Internet Engineering Task Force (IETF)
Request for Comments: 7837
Category: Experimental
ISSN: 2070-1721

S. Krishnan
Ericsson
M. Kuehlewind
ETH Zurich
B. Briscoe
Simula Research Laboratory
C. Ralli
Telefonica

May 2016

IPv6 Destination Option for Congestion Exposure (ConEx)

Abstract

Congestion Exposure (ConEx) is a mechanism by which senders inform the network about the congestion encountered by packets earlier in the same flow. This document specifies an IPv6 destination option that is capable of carrying ConEx markings in IPv6 datagrams.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Congestion Exposure (ConEx) [RFC7713] is a mechanism by which senders inform the network about the congestion encountered by packets earlier in the same flow. This document specifies an IPv6 destination option [RFC2460] that can be used for performing ConEx markings in IPv6 datagrams.

This document specifies the ConEx wire protocol in IPv6. The ConEx information can be used by any network element on the path to, for example, do traffic management or egress policing. Additionally, this information will potentially be used by an audit function that checks the integrity of the sender's signaling. Further, each transport protocol that supports ConEx signaling will need to precisely specify when the transport sets ConEx markings (e.g., the behavior for TCP is specified in [RFC7786]).

This document specifies ConEx for IPv6 only. Due to space limitations in the IPv4 header and the risk of options that might be stripped by a middlebox in IPv4, the primary goal of the working group was to specify ConEx in IPv6 for experimentation.

This specification is experimental to allow the IETF to assess whether the decision to implement the ConEx Signal as a destination option fulfills the requirements stated in this document, as well as to evaluate the proposed encoding of the ConEx Signals as described in [RFC7713].

The duration of this experiment is expected to be no less than two years from publication of this document as infrastructure is needed to be set up to determine the outcome of this experiment. Experimenting with ConEx requires IPv6 traffic. Even though the amount of IPv6 traffic is growing, the traffic mix carried over IPv6 is still very different than over IPv4. Therefore, it might take longer to find a suitable test scenario where only IPv6 traffic is managed using ConEx.

2. Conventions Used in This Document

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].

3. Requirements for the Coding of ConEx in IPv6

A set of requirements for an ideal concrete ConEx wire protocol is given in [RFC7713]. The ConEx working group recognized that it will be difficult to find an encoding in IPv6 that satisfies all requirements. The choice in this document to implement the ConEx information in a destination option aims to satisfy those requirements that constrain the placement of ConEx information:

R-1: The marking mechanism needs to be visible to all ConEx-capable nodes on the path.

R-2: The mechanism needs to be able to traverse nodes that do not understand the markings. This is required to ensure that ConEx can be incrementally deployed over the Internet.

R-3: The presence of the marking mechanism should not significantly alter the processing of the packet. This is required to ensure that ConEx-Marked packets do not face any undue delays or drops due to a badly chosen mechanism.

R-4: The markings should be immutable once set by the sender. At the very least, any tampering should be detectable.

Based on these requirements, four solutions to implement the ConEx information in the IPv6 header have been investigated: hop-by-hop options, destination options, using IPv6 header bits (from the flow label), and new extension headers. After evaluating the different solutions, the ConEx working group concluded that the use of a destination option would best address these requirements.

Hop-by-hop options would have been the best solution for carrying ConEx markings if they had met requirement R-3. There is currently some work ongoing in the 6MAN working group to address this very issue [HBH-HEADER]. This new behavior would address R-3 and would make hop-by-hop options the preferred solution for carrying ConEx markings.

Choosing to use a destination option does not necessarily satisfy the requirement for on-path visibility, because it can be encapsulated by additional IP header(s). Therefore, ConEx-aware network devices, including policy or audit devices, might have to follow the chaining (extension-) headers into inner IP headers to find ConEx information. This choice was a compromise between fast-path performance of ConEx-aware network nodes and visibility, as discussed in Section 5.

Please note that the IPv6 specification [RFC2460] does not require or expect intermediate nodes to inspect destination options such as the

ConEx Destination Option (CDO). This implies that ConEx-aware intermediate nodes following this specification need updated extension header processing code to be able read the destination options.

4. ConEx Destination Option (CDO)

The CDO is a destination option that can be included in IPv6 datagrams that are sent by ConEx-aware senders in order to inform ConEx-aware nodes on the path about the congestion encountered by packets earlier in the same flow or the expected risk of encountering congestion in the future. The CDO does not have any alignment requirements.

```
 0                   1                   2
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Option Type  | Option Length |X|L|E|C|  res  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: ConEx Destination Option Layout

Option Type

8-bit identifier of the type of option. Set to the value 30 (0x1E) allocated for experimental work.

Option Length

8-bit unsigned integer. The length of the option in octets (excluding the Option Type and Option Length fields). Set to the value 1.

X Bit

When this bit is set, the transport sender is using ConEx with this packet. If it is not set, the sender is not using ConEx with this packet.

L Bit

When this bit is set, the transport sender has experienced a loss.

E Bit

When this bit is set, the transport sender has experienced congestion signaled using Explicit Congestion Notification (ECN) [RFC3168].

C Bit

When this bit is set, the transport sender is building up congestion credit in the audit function.

Reserved (res)

These four bits are not used in the current specification. They are set to zero by the sender and are ignored by the receiver.

All packets sent over a ConEx-capable TCP connection or belonging to the same ConEx-capable flow MUST carry the CDO. The chg bit (the third-highest-order bit) in the CDO Option Type field is set to zero, meaning that the CDO option is immutable. Network devices with ConEx-aware functions read the flags, but all network devices MUST forward the CDO unaltered.

The CDO SHOULD be placed as the first option in the Destination Option header before the AH [RFC4302] and/or Encapsulating Security Payload (ESP) [RFC4303] (if present). The IPsec Authentication Header (AH) MAY be used to verify that the CDO has not been modified.

If the X bit is zero, all the other three bits are undefined and thus MUST be ignored and forwarded unchanged by network nodes. The X bit set to zero means that the connection is ConEx-capable but that this packet MUST NOT be counted when determining ConEx information in an audit function. This can be the case if no congestion feedback is (currently) available, e.g., in TCP if one endpoint has been receiving data but sending nothing but pure ACKs (no user data) for some time. This is because pure ACKs do not advance the sequence number, so the TCP endpoint receiving them cannot reliably tell whether any have been lost due to congestion. Pure TCP ACKs cannot be ECN-marked either [RFC3168].

If the X bit is set, any of the other three bits (L, E, or C) might be set. Whenever one of these bits is set, the number of bytes carried by this IP packet (including the IP header that directly encapsulates the CDO and everything that IP header encapsulates) SHOULD be counted to determine congestion or credit information. In IPv6, the number of bytes can easily be calculated by adding the number 40 (length of the IPv6 header in bytes) to the value present in the Payload Length field in the IPv6 header.

The credit signal represents potential for congestion. If a congestion event occurs, a corresponding amount of credit is consumed as outlined in [RFC7713]. A ConEx-enabled sender SHOULD, therefore, signal sufficient credit in advance of any congestion event to cover the (estimated maximum) amount of lost or CE-marked bytes that could occur in such a congestion event. This estimation depends on the heuristics used and aggressiveness of the sender when deciding the appropriate sending rate (congestion control). Note that the maximum congestion risk is that all packets in flight get lost or CE-marked; therefore, this would be the most conservative estimation for the congestion risk. After a congestion event, if the sender intends to take the same risk again, it just needs to replace the consumed credit as non-consumed credit does not expire. For the case of TCP, this is described in detail in [RFC7786].

If the L or E bit is set, a congestion signal in the form of a loss or an ECN mark, respectively, was previously experienced by the same connection.

In principle, all of these three bits (L, E, or C) might be set in the same packet. In this case, the packet size MUST be counted once for each respective ConEx information counter.

If a network node extracts the ConEx information from a connection, it is expected to hold this information in bytes, e.g., comparing the total number of bytes sent with the number of bytes sent with ConEx congestion marks (L or E) to determine the current whole path congestion level. Therefore, a ConEx-aware node that processes the CDO MUST use the Payload Length field of the preceding IPv6 header for byte-based counting. When a ratio is measured and equally sized packets can be assumed, counting the number of packets (instead of the number of bytes) should deliver the same result. But an audit function must be aware that this estimation can be quite wrong if, for example, different sized packed are sent; thus, it is not reliable.

All remaining bits in the CDO are reserved for future use (which are currently the last four bits of the eight bit option space). A ConEx sender SHOULD set the reserved bits in the CDO to zero. Other nodes MUST ignore these bits and ConEx-aware intermediate nodes MUST forward them unchanged, whatever their values. They MAY log the presence of a non-zero Reserved field.

The CDO is only applicable on unicast or anycast packets (for reasoning, see the note regarding item J on multicast at the end of Section 3.3 of [RFC7713]). A ConEx sender MUST NOT send a packet with the CDO to a multicast address. ConEx-capable network nodes MUST treat a multicast packet with the X flag set the same as an equivalent packet without the CDO, and they SHOULD forward it unchanged.

As stated in [RFC7713] (see Section 3.3, item N on network-layer requirements), protocol specs should describe any warning or error messages relevant to the encoding. There are no warnings or error messages associated with the CDO.

5. Implementation in the Fast Path of ConEx-Aware Routers

The ConEx information is being encoded into a destination option so that it does not impact forwarding performance in the non-ConEx-aware nodes on the path. Since destination options are not usually processed by routers, the existence of the CDO does not affect the fast-path processing of the datagram on non-ConEx-aware routers, i.e., they are not pushed into the slow path towards the control plane for exception processing.

ConEx-aware nodes still need to process the CDO without severely affecting forwarding. For this to be possible, the ConEx-aware routers need to quickly ascertain the presence of the CDO and process the option if it is present. To efficiently perform this, the CDO needs to be placed in a fairly deterministic location. In order to facilitate forwarding on ConEx-aware routers, ConEx-aware senders that send IPv6 datagrams with the CDO SHOULD place the CDO as the first destination option in the Destination Option header.

6. Tunnel Processing

As with any destination option, an ingress tunnel endpoint will not normally copy the CDO when adding an encapsulating outer IP header. In general, an ingress tunnel SHOULD NOT copy the CDO to the outer header as this would change the number of bytes that would be counted. However, it MAY copy the CDO to the outer header in order to facilitate visibility by subsequent on-path ConEx functions if the configuration of the tunnel ingress and the ConEx nodes is coordinated. This trades off the performance of ConEx functions against that of tunnel processing.

An egress tunnel endpoint SHOULD ignore any CDO in the outer header on decapsulation of an outer IP header. The information in any inner CDO will always be considered correct, even if it differs from any outer CDO. Therefore, the decapsulator can strip the outer CDO without comparison to the inner. A decapsulator MAY compare the two and MAY log any case where they differ. However, the packet MUST be forwarded irrespective of any such anomaly, given an outer CDO is only a performance optimization.

A network node that assesses ConEx information SHOULD search for encapsulated IP headers until a CDO is found. At any specific network location, the maximum necessary depth of search is likely to be the same for all packets between a given set of tunnel endpoints.

7. Compatibility with Use of IPsec

A network-based attacker could alter ConEx information to fool an audit function in a downstream network into discarding packets. If the endpoints are using the IPsec Authentication Header (AH) [RFC2460] to detect alteration of IP headers along the path, AH will also detect alteration of the CDO header. Nonetheless, AH protection will rarely need to be introduced for ConEx, because attacks by one network on another are rare if they are traceable. Other known attacks from one network on another, such as TTL expiry attacks, are more damaging to the innocent network (because the ConEx audit discards silently) and less traceable (because TTL is meant to change, whereas CDO is not).

Section 4 specifies that the CDO is placed in the Destination Option header before the AH and/or ESP headers so that ConEx information remains in the clear if ESP is being used to encrypt other transmitted information in transport mode [RFC4301]. In general, a Destination Option header inside an IPv6 packet can be placed in two possible positions, either before the Routing header or after the ESP/AH headers as described in Section 4.1 of [RFC2460]. If the CDO was placed in the latter position and an ESP header was used with encryption, ConEx-aware intermediate nodes would not be able to view and interpret the CDO, effectively rendering it useless.

The IPv6 protocol architecture currently does not provide a mechanism for new headers to be copied to the outer IP header. Therefore, if IPsec encryption is used in tunnel mode, ConEx information cannot be accessed over the extent of the ESP tunnel.

The destination IP stack will not usually process the CDO; therefore, the sender can send a CDO without checking if the receiver will understand it. The CDO MUST still be forwarded to the destination IP stack, because the destination might check the integrity of the whole packet, irrespective of whether it understands ConEx.

8. Mitigating Flooding Attacks by Using Preferential Drop

The ideas in this section are aspirational, not being essential to the use of ConEx for more general traffic management. However, once CDO information is present, the CDO header could optionally also be used in the data plane of any IP-aware forwarding node to mitigate flooding attacks.

Please note that ConEx is an experimental protocol and that any kind of mechanism that reacts to information provided by the ConEx protocol needs to be evaluated in experimentation as well. This is also true, or especially true, for the preferential drop mechanism described below.

Dropping packets preferentially that are not ConEx-capable or do not carry a ConEx mark can be beneficial to mitigate flooding attacks as ConEx-Marked packets can be assumed to be already restricted by a ConEx ingress policer as further described in [RFC7713]. Therefore, the following ConEx-based preferential dropping scheme is proposed:

If a router queue experiences a very high load so that it has to drop arriving packets, it MAY preferentially drop packets within the same DiffServ Per-Hop Behavior (PHB) using the preference order given in Table 1 (1 means drop first). Additionally, if a router implements preferential drop based on ConEx, it SHOULD also support ECN marking. Even though preferential dropping can be difficult to implement on some hardware, if nowhere else, routers at the egress of a network SHOULD implement preferential drop based on ConEx markings (stronger than the MAY above).

	Preference
Not-ConEx or no CDO	1 (drop first)
X (but not L,E or C)	2
X and L,E or C	3

Table 1: Drop Preference for ConEx Packets

A flooding attack is inherently about congestion of a resource. As load focuses on a victim, upstream queues grow, requiring honest sources to pre-load packets with a higher fraction of ConEx marks.

If ECN marking is supported by downstream queues, preferential dropping provides the most benefits because, if the queue is so congested that it drops traffic, it will be CE-marking 100% of any forwarded traffic. Honest sources will therefore be sending 100% ConEx E-marked packets (and subject to rate-limiting at an ingress policer).

Senders under malicious control can either do the same as honest sources and be rate-limited at ingress, or they can understate congestion and not set the E bit.

If the preferential drop ranking is implemented on queues, these queues will reserve E/L-marked traffic until last. So, the traffic from malicious sources will all be automatically dropped first. Either way, malicious sources cannot send more than honest sources.

Therefore, ConEx-based preferential dropping as described above discriminates against attack traffic if done as part of the overall policing framework as described in [RFC7713].

9. Security Considerations

[RFC7713] describes the overall audit framework for assuring that ConEx markings truly reflect actual path congestion and [CONEX-AUDIT] provides further details on the handling of audit signals. This section focuses purely on the security of the encoding chosen for ConEx markings.

The CDO Option Type is defined with a chg bit set to zero as described in Section 4. If IPsec AH is used, a zero chg bit causes AH to cover the CDO option so that its end-to-end integrity can be verified, as explained in Section 4.

This document specifies that the Reserved field in the CDO must be ignored and forwarded unchanged even if it does not contain all zeroes. The Reserved field is also required to sit outside the Encapsulating Security Payload (ESP), at least in transport mode (see Section 7). This allows the sender to use the Reserved field as a 4-bit-per-packet covert channel to send information to an on-path node outside the control of IPsec. However, a covert channel is only a concern if it can circumvent IPsec in tunnel mode and, in the tunnel mode case, ESP would close the covert channel as outlined in Section 7.

10. IANA Considerations

The IPv6 ConEx destination option is used for carrying ConEx markings. This document uses the experimental option type 0x1E (as assigned in IANA's "Destination Options and Hop-by-Hop Options" registry) with the act bits set to 00 and the chg bit set to 0 for realizing this option. No further allocation action is required from IANA at this time.

11. References

11.1. 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, <http://www.rfc-editor.org/info/rfc2119>.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, DOI 10.17487/RFC2460, December 1998, <http://www.rfc-editor.org/info/rfc2460>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <http://www.rfc-editor.org/info/rfc3168>.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <http://www.rfc-editor.org/info/rfc4301>.

[RFC4302] Kent, S., "IP Authentication Header", RFC 4302, DOI 10.17487/RFC4302, December 2005, <http://www.rfc-editor.org/info/rfc4302>.

[RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <http://www.rfc-editor.org/info/rfc4303>.

[RFC7713] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts, Abstract Mechanism, and Requirements", RFC 7713, DOI 10.17487/RFC7713, December 2015, <http://www.rfc-editor.org/info/rfc7713>.

11.2. Informative References

[CONEX-AUDIT] Wagner, D. and M. Kuehlewind, "Auditing of Congestion Exposure (ConEx) signals", Work in Progress, draft-wagner-conex-audit-02, April 2016.

[HBH-HEADER] Baker, F., "IPv6 Hop-by-Hop Options Extension Header", Work in Progress, draft-ietf-6man-hbh-header-handling-03, Marcy 2016.

[RFC7786] Kuehlewind, M., Ed. and R. Scheffenegger, "TCP Modifications for Congestion Exposure (ConEx)", RFC 7786, DOI 10.17487/RFC7786, May 2016, <http://www.rfc-editor.org/info/rfc7786>.

Acknowledgements

The authors would like to thank David Wagner, Marcelo Bagnulo, Ingemar Johansson, Joel Halpern, John Leslie, Martin Stiemerling, Robert Sparks, Ron Bonica, Brian Haberman, Kathleen Moriarty, Bob Hinden, Ole Troan, and Brian Carpenter for the discussions that made this document better.

Authors' Addresses

Suresh Krishnan
Ericsson
8400 Blvd Decarie
Town of Mount Royal, Quebec
Canada

Email: suresh.krishnan@ericsson.com

Mirja Kuehlewind
ETH Zurich

Email: mirja.kuehlewind@tik.ee.ethz.ch

Bob Briscoe
Simula Research Laboratory

Email: ietf@bobbriscoe.net
URI: http://bobbriscoe.net/

Carlos Ralli Ucendo
Telefonica

Email: ralli@tid.es