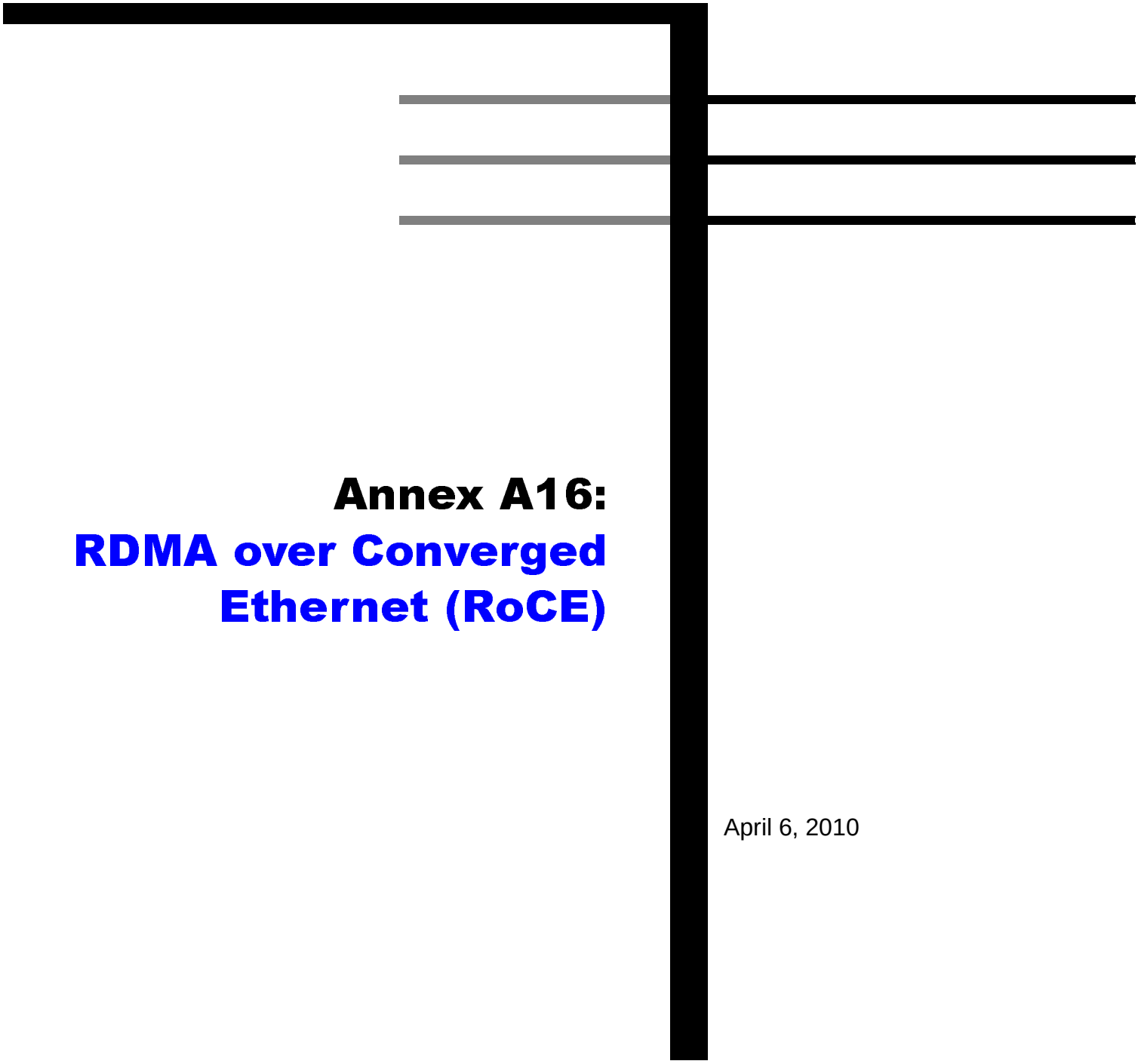
Supplement to InfiniBand™ Architecture Specification Volume 1 Release 1.2.1



Annex A16: RDMA over Converged Ethernet (RoCE)

April 6, 2010

Table 0 Revision History

Revision	Date	
1.0	04/06/2010	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 A16: RDMA OVER CONVERGED ETHERNET (ROCE)

A16.1 INTRODUCTION

This document is an annex to release 1.2.1 of Volume 1 of the InfiniBand Architecture, herein referred to as the base specification. This annex enables the use of alternate, non-IB defined, link and physical layers. The name given to describe this ability is RDMA over Converged Ethernet ('RoCE', pronounced as 'Rocky'). There is no particular significance to the given name.

Support for RoCE is optional, but if it is supported it must be done in accordance with the requirements stated in this annex.

A16.2 OVERVIEW

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 including reliable and unreliable service, connected and unconnected service, atomics, multicast and others. This is achieved through a layered architecture that specifies the first four layers of the OSI reference model network stack including the physical, link, network and transport layers as well as an accompanying management framework for the InfiniBand network. 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.

This annex specifies a way to provide these same transport services over a non-InfiniBand network. The definition preserves the IB software interface and IB transport services and protocol and describes their use when run on an Ethernet network. The specification assumes common management practices of the underlying Ethernet network and does not apply the IB management infrastructure, e.g. SM, SA, and so on.

The InfiniBand transport protocol was architected to operate on a layer 2 fabric which does not routinely drop packets even though the protocol contains capabilities to allow it to recover from some number of lost packets. Although this annex does not specifically require a 'lossless' Ethernet fabric, it is likely that a port implementing this annex which is connected to an Ethernet fabric will operate more efficiently (e.g. with a higher ratio of delivered payload to offered payload) if the underlying layer 2 fabric does not routinely drop packets.

A16.2.1 ARCHITECTURAL GOALS

RoCE has the following architectural goals:

- Provide RDMA service over an Ethernet Layer 2 network
- Maintain compliance with the existing verbs definition and accompanying implementations.
- Maintain the existing memory management paradigms as described and defined in the base specification.
- Maintain the existing InfiniBand transport architecture, including RC, UC, UD, RD and XRC services, as well as Atomic operations.

A given end node may support multiple ports; most modern implementations in fact support two or more ports. This annex does not create a requirement that all ports on a given endnode support RoCE, thus a hybrid implementation of an endnode may exist with one or more ports connected to an Ethernet network and supporting RoCE and one or more ports connected to an IB network and supporting standard InfiniBand.

The port numbering rules defined in the base specification continue to apply. RoCE ports are included in the port number space.

A16.3 ROCE ON-THE-WIRE FORMATS

A16.3.1 PACKET HEADER LAYOUT

A RoCE packet has an identical layout and definition to an existing IB Packet as defined in Chapter 5 of the base specification, with the following exceptions:

- Every RoCE packet contains a GRH
- Every RoCE packet contains a MAC header in place of the IB LRH. The format and definition of the MAC header is outside the scope of this annex.
- Every RoCE packet contains an FCS as defined for Ethernet frames. The format and definition of this FCS is beyond the scope of this annex.

Conceptually, a MAC header has the same fundamental purpose as the IB LRH, which is to provide enough information about the packet source and destination to enable the underlying network, Ethernet in this case, to successfully switch the packet through the subnet such that it is delivered to the appropriate destination endpoint.

CA16-1: Packets generated by a RoCE port shall conform to the packet structure defined in Figure 1 RoCEE Packet Format on page 3.

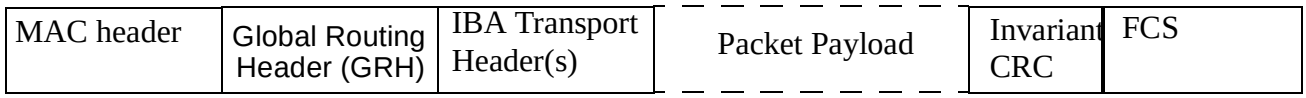


Figure 1 RoCE Packet Format

CA16-2: The fields comprising the MAC header are not defined in this annex, except that the value in the EtherType field shall be 0x8915, and except that the DMAC and SMAC fields shall be filled with the respective MAC addresses of the L2 destination and L2 source ports for the RoCE packet.

CA16-3: A RoCE packet shall not contain a Variant CRC. A RoCE packet shall contain an FCS appropriate to the Ethernet network being used. The definition of this FCS is outside the scope of this annex.

CA16-4: Each RoCE packet shall contain a valid GRH.

A16.3.2 ICRC CALCULATION

The Invariant CRC (ICRC) is calculated exactly as described in the base specification for the case where the GRH is present. Even though the LRH is not present in a RoCE packet, the CRC calculation includes 64 bits of '1' exactly as described in the base specification. The MAC header is not included in the ICRC calculation.

A16.4 MANAGEMENT CONSIDERATIONS

The InfiniBand Architecture specifies a number of management classes including Subnet Management, Communication (connection) Management, Performance Management, Device Management, Baseboard Management, SNMP Tunneling, Vendor specific, Application specific classes and Congestion Control. With the exception of Communication Management, the balance of these management classes are outside the scope of this annex.

The Subnet Management class and the Subnet Administration class are deprecated for subnets conforming to the annex. Instead, a subnet conforming to this annex is assumed to use standard Ethernet management practices for Layer 2 address assignment, fabric topology discovery and switch configuration.

Changes to the Communication Management class are addressed in section A16.5 RoCEE Addressing on page 6.

Compliance statement C16-1 in Section16.1 Performance Management of the base specification requires an InfiniBand CA, Switch or Router to make performance and error statistics available through a Performance Management Agent. This requirement does not apply to RoCE ports.

There are two verbs which contain Capability Mask bits which are a fundamental part of the mechanism for defining how certain management classes are discovered. As will be seen below in sections A16.9.1 Query HCA on page 11 and section A16.9.2 Modify HCA on page 11 describing changes to the Query HCA and Modify HCA verbs, certain of these capability bits are no longer applicable with respect to RoCE. For example, the IsSM, IsSMDisabled, IsSNMPTunnelingSupported and IsClientReregistrationSupported capability mask bits are all eliminated. On the other hand there are two capability mask bits, IsDeviceManagementSupported and IsVendorClassSupported that are preserved for possible future use, even though the device management class and the vendor class are not currently defined in this annex.

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 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).

A16.4.1 SUBNET MANAGEMENT

This annex enables the InfiniBand transport to be run directly on top of a non-IB network. Since InfiniBand's subnet management and its accompanying subnet administration architecture are solely concerned with Layer 2 address assignment, fabric topology discovery and switch configuration, it follows that definition of subnet management is outside the scope of this annex. Although there are no compliance statements contained within this annex mandating it, it is assumed by this annex that standard Ethernet-defined fabric management methods are applied to configuring and managing the Layer 2 subnet, including endpoint Layer 2 address assignment, fabric topology discovery and switch configuration.

CA16-5: A system conforming to this annex shall not be required to have a SM per subnet.

CA16-6: A RoCE port designed to be compliant with this annex shall not be required to support the functionality characterized as a Subnet Management Agent (SMA).

A16.4.2 SUBNET ADMINISTRATION

Subnet administration is a feature of an IB subnet designed to provide a central repository for information concerning the fabric and its configuration. Although described separately, the base specification considers Subnet Administration (SA) to be a component part of subnet management. Thus, for systems designed to this annex, there are no specific requirements for subnet administration.

CA16-7: A system conforming to this annex shall not be required to provide an SA for any subnets.

Note that the SA has a few other functions beyond providing subnet configuration information, such as storing application ServiceRecords. By not requiring an SA, this annex also does not require that these other services be provided.

A16.4.3 MANAGEMENT INTERFACES

As defined in the base specification, a special Queue Pair, QP0 is defined solely for communication between subnet manager(s) and subnet management agents. Since such an IB-defined subnet management architecture is outside the scope of this annex, it follows that there is also no requirement that a port which conforms to this annex be associated with a QP0. Thus, for end nodes designed to conform to this annex, the concept of QP0 is undefined and unused for any port connected to an Ethernet network.

CA16-8: A packet arriving at a RoCE port containing a BTH with the destination QP field set to QP0 shall be silently dropped.

A16.4.4 MANAGEMENT MESSAGING

Section 13.4.6 Management Messaging in the base specification discusses requirements necessary for agents using MADs for messaging.

CA16-9: The default value for RespTimeValue shall be 19. The default value for SubnetTimeout shall be 18. These default values can be modified by Ethernet management practices.

The default SubnetTimeout value can be used as an upper bound estimate of InfiniBand PacketLifeTime, should no other information be available.

A MAD responder uses a reversible path to return a response packet to the original sender using only header information present in the request packet. The method for creating a packet to follow a reversible path through an InfiniBand fabric is specified in Section 13.5.4 (Response Generation and Reversible Paths) in the base specification. Constructing a re-

sponse packet to follow a reversible path for a request packet received on a RoCE port is very similar. The entire process is as follows:

- The MAC header is generated through Ethernet-defined methods.
- The Source QP of the received packet is used as the Destination QP in the BTH of the response packet.
- The Destination QP of the received packet is used as the Source QP in the DETH of the response packet.
- The responder's P_Key used to match the P_Key in the received packet is used as the P_Key in the response packet.
- The Rate and MTU values are obtained using Ethernet management practices
- The Q_Key in the DETH of the response packet is set to the value of the Q_Key of the received packet.
- The SGID of the received packet is used as the DGID in the GRH of the response packet.
- The DGID of the received packet is used as the SGID in the GRH of the response packet.
- The FlowLabel and TrafficClass are copied from the GRH in the received packet.
- HopLimit in the GRH is set to 0xFF.
- Fields not otherwise specified in this section are filled in according to the requirements of the transport service.

A16.5 RoCE Addressing

A16.5.1 Address Assignment and Resolution

Layer 2 local addresses (i.e. SMAC, DMAC), and the methods by which those addresses are assigned, are outside the scope of this annex.

The means for resolving a GID to a local port address (i.e. SMAC or DMAC) are outside the scope of this annex. It is assumed that standard Ethernet mechanisms, such as ARP or Neighbor Discovery are used to maintain an appropriate address cache for RoCE ports.

A16.5.2 Connection Management

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

Communication Managers maintain the Connection State information during the lifetime of a connection, except for the Remote CM LID, which it is not required to maintain.

A16.5.2.1 REQ MESSAGE CONTENTS

CA16-10: When a connection is being established between RoCE ports, the Primary Local Port LID, Primary Remote Port LID, Alternate Local Port LID and Alternate Remote Port LID fields of the REQ message shall be Reserved and Ignored. The value of these fields shall not be checked or validated by a recipient of a REQ message.

A16.5.2.2 REJ MESSAGE CONTENTS

CA16-11: When a connection is being established between RoCE 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

A16.5.2.3 LAP MESSAGE CONTENTS

CA16-12: When alternate paths are being established between RoCE ports, the Alternate Local Port LID and Alternate Remote Port LID fields are Reserved and Ignored. The value of these fields shall not be checked or validated by a recipient of a LAP message.

A16.5.2.4 APR MESSAGE CONTENTS

CA16-13: When alternate paths are being established between RoCE ports, the following APR status code shall not be sent as part of an APR message:

- code 7: Proposed Alternate Remote Port LID rejected.

A16.6 LINK LAYER CONSIDERATIONS

The InfiniBand Architecture defines several available transport services plus two encapsulations called Raw services. A Raw service is defined as one which does not use the InfiniBand transport. The base specification identifies a Raw packet as one containing a zero in the msb of the Link Next Header field of the LRH. Hence, the Raw services as defined in the base specification are provided by the link layer. Since RoCE does not use the InfiniBand-defined link layer, there is likewise no concept of RAW service, as defined in the base specification.

CA16-14: An implementation of an endnode claiming conformance to this annex shall not support the concept of a Raw service on a RoCE port.

A16.7 TRANSPORT CONSIDERATIONS

CA16-15: An end node containing a RoCE port which claims compliance to this annex shall be compliant with the InfiniBand transport as defined in Chapter 9 of the base specification, with the exceptions contained in this section.

Tables 60, 62 in the Transport Layer chapter (Chapter 9) of the base specification and Table 64 in the Software Transport Interface chapter (Chapter 10) of the base specification list the sources of various header fields and parameters of interest to the transport, or the means by which the values for these fields are computed. Each of the items relevant to the LRH is sourced either from an inbound packet (denoted in the table as 'link'), or is contained as part of the QP context or is defined by the WQE, or is computed from one of the above. Since RoCE does not define a link layer, the lines of these tables that refer to the LRH are no longer applicable for any RoCE port.

In addition, Table 64 contains a column labeled Raw IP and a column labeled Raw ET. These columns are not applicable for a port claiming compliance to this annex since Raw service is not supported by RoCE ports.

A16.7.1 TRANSPORT HEADER CHECKS

The base specification requires the transport layer to conduct a comprehensive set of checks of various header fields. These checks are designed to ensure that packets are delivered correctly to the destined endpoint port and to the correct QP within that endpoint. Since the base specification does not require the use of the GRH except in certain circumstances, the transport header validation protocol requires careful examination of the LRH (Local Route Header) to verify that an incoming packet has arrived at its proper destination.

RoCE packets are always required to contain a GRH. Included as part of the GRH are a source and destination GID (SGID, DGID). These values are end-to-end invariant and serve to uniquely identify the source and destination of a RoCE packet. The existing transport header checking validation protocol requires these fields to be carefully examined for all cases in which the GRH is present.

As specified above, RoCE packets do not contain an LRH, thus the portions of the transport header validation protocol devoted to verifying the fields of the LRH are eliminated for packets arriving on an endnode port that conforms to this annex.

A16.7.2 LOOPBACK

A RoCE port supports loopback via two methods. A loopback can be created either through a self-addressed packet (i.e. SGID == DGID) or through the use of the optional loopback indicator. As is specified in the base specification, the Loopback GID (0:0:0:0:0:0:0:1) is reserved for Raw services, which are not defined for RoCE. Thus, a packet which contains either an SGID or a DGID set to the loopback GID will be silently dropped by the transport in accordance with the base specification.

A16.7.3 STATIC RATE CONTROL

CA16-15.1.1: A RoCE port shall support static rate control.

Static rate control is described in section 9.11 of the Transport Layer chapter of the base specification. Static rate control ensures that a faster source port does not emit packets at a rate greater than a slower destination port's ability to consume them or a slower intermediate link's ability to deliver them. Mechanisms for determining the speed of various paths through an Ethernet fabric are outside the scope of this annex. It is up to an individual implementation to support Interpacket Delay (IPD) values appropriate for the fabric to which the port is attached.

A16.7.4 CONGESTION CONTROL

Congestion control as defined in Annex A10 of the base specification requires the detection of congestion in the switches on a VL basis. Once detected, the switch notifies the channel adapter which is the target of the packet using the FECN bit in the BTH. The channel adapter which is the target of the packet, in turn, instructs the source channel adapter to reduce its packet injection rate. The necessary thresholds and injection rates are configured by an InfiniBand Congestion Control Manager. Since the underlying mechanisms are based on InfiniBand switches and VLs, neither of which exist in an Ethernet fabric, this congestion control mechanism is not applicable to an Ethernet fabric. 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 RoCE ports.

CA16-16: 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 RoCE port.

In any case, the B (BECN) and F(FECN) bits are not covered by the invariant CRC.

A16.7.5 QOS MANAGEMENT

QoS Management as defined in Annex A13 of the base specification makes use of InfiniBand capabilities such as VLs, VL arbitration and SL-to-VL mapping. It also relies on InfiniBand-defined performance metrics such as PortXmitData(VL), PortRecvData(SL) and PortXmitQueue(VL). Further, the InfiniBand QoS Manager is tightly coupled with Subnet Administration (SA) and is used to filter path requests based on QoS requirements. These underlying mechanisms and relationships are not applicable to Ethernet fabrics. 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 RoCE ports.

A16.8 SOFTWARE TRANSPORT INTERFACE

The addressing structure used by a verbs consumer to access transport services through the software interface is called an address vector. In accordance with the principles of a layered architecture, a verbs consumer refers to both sources and destinations via a GID (Global ID).

A verbs consumer using a RoCE network relies strictly on so-called Layer 3 addressing (GIDs); layer 2 addresses (e.g. subnet local identifiers) are not passed across the verbs interface. Therefore, the concepts of a Base LID, LMC and Path bits do not apply to an implementation which conforms to this annex.

CA16-17: When accessing the services of a RoCE verbs provider, the source and destination identifiers contained in the address vector shall consist of GIDs; the address vector shall not contain layer 2 references (e.g. local addresses). Layer 2 references include source and destination local identifiers and LID Path Bits.

A16.8.1 SL, VL MANAGEMENT

InfiniBand defines an end-to-end service level, called SL. End-to-end service levels are represented hop-to-hop by Virtual Lanes (VLs). A virtual lane is implemented as a set of physical resources in an endnode's IB ports and in an InfiniBand switch. A verbs consumer accessing an IB port specifies its desired level of service by setting the SL. At each hop, the SL is mapped to a particular VL. The InfiniBand architecture supports up to 16 SLs, numbered zero through 15. The number of VLs supported by any given IB port or switch is a vendor differentiation item. VLs are defined at the InfiniBand link layer (layer 2); they do not exist on an Ethernet fabric.

Ethernet provides a construct, called a Priority Level which corresponds conceptually to InfiniBand's SLs. Eight priorities, numbered zero through seven are supported. As in InfiniBand, a verbs consumer accessing a RoCE port specifies its desired service level, which is then mapped to a given Ethernet Priority. The default mapping is as follows:

SL 0-7 are mapped directly to Priorities 0-7, respectively

SL 8-15 are reserved.

CA16-18: An attempt to use an Address Vector for a RoCE port containing a reserved SL value shall result in the Invalid Address Vector verb result.

A16.8.2 PARTITIONING

Methods to populate the P_Key table associated with a RoCE 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 of course validated for an inbound packet as required by the packet header validation protocols defined in Chapter 9 of the base specification.

A16.9 VERBS CONSIDERATIONS

The following sections specify modifications required to any verbs, for example to eliminate references to a local address, or to force the verb to provide a global identifier even for locally routed packets.

A16.9.1 QUERY HCA

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

- The Base LID & LMC fields are unused and are ignored.
- The maximum number of virtual lanes supported by this HCA is unused and is ignored.
- The Optional InitTypeReply value is unused and is ignored.
- The Subnet Manager address information for each RoCE port of this HCA is unused and is ignored.
- The CapabilityMask bits IsSM, IsSMDisabled, IsSNMPTunnelingSupported, IsClientReregistrationSupported are unused and are ignored.
- A new attribute is added to the Port Attribute list (one for each port on this HCA) indicating the port type; whether the port is a RoCE port or an IB port.

A16.9.2 MODIFY HCA

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

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

A16.9.3 CREATE ADDRESS HANDLE

The Address Vector of the Input Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID field is unused and are reserved.
- The Source GID index is set for all destinations (global and local)
- The Destination GID is set for all destinations (global and local).
- The Source Path Bits are unused and are reserved.

- The Send Global Routing Header Flag is set, regardless of the selected transport service.

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 RoCE port.

A16.9.4 MODIFY ADDRESS HANDLE

The Address Vector of the Input Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID field is unused and is reserved.
- The Source GID index is set for all destinations
- The Destination GID is set for all destinations
- The Source Path Bits are unused and are reserved.
- The Send Global Routing Header Flag is set, regardless of the selected transport service.

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 RoCE port.

A16.9.5 QUERY ADDRESS HANDLE

The Address Vector Output Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is ignored.
- The Source GID index is included for all destinations
- The Destination GID is set for all destinations
- The Source Path bits are unused and are ignored.
- The Send Global Routing Header Flag is set.

A16.9.6 MODIFY QUEUE PAIR / MODIFY XRC INITIATOR QP / MODIFY XRC TARGET QP

The Address Vector Input Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is reserved.
- The Source Path Bits are unused and are reserved.
- The Source GID Index is set for all destinations (global and local)
- The Destination GID is set for all destinations (global and local)
- The Send Global Routing Header Flag is set.

The Alternate Path Address Information Input Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is reserved.
- The Source Path Bits are unused and are reserved.
- The Source GID Index is set for all destinations (global and local)
- The Destination GID is set for all destinations (global and local)
- The Send Global Routing Header Flag is set.

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 RoCE port.

A16.9.7 QUERY QUEUE PAIR / QUERY XRC INITIATOR QP / QUERY XRC TARGET QP

The Primary Address Vector Output Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is ignored
- The Source Path Bits are unused and are ignored.
- The Source GID Index is returned for all destinations (global and local).
- The Destination GID is returned for all destinations (global and local).
- The Send Global Routing Header Flag is set.

The Alternate Path Address Information Output Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is ignored
- The Source Path Bits are unused and are ignored.
- The Source GID Index is returned for all destinations (global and local).
- The Destination GID is returned for all destinations (global and local).
- The Send Global Routing Header Flag is set.

A16.9.8 MODIFY EE CONTEXT ATTRIBUTES

The Primary Address Vector Input Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is reserved.
- The Source Path Bits are unused and are reserved.
- The Source GID Index is set for all destinations (global and local).
- The Destination GID is set for all destinations (global and local).
- The Send Global Routing Header Flag is set

The Alternate Path Address Vector Input Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is reserved.
- The Source Path Bits are unused and are reserved.
- The Source GID Index is set for all destinations (global and local).
- The Destination GID is set for all destinations (global and local).
- The Send Global Routing Header Flag is set.

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 RoCE port.

A16.9.9 QUERY EE CONTEXT

The Primary Address Vector Output Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is ignored
- The Source Path Bits are unused and are ignored.
- The Source GID Index is set for all destinations (global and local)
- The Destination GID is set for all destinations (global and local).
- The Send Global Routing Header Flag is set.

The Alternate Path Address Information Output Modifier for this verb when associated with a RoCE port is changed as follows:

- The Destination LID is unused and is ignored
- The Source Path Bits are unused and are ignored.
- The Source GID Index is set for all destinations (global and local)
- The Destination GID is set for all destinations (global and local).
- The Send Global Routing Header Flag is set.

A16.9.10 ATTACH QP TO MULTICAST GROUP

If the QP is associated with a RoCE port, the Input Modifiers for this verb are changed as follows:

- The Multicast group MLID is unused and is ignored.

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

- “Invalid multicast MLID” is removed as a valid Verb Result

A16.9.11 DETACH QP FROM MULTICAST GROUP

If the QP is associated with a RoCE port, the Input Modifiers for this verb are changed as follows:

- The Multicast group MLID is unused and is ignored.

If the QP is associated with a RoCE port, the Output Modifiers for this verb are changed as follows:

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

A16.9.12 POLL FOR COMPLETION

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

- If the remote port is a RoCE port, the remote port address and QP information returned for datagram services (shown in Table 97 of the base specification) do not include the 16-bit SLID or the DLID path bits.

A16.9.13 GET SPECIAL QP

For a given HCA port, the Get Special QP verb returns the handle for the following special QP types: SMI QP (QP0), GSI QP (QP1), Raw IPv6 and Raw Ethertype. Compliance statement C11-13 in Section 11.2.5 Get Special QP in the base specification requires that the verb support QP0 and QP1. The optional compliance statement o11-1 requires the verb to support the Raw Datagram types if the HCA supports them. Since there is no QP0 associated with a RoCE port, and since a RoCE port does not support either Raw datagram type, these compliance statements do not apply to a RoCE port with respect to QP0 or Raw Datagram QPs.

A16.9.14 POST SEND REQUEST

The Post Send Request verb is modified to eliminate Raw as one of the possible service types. Table 93 in the base document is modified to effectively eliminate the last row of the table governing the Raw service type.

CA16-19: A QP associated with a RoCE port does not support any operations for the Raw service type.

A16.9.15 UNAFFILIATED ASYNCHRONOUS EVENTS

The base specification describes an optional Client Reregistration Event generated by the CI when the SMA receives this request from the SM.

CA16-20: A RoCE port shall not support Client Reregistration.

The base specification describes an optional Port Change event which is triggered when one of a number of port attributes change.

CA16-21: A RoCE port shall not support the optional Port Change Event.

A16.10 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 RoCE ports. This section describes the exceptions.

A16.10.1 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, Ethernet fabrics and the RoCE ports that are attached to them do not support InfiniBand Subnet Management. Therefore, compliance statement C17-7 does not apply to RoCE ports.

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

A16.10.2 LOCALLY ROUTED PACKETS

Compliance statement C17-9 states that a CA shall be able to source and sink locally routed packets, which are described as those containing no GRH. However, consistent with other sections of this annex, each RoCE packet is required to contain a GRH. Thus, C17-9 is replaced by the following compliance statement.

CA16-22: A CA shall be able to source and sink locally routed packets, including those containing a GRH in the case of a RoCE packet.

A16.10.3 BACKPRESSURE, 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 RoCE ports.

A16.10.4 INBOUND PACKET CHECKING

Compliance statement C17-21 requires a CA to check for link, network and transport layer errors in all incoming packets. However, a RoCE port does not implement either the link or network layers. Thus, for CAs which claim compliance to this annex, C17-21 is replaced as follows:

CA16-23: The CA shall check for link, network and transport layer errors in all incoming packets received on an InfiniBand port. For incoming

packets received on a RoCE port, the CA shall check for transport layer errors.

A16.10.5 SUPPORT FOR QP0

Compliance statement C17-24 requires each port of a CA to support QP0. This requirement is removed for any RoCE ports supported by a CA.