OPSAWG
Internet-Draft
Intended status: Standards Track
Expires: 31 August 2026

C. Lin
H. Zhang
New H3C Technologies

27 February 2026

Export of RoCEv2 Base Transport Header (BTH) Information Using IP Flow Information Export (IPFIX)
draft-lin-opsawg-ipfix-rocev2-00

Abstract

This document defines a new set of IP Flow Information Export (IPFIX) Information Elements (IEs) for exporting Base Transport Header (BTH) information for RDMA over Converged Ethernet version 2 (RoCEv2) traffic. These extensions enable network monitoring systems to collect and analyze the characteristics of RDMA traffic widely used in high-performance computing, storage, and artificial intelligence applications.

Table of Contents

1. Introduction 2
2. Terminology 3
3. New IPFIX Information Elements for RoCEv2 BTH 3
4. Operational Considerations 4
5. Security Considerations 4
6. IANA Considerations 5
6.1. New IPFIX IEs for RoCEv2 BTH 5
6.1.1. rdmaOpCode 5
6.1.2. rdmaPartitionKey 6
6.1.3. rdmaDestinationQP 6
6.1.4. rdmaSourceQP 6
6.1.5. rdmaPacketSequenceNumber 7
6.1.6. rdmaBTHFlags1 7
6.1.7. rdmaBTHFlags2 8
6.1.8. rdmaBTHFlags3 8
7. References 9
7.1. Normative References 9
7.2. Informative References 9
Authors' Addresses 10

1. Introduction

Remote Direct Memory Access (RDMA) [RFC5040] is a network technology that allows a computer to read from or write to the memory of another computer directly, without involving the operating system. This zero-copy and kernel-bypass feature greatly reduces CPU overhead and communication latency. InfiniBand [IBTA-SPEC] and RDMA over Converged Ethernet (RoCE) are two mainstream RDMA implementations that bypass the operating system kernel and achieve zero-copy data transfer.

RoCE technology has become a key component of high-performance data center networks, especially in low-latency, high-throughput scenarios such as artificial intelligence training, distributed storage, and financial transactions. RDMA over Converged Ethernet version 2 (RoCEv2) runs on top of UDP (port 4791) and inherits the transport layer protocol of the InfiniBand Architecture (IBA).

The existing IPFIX [RFC7011] standard lacks the ability to monitor specific fields of RoCEv2, which limits the ability of network operators to perform in-depth analysis, troubleshooting, and performance optimization of RDMA traffic. To close this gap, this document defines a new set of Information Elements (IEs) to carry RoCEv2 BTH key fields.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

The following terms are used as defined in [RFC7011]:

* IPFIX

* IPFIX Information Elements

The following terms are used in this document:

* QP: Queue Pair, a communication endpoint in RDMA architecture.

* BTH: Base Transport Header, the RDMA transport header containing QP information.

* P_Key: Partition Key.

* PSN: Packet Sequence Number.

3. New IPFIX Information Elements for RoCEv2 BTH

This section defines new IPFIX IEs for exporting RoCEv2 BTH key fields.

rdmaOpCode
: The RoCEv2 BTH OpCode field, which indicates the type of the InfiniBand Architecture data packet.

rdmaPartitionKey
: The RoCEv2 BTH Partition Key field, which identifies the logical partition to which the data packet belongs.

rdmaDestinationQP
: The RoCEv2 BTH Destination Queue Pair (QP) field, indicating the work QP number at the destination.

rdmaSourceQP
The RoCEv2 BTH Source Queue Pair (QP) field (if present), indicating the work QP number at the Source.

rdmaPacketSequenceNumber
The RoCEv2 BTH Packet Sequence Number (PSN) field, which is used to detect whether data packets are lost or duplicated, ensuring the reliability and orderliness of transmission.

rdmaBTHFlags1
The SE, M, Pad, and TVer fields of The RoCEv2 BTH. The size of this Information Element is 1 octet.

rdmaBTHFlags2
The F/R, and B/R fields of The RoCEv2 BTH. The size of this Information Element is 1 octet.

rdmaBTHFlags3
The A field of The RoCEv2 BTH. The size of this Information Element is 1 octet.

4. Operational Considerations

The exporter needs to parse RoCEv2 BTH information, which may affect the exporter’s performance. Implementers SHOULD consider measures to mitigate this impact, such as sampling rate limiting or hardware acceleration.

5. Security Considerations

The Security Considerations for IPFIX [RFC7011] apply to this document as well.

When exporting RDMA BTH information across security domains, to prevent information security risks caused by the leakage of sensitive network topology details such as QP numbers, it is recommended to implement comprehensive protection measures, such as using the encrypted transmission options of the IPFIX framework (such as DTLS [RFC9147]) to ensure the confidentiality and integrity of data during transmission.

To defend against Denial-of-Service (DOS) attacks that may be caused by maliciously crafted RoCEv2 packets and to prevent the exporter from being overloaded by processing a large amount of invalid traffic, it is recommended to adopt certain protection strategies, such as configuring reasonable rate limiting policies to prevent the monitoring system from being overwhelmed by massive data packets and to ensure its stable operation.

6. IANA Considerations

6.1. New IPFIX IEs for RoCEv2 BTH

This document specifies new IPFIX IEs to enable export of RoCEv2 BTH key fields along with other flow information. This document requests IANA to add these IPFIX IEs to the "IPFIX Information Elements" registry available at [IANA-IPFIX].

Table 1 lists the new IPFIX IEs for RoCEv2 BTH:

Element ID	Name	Reference
TBD1	rdmaOpCode	This document
TBD2	rdmaPartitionKey	This document
TBD3	rdmaDestinationQP	This document
TBD4	rdmaSourceQP	This document
TBD5	rdmaPacketSequenceNumber	This document
TBD6	rdmaBTHFlags1	This document
TBD7	rdmaBTHFlags2	This document
TBD8	rdmaBTHFlags3	This document

Table 1: New IEs in the "IPFIX Information Elements" Registry

6.1.1. rdmaOpCode

Name: rdmaOpCode

Element ID: TBD1

Description: The RoCEv2 BTH OpCode field, which indicates the type of the InfiniBand Architecture data packet.

Abstract Data Type: unsigned8

Data Type Semantics: identifier

Status: current

Reference: [this document]

6.1.2. rdmaPartitionKey

Name: rdmaPartitionKey

Element ID: TBD2

Description: The RoCEv2 BTH Partition Key field, which identifies the logical partition to which the data packet belongs.

Abstract Data Type: unsigned16

Data Type Semantics: identifier

Status: current

Reference: [this document]

6.1.3. rdmaDestinationQP

Name: rdmaDestinationQP

Element ID: TBD3

Description: The RoCEv2 BTH Destination Queue Pair (QP) field, indicating the work QP number at the destination. The actual effective bits are 24 bits, stored in the lower 24 bits of the 32-bit field, and the higher 8 bits should be 0.

Abstract Data Type: unsigned32

Data Type Semantics: identifier

Status: current

Reference: [this document]

6.1.4. rdmaSourceQP

Name: rdmaSourceQP

Element ID: TBD4

Description: The RoCEv2 BTH Source Queue Pair (QP) field (if present), indicating the work QP number at the Source. The actual effective bits are 24 bits, stored in the lower 24 bits of the 32-bit field, and the higher 8 bits should be 0.

Abstract Data Type: unsigned32

Data Type Semantics: identifier

Status: current

Reference: [this document]

6.1.5. rdmaPacketSequenceNumber

Name: rdmaPacketSequenceNumber

Element ID: TBD5

Description: The RoCEv2 BTH Packet Sequence Number (PSN) field, which is used to detect whether data packets are lost or duplicated, ensuring the reliability and orderliness of transmission. The actual effective bits are 24 bits, stored in the lower 24 bits of the 32-bit field, and the higher 8 bits should be 0.

Abstract Data Type: unsigned32

Data Type Semantics: default

Status: current

Reference: [this document]

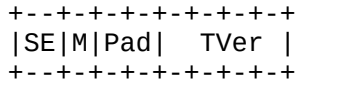
6.1.6. rdmaBTHFlags1

Name: rdmaBTHFlags1

Element ID: TBD6

Description: The SE, M, Pad, and TVer fields of The RoCEv2 BTH.

The size of this Information Element is 1 octet.



Bits 0: Solicited Event (SE) field.
Bits 1: Migration Request (M) field.
Bits 2-3: Pad Count (Pad) field.
Bits 4-7: Transport Header Version (TVer) field.

Abstract Data Type: unsigned8

Data Type Semantics: default

Status: current

Reference: [this document]

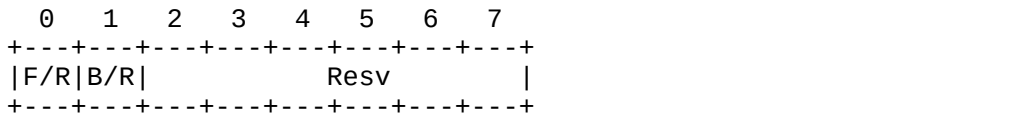
6.1.7. rdmaBTHFlags2

Name: rdmaBTHFlags2

Element ID: TBD7

Description: The F/R, and B/R fields of The RoCEv2 BTH.

The size of this Information Element is 1 octet.



Bits 0: Forward Explicit Congestion Notification (FECN)/Res1 (F/R) field.
Bits 1: Backward Explicit Congestion Notification (BECN)/Res1 (B/R) field.
Bits 2-7: Reserved field.

Abstract Data Type: unsigned8

Data Type Semantics: flags

Status: current

Reference: [this document]

6.1.8. rdmaBTHFlags3

Name: rdmaBTHFlags3

Element ID: TBD8

Description: The A field of The RoCEv2 BTH.

The size of this Information Element is 1 octet.

Bits 0: Acknowledge Request (A) field.
Bits 1-7: Reserved field.

Abstract Data Type: unsigned8

Data Type Semantics: flags

Status: current

Reference: [this document]

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC5040] Recio, R., Metzler, B., Culley, P., Hilland, J., and D. Garcia, "A Remote Direct Memory Access Protocol Specification", RFC 5040, DOI 10.17487/RFC5040, October 2007, <https://www.rfc-editor.org/info/rfc5040>.

[RFC7011] Claise, B., Ed., Trammell, B., Ed., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, DOI 10.17487/RFC7011, September 2013, <https://www.rfc-editor.org/info/rfc7011>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

7.2. Informative References

[IANA-IPFIX] "IP Flow Information Export (IPFIX) Entities", n.d., <https://www.iana.org/assignments/ipfix/ipfix.xhtml>.

[IBTA-SPEC] InfiniBand Trade Association, "InfiniBand Architecture Specification", InfiniBand Architecture Specification Volume 1-2, Release 1.6, December 2023, <https://www.infinibandta.org/ibta-specification/>.

[RFC9147] Rescorla, E., Tschofenig, H., and N. Modadugu, "The Datagram Transport Layer Security (DTLS) Protocol Version 1.3", RFC 9147, DOI 10.17487/RFC9147, April 2022, <https://www.rfc-editor.org/info/rfc9147>.

Authors' Addresses

Changwang Lin
New H3C Technologies
Beijing
China
Email: linchangwang.04414@h3c.com

Haiyang Zhang
New H3C Technologies
Beijing
China
Email: zhang.haiyangA@h3c.com