Workgroup: 6man
Internet-Draft:
draft-bonica-6man-comp-rtg-hdr-27
Published: 12 November 2021
Intended Status: Standards Track
Expires: 16 May 2022
Authors: R. Bonica, Juniper Networks
Y. Kamite, NTT Communications Corporation
A. Alston, Liquid Telecom
D. Henriques, Liquid Telecom
L. Jalil, Verizon

The IPv6 Compact Routing Header (CRH)

Abstract

This document defines two new Routing header types. Collectively, they are called the Compact Routing Headers (CRH). Individually, they are called CRH-16 and CRH-32.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

IPv6 [RFC8200] source nodes use Routing headers to specify the path that a packet takes to its destination. The IETF has defined several Routing header types [IANA-RH]. This document defines two new Routing header types. Collectively, they are called the Compact Routing Headers (CRH). Individually, they are called CRH-16 and CRH-32.

The CRH allows IPv6 source nodes to specify the path that a packet takes to its destination. The CRH:

* Can be encoded in relatively few bytes.

* Is designed to operate within a network domain. (See Section 9).

The following are reasons for encoding the CRH in as few bytes as possible:

* Many ASIC-based forwarders copy headers from buffer memory to on-chip memory. As header sizes increase, so does the cost of this copy.

* Because Path MTU Discovery (PMTUD) [RFC8201] is not entirely reliable, many IPv6 hosts refrain from sending packets larger than the IPv6 minimum link MTU (i.e., 1280 bytes). When packets are small, the overhead imposed by large Routing Headers is excessive.

2. Requirements Language

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.

3. The Compressed Routing Headers (CRH)

Both CRH versions (i.e., CRH-16 and CRH-32) contain the following fields:

* Next Header - Defined in [RFC8200].

* Hdr Ext Len - Defined in [RFC8200].

* Routing Type - Defined in [RFC8200]. Value TBD by IANA. (For CRH-16, the suggested value is 5. For CRH-32, the suggested value is 6.)

* Segments Left - Defined in [RFC8200].

* Type-specific Data - Described in [RFC8200].

In the CRH, the Type-specific data field contains a list of Segment Identifiers (SIDs). Each SID represents both of the following:

* A segment of the path that the packet takes to its destination.

* An entry in the CRH Forwarding Information Base (CRH-FIB) (Section 4).

SIDs are listed in reverse order. So, the first SID in the list represents the final segment in the path. Because segments are listed in reverse order, the Segments Left field can be used as an index into the SID list. In this document, the "current SID" is the SID list entry referenced by the Segments Left field.

The first segment in the path can be omitted from the list. See Appendix A for examples.

In the CRH-16 (Figure 1), each SID is encoded in 16-bits. In the CRH-32 (Figure 2), each SID is encoded in 32-bits.

In all cases, the CRH MUST end on a 64-bit boundary. So, the Type-specific data field MUST be padded with zeros if the CRH would otherwise not end on a 64-bit boundary.

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |  Next Header  |  Hdr Ext Len  | Routing Type  | Segments Left |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |             SID[0]            |          SID[1]               |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-|
 |                           .........
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```

Figure 1: CRH-16

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |  Next Header  |  Hdr Ext Len  | Routing Type  | Segments Left |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 +                              SID[0]                           +
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 +                              SID[1]                           +
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 //                                                              //
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 +                              SID[n]                           +
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: CRH-32

4. The CRH Forwarding Information Base (CRH-FIB)

Each SID identifies a CRH-FIB entry.

Each CRH-FIB entry contains:

- *An IPv6 address (optional).
- *A topological function.
- *Arguments for the topological function (optional).
- *Flags.

*A service function (optional).

*Arguments for the service function (optional).

The IPv6 address can represent either:

*An interface on the next segment endpoint.

*An SRv6 SID [RFC8986], instantiated on the next segment endpoint.

The first ten bits of the IPv6 address MUST NOT be fe00. That prefix is reserved for link-local [RFC6890] addresses.

The topological function specifies how the processing node forwards the packet to the next segment endpoint. The following are examples:

*Forward the packet through the least-cost path to the next segment endpoint.

*Forward the packet through a specified interface.

*Encapsulate the packet in another IPv6 header of any type (e.g., MPLS, IPv6) and forward either through the least cost path or a specified interface.

*Recycle the packet, as if the node had forwarded to one of its own interfaces. When recycling is complete, process the next SID. See Appendix B for a packet recycling use-case.

Some topological functions require parameters. For example, a topological function might require a parameter that identifies the interface through which the packet should be forwarded.

The following flags are defined:

*The PSP flag indicates whether the penultimate segment endpoint (i.e., the node that sets Segments Left to 0) MAY remove the CRH.

*The OAM flag indicates whether the processing node should invoke OAM procedures for which it is configured.

The service function is optional. If present, it invokes a node specific procedure. The following are examples of node specific procedures:

*Emit telemetry.

*Subject the packet's payload to a firewall rule.

* Replicate the packet, forwarding one copy and retaining the other for sampling, analysis, or other purposes.

Node specific procedures are not subject to standardization. A node can support any number of node specific procedures and associate them with any SIDs.

Some service functions require parameters. For example, an instruction to emit telemetry might require an IP address to which telemetry should be sent.

The CRH-FIB can be populated:

* By an operator, using a Command Line Interface (CLI).
* By a controller, using the Path Computation Element (PCE) Communication Protocol (PCEP) [RFC5440] or the Network Configuration Protocol (NETCONF) [RFC6241].
* By a distributed routing protocol [ISO10589-Second-Edition], [RFC5340], [RFC4271].

5. Processing Rules

The following rules describe CRH processing:

* If Segments Left equals 0, skip over the CRH and process the next header in the packet.
* If Hdr Ext Len indicates that the CRH is larger than the implementation can process, discard the packet and send an ICMPv6 [RFC4443] Parameter Problem, Code 0, message to the Source Address, pointing to the Hdr Ext Len field.
* Compute L, the minimum CRH length (Section 5.1).
* If L is greater than Hdr Ext Len, discard the packet and send an ICMPv6 Parameter Problem, Code 0, message to the Source Address, pointing to the Segments Left field.
* Decrement Segments Left.
* Search for the current SID in the CRH-FIB. In this document, the "current SID" is the SID list entry referenced by the Segments Left field.
* If the search does not return a CRH-FIB entry, discard the packet and send an ICMPv6 Parameter Problem, Code 0, message to the Source Address, pointing to the current SID.

* If Segments Left is greater than 0 and the CRH-FIB entry contains a multicast address, discard the packet and send an ICMPv6 Parameter Problem, Code 0, message to the Source Address, pointing to the current SID.

* If present, copy the IPv6 address from the CRH-FIB entry to the Destination Address field in the IPv6 header.

* Decrement the IPv6 Hop Limit.

* If the CRH-FIB entry contains a service function, execute it.

* If Segments Left is equal to zero, and the PSP flag in the CRH-FIB entry is set, execute the CRH removal procedure (Section 5.2).

* Submit the packet, its topological function and its parameters to the IPv6 module. See NOTE.

NOTE: By default, the IPv6 module determines the next-hop and forwards the packet. However, the topological function may elicit another behavior. For example, the IPv6 module may forward the packet through a specified interface.

5.1. Computing Minimum CRH Length

The algorithm described in this section accepts the following CRH fields as its input parameters:

* Routing Type (i.e., CRH-16 or CRH-32).

* Segments Left.

It yields L, the minimum CRH length. The minimum CRH length is measured in 8-octet units, not including the first 8 octets.

<CODE BEGINS>

```
switch(Routing Type) {
    case CRH-16:
        if (Segments Left <= 2)
            return(0)
        sidsBeyondFirstWord = Segments Left - 2;
        sidPerWord = 4;
    case CRH-32:
        if (Segments Left <= 1)
            return(0)
        sidsBeyondFirstWord = Segments Left - 1;
        sidsPerWord = 2;
    case default:
        return(0xFF);
    }

words = sidsBeyondFirstWord div sidsPerWord;
if (sidsBeyondFirstWord mod sidsPerWord)
    words++;

return(words)
```

<CODE ENDS>

5.2. CRH Removal Procedure

The processing node SHOULD execute the following procedure, if it is capable of doing so:

- *Update the Next Header field in the header preceding the CRH using a value taken from the Next Header field in the CRH.

- *Decrease the Payload Length filed in the IPv6 header by 8*(x+1), where value of x is equal to the value of the Hdr Ext Len field in the CRH.

- *Remove the CRH from the IPv6 header chain.

6. Mutability

In the CRH, the Segments Left field is mutable. All remaining fields are immutable.

7. Applications And SIDs

A CRH contains one or more SIDs. Each SID is processed by exactly one node.

Therefore, a SID is not required to have domain-wide significance. Applications can:

* Allocate SIDs so that they have domain-wide significance.

* Allocate SIDs so that they have node-local significance.

8. Management Considerations

PING and TRACEROUTE [RFC2151] both operate correctly in the presence of the CRH.

9. Security Considerations

Networks that process the CRH MUST NOT accept packets containing the CRH from untrusted sources. Their border routers SHOULD discard packets that satisfy the following criteria:

* The packet contains a CRH

* The Segments Left field in the CRH has a value greater than 0

* The Destination Address field in the IPv6 header represents an interface that resides inside of the network.

Many border routers cannot filter packets based upon the Segments Left value. These border routers MAY discard packets that satisfy the following criteria:

* The packet contains a CRH

* The Destination Address field in the IPv6 header represents an interface that resides inside of the network.

10. Implementation and Deployment Status

Juniper Networks has produced experimental implementations of the CRH on:

* A LINUX-based software platform

* The MX-series (ASIC-based) router

Liquid Telecom has deployed the CRH, on a limited basis, in their network. Other experimental deployments are in progress.

11. IANA Considerations

This document makes the following registrations in the "Internet Protocol Version 6 (IPv6) Parameters" "Routing Types" subregistry maintained by IANA:

Value	Description	Reference
5	CRH-16	This document
6	CRH-32	This document

12. Acknowledgements

Thanks to Dr. Vanessa Ameen, Fernando Gont, Naveen Kottapalli, Joel Halpern, Tony Li, Gerald Schmidt, Nancy Shaw, Ketan Talaulikar, and Chandra Venkatraman for their contributions to this document.

13. Contributors

Gang Chen

Baidu

No.10 Xibeiwang East Road Haidian District

Beijing 100193 P.R. China

Email: phdgang@gmail.com

Yifeng Zhou

ByteDance

Building 1, AVIC Plaza, 43 N 3rd Ring W Rd Haidian District

Beijing 100000 P.R. China

Email: yifeng.zhou@bytedance.com

Gyan Mishra

Verizon

Silver Spring, Maryland, USA

Email: hayabusagsm@gmail.com

14. References

14.1. 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>.

[RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/info/rfc4443>.

[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>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

[RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <https://www.rfc-editor.org/info/rfc8201>.

14.2. Informative References

[IANA-RH] IANA, "Routing Headers", <https://www.iana.org/assignments/ipv6-parameters/ipv6-parameters.xhtml#ipv6-parameters-3>.

[ISO10589-Second-Edition] International Organization for Standardization, ""Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, Second Edition,", November 2001.

[RFC2151] Kessler, G. and S. Shepard, "A Primer On Internet and TCP/IP Tools and Utilities", FYI 30, RFC 2151, DOI 10.17487/RFC2151, June 1997, <https://www.rfc-editor.org/info/rfc2151>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI

10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <https://www.rfc-editor.org/info/rfc6241>.

[RFC6890] Cotton, M., Vegoda, L., Bonica, R., Ed., and B. Haberman, "Special-Purpose IP Address Registries", BCP 153, RFC 6890, DOI 10.17487/RFC6890, April 2013, <https://www.rfc-editor.org/info/rfc6890>.

[RFC8986] Filsfils, C., Ed., Camarillo, P., Ed., Leddy, J., Voyer, D., Matsushima, S., and Z. Li, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, DOI 10.17487/RFC8986, February 2021, <https://www.rfc-editor.org/info/rfc8986>.

Appendix A. CRH Processing Examples

This appendix demonstrates CRH processing in the following scenarios:

* The SID list contains one entry for each segment in the path (Appendix A.1).

* The SID list omits the first entry in the path (Appendix A.2).

```
 -----------                  -----------                  -----------
|Node: S    |                |Node: I1   |                |Node: I2   |
|Loopback:  |----------------|Loopback:  |----------------|Loopback:  |
|2001:db8::a|                |2001:db8::1|                |2001:db8::2|
 -----------                  -----------                  -----------
      |                                                          |
      |                       -----------                        |
      |                      |Node: D    |                       |
       ----------------------|Loopback:  |----------------------
                             |2001:db8::b|
                              -----------
```

Figure 3: Reference Topology

Figure 3 provides a reference topology that is used in all examples.

SID	IPv6 Address	Forwarding Method
2	2001:db8::2	Least-cost path
11	2001:db8::b	Least-cost path

Table 1: Node SIDs

Table 1 describes two entries that appear in each node's CRH-FIB.

A.1. The SID List Contains One Entry For Each Segment In The Path

In this example, Node S sends a packet to Node D, via I2. In this example, I2 appears in the CRH segment list.

As the packet travels from S to I2:	
Source Address = 2001:db8::a	Segments Left = 1
Destination Address = 2001:db8::2	SID[0] = 11
	SID[1] = 2

Table 2

As the packet travels from I2 to D:	
Source Address = 2001:db8::a	Segments Left = 0
Destination Address = 2001:db8::b	SID[0] = 11
	SID[1] = 2

Table 3

A.2. The SID List Omits The First Entry In The Path

In this example, Node S sends a packet to Node D, via I2. In this example, I2 does not appear in the CRH segment list.

As the packet travels from S to I2:	
Source Address = 2001:db8::a	Segments Left = 1
Destination Address = 2001:db8::2	SID[0] = 11

Table 4

As the packet travels from I2 to D:	
Source Address = 2001:db8::a	Segments Left = 0
Destination Address = 2001:db8::b	SID[0] = 11

Table 5

Appendix B. A Packet Recycling Use-Case

```
             Network A--D1
            /      \      /
           /        \    /
          /          \  /
S   ----  P            +
          \         /  \
           \       /    \
            \     /      \
             Network B---DN
```

Figure 4: Packet Recycling Use-case

In Figure 4:

* The SR domain contains Node S, Node P, and a set of destination nodes (D1 through DN)
* S is connected to P
* P is connected to Network A and to Network B. Neither of these networks are SR-capable.
* The destination nodes connect to both Network A and Network B

S needs to reach each destination node through two SR paths. One SR path traverses Network A while the other traverses Network B.

Uncompressed SRv6 can encode this SR Path in two segments,with one segment instantiated on P and the other on the destination. To support this strategy, P instantiates two END.X SIDs (one per network).

CRH compressed SRv6 can encode this SR Path in two or three segments. When it encodes the path in two segments, one segment instantiated on P and the other on the destination. To support this strategy, P instantiates 2*N SIDs (one per network per destination). When CRH compressed SRv6 encodes the path in three segments, two segments are instantiated on P and the other on the destination. The first segment on P updates the IPv6 Destination address without forwarding the packet, while the other segment on P forwards the packet without updating the IPv6 destination address. To support this strategy, P instantiates 2+N SIDs (one per network and one per destination).

Authors' Addresses

Ron Bonica
Juniper Networks
2251 Corporate Park Drive
Herndon, Virginia 20171

United States of America

Email: rbonica@juniper.net

Yuji Kamite
NTT Communications Corporation
3-4-1 Shibaura, Minato-ku,
108-8118
Japan

Email: y.kamite@ntt.com

Andrew Alston
Liquid Telecom
Nairobi
Kenya

Email: Andrew.Alston@liquidtelecom.com

Daniam Henriques
Liquid Telecom
Johannesburg
South Africa

Email: daniam.henriques@liquidtelecom.com

Luay Jalil
Verizon
Richardson, Texas
United States of America

Email: luay.jalil@one.verizon.com