that the port is of RoCEv2 type
- A new "RoCE Supported" capability bit shall be added to the Port Attributes list. This capability bit applies exclusively to ports of the new "RoCEv2" type. When set, it denotes that the port is capable of operating simultaneously in RoCEv2 and RoCE modes. See Section 17.8, "Interoperability with RoCE Endnodes," on page 18.

A17.5.2 MODIFY HCA

The Port Attribute List Input Modifier for this verb when associated with a RoCEv2 port shall be changed as follows:

- The CapabilityMask bits IsSM, IsSNMPTunnelingSupported are unused and are reserved.

A17.5.3 CREATE/MODIFY/QUERY ADDRESS HANDLE

The Address Vector component shall be modified in accordance with Section 17.4.2, "Address Vector," on page 10.

For CREATE/MODIFY ADDRESS HANDLE, the Invalid Address Vector error may be returned due to the use of a reserved SL value (SL 8-15 are reserved) when the QP is associated with a RoCEv2 port.

A17.5.4 MODIFY/QUERY QUEUE PAIR / MODIFY/QUERY XRC TARGET QP

The Address Vector and Alternate Path components shall be modified in accordance with Section 17.4.2, "Address Vector," on page 10.

For MODIFY QUEUE PAIR / MODIFY XRC TARGET QP, the Invalid Address Vector error may be returned due to the use of a reserved SL value (SL 8-15 are reserved) when this QP is associated with a RoCEv2 port.

A17.5.5 MODIFY/QUERY EE CONTEXT

The Primary Address Vector and Alternate Path components shall be modified in accordance with Section 17.4.2, "Address Vector," on page 10.

For MODIFY EE CONTEXT, the Invalid Address Vector error may be returned due to the use of a reserved SL value (SL 8-15 are reserved) when this EE CONTEXT is associated with a RoCEv2 port.

A17.5.6 ATTACH/DETACH QP TO/FROM MULTICAST GROUP

If the QP is associated with a RoCEv2 port, the Input Modifiers for this verb shall be changed as follows:

- The Multicast group MLID is unused and shall be ignored.

The Output Modifiers for this verb when associated with a RoCE port shall be changed as follows:

- "Invalid multicast MLID" is removed as a valid Verb Result

A17.5.7 POLL FOR COMPLETION

The output modifier of the Poll for Completion shall be changed as follows:

- If the port is a RoCEv2 port, the remote port address and QP information returned for datagram services (shown in Table 97 of the base specification) shall be modified in accordance with Section 17.4.5.1, "UD Completion Queue Entries (CQEs)," on page 12

A17.5.8 GET SPECIAL QP

Since there is no QP0 associated with a RoCEv2 port, and since a RoCEv2 port does not support either Raw datagram type, for a RoCEv2 port, Get Special QP only applies to the GSI QP (QP1).

Thus compliance statement C11-13 in Section 11.2.5 of the base specification does not apply to a RoCEv2 port with respect to QP0, and compliance statement o11-1 does not apply. An attempt to call GET SPECIAL QP on a RoCEv2 port for a QP other than QP1 shall return an "Invalid Special QP Type" error.

A17.5.9 POST SEND REQUEST

The Post Send Request verb shall be modified to eliminate Raw as one of the possible service types. The "Operation Type Matrix" under the POST SEND REQUEST verb in the base document is modified to effectively eliminate the row of the table governing the Raw service type.

A17.5.10 UNAFFILIATED ASYNCHRONOUS EVENTS

CA17-31: A RoCEv2 port shall not support Client Reregistration.

CA17-32: A RoCEv2 port shall not support the optional Port Change Event

A17.6 INFINIBAND MANAGEMENT CONSIDERATIONS

The following management classes specified in the InfiniBand Architecture as well as their associated normative statements are not applicable to RoCEv2 ports: Subnet Management, Subnet Administration, Performance Management, Device Management, Baseboard Management, SNMP Tunneling, Vendor specific, Application specific classes, Congestion Control, Boot Management and BIS. Instead, RoCEv2 ports are attached to IP subnets that are managed using common Ethernet/IP management practices and standards that are out of scope of this specification.

In the same spirit, the special Queue Pair, QP0 that is defined solely for communication between subnet manager(s) and subnet management agents is not relevant for RoCEv2 ports.

CA17-33: A packet arriving at a RoCEv2 port containing a BTH with the destination QP field set to QP0 shall be silently dropped.

It is expected that there is no InfiniBand management communication between an Ethernet and an InfiniBand management domain. Therefore, any InfiniBand method/attribute combination that refers to a RoCE port may return error code 7 in the “Code for invalid field” of the MAD Common Status field (One or more fields in the attribute or attribute modifier contains an invalid value. Invalid values include reserved values and values that exceed architecturally defined limits).

A17.6.1 COMMUNICATION MANAGEMENT

RoCEv2 utilizes the InfiniBand Architecture Communication Management protocol as defined in Chapter 12 of the base specification. Modifications to the specific MADs required to eliminate references to local addresses are contained in this section.

Communication Management packets for RoCEv2 connections are regular RoCEv2 packets of the same type (IPv4 vs. IPv6) as the intended connection.

Communication Management packets for RoCE connections involving RoCEv2 ports that are also capable of generating and processing RoCE packets (Section 17.5.1, “QUERY HCA,” on page 13), are RoCE packets.

A17.6.1.1 REQ MESSAGE

CA17-34: When a connection is being established between RoCEv2 ports, the Primary Local Port LID, Primary Remote Port LID, Alternate Local Port LID and Alternate Remote Port LID fields of the REQ message are Reserved and shall be ignored on receipt. The value of these fields shall not be checked or validated by a recipient of a REQ message.

A17.6.1.2 REJ MESSAGE

CA17-35: When a connection is being established between RoCEv2 ports, the following reject reason codes shall not be sent as part of a REJ message:

- code 13: Primary Remote Port LID rejected
- code 19: Alternate Remote Port LID rejected

A17.6.1.3 LAP MESSAGE

CA17-36: When alternate paths are being established between RoCEv2 ports, the Alternate Local Port LID and Alternate Remote Port LID fields are Reserved and shall be ignored on receipt. The value of these fields shall not be checked or validated by a recipient of a LAP message.

A17.6.1.4 APR MESSAGE

CA17-37: When alternate paths are being established between RoCEv2 ports, the following APR status code shall not be sent as part of an APR message:

- code 7: Proposed Alternate Remote Port LID rejected.

A17.6.1.5 SAP MESSAGE

CA17-38: The value of the Alternate Local Port LID is reserved and shall be ignored on receipt.

A17.7 CHANNEL ADAPTERS

The base specification defines specific hardware entities such as channel adapters and switches which implement all layers of the InfiniBand Architecture including the InfiniBand-defined physical and link layers. Chapter 17 of the base specification sets forth specific requirements for a channel adapter. Most of the compliance statements contained in Chapter 17 of the base specification apply to a CA which supports one or more RoCEv2 ports. This section describes the exceptions.

A17.7.1 LOADING THE P_KEY TABLE

Compliance statement C17-7 describes requirements for setting the P_Key table based on an assumption that the P_Key table is set directly by a Subnet Manager. However, RoCEv2 ports do not support InfiniBand Subnet Management. Therefore, compliance statement C17-7 does not apply to RoCEv2 ports.

Methods for setting the P_Key table associated with a RoCEv2 port are not defined in this specification, except for the requirements for a default P_Key described elsewhere in this annex.

A17.7.2 LOCALLY ROUTED PACKETS

Compliance statement C17-9 refers to locally routed packets which don't exist with RoCEv2 Thus, C17-9 is not applicable to RoCEv2 ports.

A17.7.3 BACKPRESSURE AND DEADLOCK PREVENTION

Compliance statements C17-19 and C17-20 place specific limitations on a CA's ability to apply backpressure. For a CA claiming compliance to this annex, these requirements do not apply to RoCEv2 ports.

A17.7.4 INBOUND PACKET CHECKING

Compliance statement C17-21 is replaced as follows:

CA17-39: For RoCEv2 ports, the CA shall check for transport layer errors in all incoming packets

A17.7.5 SUPPORT FOR QP0

Compliance statement C17-24 is not applicable to RoCEv2 ports.

A17.8 INTEROPERABILITY WITH ROCE ENDNODES

RoCEv2 and RoCE Ports can interoperate with each other. For this purpose, RoCEv2 ports are optionally capable of generating and processing RoCE packets. (Section 17.5.1, "QUERY HCA," on page 13)

Protocol selection (RoCE vs RoCEv2) is controlled through the GID type attribute in the corresponding entry of the RoCEv2 Port GID table (See Section 17.4.3, "Port GID Table," on page 11).

CA17-40: Connected QPs on RoCEv2 ports that support RoCE shall operate in the mode (RoCE or RoCEv2) as determined by the GID type attribute in the port GID Table entry configured for the QP (MODIFY_QP).

CA17-41: For inbound packets subject to DGID checks, as mandated by the transport protocol (InfiniBand Specification Vol.1 Rev 1.2.1 Section 9.6.1.2.3), if the GID type of the Port GID table entry against which the DGID check is performed doesn't match the protocol of the received packet the check shall be considered as failed.

CA17-42: UD QPs on RoCEv2 ports that support RoCE shall generate packets in the mode (RoCE or RoCEv2) as determined by the GID type attribute in the port GID Table entry referenced by the Address Vector (AV) in the WQE.

CA17-43: UD QPs on RoCEv2 ports that support RoCE shall accept as valid both RoCE and RoCEv2 packets.

A17.9 RoCEv2 Network Considerations

A17.9.1 Lossless Network

As with RoCE, the underlying networks for RoCEv2 should be configured as lossless. In this context, lossless doesn't mean that packets are absolutely never lost. Moreover, the Transport Protocol in RoCEv2 includes an end-to-end reliable delivery mechanism with built-in packet retransmission logic. This logic is typically implemented in HW and is triggered to recover from lost packets without the need for intervention by the software stack. The requirement for an underlying lossless network is aimed at preventing RoCEv2 packet drops as a result of contention in the fabric.

In Data Center Ethernet networks, this kind of lossless behavior is typically achieved through the use of Link-Layer Flow-Control. IEEE 802.1Qbb specifies per-priority link-layer flow-control for Ethernet Networks and is ideally suited for RoCEv2 traffic. Other methods to attain lossless network behavior besides 802.1Qbb may also be used.

In order to ensure end-to-end lossless behavior for RoCEv2 packets that traverse multiple L2 subnets, the intervening L3 routers should be configured to generate an L2 priority for RoCEv2 packets that is set to a value configured for "Link Layer Flow Control (802.1Qbb) enabled" on the subnet where the packet is about to be injected. This is normally obtained through regular configuration mechanisms of the L3 Routers that control the mapping of the "Class of Service" for traversing packets.

This lossless recommendation does not impose a constraint on non-RoCEv2 traffic that could coexist with RoCEv2 on a converged network scenario. Such traffic should be configured to use a distinct set of priorities where Link Level Flow Control could be disabled. Network devices (L2 switches and L3 routers) typically allocate separate queues and buffer resources to traffic on distinct priorities. This creates an effective isolation that decouples the RoCEv2 lossless traffic from the other flows (e.g. TCP) that usually operate in lossy mode.

A17.9.2 RoCEv2 QoS

RoCEv2 traffic can take advantage of IP/Ethernet L3/L2 QoS. Given some of the most prevalent use cases for RDMA technology (e.g. low latency, high bandwidth), the use of QoS becomes particularly relevant in a converged environment where RoCEv2 traffic shares the underlying network with other TCP/UDP packets. In this regard, RoCEv2 traffic is no different than other IP flows: QoS is achieved through proper configuration of relevant mechanisms in the fabric such as the Enhanced Transmission Selection (ETS) defined in 802.1Qaz. Packet/flow identification follows standard practices of IP/Ethernet networks (i.e. DSCP/802.1Q) and is controlled through the Traffic Class parameter in the Address Vector - see Section 17.4.2, "Address Vector," on page 10.

A17.9.3 RoCEv2 Congestion Management

RoCEv2 Congestion Management (RCM) provides the capability to avoid congestion hot spots and optimize the throughput of the fabric. With RCM, incipient congestion in the fabric is reported back to the sources of traffic that in turn react by throttling down their injection rates thus preventing the negative effects of fabric buffer saturation and increased queuing delays.

Congestion Management is also relevant for co-existing TCP/UDP/IP traffic. However, assuming the intended use of a distinct set of priorities for RoCEv2 and the other traffic, each set of priorities having a bandwidth allocation[12], the effects of congestion and the reaction (or lack of it) shouldn't impact one another.

For signaling of congestion, RCM relies on the mechanism defined in RFC3168 (ECN). Upon congestion that involves RoCEv2 traffic, network devices mark the packets using the ECN field in the IP header Section 17.3.1.1.3, "Explicit Congestion Notification (ECN)," on page 5 and Section 17.3.1.2.2, "Explicit Congestion Notification (ECN)," on page 6. This congestion indication is interpreted by destination end-nodes in the spirit of the FECN congestion indication flag of the Base Transport Header (BTH). In other words, as ECN marked packets arrive to their intended destination, the congestion notification is reflected back to the source which in turn reacts by rate limiting the packet injection for the QP in question.

RCM is optional normative behavior. RoCEv2 HCAs that implement RCM shall follow the rules specified in this section:

CA17-44: If RoCEv2 Congestion Management is supported, upon receiving a valid RoCEv2 packet with a value of '11 in its IP.ECN field the HCA shall generate a RoCEv2 CNP formatted as shown in Figure 6 on page 21 directed to the source of the received packet. The HCA may choose to send a single CNP for multiple such ECN marked packets on a given QP.

CA17-45: If RoCEv2 Congestion Management is supported, upon reception of a RoCEv2 CNP the HCA shall reduce the rate of injection for the QP indicated in the RoCEv2 CNP. The amount of rate change is determined by a rate reduction parameter, whose configuration is outside the scope of this specification.

CA17-46: If RoCEv2 Congestion Management is supported, the HCA should increase the injection rate on a QP when a configurable amount of elapsed time and/or a configurable number of bytes have been transmitted on that QP since the reception of the most recent RoCEv2 CNP for

12. IEEE 802.1Qaz ETS or similar mechanisms

that QP. The configuration for the amount of time, number of bytes transmitted and rate of increase are outside the scope of this specification.

The RoCEv2 Congestion Notification Packet format is shown in Figure 6.

MAC Header
IPv4/IPv6 Header
UDP Header
BTH DestQP set to QPN for which the RoCEv2 CNP is generated Opcode set to b'10000001 PSN set to 0 SE set to 0 M set to 0 P_Key set to the same value as in the BTH of the ECN packet marked
(16 bytes) - Reserved. MUST be set to 0 by sender. Ignored by receiver
ICRC
FCS

Figure 6 RoCEv2 CNP Format

A17.9.4 ECMP FOR ROCEV2

Data Center IP networks usually implement path selection mechanisms for load balancing and improved utilization of the fabric topology. Equal Cost Multiple Paths (ECMP) is one prevalent method to achieve this goal. For a given packet, L3 Routers select among the possible different paths using a hash on some of the packet fields. The choice is aimed at allowing multiple paths while preserving the ordering requirements of individual flows.

RoCEv2 packets carry an opaque flow identifier in their UDP Source Port field Section 17.3.2.1, "Source Port," on page 7 which is part of said hash for UDP packets. Consequently, RoCEv2 endnodes set this field so that packets in a sequence that has ordering constraints (e.g. packets from a connected QP) will all carry a constant value. For packets that have no ordering constraints with respect to each other, the UDP Source Port field can be set to different values.