TRILL (TRansparent Interconnection of Lots of Links): ECN (Explicit Congestion Notification) Support
<draft-ietf-trill-ecn-support-07.txt>

Abstract

Explicit congestion notification (ECN) allows a forwarding element to notify downstream devices, including the destination, of the onset of congestion without having to drop packets. This can improve network efficiency through better congestion control without packet drops. This document extends ECN to TRILL (TRansparent Interconnection of Lots of Links) switches, including integration with IP ECN, and provides for ECN marking in the TRILL Header Extension Flags Word (see RFC 7179).

Status of This Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Distribution of this document is unlimited. Comments should be sent to the TRILL working group mailing list <trill@ietf.org>.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html. The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.
# Table of Contents

1. Introduction ............................................. 3  
   1.1 Conventions used in this document ...................... 4  

2. The ECN Specific Extended Header Flags .................... 6  

3. ECN Support ............................................. 7  
   3.1 Ingress ECN Support .................................... 7  
   3.2 Transit ECN Support .................................... 7  
   3.3 Egress ECN Support .................................... 8  
      3.3.1 Non-ECN Egress RBridges .......................... 8  
      3.3.2 ECN Egress RBridges .............................. 8  

4. TRILL Support for ECN Variants .......................... 11  
   4.1 Pre-Congestion Notification (PCN) ..................... 11  
   4.2 Low Latency, Low Loss, Scalable Throughput (L4S) ...... 12  

5. IANA Considerations ..................................... 13  

6. Security Considerations .................................. 14  

7. Acknowledgements ....................................... 14  

Normative References ...................................... 15  
Informative References .................................... 16  

Appendix A. TRILL Transit RBridge Behavior to Support L4S. 17  

Authors' Addresses ......................................... 19
Explicit congestion notification (ECN [RFC3168] [RFC8311]) allows a forwarding element (such as a router) to notify downstream devices, including the destination, of the onset of congestion without having to drop packets. This can improve network efficiency through better congestion control without packet drops. The forwarding element can explicitly mark a proportion of packets in an ECN field instead of dropping the packet. For example, a two-bit field is available for ECN marking in IP headers.

Figure 1. Example Path Forwarding Nodes
In [RFC3168], it was recognized that tunnels and lower layer protocols would need to support ECN, and ECN markings would need to be propagated, as headers were encapsulated and decapsulated. [ECNencapGuide] gives guidelines on the addition of ECN to protocols like TRILL (TRansparent Interconnection of Lots of Links) that often encapsulate IP packets, including propagation of ECN from and to IP.

In the figure above, assuming IP traffic, RB1 is an encapsulator and RB9 a decapsulator. Traffic from Source to RB1 might or might not get marked as having experienced congestion in forwarding elements, such as X, before being encapsulated at ingress RB1. Any such ECN marking is encapsulated with a TRILL Header [RFC6325].

This document specifies how ECN marking in traffic at the ingress is copied into the TRILL Extension Header Flags Word and requires such copying for IP traffic. It also enables congestion marking by a congested RBridge such as RBn or RB1 above in the TRILL Header Extension Flags Word [RFC7179].

D. Eastlake & B. Briscoe

INTERNET-DRAFT

TRILL ECN Support

At RB9, the TRILL egress, it specifies how any ECN markings in the TRILL Header Flags Word and in the encapsulated traffic are combined so that subsequent forwarding elements, such as Y and the Destination, can see if congestion was experienced at any previous point in the path from Source.

A large part of the guidelines for adding ECN to lower layer protocols [ECNencapGuide] concerns safe propagation of congestion notifications in scenarios where some of the nodes do not support or understand ECN. Such ECN ignorance is not a major problem with RBrigdes using this specification because the method specified assures that, if an egress RBridge is ECN ignorant (so it cannot further propagate ECN) and congestion has been encountered, the egress RBridge will at least drop the packet and this drop will itself indicate congestion to end stations.

1.1 Conventions used in this document

The terminology and acronyms defined in [RFC6325] are used herein with the same meaning.
In this document, "IP" refers to both IPv4 and IPv6.

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

Acronyms:

AQM - Active Queue Management
CCE - Critical Congestion Experienced
CE - Congestion Experienced
CIte - Critical Ingress-to-Egress
ECN - Explicit Congestion Notification
ECT - ECN Capable Transport
L4S - Low Latency, Low Loss, Scalable throughput
NCHbH - Non-Critical Hop-by-Hop
NCCE - Non-Critical Congestion Experienced

Not-ECT - Not ECN-Capable Transport
PCN - Pre-Congestion Notification
2. The ECN Specific Extended Header Flags

The extension header fields for explicit congestion notification (ECN) in TRILL are defined as a two-bit TRILL-ECN field and a one-bit Critical Congestion Experienced (CCE) field in the 32-bit TRILL Header Extension Flags Word [RFC7780].
These fields are shown in Figure 2 as "ECN" and "CCE". The TRILL-ECN field consists of bits 12 and 13, which are in the range reserved for non-critical hop-by-hop (NCHbH) bits. The CCE field consists of bit 26, which is in the range reserved for Critical Ingress-to-Egress (CItE) bits. The CRItE bit is the critical Ingress-to-Egress summary bit and will be one if and only if any of the bits in the CItE range (21-26) are one or there is a critical feature invoked in some further extension of the TRILL Header after the Extension Flags Word. The other bits and fields shown in Figure 2 are not relevant to ECN. See [RFC7780], [RFC7179], and [IANAFlag] for the meaning of these other bits and fields.

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Crit.|  CHbH   |   NCHbH   |CRSV | NCRSV |   CItE    |  NCItE  |
|.....|.........|...........|.....|.......|...........|.........|
|C|C|C|       |C|N|     |   |     |       |         |C|Ext|     |
|R|R|R|       |R|C|     |ECN| Ext |       |         |C|Clr|     |
|H|I|R|       |C|C|     |   | Hop |       |         |E|   |     |
|b|t|s|       |A|A|     |   | Cnt |       |         | |   |     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

**Figure 2. The TRILL-ECN and CCE TRILL Header Extension Flags Word Fields**

Table 1 shows the meaning of the codepoints in the TRILL-ECN field. The first three have the same meaning as the corresponding ECN field codepoints in the IP header as defined in [RFC3168]. However, codepoint 0b11 is called Non-Critical Congestion Experienced (NCCE) to distinguish it from Congestion Experienced in IP.

<table>
<thead>
<tr>
<th>Binary</th>
<th>Name</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>Not-ECT</td>
<td>Not ECN-Capable Transport</td>
</tr>
<tr>
<td>01</td>
<td>ECT(1)</td>
<td>ECN-Capable Transport (1)</td>
</tr>
<tr>
<td>10</td>
<td>ECT(0)</td>
<td>ECN-Capable Transport (0)</td>
</tr>
<tr>
<td>11</td>
<td>NCCE</td>
<td>Non-Critical Congestion Experienced</td>
</tr>
</tbody>
</table>

**Table 1. TRILL-ECN Field Codepoints**
3. ECN Support

This section specifies interworking between TRILL and the original standardized form of ECN in IP [RFC3168].

The subsections below describe the required behavior to support ECN at TRILL ingress, transit, and egress. The ingress behavior occurs as a native frame is encapsulated with a TRILL Header to produce a TRILL Data packet. The transit behavior occurs in all RBridges where TRILL Data packets are queued, usually at the output port. The egress behavior occurs where a TRILL Data packet is decapsulated and output as a native frame through an RBridge port.

An RBridge that supports ECN MUST behave as described in the relevant subsections below, which correspond to the recommended provisions in Section 3 and Sections 5.1-5.4 of [ECNencapGuide]. Nonetheless, the scheme is designed to safely propagate some form of congestion notification even if some RBridges in the path followed by a TRILL Data packet support ECN and others do not.

3.1 Ingress ECN Support

The behavior at an ingress RBridge is as follows:

- When encapsulating an IP frame, the ingress RBridge MUST:
  - set the F flag in the main TRILL header [RFC7780];
  - create a Flags Word as part of the TRILL Header;
  - copy the two ECN bits from the IP header into the TRILL-ECN field (Flags Word bits 12 and 13)
  - ensure the CCE flag is set to zero (Flags Word bit 26).

- When encapsulating a frame for a non-IP protocol, where that protocol has a means of indicating ECN that is understood by the ingress RBridge, it MUST follow the guidelines in Section 5.3 of [ECNencapGuide] to add a Flags Word to the TRILL Header. For a non-IP protocol with a similar ECN field to IP, this would be achieved by copying into the TRILL-ECN field from the encapsulated native frame.

3.2 Transit ECN Support

The transit behavior, shown below, is required at all RBridges where TRILL Data packets are queued, usually at the output port.

- An RBridge that supports ECN MUST implement some form of active
queue management (AQM) according to the guidelines of [RFC7567]. The RBridge detects congestion either by monitoring its own queue depth or by participating in a link-specific protocol.

- If the TRILL Header Flags Word is present, whenever the AQM algorithm decides to indicate congestion on a TRILL Data packet it MUST set the CCE flag (Flags Word bit 26).

- If the TRILL header Flags Word is not present, to indicate congestion the RBridge will either drop the packet or it MAY do all of the following instead:
  
  + set the F flag in the main TRILL header;
  + add a Flags Word to the TRILL Header;
  + set the TRILL-ECN field to Not-ECT (00);
  + and set the CCE flag and the Ingress-to-Egress critical summary bit (CRIbE).

Note that a transit RBridge that supports ECN does not refer to the TRILL-ECN field before signaling CCE in a packet. It signals CCE irrespective of whether the packet indicates that the transport is ECN-capable. The egress decapsulation behavior (described next) ensures that a CCE indication is converted to a drop if the transport is not ECN-capable.

### 3.3 Egress ECN Support

#### 3.3.1 Non-ECN Egress RBridges

If the egress RBridge does not support ECN, that RBridge will ignore bits 12 and 13 of any Flags Word that is present, because it does not contain any special ECN logic. Nonetheless, if a transit RBridge has set the CCE flag, the egress will drop the packet. This is because drop is the default behavior for an RBridge decapsulating a Critical Ingress-to-Egress flag when it has no specific logic to understand it. Drop is the intended behavior for such a packet, as required by Section 5.4 of [ECNencapGuide].
3.3.2 ECN Egress R Bridges

If an RBridge supports ECN, for the two cases of an IP and a non-IP inner packet, the egress behavior is as follows:

Decapsulating an inner IP packet: The RBridge sets the ECN field of the outgoing native IP packet using Table 3. It MUST set the ECN field of the outgoing IP packet to the codepoint at the intersection of the row for the arriving encapsulated IP packet and the column for 3-bit ECN codepoint in the arriving outer TRILL Data packet TRILL Header. If no TRILL Header Extension Flags Word is present, the 3-bit ECN codepoint is assumed to be all zero bits.

The name of the TRILL 3-bit ECN codepoint is defined using the combination of the TRILL-ECN and CCE fields in Table 2. Specifically, the TRILL 3-bit ECN codepoint is called CE if either NCCE or CCE is set in the TRILL Header Extension Flags Word. Otherwise it has the same name as the 2-bit TRILL-ECN codepoint.

In the case where the TRILL 3-bit ECN codepoint indicates congestion experienced (CE) but the encapsulated native IP frame indicates a not ECN-capable transport (Not-ECT), it can be seen that the RBridge MUST drop the packet. Such packet dropping is necessary because a transport above the IP layer that is not ECN-capable will have no ECN logic, so it will only understand dropped packets as an indication of congestion.

Decapsulating a non-IP protocol frame: If the frame has a means of indicating ECN that is understood by the RBridge, it MUST follow the guidelines in Section 5.4 of [ECNencapGuide] when setting the ECN information in the decapsulated native frame. For a non-IP protocol with a similar ECN field to IP, this would be achieved by combining the information in the TRILL Header Flags Word with the encapsulated non-IP native frame, as specified in Table 3.

+------------+-----+---------------------+
| TRILL-ECN  | CCE | Arriving TRILL 3-bit|
+------------+-----+---------------------+
<table>
<thead>
<tr>
<th>Inner</th>
<th>Arriving TRILL 3-bit ECN Codepoint Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Header</td>
<td>Not-ECT</td>
</tr>
<tr>
<td></td>
<td>Not-ECT</td>
</tr>
<tr>
<td></td>
<td>ECN</td>
</tr>
<tr>
<td></td>
<td>CE</td>
</tr>
</tbody>
</table>

Table 3. Egress ECN Behavior

An asterisk in the above table indicates a combination that is currently unused in all variants of ECN marking (see Section 4) and therefore SHOULD be logged.

With one exception, the mappings in Table 3 are consistent with those for IP-in-IP tunnels [RFC6040], which ensures backward compatibility with all current and past variants of ECN marking (see Section 4). It also ensures forward compatibility with any future form of ECN marking that complies with the guidelines in [ECNencapGuide], including cases where ECT(1) represents a second level of marking severity below CE.

The one exception is that the drop condition in Table 3 need not be
logged because, with TRILL, it is the result of a valid combination of events.

4. TRILL Support for ECN Variants

This section is informative, not normative; it discusses interworking between TRILL and variants of the standardized form of ECN in IP [RFC3168]. See also [RFC8311].

The ECN wire protocol for TRILL (Section 2) and the ingress (Section 3.1) and egress (Section 3.3) ECN behaviors have been designed to support the other known variants of ECN, as detailed below. New variants of ECN will have to comply with the guidelines for defining alternative ECN semantics [RFC4774]. It is expected that the TRILL ECN wire protocol is generic enough to support such potential future variants.

4.1 Pre-Congestion Notification (PCN)
The PCN wire protocol [RFC6660] is recognized by the use of a PCN-compatible Diffserv codepoint in the IP header and a non-zero IP-ECN field. For TRILL or any lower layer protocol, equivalent traffic classification codepoints would have to be defined, but that is outside the scope of the current document.

The PCN wire protocol is similar to ECN, except it indicates congestion with two levels of severity. It uses:

- 11 (CE) as the most severe, termed the Excess-traffic-marked (ETM) codepoint
- 01 ECT(1) as a lesser severity level, termed the Threshold-Marked (ThM) codepoint. (This difference between ECT(1) and ECT(0) only applies to PCN, not to the classic ECN support specified for TRILL in this document before Section 4.)

To implement PCN on a transit RBridge would require a detailed specification. But in brief:

- the TRILL Critical Congestion Experienced (CCE) flag would be used for the Excess-Traffic-Marked (ETM) codepoint;
- ECT(1) in the TRILL-ECN field would be used for the Threshold-Marked codepoint.

Then the ingress and egress behaviors defined in Section 3 would not need to be altered to ensure support for PCN as well as ECN.

4.2 Low Latency, Low Loss, Scalable Throughput (L4S)

L4S is currently on the IETF's experimental track. An outline of how a transit TRILL RBridge would support L4S [ECNL4S] is given in Appendix A.
5. IANA Considerations
IANA is requested to update the TRILL Extended Header Flags registry by replacing the lines for bits 9-13 and for bits 21-26 with the following:

<table>
<thead>
<tr>
<th>Bits</th>
<th>Purpose</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-11</td>
<td>available non-critical hop-by-hop flags</td>
<td>[this doc]</td>
</tr>
<tr>
<td>12-13</td>
<td>TRILL-ECN (Explicit Congestion Notification)</td>
<td>[this doc]</td>
</tr>
<tr>
<td>21-25</td>
<td>available critical ingress-to-egress flags</td>
<td>[this doc]</td>
</tr>
<tr>
<td>26</td>
<td>Critical Congestion Experienced (CCE)</td>
<td>[this doc]</td>
</tr>
</tbody>
</table>

D. Eastlake & B. Briscoe

[Page 13]
6. Security Considerations

TRILL support of ECN is a straightforward combination of previously specified ECN and TRILL with no significant new security considerations.

For ECN tunneling security considerations, see [RFC6040].

For general TRILL protocol security considerations, see [RFC6325].

7. Acknowledgements

The helpful comments of Loa Andersson and Adam Roach are hereby acknowledged.

This document was prepared with basic NROFF. All macros used were defined in the source file.
Normative References


Informative References


[IANAthFlags] - IANA TRILL Extended Header word flags: http://www.iana.org/assignments/trill-parameters/trill-parameters.xhtml#extended-header-flags


Appendix A. TRILL Transit RBridge Behavior to Support L4S

The specification of the Low Latency, Low Loss, Scalable throughput (L4S) wire protocol for IP is given in [ECNL4S]. L4S is one example of the ways TRILL ECN handling may evolve [RFC8311]. It is similar to the original ECN wire protocol for IP [RFC3168], except:

- An AQM that supports L4S classifies packets with ECT(1) or CE in the IP header into an L4S queue and a "Classic" queue otherwise.

- The meaning of CE markings applied by an L4S queue is not the same as the meaning of a drop by a "Classic" queue (contrary to the original requirement for ECN [RFC3168]). Instead, the likelihood that the Classic queue drops packets is defined as the square of the likelihood that the L4S queue marks packets (e.g., when there is a drop probability of 0.0009 (0.09%) the L4S marking probability will be 0.03 (3%)).

This seems to present a problem for the way that a transit TRILL RBridge defers the choice between marking and dropping to the egress. Nonetheless, the following pseudocode outlines how a transit TRILL RBridge can implement L4S marking in such a way that the egress
behavior already described in Section 3.3 for Classic ECN [RFC3168] will produce the desired outcome.

/* p is an internal variable calculated by any L4S AQM
 * dependent on the delay being experienced in the Classic queue.
 * bit13 is the least significant bit of the TRILL-ECN field */

% On TRILL transit
if (bit13 == 0 ) {
    % Classic Queue
    if (p > max(random(), random()) )
        mark(CCE)                         % likelihood: p^2
} else {
    % L4S Queue
    if (p > random() ) {
        if (p > random() )
            mark(CCE)                      % likelihood: p^2
        else
            mark(NCCE)                     % likelihood: p - p^2
    }
}

With the above transit behavior, an egress that supports ECN (Section 3.3) will drop packets or propagate their ECN markings depending on whether the arriving inner header is from a non-ECN-capable or ECN-capable transport.

D. Eastlake & B. Briscoe

Internet-Draft TRILL ECN Support

Even if an egress has no L4S-specific logic of its own, it will drop packets with the square of the probability that an egress would if it did support ECN, for the following reasons:

- Egress with ECN support:
  + L4S: propagates both the Critical and Non-Critical CE marks (CCE & NCCE) as a CE mark.
    Likelihood: p^2 + p - p^2 = p
  + Classic: Propagates CCE marks as CE or drop, depending on inner.
Likelihood: $p^2$

o Egress without ECN support:

  + L4S: does not propagate NCCE as a CE mark, but drops CCE marks.
    Likelihood: $p^2$
  
  + Classic: drops CCE marks.
    Likelihood: $p^2$
Copyright and IPR Provisions

Copyright (c) 2018 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License. The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions. For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.