ConEx Working Group
Internet-Draft
Intended status: Experimental
Expires: April 17, 2015

S. Krishnan
Ericsson
M. Kuehlewind
ETH Zurich
C. Ralli
Telefonica

October 14, 2014

IPv6 Destination Option for ConEx
draft-ietf-conex-destopt-07

Abstract

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

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on April 17, 2015.

Table of Contents

1. Introduction 2
2. Conventions used in this document 3
3. Requirements for the coding of ConEx in IPv6 3
4. ConEx Destination Option (CDO) 4
5. Implementation in the fast path of ConEx-aware routers . . . 6
6. Tunnel Processing 7
7. Compatibility with use of IPsec 7
8. Mitigating flooding attacks by using preferential drop . . . 8
9. Acknowledgements 9
10. Security Considerations 9
11. IANA Considerations 9
12. References 10
12.1. Normative References 10
12.2. Informative References 10
Authors' Addresses 10

1. Introduction

ConEx [I-D.ietf-ConEx-abstract-mech] 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. The ConEx information can be used by any network element on the path to e.g. 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 specify precisely when the transport sets ConEx markings (e.g. the behavior for TCP is specified in [ID.conex-tcp-modifications]).

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 [I-D.ietf-ConEx-abstract-mech].

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. Given ConEx is only chartered for IPv6, 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 requirement for an ideal concrete ConEx wire protocol is given in [I-D.ietf-ConEx-abstract-mech]. In the ConEx working group is was 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.

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 bury 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 Section 5.

4. ConEx Destination Option (CDO)

The ConEx Destination Option (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 has an alignment requirement of (none).

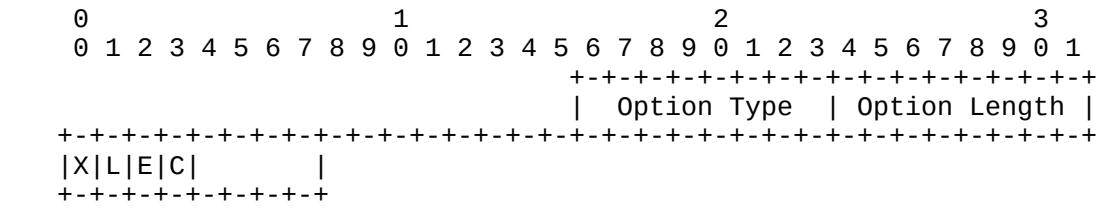

Figure 1: ConEx Destination Option Layout

Option Type

8-bit identifier of the type of option. The option identifier for the ConEx destination option will be allocated by the IANA.

Option Length

8-bit unsigned integer. The length of the option (excluding the Option Type and Option Length fields). This field MUST be 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 ECN-signaled congestion.

C Bit

When this bit is set, the transport sender is building up

congestion credit in the audit function.

Reserved

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

All packets sent over a ConEx-capable connection MUST carry the CDO. The CDO is immutable. Network devices with ConEx-aware functions read the flags, but all network devices MUST forward the CDO unaltered.

CDO MUST be placed as the first option in the destination option header before the AH and/or ESP (if present). IPsec Authentication Header (AH) MAY be used to verify that the CDO has not been modified.

If the X bit is zero all other three bits are undefined and thus should be ignored and forwarded unchanged by network nodes. The X bit set to zero means that the connection is ConEx-capable but 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, C) MAY 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.

A transport sends credits prior to the occurrence of congestion (loss or ECN-CE marks) and the amount of credits should cover the congestion risk. Note, the maximum congestion risk is that all packets in flight get lost or ECN marked.

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

In principle all of these three bits (L, E, C) MAY be set in the same packet. In this case the packet size MUST be accounted more than 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, E) to determine the current whole path congestion level. For ConEx-aware node processing, the CDO MUST use the Payload length field of the preceding IPv6 header for byte-based accounting. 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 a network node must be aware that this estimation can be quite wrong, if e.g. different sized packed are sent and thus it is not reliable.

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.

It might be possible to implement a proxy for a ConEx sender, as long as it is located where receiver feedback is always visible. A ConEx proxy MUST NOT introduce a CDO header into a packet already carrying one and it MUST NOT alter the information in any existing CDO header. However, it can add a CDO header to any packets without one, taking care not to disrupt any integrity or authentication mechanisms.

The CDO is only applicable on unicast or anycast packets (see [I-D.ietf-ConEx-abstract-mech] for reasoning). 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, but they SHOULD forward it unchanged.

There are no warning 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.

The 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 MUST place the CDO as the first destination option in the destination options header.

6. Tunnel Processing

As with any destination option, an ingress tunnel endpoint will not natively 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 changed the number of bytes that would be counted. However, it MAY copy the CDO to the outer in order to facilitate visibility by subsequent on-path ConEx functions if the configuration of the tunnel ingress and the ConEx nodes is co-ordinated. This trades off the performance of ConEx functions against that of tunnel processing.

An egress tunnel endpoint SHOULD ignore any CDO 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.

7. Compatibility with use of IPsec

If the transport network cannot be trusted, IPsec Authentication should be used to ensure integrity of the ConEx information. If an attacker would be able to remove the ConEx marks, this could cause an audit device to penalize the respective connection, while the sender cannot easily detect that ConEx information is missing.

In IPv6 a Destination Option header can be placed in two possible position in the order of possible headers, either before the Routing header or after the Encapsulating Security Payload (ESP) header [RFC2460]. As the CDO is placed in the destination option header before the AH and/or ESP, it is not encrypted in transport mode

[RFC4301]. Otherwise, if the CDO were placed in the latter position and an ESP header were used, the CDO would also be encrypted and could not be interpreted by ConEx-aware devices.

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.

8. Mitigating flooding attacks by using preferential drop

This section is aspirational, and not critical 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.

If a router queue experiences very high load so that it has to drop arriving packets, it MAY preferentially drop packets within the same Diffserv 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. Preferential dropping can be difficult to implement on some hardware, but if feasible it would discriminate against attack traffic if done as part of the overall policing framework as described in [I-D.ietf-ConEx-abstract-mech]. 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 preserve 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.

9. Acknowledgements

The authors would like to thank Marcelo Bagnulo, Bob Briscoe, Ingemar Johansson, Joel Halpern and John Leslie for the discussions that led to this document.

Special thanks to Bob Briscoe who contributed text and analysis work on preferential dropping.

10. Security Considerations

[I-D.ietf-ConEx-abstract-mech] describes the overall audit framework for assuring that ConEx markings truly reflect actual path congestion. This section focuses purely on the security of the encoding chosen for ConEx markings.

The chg bit in the CDO option type field is set to zero, meaning that the CDO option is immutable. 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 encrypting security payload (ESP), at least in transport mode (see Section 7). This allows the sender to use the Reserved field as a 28-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.

11. IANA Considerations

This document defines a new IPv6 ConEx destination option for carrying ConEx markings. IANA is requested to assign a new destination option type in the Destination Options registry maintained at http://www.iana.org/assignments/ipv6-parameters <TBA1> ConEx Destination Option [RFCXXXX] The act bits for this option need to be 00. 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.

12. References

12.1. Normative References

[I-D.ietf-ConEx-abstract-mech] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-ConEx-abstract-mech (work in progress), July 2011.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC4302] Kent, S., "IP Authentication Header", RFC 4302, December 2005.

[RFC6789] Briscoe, B., Woundy, R., and A. Cooper, "Congestion Exposure (ConEx) Concepts and Use Cases", RFC 6789, December 2012.

12.2. Informative References

[RFC2401] Kent, S. and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

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

Carlos Ralli Ucendo
Telefonica

Email: ralli@tid.es