Reviewed by Bob Briscoe, 11 Aug 2014
Technical or normative points highlighted in yellow.

```
ConEx Working Group                                          S. Krishnan
Internet-Draft                                                  Ericsson
Intended status: Experimental                             M. Kuehlewind
Expires: August 18, 2014                      IKR University of Stuttgart
                                                               C. Ucendo
                                                              Telefonica
                                                       February 14, 2014
```

Deleted: Standards Track

# IPv6 Destination Option for ConEx
draft-ietf-conex-destopt-06

## 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

## Copyright Notice

Table of Contents

## 1. Introduction

ConEx [CAM] 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 ConEx-capable transport protocols can use for performing ConEx marking in IPv6 datagrams.

**Deleted:** be¶

**Deleted:** d

**Deleted:** s

This document solely specifies the ConEx wire protocol for use by any end-to-end transport protocol. Each transport protocol will need an update to specify precisely when a transport sets the various ConEx markings (e.g. the behaviour for TCP is specified in [ID.conex-tcp-modifications]).

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.

### 1.1 Experiment Goals

Initially ConEx is specified experimentally so that the IETF can assess whether the compromises necessary in its design where the right ones.

Duration of Experiment: No less than 2 years from publication of this document will be needed to set up the infrastructure needed 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.

Criteria for Success: The protocol will be deemed successful if, in the opinion of XXX? {the IETF’s Transport Area Director?}, it satisfies the requirements in [CAM], allowing it to be audited and trusted for use by traffic management functions, with minimal impact on performance.

## 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

[CAM] describes the ConEx mechanism in abstract terms and sets requirements for an ideal concrete protocol, but recognizes that it will be difficult to satisfy every requirement. This specification satisfies most of those requirements and gives justifications where compromises have had to be made.
The requirements in [CAM} relevant to wire protocol design are paraphrased below:

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.

R-5: The ConEx signals for packet loss and ECN marking SHOULD have distinct encodings.

R-6: Additionally there SHOULD be an auditable ConEx Credit signal.

**Deleted:** ¶

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 wg concluded that the use of a destination option would best address the requirements.

**Deleted:** only

**Deleted:** fulfil

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 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 and visibility, as discussed in 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 in the same flow. The CDO has an alignment requirement of (none).

**Comment [BB1]:** Not necessarily earlier (e.g. Credit). OK, not necessarily actually ‘encountered’ either, but...

**Deleted:** earlier

```
 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
                                +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                                |  Option Type  | Option Length |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|X|L|E|C|                       Reserved                        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

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 4.

X Bit

**Comment [BB2]:** Shouldn't these be called flags, not bits?

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.  ConEx-aware functions read the flags, but all network devices MUST forward the CDO field unaltered. IPSeC Authentication Header (AH) may be used to verify that the CDO has not been modified.

**Deleted:** N

**Deleted:** SHOULD

**Deleted:** only read

**Deleted:** flags

**Deleted:**    If the X bit is zero all other three bits are undefined and thus¶
   should be ignored.  The X bit set to zero means that the connection¶
   is ConEx-capable but this packet SHOULD NOT be accounted to determine¶
   ConEx information in an audit function.  This can be the case for¶
   e.g. pure control packets not carrying any user data.  As an example¶
   in TCP pure ACKs are usually not ECN-capable and TCP does not have an¶

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 immediately preceding the CDO) SHOULD be accounted for determining 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 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 being 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.

The CDO is only applicable on unicast or anycast packets (see [CAM] for reasoning). A ConEx sender MUST NOT send a packet with X = 1 (ConEx-capable) to a multicast address, and it SHOULD NOT even include the CDO in such a packet. ConEx-capable network nodes MUST treat a multicast packet with the X flag set the same as an equivalent packet without the CDO. They MAY log the anomaly.

If the X bit is zero all the other three bits are undefined and thus should be ignored by the receiver 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. ConEx-capable network nodes MUST treat a packet with the X flag cleared the same as an equivalent packet without the ConEx destination option. They MAY log the anomaly. ~~This can be the case for e.g. pure control packets not carrying any user data. As an example~~

~~in TCP pure ACKs are usually not ECN-capable and TCP does not have a mechanism to announce the loss of a pure ACK to the sender. Thus~~

**Deleted:** mechanism to announce the lost of a pure ACK to the sender. Thus¶ congestion information about ACKs are not available at the sender.¶ ¶

**Deleted:** all

**Deleted:** other

**Deleted:** C

**Deleted:** are sent previous

**Deleted:** ¶

**Deleted:** i

**Deleted:** the

**Comment [BB3]:** byte-wise is ambiguous. It sounds like it means in byte order, not just in bytes.

**Deleted:** this node

**Deleted:** usually supposed

**Deleted:** -wise

**Deleted:** ¶

**Deleted:** the ac

**Deleted:** of

**Deleted:** d

**Comment [BB4]:** Shifted para down.

**Deleted:** SHOULD

**Deleted:** ac

**Deleted:** to

**Deleted:** e¶

**Deleted:** in an audit function

**Deleted:** n

**Deleted:** t

~~congestion information about ACKs are not available at the sender.~~

A ConEx-capable network node MUST forward unchanged any CDO with length other than 4, and treat the packet the same as it would treat an equivalent packet without CDO. It MAY log the anomaly.

A ConEx sender MUST 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.

## 5. Implementation in the fast path of ConEx-aware routers

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

**Comment [BB5]:** The ConEx info isn't about the congestion info on the ACK itself; it's about the packet in the same direction as the pure ACK, but 1RTT earlier.

I suggest these sentences are removed (see email discussion).

The only need I can see for the X-flag is if the Reserved field gets used in future for something in addition to ConEx. Then there would be a need to identify packets that are not ConEx-capable but still carry the CDO option (for the new reason).

**Deleted:** SHOULD

**Deleted:** SHOULD not interpret

**Deleted:** The

**Deleted:** the

**Deleted:** the

**Deleted:** .

**Deleted:** T

**Deleted:** towards the control plane

**Deleted:** ¶

ConEx-aware nodes still need to process the CDO without severely affecting forwarding. For this to be possible, 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 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 options header. (This is not stated as a 'MUST', because some future destination option might need to be placed first for functional rather than just performance reasons.)

**Deleted:** The

**Deleted:** the

**Deleted:** fairly

**Deleted:** who

**Deleted:** MUST

## 6. Configuration and Management

**Comment [BB6]:** Required by abstract-mech

Once ConEx is implemented in a transport protocol it requires no configuration. For experimental purposes, it would be appropriate to be able to switch ConEx on or off at each sender, on a per-transport protocol basis.

There are no warning or error messages associated with the CDO.

## 7. Compatibility with use of IPsec

**Deleted:** 6

In IPsec transport mode no action needs to be taken as the CDO is visible to the network. When accounting for ConEx information, the size of the Authentication Header (AH) SHOULD NOT be accounted for as this information has been added later. In the IPsec Tunnel model the CDO SHOULD be copied to the outer IP header as this information is end-to-end. Only the payload of the outer IP header minus the AH SHOULD be counted.

**Comment [BB7]:** I think we ought to say AH MUST be accounted for, because it makes the rules as to what is included cleaner. Otherwise, what about other destination options added after CDO?

**Comment [BB8]:** About the whole section:

Visibility is only a problem if both ESP and tunnel mode are used. If ESP is used in tunnel mode, CDO MUST be copied to the outer first.

Also, in tunnel mode, CDO MAY be copied to the outer, which I have suggested as a more general optimisation under general tunnel processing (whether IPsec or non-IPsec).

**Comment [BB9]:** I don't agree with this exception, because you cannot tell that a packet is in tunnel mode just by looking at it, so a ConEx function won't know when to apply this exception.

IMO, we should keep to the single rule that the header size for ConEx includes the IP header immediately before the CDO (and any other headers between).

If the transport network cannot be trusted, 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.

**Deleted:** ac

**Deleted:**

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.

## 8. Tunnel Processing

**Comment [BB10]:** abstract-mech requires tunnel processing to be defined.

This is my proposal. But we could consider a simple "MUST NOT copy CDO" approach instead.

As with any destination option, an ingress tunnel endpoint SHOULD NOT copy the CDO when adding an encapsulating outer IP header. However, it MAY copy the CDO to the outer in order to facilitate visibility by subsequent on-path ConEx functions. 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 optimisation.

**Comment [BB11]:** Will a DO confuse some tunnel endpoints, perhaps causing them to discard such packets?

## 9. Mitigating flooding attacks by using preferential drop

**Deleted:** 7

**Deleted:** DDoS mitigation

This section is aspirational, and not critical to the use of ConEx for more general traffic management.

The only nodes other than the sender that need to parse the CDO are ConEx-aware policy or audit devices. 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 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 [RFC6789]. If nowhere else, routers at the egress of a network SHOULD implement preferential drop (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 (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 flag. 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.

**Deleted:** the

**Deleted:** the

**Deleted:** and therefore being

**Deleted:** ed

**Deleted:** the

## 8. 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.

## 9. Security Considerations

This document does not bring up any new security issues concerning the overall ConEx integrity and traffic management framework that have not already been discussed in [CAM] In particular, [CAM] introduces the audit framework necessary to ensure that a sender has to set the CDO flags to reflect actual congestion experienced.

All undefined and invalid combinations of fields, including reserved fields, have been identified in Section 4, and appropriate behaviours have been assigned.

Protocol interaction with IPsec is addressed in Section 6.

Section 8 on using ConEx against flooding attacks is aspirational. If this proves not to be effective, it does not invalidate the use of ConEx for more general traffic management.

## 10. 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 and the chg bit needs to be 0.

**Comment [BB12]:** Surely 10 would cause any node (including any receiver) not recognising the new option to discard the packet and send an ICMP parameter problem message?

Even if dest opts are only checked by the receiver, there is no need for a receiver to have to understand ConEx so we don't want the packet discarded just before it reaches the destination transport.

**Deleted:** 10

## 11. Normative References

[CAM] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-ConEx-abstract-mech-05 (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.

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

[ID.conex-tcp-modifications] draft-ietf-conex-tcp-modifications...

## Authors' Addresses

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

Email: suresh.krishnan@ericsson.com

Mirja Kuehlewind
IKR University of Stuttgart

Email: mirja.kuehlewind@ikr.uni-stuttgart.de

Carlos Ralli Ucendo
Telefonica

Email: ralli@tid.es