

conex  
Internet-Draft  
Intended status: Informational  
Expires: April 3, 2011

Lei. zhu  
Huawei Technologies  
September 30, 2010

Requirements of Localised Congestion Notification  
draft-lei-ecn-localised-congestion-notification-01

Abstract

This document introduces analyzes of ECN(Explicit Congestion Notification)in case of congestion of local links. 3GPP adopts and specifies shared channel for multiple user equipments in a cell. Other last mile access systems (e.g. xDSL access and Frame Relay access) have to handle congestion since bandwidth multiplexing is most likely considered for reducing unnecessary cost. Therefore, congestion in access network is to inform user equipments in case of traffic congestion utilizing congestion notification concepts and similar mechanisms.

Besides the argument on the congestions of access networks, this document is also to introduce the use case of Conex which might be useful for the potential Conex proposals to notify and address the congestion of wireless access network further.

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 3, 2011.

Copyright Notice

Copyright (c) 2010 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.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

## Table of Contents

|   |   |
|---|---|
| 1. Introduction . . . . .                     | 4 |
| 2. Terminology . . . . .                      | 5 |
| 3. Definition . . . . .                       | 5 |
| 4. Requirements . . . . .                     | 5 |
| 5. Use case . . . . .                         | 6 |
| 5.1. Access Congestion . . . . .              | 6 |
| 5.2. Generated Traffic from Network . . . . . | 7 |
| 6. Security Considerations . . . . .          | 7 |
| 7. IANA Considerations . . . . .              | 7 |
| 8. Acknowledgments . . . . .                  | 7 |
| 9. Normative References . . . . .             | 7 |
| Author's Address . . . . .                    | 8 |

## 1. Introduction

ECN [RFC3168], a proposed standard document, defines the ECN field in the IP header, and specifies the semantics for the code points for the ECN field. These are all unicast protocols which negotiate the use of ECN during the initial connection establishment handshake (supporting incremental deployment, and checking if ECN marked packets pass all middleboxes on the path).

Basically, ECN is an end to end solution involving endpoints and intermediate entities in the transport path. The ECN solution requires that terminals and intermediate nodes (e.g. router and firewall) in the path to support ECN functions. Otherwise, the whole solution does not fulfill the requirements to mechanisms of ECN. The ECN also updates the definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers with optional ECN fields. The ECN mechanism originally includes the associations with congestion control and avoidance algorithms of TCP [RFC0793]. The problem is that this end to end solution highly relies on the supports of intermediate node in the path. Since, there exist some middleboxes (firewalls, load balancers, or intrusion detection systems) in the Internet that either drop a TCP SYN packet configured to negotiate ECN, or respond with a RST.

Besides the congestion appears in backbones of Tier 1 or Tier 2 service providers, the most of traffic congestion seem happen in access network due to a lot of reasons. For example, the ADSL multiplexer is likely to multiplex traffic, but multiplexer would not avoid congestion in particular periods (e.g. rush hour of transportation system). Another example is 3GPP LTE architecture which specifies shared logic channel in particular cell with very high capacity. However, even though terminals use low class traffic (e.g the traffic class only insure best effort service.), the capacity of the shared channel could be exhausted by applications which generates huge number of traffic. In general, some problems could happy or have appeared in network, all indicate that the congestion of access network is likely critical issue to usages of congestion notifications in next generation networks.

Changes from previous drafts (to be removed by the RFC Editor):

From -00 to -01:

The revision 01 is to introduce some description of use case on the issues that the Conex proposal might be useful to notify congestion of wireless access network. The use case described in this document is also expected to be considered as an input of Conex mechanism discussion and to be address in future.

Argument in "Introduction and problem statement" part of 00 version could be kept in 01 version.

The definition and particular description of congestion issues of wireless access network might be supplied based on the draft-moncaster-conex-concepts-uses-01. Those supplements are expected to be adopting in proper working group document.

The use case describing Conex usage of wireless access network is added in the document.

A supplied use case describe needs to consider the congestion control for Conex proposal in case that traffic is generated from network.

## 2. Terminology

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. Definition

Access Congestion: Access Congestion is due to the multiplexing function of access network, and is the congestion situation of access network which link layer resource of access network is not sufficient to transport downstream or upstream data to the endpoints. Access Congestion would be located at network ingress and egress. Mobile broadband access networks may not use RED algorithm as the criteria to detect congestion.

## 4. Requirements

In this section, the author tries to introduce requirements of localized ECN mechanism.

The goals of localized ECN are:

- a. Provide a localized congestion notification mechanism extending the scope of ECN solution to notify congestion only happened in access network;
- b. The localized ECN mechanism should support notification of congestion at access network in spite of intermediate supports in the whole path;

c. The localized congestion notification mechanism should be compliant with the ECN solution, and support the back compatibility to ECN solution.

Therefore, the concluded requirements are,

the localized ECN mechanism:

- a. MUST be able to notify congestion of access network to the local senders which contributes to the uplink congestion.
- b. MUST be able to notify congestion of access network to the local receivers which receive data contributing to the downlink congestion.
- c. MUST be able to notify the uplink congestion and downlink congestion of access network to senders and receivers.

## 5. Use case

### 5.1. Access Congestion

Access Congestion happens in the network ingress and egress. Access Congestion is mainly due to the access network transmission capacity of transmission in a specific period of time can not satisfy the transportation needs of upstream or downstream traffic. In addition to the existing fixed access network, e.g. the different types of last mile access approaches, including xDSL, optical access, more legacy TDM, ISDN access networks, it also includes mobile broadband access network. Especially for the mobile broadband access networks, at air interface, radio resource controllers need to execute scheduling or priority handling functions to enable the flow of different users and applications in the empty transmission, with the congestion detection and discovery mechanisms different from AQM mechanisms. The Conex mechanism is to reflect the measurement of the mobile broadband network as the network ingress or egress of the conex enabled network, and might require the supports of link layer of such network.

The RLC layer retransmission by HARQ mechanism and timer in the data link layer of mobile broadband network ensure reliable data transmission services. From the implementation point of view, the access network entities, such as RNC (Radio Network Controller) may support the buffering mechanism which could be able to calculate the change in the length of the buffer entry and exit that can be estimated network congestion state. So, Conex is expected to make congestion event according to ECN RFC3168 in addition to other provisions of the RED algorithm to obtain the parameters of traffic

congestion. Mechanism in the Conex eventually hopes to influence the application of business data generated.

## 5.2. Generated Traffic from Network

The reasons generating more traffic congestion in the networks are complicated. But, end users get content from the network in most cases. The endpoints have rare opportunities to send huge number of traffic in real network. For example, HTTP based applications, service data are sent from the original server. In other word, the content is transferred to the end points request the traffic.

According to the ECN concepts, ECN parameter in IP packet header will be sent to receiver, while the ECT flag in the TCP ACK is sent to the sender. In the example of the HTTP based applications if the source can reduce the transmission of business data is a pretty well situation. On the other hand, Conex solution hopes controlling the amount of traffic generated in the state of network congestion through economic means. In these scenarios, existing ECN solution does not completely and surely satisfy, because traffic resulting in congestion indication is always transferred to the sender.

In the real networks, operators are able to monitor data flow and a user, or subscription of the user, and associate them with policy control functions. If congestion is expected to be controlled by economic means, Conex mechanism can be used to further support to the above scenarios.

## 6. Security Considerations

TBD

## 7. IANA Considerations

There have been no IANA considerations so far in this document.

## 8. Acknowledgments

TBD

## 9. Normative References

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

Author's Address

Lei Zhu  
Huawei Technologies  
Huawei Building, Xinxu Road No.3  
Haidian District, Beijing 100085  
P. R. China

Phone: +86-10-82836301  
Email: Lei.zhu@huawei.com





Congestion Exposure (ConEx)  
Working Group  
Internet-Draft  
Intended status: Informational  
Expires: April 22, 2011

M. Mathis  
Google  
B. Briscoe  
BT  
October 19, 2010

Congestion Exposure (ConEx) Concepts and Abstract Mechanism  
draft-mathis-conex-abstract-mech-00

Abstract

This document describes an abstract mechanism by which senders inform the network about the congestion encountered by packets earlier in the same flow. Today, the network may signal congestion to the receiver by ECN markings or by dropping packets, and the receiver may pass this information back to the sender in transport-layer feedback. The mechanism to be developed by the ConEx WG will enable the sender to also relay this congestion information back into the network in-band at the IP layer, such that the total level of congestion is visible to all IP devices along the path, from where it could, for example, be provided as input to traffic management.

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 22, 2011.

Copyright Notice

Copyright (c) 2010 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.

## Table of Contents

|  |    |
|--|----|
| 1. Introduction . . . . .                      | 3  |
| 1.1. Terminology . . . . .                     | 4  |
| 2. Requirements for the ConEx Signal . . . . . | 5  |
| 3. Representing Congestion Exposure . . . . .  | 7  |
| 3.1. Strawman Encoding . . . . .               | 7  |
| 3.2. ECN Based Encoding . . . . .              | 8  |
| 3.2.1. ECN Changes . . . . .                   | 8  |
| 3.3. Abstract Encoding . . . . .               | 9  |
| 3.3.1. Independent Bits . . . . .              | 9  |
| 3.3.2. Codepoint Encoding . . . . .            | 9  |
| 4. Congestion Exposure Components . . . . .    | 10 |
| 4.1. Modified Senders . . . . .                | 10 |
| 4.2. Receivers (Optionally Modified) . . . . . | 10 |
| 4.3. Audit . . . . .                           | 10 |
| 4.4. Policy Devices . . . . .                  | 11 |
| 4.4.1. Congestion Policers . . . . .           | 12 |
| 4.4.2. Other Policy Devices . . . . .          | 12 |
| 5. IANA Considerations . . . . .               | 12 |
| 6. Security Considerations . . . . .           | 12 |
| 7. Conclusions . . . . .                       | 13 |
| 8. Acknowledgements . . . . .                  | 13 |
| 9. Comments Solicited . . . . .                | 13 |
| 10. References . . . . .                       | 13 |
| 10.1. Normative References . . . . .           | 13 |
| 10.2. Informative References . . . . .         | 13 |

## 1. Introduction

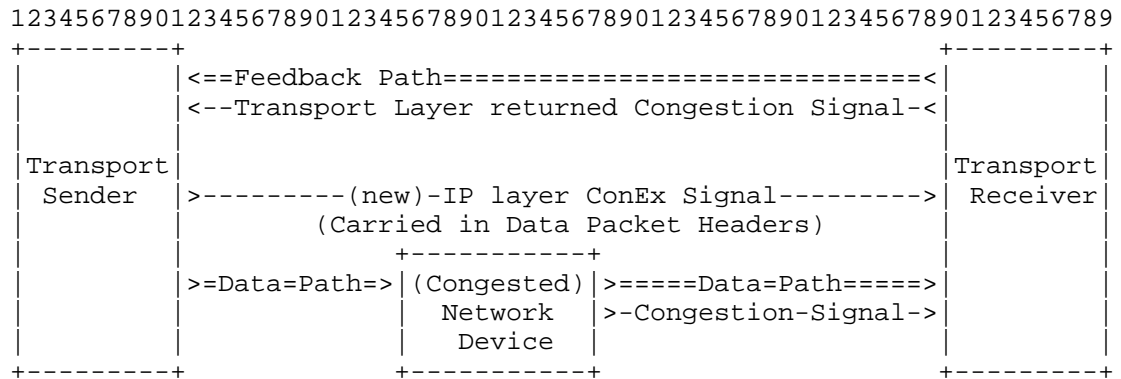
One of the required functions of a transport protocol is controlling congestion in the network. There are three techniques in use today for the network to signal congestion to a transport:

- o The most common congestion signal is packet loss. When congested, the network simply discards some packets either as part of an explicit control function [RFC2309] or as the consequence of a queue overflow or other resource starvation. The transport receiver detects that some data is missing and signals such through transport acknowledgments to the transport sender (e.g. TCP SACK options). The sender performs the appropriate congestion control rate reduction (e.g. [RFC5681] for TCP) and, if it is a reliable transport, it retransmits the missing data.
- o If the transport supports explicit congestion notification (ECN) [RFC3168] or pre-congestion notification (PCN) [RFC5670], the transport sender indicates this by setting an ECN-capable transport (ECT) codepoint in every packet. Network devices can then explicitly signal congestion to the receiver by setting ECN bits in the IP header of such packets. The transport receiver communicates these ECN signals back to the sender, which then performs the appropriate congestion control rate reduction.
- o Some experimental transport protocols and TCP variants [Vegas] sense queuing delays in the network and reduce their rate before the network has to signal congestion using loss or ECN. A purely delay-sensing transport will tend to be pushed out by other competing transports that do not back off until they have driven the queue into loss. Therefore, modern delay-sensing algorithms use delay in some combination with loss to signal congestion (e.g. LEDBAT [I-D.ietf-ledbat-congestion], Compound [I-D.sridharan-tcpm-ctcp]). In the rest of this document, we will confine the discussion to concrete signals of congestion such as loss and ECN. We will not discuss delay-sensing further, because it can only avoid these more concrete signals of congestion in some circumstances.

In all cases the congestion signals follow the route indicated in Figure 1. A congested network device sends a signal in the data stream on the forward path to the transport receiver, the receiver passes it back to the sender through transport level feedback, and the sender makes some congestion control adjustment.

This document proposes to extend the capabilities of the Internet protocol suite with the addition of a ConEx Signal that, to a first approximation, relays the congestion information from the transport

sender back through the internetwork layer. That signal is shown in Figure 1. It would be visible to all internetwork layer devices along the forward (data) path and is intended to support a number of new policy-controlled mechanisms that might be used to manage traffic.



Not shown are policy devices along the data path that observe the ConEx Signal, and use the information to monitor or manage traffic. These are discussed in Section 4.4.

Figure 1

### 1.1. Terminology

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 RFC 2119 [RFC2119].

ConEx signals in IP packet headers from the sender to the network  
 {ToDo: These are placeholders for whatever words we decide to use}:

Not-ConEx: The transport is not ConEx-capable

ConEx-Capable: The transport is ConEx-Capable. This is the opposite of Not-ConEx and implies one of the following signals

Re-Echo-Loss: (aka Purple) The transport has experienced a loss

Re-Echo-ECN: (aka Black) The transport has experienced an ECN mark

Credit: (aka Green) The transport is building up credit to allow for any future delay in expected ConEx signals

ConEx-Not-Marked: The transport is ConEx-capable but is signaling none of Re-Echo-Loss, Re-Echo-ECN or Credit

ConEx-Marked: At least one of Re-Echo-Loss, Re-Echo-ECN or Credit.

## 2. Requirements for the ConEx Signal

Ideally, all the following requirements would be met by a Congestion Exposure Signal. However it is already known that some compromises will be necessary, therefore all the requirements are expressed with the keyword 'SHOULD' rather than 'MUST'. The only mandatory requirement is that a concrete protocol description MUST give sound reasoning if it chooses not to meet any of these requirements:

- a. The ConEx Signal SHOULD be visible to internetwork layer devices along the entire path from the transport sender to the transport receiver. Equivalently, it SHOULD be present in the IPv4 or IPv6 header, and in the outermost IP header if using IP in IP tunneling. The ConEx Signal SHOULD be immutable once set by the transport sender. A corollary of these requirements is that existing (legacy) networking gear SHOULD pass the Congestion Exposure Signal silently without modification.
- b. The ConEx Signal SHOULD be useful under only partial deployment. A minimal deployment SHOULD only require changes to transport senders. Furthermore, partial deployment SHOULD create incentives for additional deployment, both in terms of enabling ConEx on more devices and adding richer features to existing devices. Nonetheless, ConEx deployment need never be universal, and it is anticipated that some hosts and some transports may never support the ConEx Protocol and some networks may never use the ConEx Signals.
- c. The ConEx Signal SHOULD be accurate. In potentially hostile environments such as the public Internet, it SHOULD be possible for techniques to be deployed to audit the Congestion Exposure Signal by comparing it to the actual congestion signals on the forward data path. The auditing mechanism must have a capability for providing sufficient disincentives against misreported congestion, such as by throttling traffic that reports less congestion than it is actually experiencing.
- d. The ConEx Signal SHOULD be timely. There will be a delay between the time when an auditing device sees an actual congestion signal

and when it sees the subsequent Congestion Exposure Signal from the sender. The minimum delay will be one round trip, but it may be much longer depending on the transport's choice of feedback delay (consider RTCP [RFC3550] for example). It is not practical to expect auditing devices in the network to make allowance for such feedback delays. Instead, the sender SHOULD be able to send ConEx signals in advance, as 'credit' for any audit device to hold as a balance against the risk of congestion during the feedback delay. This design choice simplifies auditing devices and correctly makes the transport responsible for both minimizing feedback delay and minimizing sharp increases in packets in flight that would risk causing excessive congestion to others. This issue is discussed in more detail in Section 4.3.

It is important to note that the auditing requirement implies a number of additional constraints: The basic auditing technique is to count both actual congestion signals and ConEx Signals someplace along the data path:

- o For congestion signaled by ECN, auditing is most accurate when located near the transport receiver. Within any flow or aggregate of flows, the total volume of ECN marked data seen near the receiver should always be equal to or less than the volume of data tagged with ConEx Signals.
- o For congestion signaled by loss, totally accurate auditing is not believed to be possible in the general case, because it involves a network node detecting the absence of some packets, when it cannot necessarily see the transport protocol sequence numbers and when the missing packets might simply be taking a different route. But there are common cases where sufficient audit accuracy should be possible:
  - \* For non-IPsec traffic conforming to standard TCP sequence numbering on a single path, an auditor could detect losses by observing both the original transmission and the retransmission after the loss. Such auditing would be most accurate near the sender.
  - \* For networks designed so that losses predominantly occur under the management of one IP-aware node on the path, the auditor could be located at this bottleneck. It could simply compare ConEx Signals with actual local losses. This is a good model for most consumer access networks and audit accuracy could well be sufficient even if losses occasionally occurred at other nodes in the network, such as border gateways (see Section 4.3 for details).

Given that loss-based and ECN-based ConEx might sometimes be best audited at different locations, having distinct encodings would widen the design space for the auditing function.

### 3. Representing Congestion Exposure

Most protocol specifications start with a description of packet formats and codepoints with their associated meanings. This document does not: It is already known that choosing the encoding for the ConEx Signal is likely to entail some engineering compromises that have the potential to reduce the protocol's usefulness in some settings. Rather than making these engineering choices prematurely, this document side-steps the encoding problem by describing an abstract representation of ConEx Signals. All of the elements of the protocol can be defined in terms of this abstract representation. Most important, the preliminary use cases for the protocol are described in terms of the abstract representation in companion documents [I-D.conex-concepts-uses].

Once we have some example use cases we can evaluate different encoding schemes. Since these schemes are likely to include some conflated code points, some information will be lost resulting in weakening or disabling some of the algorithms and eliminating some use cases.

The goal of this approach is to be as complete as possible for discovering the potential usage and capabilities of the ConEx protocol, so we have some hope of making optimal design decisions when choosing the encoding.

#### 3.1. Strawman Encoding

As an aid to the reader, it might be helpful to describe a naive strawman encoding of the ConEx protocol described solely in terms of TCP: set the Reserved bit in the IPv4 header (bit 48 counting from zero [RFC0791])--aka the "evil bit" [RFC3514]) on all retransmissions or once per ECN signaled window reduction. Clearly network devices along the forward path can see this bit and act on it. For example they can count marked and unmarked packets to estimate the congestion levels along the path.

However, the IESG has chartered the ConEx working group to establish that there is sufficient demand for an IPv6 ConEx protocol before using the last available bit in the IPv4 header. Furthermore this encoding, by itself, does not sufficiently support partial deployment or strong auditing and might motivate users and/or applications to misrepresent the congestion that they are causing.



Nonetheless, this strawman encoding does present a clear mental model of how the ConEx protocol might function under various uses.

### 3.2. ECN Based Encoding

Ideally ConEx and ECN are orthogonal signals and SHOULD be entirely independent. However, given the limited number of header bit and/or code points, these signals may have to share code points, at least partially.

The re-ECN specification [I-D.briscoe-tsvwg-re-ecn-tcp] presents an implementation of ConEx that is tightly integrated with the encoding of ECN in the IP header. The central theme of this work is an audit mechanism that can provide sufficient disincentives against misrepresenting congestion [I-D.briscoe-tsvwg-re-ecn-motiv], which is analyzed extensively in Briscoe's PhD dissertation [Refb-dis].

Re-ECN is a good example of one chosen set of compromises attempting to meet the requirements of Section 2. However, the present document takes a step back, aiming to state the ideal requirements in order to allow the Internet community to assess whether other compromises are possible.

In particular, different incremental deployment choices may be desirable to meet the partial deployment requirement of Section 2. Re-ECN requires the receiver to be at least ECN-capable as well as requiring an update to the sender. Although ConEx will inherently require change at the sender, it would be preferable if it could work, even partially, with any receiver.

The chosen ConEx protocol certainly must not require ECN to be deployed in any network. In this respect re-ECN is already a good example--it acts perfectly well as a loss-based ConEx protocol if the loss-based audit techniques in Section 4.3 are used. However, it would still be desirable to avoid the dependence on an ECN receiver.

For a tutorial background on Re-Feedback techniques, see [Re-fb, FairerFaster].

#### 3.2.1. ECN Changes

Although the re-ECN protocol requires no changes to the network side of the ECN protocol, it is important to note that it does propose some relatively minor modifications to the host-to-host aspects of the ECN protocol specified in RFC 3168. They include: redefining the ECT(1) code point (the change is consistent with RFC3168 but requires deprecating the experimental ECN nonce [RFC3540]); modifications to the ECN negotiations carried on the SYN and SYN-ACK; and using a

different state machine to carry ECN signals in the transport acknowledgments from the Receiver to the Sender. This last change permits the transport protocol to carry multiple congestion signals per round trip, and greatly simplifies accurate auditing.

All of these adjustments to RFC 3168 may also be needed in a future standardized ConEx protocol. There will need to be very careful consideration of any proposed changes to ECN or other existing protocols, because any such changes increase the cost of deployment.

### 3.3. Abstract Encoding

The ConEx protocol could take one of two different encodings: independently settable bits or an enumerated set of mutually exclusive codepoints.

In both cases, the amount of congestion is signaled by the volume of marked data--just as the volume of lost data or ECN marked data signals the amount of congestion experienced. Thus the size of each packet carrying a ConEx Signal is significant.

#### 3.3.1. Independent Bits

This encoding involves flag bits, each of which the sender can set independently to indicate to the network one of the following four signals:

ConEx (Not-ConEx) The transport is (or is not) using ConEx with this packet (the protocol MUST be arranged so that legacy transport senders implicitly send Not-ConEx)

Re-Echo-Loss (Not-Re-Echo-Loss) The transport has (or has not) experienced a loss

Re-Echo-ECN (Not-Re-Echo-ECN) The transport has (or has not) experienced ECN signaled congestion

Credit (Not-Credit) The transport is (or is not) building up congestion credit (see Section 4.3 on audit devices)

#### 3.3.2. Codepoint Encoding

This encoding involves signaling one of the following five codepoints:

ENUM {Not-ConEx, ConEx, Re-Echo-Loss, Re-Echo-ECN, Credit}

Each named codepoint has the same meaning as in the encoding using

independent bits (Section 3.3.1). The use of any one codepoint implies the negative of all the others, except the last three codepoints (Re-Echo-Loss, Re-Echo-ECN and Credit) obviously also imply ConEx is supported.

Inherently, the semantics of most of the enumerated codepoints are mutually exclusive. 'Credit' is the only one that might need to be used in combination with either Re-Echo-Loss or Re-Echo-ECN, but even that requirement is questionable. It must not be forgotten that the enumerated encoding loses the flexibility to signal these two combinations, whereas the encoding with four independent bits is not so limited. Alternatively two extra codepoints could be assigned to these two combinations of semantics.

#### 4. Congestion Exposure Components

{ToDo: Picture of the components, similar to that in the last slideset about conex-concepts-uses?}

##### 4.1. Modified Senders

The sending transport needs to be modified to send Congestion Exposure Signals in response to congestion feedback signals.

##### 4.2. Receivers (Optionally Modified)

The receiving transport may already feedback sufficiently useful signals to the sender so that it does not need to be altered.

However, a TCP receiver feeds back ECN congestion signals no more than once within a round trip. The sender may require more precise feedback from the receiver otherwise it will appear to be understating its ConEx Signals (see Section 3.2.1).

Ideally, ConEx should be added to a transport like TCP without mandatory modifications to the receiver. But an optional modification to the receiver could be recommended for precision. This was the approach taken when adding re-ECN to TCP [I-D.briscoe-tsvwg-re-ecn-tcp].

##### 4.3. Audit

To audit ConEx Signals against actual losses an auditor could use one of the following techniques:

TCP-specific approach: The auditor could monitor TCP flows or aggregates of flows, only holding state on a flow if it first sends a Credit or a Re-Echo-Loss marking. The auditor could detect retransmissions by monitoring sequence numbers. It would assure that (volume of retransmitted data)  $\leq$  (volume of data marked Re-Echo-Loss). Traffic would only be auditable in this way if it conformed to the standard TCP protocol and the IP payload was not encrypted (e.g. with IPsec).

Predominant bottleneck approach: Unlike the above TCP-specific solution, this technique would work for IP packets carrying any transport layer protocol, and whether encrypted or not. But it only works well for networks designed so that losses predominantly occur under the management of one IP-aware node on the path. The auditor could then be located at this bottleneck. It could simply compare ConEx Signals with actual local losses. Most consumer access networks are design to this model, e.g. the radio network controller (RNC) in a cellular network or the broadband remote access server (BRAS) in a digital subscriber line (DSL) network.

The accuracy of an auditor at one predominant bottleneck might still be sufficient, even if losses occasionally occurred at other nodes in the network (e.g. border gateways). Although the auditor at the predominant bottleneck would not always be able to detect losses at other nodes, transports would not know where losses were occurring either. Therefore any transport would not know which losses it could cheat on without getting caught, and which ones it couldn't.

To audit ConEx Signals against actual ECN markings or losses, the auditor could work as follows: monitor flows or aggregates of flows, only holding state on a flow if it first sends a Credit or either Re-Echo marking. Count the number of bytes marked with Credit or Re-Echo-ECN. Separately count the number of bytes marked with ECN. Use Credits to assure that  $\#ECN \leq \#Re-Echo-ECN + \#Credit$ , even though the Re-Echo-ECN markings are delayed by at least one RTT.

#### 4.4. Policy Devices

Policy devices are characterised by a need to be configured with a policy related to the users or neighboring networks being served. In contrast, the auditing devices referred to in the previous section primarily enforce compliance with the ConEx protocol and do not need to be configured with any client-specific policy.

#### 4.4.1. Congestion Policers

Note that a congestion policer can be implemented in a very similar way to a bit-rate policer, but its effect is focused solely on traffic causing congestion downstream, not on all traffic just in case it causes congestion.

It monitors all ConEx traffic entering a network, or some identifiable subset. Using ConEx signals, it measures the amount of congestion being caused by this traffic. If this exceeds a policy-configured 'congestion-bit-rate' the congestion policer will limit all the monitored ConEx traffic. A congestion policer can be implemented by a simple token bucket. But unlike a bit-rate policer, it only removes tokens when forwarding packets that a ConEx marked. See [CongPol] for details.

#### 4.4.2. Other Policy Devices

Other policy devices that use ConEx signaling might traffic traffic based on ConEx Signals in much the same way as the monitoring element of a Congestion Policer. But the resulting action could be different. It might re-route traffic or downgrade the class of service.

It might do nothing directly to the traffic, but instead report measurements of ConEx Signals to systems designed to control congestion indirectly. For instance the measurements might be used to trigger penalty clauses in contracts, to levy charges between networks based on congestion or simply to notify customers who cause excessive congestion.

an auditing device only needs to enforce protocol compliance, it does not need to reflect any policy.

### 5. IANA Considerations

This memo includes no request to IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

### 6. Security Considerations

Significant parts of this whole document are about the auditability of ConEx Signals, in particular Section 4.3.

## 7. Conclusions

{ToDo:}

## 8. Acknowledgements

This document was improved by review comments from Toby Moncaster.

## 9. Comments Solicited

Comments and questions are encouraged and very welcome. They can be addressed to the IETF Congestion Exposure (ConEx) working group mailing list <conex@ietf.org>, and/or to the authors.

## 10. References

### 10.1. Normative References

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

### 10.2. Informative References

[CongPol] Jacquet, A., Briscoe, B., and T. Moncaster, "Policing Freedom to Use the Internet Resource Pool", Proc ACM Workshop on Re-Architecting the Internet (ReArch'08), December 2008, <<http://bobbriscoe.net/projects/refb/#polfree>>.

[FairerFaster] Briscoe, B., "A Fairer, Faster Internet Protocol", IEEE Spectrum Dec 2008:38--43, December 2008, <<http://bobbriscoe.net/projects/refb/#fairfastip>>.

[I-D.briscoe-tsvwg-re-ecn-motiv] Briscoe, B., Jacquet, A., Moncaster, T., and A. Smith, "Re-ECN: A Framework for adding Congestion Accountability to TCP/IP", draft-briscoe-tsvwg-re-ecn-tcp-motivation-01 (work in progress), September 2009.

- [I-D.briscoe-tsvwg-re-ecn-tcp]      Briscoe, B., Jacquet, A., Moncaster, T., and A. Smith, "Re-ECN: Adding Accountability for Causing Congestion to TCP/IP", draft-briscoe-tsvwg-re-ecn-tcp-08 (work in progress), September 2009.
- [I-D.conex-concepts-uses]      Briscoe, B., Woundy, R., Moncaster, T., and J. Leslie, "ConEx Concepts and Use Cases", draft-moncaster-conex-concepts-uses-01 (work in progress), July 2010.
- [I-D.ietf-ledbat-congestion]      Shalunov, S. and G. Hazel, "Low Extra Delay Background Transport (LEDBAT)", draft-ietf-ledbat-congestion-02 (work in progress), July 2010.
- [I-D.sridharan-tcpm-ctcp]      Sridharan, M., Tan, K., Bansal, D., and D. Thaler, "Compound TCP: A New TCP Congestion Control for High-Speed and Long Distance Networks", draft-sridharan-tcpm-ctcp-02 (work in progress), November 2008.
- [RFC0791]      Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.
- [RFC2309]      Braden, B., Clark, D., Crowcroft, J., Davie, B., Deering, S., Estrin, D., Floyd, S., Jacobson, V., Minshall, G., Partridge, C., Peterson, L., Ramakrishnan, K., Shenker, S., Wroclawski, J., and L. Zhang, "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, April 1998.
- [RFC3168]      Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.
- [RFC3514]      Bellovin, S., "The Security Flag in the IPv4 Header", RFC 3514, April 2003.

- [RFC3540] Spring, N., Wetherall, D., and D. Ely, "Robust Explicit Congestion Notification (ECN) Signaling with Nonces", RFC 3540, June 2003.
- [RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.
- [RFC5670] Eardley, P., "Metering and Marking Behaviour of PCN-Nodes", RFC 5670, November 2009.
- [RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, September 2009.
- [Re-fb] Briscoe, B., Jacquet, A., Di Cairano-Gilfedder, C., Salvatori, A., Soppera, A., and M. Koyabe, "Policing Congestion Response in an Internetwork Using Re-Feedback", ACM SIGCOMM CCR 35(4)277--288, August 2005, <<http://www.acm.org/sigs/sigcomm/sigcomm2005/techprog.html#session8>>.
- [Refb-dis] Briscoe, B., "Re-feedback: Freedom with Accountability for Causing Congestion in a Connectionless Internetwork", UCL PhD Dissertation, 2009, <<http://bobbriscoe.net/projects/refb/#refb-dis>>.
- [Vegas] Brakmo, L. and L. Peterson, "TCP Vegas: End-to-End Congestion Avoidance on a Global Internet", IEEE Journal on Selected Areas in Communications 13(8)1465--80, October 1995, <<http://ieeexplore.ieee.org/iel1/49/9740/00464716.pdf?arnumber=464716>>.



Authors' Addresses

Matt Mathis  
Google

Phone:  
Fax:  
EMail: mattmathis at google.com  
URI:

Bob Briscoe  
BT  
B54/77, Adastral Park  
Martlesham Heath  
Ipswich IP5 3RE  
UK

Phone: +44 1473 645196  
EMail: bob.briscoe@bt.com  
URI: <http://bobbbriscoe.net/>



Conex Group  
Internet Draft  
Intended Status: Informational  
Expires: April 17, 2011

D. McDysan  
Verizon

October 17, 2010

## Proposed Additional Use Cases for Congestion Exposure

draft-mcdysan-conex-other-usecases-00.txt

### Status of this Memo

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

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/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 17, 2011.

### Copyright Notice

Copyright (c) 2010 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.

## Abstract

This draft proposes some use cases for inclusion in the conex Working group charter's deliverable for an informational RFC covering use case description. These use cases are in addition to and/or complement those described in [UseCases], and focus on forms of congestion exposure that involve resources other than queues and timeframes other than real-time.

## Table of Contents

|  |   |
|--|---|
| 1. Introduction.....                                     | 2 |
| 2. Conventions used in this document.....                | 2 |
| 2.1. Acronyms.....                                       | 3 |
| 2.2. Terminology.....                                    | 3 |
| 3. Motivation and Background.....                        | 3 |
| 4. Proposed Use Cases.....                               | 4 |
| 4.1. Inequity of Heavy versus Light Users.....           | 4 |
| 4.2. Usage Tier/ Volume Feedback.....                    | 4 |
| 4.3. Feedback on Time of Day, Day of Week Charging.....  | 5 |
| 4.4. Recharging for Implementing Congestion Pricing..... | 6 |
| 5. Security Considerations.....                          | 6 |
| 6. IANA Considerations.....                              | 6 |
| 7. References.....                                       | 6 |
| 7.1. Normative References.....                           | 6 |
| 7.2. Informative References.....                         | 7 |
| 8. Acknowledgments.....                                  | 7 |

## 1. Introduction

This draft proposes some use cases for inclusion in the conex Working group charter's deliverable for an informational RFC covering use case description. These use cases are in addition to and/or complement those described in [UseCases], and focus on forms of congestion exposure that involve resources other than queues and timeframes other than real-time.

Section 3 provides some motivational background and a statement of problems involved with congestion pricing, with references to the presentations by experts in this area at the IETF 78 Technical Plenary in Maastricht.

Section 4 provides text for each of the above use cases in a mechanism independent manner.

## 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 RFC-2119 [RFC2119].

## 2.1. Acronyms

conex congestion exposure

## 2.2. Terminology

The following is a quote from the CONEX working group charter:

" ... develop a mechanism by which senders inform the network about the congestion encountered by previous packets on the same flow ... at the IP layer, such that the total level of congestion is visible to all IP devices along the path"

## 3. Motivation and Background

Successful adoption of experimental conex protocol(s) must address use cases that provide significant value to users, content providers and service providers. Central themes to this value proposition are incentives (i.e., congestion pricing) and the cost of providing marginal capacity [Varian, Johari].

There are three time scales over which congestion pricing can operate [Johari]: short (milliseconds to seconds), medium (minutes to hours to days) and long (months to years). Currently, the short term congestion signal is lost packets or a specific indication of congestion of a particular resource (e.g., a ECN indication for queue congestion), as stated in the conex charter. Setting congestion price as the marginal value of capacity is useful for medium timescales via traffic engineering and longer time scales via provisioning [Johari].

Some congestion exposure problems are challenging to address and are not completely addressed in [UseCases]:

- o 20% of the users generate 80% of the traffic and create unfairness with certain resource sharing [Varian]
- o Volume-based pricing makes it difficult for users to manage costs incurred, [Varian], [Briscoe]
- o Customers will pay a premium for unmetered use [Varian]
- o A form of congestion pricing is "recharging" (e.g., "free shipping") [Varian], where someone other than the end user pays for incurred congestion.
- o In general a content or service provider has hard capacity constraints at certain bottlenecks in their infrastructure (e.g., server capacity, router interface/queue service rate). Some form of adaption, such as time-shifting, route-shifting, or moderating the demand is required to adapt to these constraints [Kelly].

If conex exposes congestion without damage (e.g., loss) then many forms of adaption are feasible, as long as incentives are aligned with the signaled congestion [Kelly]. The use cases in the next section focus on forms of adaption that enable certain sets of incentives that are not completely covered in [UseCases].

#### 4. Proposed Use Cases

##### 4.1. Inequity of Heavy versus Light Users

In many networks, 20% are heavy users generating 80% of the traffic [Varian]. This means that a heavy user generates 16 times the traffic as a light user as a medium term (e.g, monthly) average. But, in a bandwidth-tiered flat priced network, heavy and light users often pay nearly the same price since pricing is based upon the short term (milliseconds) bandwidth measure of a shaper and/or a policer.

During non-peak periods, the resources of a service or content provider are underutilized and the marginal cost of capacity is small.

The access network is provisioned and traffic engineered for peak capacity of all users, and when congested, heavy users create 16 times the congestion of small users.

However, during peak use periods, a heavy user may send at near the bandwidth tier while light users may send intermittently. There is a need for a means for service providers to equitably assign costs to heavy versus light users. For example, the light users may pay less if they were charged by volume, as described in another use case.

A congestion measure of burstiness (e.g., ratio of peak rate to average rate over a longer interval than the tiered bandwidth shaper or policer) could be helpful in this use case. In general, a bursty packet flow is light (e.g., web surfing) and a non-bursty packet flow is heavy (e.g., viewing a lengthy HD video). The destination could perform this measure and feed it back to the sender. It may only be necessary to feedback this measure for heavy users, since the absence of such feedback could be inferred as an indication of a light user. The sender could insert some processed version of this feedback measure this at the IP layer so that all IP devices could be aware of whether this is a heavy or light user.

##### 4.2. Usage Tier/ Volume Feedback

Long-term (e.g., monthly) usage volume based pricing can more equitably assign costs to the prices paid, but;

- o is complex for users to keep track of usage and manage their activity to control the price they pay for access [Varian], [Briscoe],

- o does not address situation where heavy users send at a high rate, but only for a fraction of the usage measurement interval (e.g., only for a few hours or days during a month).
- o If usage counting is performed differently dependent upon the amount of congestion incurred (i.e., some form of congestion pricing as an incentive), then feedback is more important since in general users will not know when congestion is occurring, and even if they were informed this makes their usage tracking problem even more complex.

If usage volume information could be fed forward from an IP device using a conex mechanism on the path from senders to a destination, then this information could be fed back from the destination to the sender using TCP as stated in the Conex charter. Such information could include: the duration of the usage volume measurement tier (e.g., a month), the fraction of the usage tier already used, estimate of whether the user will exceed the usage tier if the historical rate to date continues, and any other information (or a pointer to such information) that would address the challenges of usage tier based pricing.

There are instances where a service/ content provider may choose to not count certain packets against the volume tier, such as when there is no congestion occurring, recharging is being done for this packet flow, and/or there is no usage counting being done for this time of day/week. A secure means to feed such information forward at the IP layer would allow for a number of different forms of counting, and hence adaption to congestion to occur.

#### 4.3. Feedback on Time of Day, Day of Week Charging

Congestion occurs when the offered load approaches that of the provisioned capacity, which often does not occur until shortly before there would be a need to provision additional capacity. Depending upon how restoration capacity is allocated by a service/content provider, congestion may only occur during peak periods when a failure is present.

Without Conex, utilization averaged over several minutes can be as high as 70 to 80% in typical network bottlenecks without loss that would reduce TCP effective throughput (i.e., goodput). A short term congestion control (sub-second to seconds) method that could increase utilization to above 90% and still avoid loss would only increase effective capacity by 10-20%.

If traffic increases at 50-75% per year, then a 10-20% increase in effective capacity means that the provisioning interval is only shifted by a few months. This handling of short term congestion use case alone may not be sufficient motivation for a service provider to deploy congestion handling measures.

However, in many points of a network, the majority of usage occurs during peak periods (e.g, a few busy hours) while much spare capacity exists off peak. The product of the spare capacity (bits/second) and the non-peak interval (seconds) that could carry traffic ranges from 2 to 10 times of the traffic carried during the peak period.

Congestion exposure and congestion pricing that enables users and content providers to time shift traffic to off-peak periods that would have otherwise been sent during peak periods can reduce provisioned capacity cost by as much as several hundred percent.

Historical time of day usage patterns could be employed to time shift traffic, but often maintenance actions are performed during the off peak periods, making the prediction of congestion using these methods less reliable. An automatic method for detection of congestion during off-peak periods is highly desirable.

#### 4.4. Recharging for Implementing Congestion Pricing

There should be a means to recharge (i.e., someone other than the receiving user pays) for usage that causes congestion during peak demand period versus that which does not [Varian].

If TCP were augmented with information related to the form of congestion, including not only short term as covered in [UseCases], but also including usage tier, Time of Day, or burstiness then sufficient information to implement sender pays (e.g., Content provider) versus receiver pays (e.g., end user) could be implemented. When the sender includes this information in the IP layer, then usage tier counting and TOD counting could be accounted for differently in IP devices in the path between sender and destination. Such an indication of recharging would need to be authenticated in some way.

#### 5. Security Considerations

Some use cases involve indications that could be spoofed or used to game counting and congestion feedback mechanisms, and therefore an authentication mechanism is needed when this information is handled at the IPv6 layer in the sender to destination direction or at the TCP layer in the destination to sender direction.

#### 6. IANA Considerations

None

#### 7. References

##### 7.1. Normative References

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



## 7.2. Informative References

[UseCases] B. Briscoe, R. Woundy, T. Moncaster, Ed., J. Leslie, Ed., "ConEx Concepts and Use Cases," draft-moncaster-conex-concepts-uses-01, Work in Progress

[Varian] Hal Varian, Google, "Congestion pricing principles," IETF 78 Technical Plenary, 29 July 2010

[Kelly] Frank Kelly, University of Cambridge, "Economic perspectives on congestion," IETF 78 Technical Plenary, 29 July 2010

[Johari] Ramesh Johari, Stanford University, "The information in congestion prices: milliseconds to years," IETF 78 Technical Plenary, 29 July 2010

[Briscoe] Bob Briscoe, BT, "Congestion Exposure," IETF 78 Technical Plenary, 29 July 2010.

## 8. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Copyright (c) 2010 IETF Trust and the persons identified as authors of the code. All rights reserved.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

This code was derived from IETF RFC [insert RFC number]. Please reproduce this note if possible.

## Authors' Addresses

Dave McDysan  
Verizon  
22001 Loudoun County PKWY  
Ashburn, VA 20147  
Email: dave.mcdysan@verizon.com



CONEX  
Internet-Draft  
Intended status: Informational  
Expires: April 28, 2011

B. Briscoe  
BT  
R. Woundy  
Comcast  
T. Moncaster, Ed.  
Moncaster.com  
J. Leslie, Ed.  
JLC.net  
October 25, 2010

ConEx Concepts and Use Cases  
draft-moncaster-conex-concepts-uses-02

Abstract

Internet Service Providers (operators) are facing problems where localized congestion prevents full utilization of the path between sender and receiver at today's "broadband" speeds. Operators desire to control this congestion, which often appears to be caused by a small number of users consuming a large amount of bandwidth. Building out more capacity along all of the path to handle this congestion can be expensive and may not result in improvements for all users so network operators have sought other ways to manage congestion. The current mechanisms all suffer from difficulty measuring the congestion (as distinguished from the total traffic).

The ConEx Working Group is designing a mechanism to make congestion along any path visible at the Internet Layer. This document describes example cases where this mechanism would be useful.

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 28, 2011.

## Copyright Notice

Copyright (c) 2010 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.

## Table of Contents

|  |    |
|--|----|
| 1. Introduction . . . . .                                | 3  |
| 2. Definitions . . . . .                                 | 5  |
| 3. Congestion Management . . . . .                       | 6  |
| 3.1. Existing Approaches . . . . .                       | 7  |
| 4. Exposing Congestion . . . . .                         | 8  |
| 4.1. ECN - a Step in the Right Direction . . . . .       | 9  |
| 5. ConEx Use Cases . . . . .                             | 9  |
| 5.1. ConEx as a basis for traffic management . . . . .   | 10 |
| 5.2. ConEx to incentivise scavenger transports . . . . . | 10 |
| 5.3. ConEx to mitigate DDoS . . . . .                    | 11 |
| 5.4. Accounting for Congestion Volume . . . . .          | 11 |
| 5.5. ConEx as a form of differential QoS . . . . .       | 12 |
| 5.6. Partial vs. Full Deployment . . . . .               | 12 |
| 6. Other issues . . . . .                                | 13 |
| 6.1. Congestion as a Commercial Secret . . . . .         | 13 |
| 6.2. Information Security . . . . .                      | 14 |
| 7. Security Considerations . . . . .                     | 15 |
| 8. IANA Considerations . . . . .                         | 16 |
| 9. Acknowledgments . . . . .                             | 16 |
| 10. References . . . . .                                 | 16 |
| 10.1. Normative References . . . . .                     | 16 |
| 10.2. Informative References . . . . .                   | 16 |
| Appendix A. ConEx Architectural Elements . . . . .       | 20 |
| A.1. ConEx Monitoring . . . . .                          | 20 |
| A.1.1. Edge Monitoring . . . . .                         | 21 |
| A.1.2. Border Monitoring . . . . .                       | 22 |
| A.2. ConEx Policing . . . . .                            | 22 |
| A.2.1. Egress Policing . . . . .                         | 23 |
| A.2.2. Ingress Policing . . . . .                        | 24 |
| A.2.3. Border Policing . . . . .                         | 25 |

## 1. Introduction

The growth of "always on" broadband connections, coupled with the steady increase in access speeds [OfCom], have caused unforeseen problems for network operators and users alike. Users are increasingly seeing congestion at peak times and changes in usage patterns (with the growth of real-time streaming) simply serve to exacerbate this. Operators want all their users to see a good service but are unable to see where congestion problems originate. But congestion results from sharing network capacity with others, not merely from using it. In general, today's "DSL" and cable-internet users cannot "cause" congestion in the absence of competing traffic. (Wireless operators and cellular internet have different tradeoffs which we will not discuss here.)

Congestion generally results from the interaction of traffic from one network operator's users with traffic from other users. The tools currently available don't allow an operator to identify which traffic contributes most to the congestion and so they are powerless to properly control it.

While building out more capacity to handle increased traffic is always good, the expense and lead-time can be prohibitive, especially for network operators that charge flat-rate feeds to subscribers and are thus unable to charge heavier users more for causing more congestion [BB-incentive]. For an operator facing congestion caused by other operators' networks, building out its own capacity is unlikely to solve the congestion problem. Operators are thus facing increased pressure to find effective solutions to dealing with the increasing bandwidth demands of all users.

The growth of "scavenger" behaviour (e.g. [LEDBAT]) helps to reduce congestion, but can actually make the problem less tractable. These users are trying to make good use of the capacity of the path while minimising their own costs. Thus, users of such services may show very heavy total traffic up until the moment congestion is detected (at the Transport Layer), but then will immediately back off. Monitoring (at the Internet Layer) cannot detect this congestion avoidance if the congestion in question is in a different domain further along the path; and must treat such users as congestion-causing users.

The ConEx working group proposes that Internet Protocol (IP) packets will carry additional ConEx information. The exact protocol details are not described in this document, but the ConEx information will be sufficient to allow any node in the network to see how much congestion is attributable to a given traffic flow. See [ConEx-Abstract-Mech] for further details.

Changes from previous drafts (to be removed by the RFC Editor):

From -01 to -02:

Updated document to take account of the new Abstract Mechanism draft [ConEx-Abstract-Mech]. PLEASE NOTE: As that draft develops, it is envisaged that more material will be able to be removed from this document, leaving this document free to concentrate on actual use cases for ConEx.

Updated the definitions section.

Removed sections on Requirements and Mechanism.

Moved section on ConEx Architectural Elements to appendix.

Minor changes throughout.

From -00 to -01:

Changed end of Abstract to better reflect new title

Created new section describing the architectural elements of ConEx Appendix A. Added Edge Monitors and Border Monitors (other elements are Ingress, Egress and Border Policers).

Extensive re-write of Section 5 partly in response to suggestions from Dirk Kutscher

Improved layout of Section 2 and added definitions of Whole Path Congestion, ConEx-Enabled and ECN-Enabled. Re-wrote definition of Congestion Volume. Renamed Ingress and Egress Router to Ingress and Egress Node as these nodes may not actually be routers.

Improved document structure. Merged sections on Exposing Congestion and ECN.

Added new section on ConEx requirements with a ConEx Issues subsection. Text for these came from the start of the old ConEx Use Cases section

Added a sub-section on Partial vs Full Deployment Section 5.6

Added a discussion on ConEx as a Business Secret Section 6.1

From draft-conex-mechanism-00 to  
draft-moncaster-conex-concepts-uses-00:

Changed filename to draft-moncaster-conex-concepts-uses.

Changed title to ConEx Concepts and Use Cases.

Chose uniform capitalisation of ConEx.

Moved definition of Congestion Volume to list of definitions.

Clarified mechanism section. Changed section title.

Modified text relating to conex-aware policing and policers (which are NOT defined terms).

Re-worded bullet on distinguishing ConEx and non-ConEx traffic in Section 5.

## 2. Definitions

In this section we define a number of terms that are used throughout the document. The key definition is that of congestion, which has a number of meanings depending on context. The definition we use in this document is based on the definition in [Padhye] where congestion is viewed as a probability that a packet will be dropped. This list of definitions is supplementary to that in [ConEx-Abstract-Mech].

**Congestion:** Congestion is a measure of the probability that a packet will be marked or dropped as it traverses a queue.

**flow:** a series of packets from a single sender to a single receiver that are treated by that sender as belonging to a single stream for the purposes of congestion control. NB in general this is not the same as the aggregate of all traffic between the sender and receiver.

**Congestion-rate:** For any granularity of traffic (packet, flow, aggregate, etc.), the instantaneous rate of traffic discarded or marked due to congestion. Conceptually, the instantaneous bit-rate of the traffic multiplied by the instantaneous congestion it is experiencing.

**Congestion-volume:** For any granularity of traffic (packet, flow, aggregate, etc.), the volume of bytes dropped or marked in a given period of time. Conceptually, congestion-rate multiplied by time.

**Upstream Congestion:** the accumulated level of congestion experienced by a traffic flow thus far along its path. In other words, at any point the Upstream Congestion is the accumulated level of congestion the traffic flow has experienced as it travels from the sender to that point. At the receiver this is equivalent to the end-to-end congestion level that (usually) is reported back to the sender.

**Downstream Congestion:** the level of congestion a flow of traffic is expected to experience on the remainder of its path. In other words, at any point the Downstream Congestion is the level of congestion the traffic flow is yet to experience as it travels from that point to the receiver.

**Ingress:** the first node a packet traverses that is outside the source's own network. In a domestic network that will be the first node downstream from the home access equipment. In an enterprise network this is the provider edge router.

**Egress:** the last node a packet traverses before reaching the receiver's network.

**ConEx-enabled:** Any piece of equipment (end-system, router, tunnel end-point, firewall, policer, etc) that complies with the core ConEx protocol, which is to be defined by the ConEx working group. By extension a ConEx-enabled network is a network whose edge nodes are all ConEx-enabled.

### 3. Congestion Management

Since 1988 the Internet architecture has made congestion management the responsibility of the end-systems. The network signals congestion to the receiver, the receiver feeds this back to the sender and the sender is expected to try and reduce the traffic it sends.

Any network that is persistently highly congested is inefficient. However the total absence of congestion is equally bad as it means there is spare capacity in the network that hasn't been used. The long-standing aim of congestion control has been to find the point where these two things are in balance.

Over recent years, some network operators have come to the view that end-system congestion management is insufficient. Because of the heuristics used by TCP, a relatively small number of end-machines can get a disproportionately high share of network resources. They have sought to "correct" this perceived problem by using middleboxes that try and reduce traffic that is causing congestion or by artificially



starving some traffic classes to create stronger congestion signals.

### 3.1. Existing Approaches

The authors have chosen not to exhaustively list current approaches to congestion management. Broadly these approaches can be divided into those that happen at Layer 3 of the OSI model and those that use information gathered from higher layers. In general these approaches attempt to find a "proxy" measure for congestion. Layer 3 approaches include:

- o Volume accounting -- the overall volume of traffic a given user or network sends is measured. Users may be subject to an absolute volume cap (e.g. 10Gbytes per month) or the "heaviest" users may be sanctioned in some manner.
- o Rate measurement -- the traffic rate per user or per network can be measured. The absolute rate a given user sends at may be limited at peak hours or the average rate may be used as the basis for inter-network billing.

Higher layer approaches include:

- o Bottleneck rate policing -- bottleneck flow rate policers aim to share the available capacity at a given bottleneck between all concurrent users.
- o DPI and application rate policing -- deep packet inspection and other techniques can be used to determine what application a given traffic flow is associated with. Operators may then use this information to rate-limit or otherwise sanction certain applications at peak hours.

All of these current approaches suffer from some general limitations. First, they introduce performance uncertainty. Flat-rate pricing plans are popular because users appreciate the certainty of having their monthly bill amount remain the same for each billing period, allowing them to plan their costs accordingly. But while flat-rate pricing avoids billing uncertainty, it creates performance uncertainty: users cannot know whether the performance of their connection is being altered or degraded based on how the network operator manages congestion.

Second, none of the approaches is able to make use of what may be the most important factor in managing congestion: the amount that a given endpoint contributes to congestion on the network. This information simply is not available to network nodes, and neither volume nor rate nor application usage is an adequate proxy for congestion volume,

because none of these metrics measures a user or network's actual contribution to congestion on the network.

Finally, none of these solutions accounts for inter-network congestion. Mechanisms may exist that allow an operator to identify and mitigate congestion in their own network, but the design of the Internet means that only the end-hosts have full visibility of congestion information along the whole path. ConEx allows this information to be visible to everyone on the path and thus allows operators to make better-informed decisions about controlling traffic.

#### 4. Exposing Congestion

We argue that current traffic-control mechanisms seek to control the wrong quantity. What matters in the network is neither the volume of traffic nor the rate of traffic: it is the contribution to congestion over time -- congestion means that your traffic impacts other users, and conversely that their traffic impacts you. So if there is no congestion there need not be any restriction on the amount a user can send; restrictions only need to apply when others are sending traffic such that there is congestion.

For example, an application intending to transfer large amounts of data could use a congestion control mechanism like [LEDBAT] to reduce its transmission rate before any competing TCP flows do, by detecting an increase in end-to-end delay (as a measure of impending congestion). However such techniques rely on voluntary, altruistic action by end users and their application providers. Operators can neither enforce their use nor avoid penalizing them for congestion they avoid.

The Internet was designed so that end-hosts detect and control congestion. We argue that congestion needs to be visible to network nodes as well, not just to the end hosts. More specifically, a network needs to be able to measure how much congestion any particular traffic expects to cause between the monitoring point in the network and the destination ("rest-of-path congestion"). This would be a new capability. Today a network can use Explicit Congestion Notification (ECN) [RFC3168] to detect how much congestion the traffic has suffered between the source and a monitoring point, but not beyond. This new capability would enable an ISP to give incentives for the use of LEDBAT-like applications that seek to minimise congestion in the network whilst restricting inappropriate uses of traditional TCP and UDP applications.

So we propose a new approach which we call Congestion Exposure. We propose that congestion information should be made visible at the IP

layer, so that any network node can measure the contribution to congestion of an aggregate of traffic as easily as straight volume can be measured today. Once the information is exposed in this way, it is then possible to use it to measure the true impact of any traffic on the network.

In general, congestion exposure gives operators a principled way to hold their customers accountable for the impact on others of their network usage and reward them for choosing congestion-sensitive applications.

#### 4.1. ECN - a Step in the Right Direction

Explicit Congestion Notification [RFC3168] allows routers to explicitly tell end-hosts that they are approaching the point of congestion. ECN builds on Active Queue Mechanisms such as random early discard (RED) [RFC2309] by allowing the router to mark a packet with a Congestion Experienced (CE) codepoint, rather than dropping it. The probability of a packet being marked increases with the length of the queue and thus the rate of CE marks is a guide to the level of congestion at that queue. This CE codepoint travels forward through the network to the receiver which then informs the sender that it has seen congestion. The sender is then required to respond as if it had experienced a packet loss. Because the CE codepoint is visible in the IP layer, this approach reveals the upstream congestion level for a packet.

Alas, this is not enough - ECN gives downstream nodes an idea of the congestion so far for any flow. This can help hold a receiver accountable for the congestion caused by incoming traffic. But a receiver can only indirectly influence incoming congestion, by politely asking the sender to control it. A receiver cannot make a sender install an adaptive codec, or install LEDBAT instead of TCP congestion-control. And a receiver cannot cause an attacker to stop flooding it with traffic.

What is needed is knowledge of the downstream congestion level, for which you need additional information that is still concealed from the network.

#### 5. ConEx Use Cases

This section sets out some of the use cases for ConEx. These use cases rely on some of the conceptual elements described in [ConEx-Abstract-Mech] and Appendix A. The authors don't claim this is an exhaustive list of use cases, nor that these have equal merit. In most cases ConEx is not the only solution to achieve these. But these use cases represent a consensus among people that have been

working on this approach for some years.

#### 5.1. ConEx as a basis for traffic management

Currently many operators impose some form of traffic management at peak hours. This is a simple economic necessity -- the only reason the Internet works as a commercial concern is that operators are able to rely on statistical multiplexing to share their expensive core network between large numbers of customers. In order to ensure all customers get some chance to access the network, the "heaviest" customers will be subjected to some form of traffic management at peak times (typically a rate cap for certain types of traffic) [Fair-use]. Often this traffic management is done with expensive flow aware devices such as DPI boxes or flow-aware routers.

ConEx offers a better approach that will actually target the users that are causing the congestion. By using Ingress or Egress Policers, an ISP can identify which users are causing the greatest Congestion Volume throughout the network. This can then be used as the basis for traffic management decisions. The Ingress Policier described in [Policing-freedom] is one interesting approach that gives the user a congestion volume limit. So long as they stay within their limit then their traffic is unaffected. Once they exceed that limit then their traffic will be blocked temporarily.

#### 5.2. ConEx to incentivise scavenger transports

Recent work proposes a new approach for QoS where traffic is provided with a less than best effort or "scavenger" quality of service. The idea is that low priority but high volume traffic such as OS updates, P2P file transfers and view-later TV programs should be allowed to use any spare network capacity, but should rapidly get out of the way if a higher priority or interactive application starts up. One solution being actively explored is LEDBAT which proposes a new congestion control algorithm that is less aggressive in seeking out bandwidth than TCP.

At present most operators assume a strong correlation between the volume of a flow and the impact that flow causes in the network. This assumption has been eroded by the growth of interactive streaming which behaves in an inelastic manner and hence can cause high congestion at relatively low data volumes. Currently LEDBAT-like transports get no incentive from the ISP since they still transfer large volumes of data and may reach high transfer speeds if the network is uncongested. Consequently the only current incentive for LEDBAT is that it can reduce self-congestion effects.

If the ISP has deployed a ConEx-aware Ingress Policier then they are

able to incentivise the use of LEDBAT because a user will be policed according to the overall congestion volume their traffic generates, not the rate or data volume. If all background file transfers are only generating a low level of congestion, then the sender has more "congestion budget" to "spend" on their interactive applications. It can be shown [Kelly] that this approach improves social welfare -- in other words if you limit the congestion that all users can generate then everyone benefits from a better service.

### 5.3. ConEx to mitigate DDoS

DDoS relies on subverting innocent end users and getting them to send flood traffic to a given destination. This is intended to cause a rapid increase in congestion in the immediate vicinity of that destination. If it fails to do this then it can't be called Denial of Service. If the ingress ISP has deployed Ingress Policers, that ISP will effectively limit how much DDoS traffic enters the 'net. If any ISP along the path has deployed Border Monitors then they will be able to detect a sharp rise in Congestion Volume and if they have Border Policers they will be able to "turn off" this traffic. If the victim of the DDoS attack is behind an Egress Monitor then their ISP will be able to detect which traffic is causing problems. If the compromised user tries to use the 'net during the DDoS attack, they will quickly become aware that something is wrong, and their ISP can show the evidence that their computer has become zombified.

DDoS is a genuine problem and so far there is no perfect solution. ConEx does serve to raise the bar somewhat and can avoid the need for some of the more draconian measures that are currently used to control DDoS. More details of this can be found in [Malice].

### 5.4. Accounting for Congestion Volume

Accountability was one of the original design goals for the Internet [Design-Philosophy]. At the time it was ranked low because the network was non-commercial and it was assumed users had the best interests of the network at heart. Nowadays users generally treat the network as a commodity and the Internet has become highly commercialised. This causes problems for operators and others which they have tried to solve and often leads to a tragedy of the commons where users end up fighting each other for scarce peak capacity.

The most elegant solution would be to introduce an Internet-wide system of accountability where every actor in the network is held to account for the impact they have on others. If Policers are placed at every Network Ingress or Egress and Border Monitors at every border, then you have the basis for a system of congestion accounting. Simply by controlling the overall Congestion Volume each

end-system or stub-network can send you ensure everyone gets a better service.

#### 5.5. ConEx as a form of differential QoS

Most QoS approaches require the active participation of routers to control the delay and loss characteristics for the traffic. For real-time interactive traffic it is clear that low delay (and predictable jitter) are critical, and thus these probably always need different treatment at a router. However if low loss is the issue then ConEx offers an alternative approach.

Assuming the ingress ISP has deployed a ConEx Ingress Policer, then the only control on a user's traffic is dependent on the congestion that user has caused. Likewise, if they are receiving traffic through a ConEx Egress Policer then their ISP will impose traffic controls (prioritisation, rate limiting, etc) based on the congestion they have caused. If an end-user (be they the receiver or sender) wants to prioritise some traffic over other traffic then they can allow that traffic to generate or cause more congestion. The price they will pay will be to reduce the congestion that their other traffic causes.

Streaming video content-delivery is a good candidate for such ConEx-mediated QoS. Such traffic can tolerate moderately high delays, but there are strong economic pressures to maintain a high enough data rate (as that will directly influence the Quality of Experience the end-user receives. This approach removes the need for bandwidth brokers to establish QoS sessions, by removing the need to coordinate requests from multiple sources to pre-allocate bandwidth, as well as to coordinate which allocations to revoke when bandwidth predictions turn out to be wrong. There is also no need to "rate-police" at the boundaries on a per-flow basis, removing the need to keep per-flow state (which in turn makes this approach more scalable).

#### 5.6. Partial vs. Full Deployment

In a fully-deployed ConEx-enabled internet, [QoS-Models] shows that ISP settlements based on congestion volume can allocate money to where upgrades are needed. Fully-deployed implies that ConEx-marked packets which have not exhausted their expected congestion would go through a congested path in preference to non-ConEx packets, with money changing hands to justify that priority.

In a partial deployment, routers that ignore ConEx markings and let them pass unaltered are no problem unless they become congested and drop packets. Since ConEx incentivises the use of lower congestion transports, such congestion drops should anyway become rare events.

ConEx-unaware routers that do drop ConEx-marked packets would cause a problem so to minimise this risk ConEx should be designed such that ConEx packets will appear valid to any node they traverse. Failing that it could be possible to bypass such nodes with a tunnel.

If any network is not ConEx enabled then the sender and receiver have to rely on ECN-marking or packet drops to establish the congestion level. If the receiver isn't ConEx-enabled then there needs to be some form of compatibility mode. Even in such partial deployments the end-users and access networks will benefit from ConEx. This will put create incentives for ConEx to be more widely adopted as access networks put pressure on their backhaul providers to use congestion as the basis of their interconnect agreement.

The actual charge per unit of congestion would be specified in an interconnection agreement, with economic pressure driving that charge downward to the cost to upgrade whenever alternative paths are available. That charge would most likely be invisible to the majority of users. Instead such users will have a contractual allowance to cause congestion, and would see packets dropped when that allowance is depleted.

Once an Autonomous System (AS) agrees to pay any congestion charges to any other AS it forwards to, it has an economic incentive to increase congestion-so-far marking for any congestion within its network. Failure to do this quickly becomes a significant cost, giving it an incentive to turn on such marking.

End users (or the writers of the applications they use) will be given an incentive to use a congestion control that back off more aggressively than TCP for any elastic traffic. Indeed they will actually have an incentive to use fully weighted congestion controls that allow traffic to cause congestion in proportion to its priority. Traffic which backs off more aggressively than TCP will see congestion charges remain the same (or even drop) as congestion increases; traffic which backs off less aggressively will see charges rise, but the user may be prepared to accept this if it is high-priority traffic; traffic which backs off not at all will see charges rise dramatically.

## 6. Other issues

### 6.1. Congestion as a Commercial Secret

Network operators have long viewed the congestion levels in their network as a business secret. In some ways this harks back to the days of fixed-line telecommunications where congestion manifested as failed connections or dropped calls. But even in modern data-centric

packet networks congestion is viewed as a secret not to be shared with competitors. It can be debated whether this view is sensible, but it may make operators uneasy about deploying ConEx. The following two examples highlight some of the arguments used:

- o An ISP buys backhaul capacity from an operator. Most operators want their customers to get a decent service and so they want the backhaul to be relatively uncongested. If there is competition, operators will seek to reassure their customers (the operators) that their network is not congested in order to attract their custom. Some operators may see ConEx as a threat since it will enable those operators to see the actual congestion in their network. On the other hand, operators with low congestion could use ConEx to show how well their network performs, and so might have an incentive to enable it.
- o Operators would like to be part of the lucrative content provision market. Currently the ISP can gain a competitive edge as it can put its own content in a higher QoS class, whereas traffic from content providers has to use the Best Effort class. The ISP may take the view that if they can conceal the congestion level in their Best Effort class this will make it harder for the content provider to maintain a good level of QoS. But in reality the Content Provider will just use the feedback mechanisms in streaming protocols such as Adobe Flash to monitor the congestion.

Of course some might say that the idea of keeping congestion secret is silly. After all, end-hosts already have knowledge of the congestion throughout the network, albeit only along specific paths, and operators can work out that there is persistent congestion as their customers will be suffering degraded network performance.

## 6.2. Information Security

make a source believe it has seen more congestion than it has

hijack a user's identity and make it appear they are dishonest at an egress policer

clear or otherwise tamper with the ConEx markings

...

{ToDo} Write these up properly...



## 7. Security Considerations

This document proposes a mechanism tagging onto Explicit Congestion Notification [RFC3168], and inherits the security issues listed therein. The additional issues from ConEx markings relate to the degree of trust each forwarding point places in the ConEx markings it receives, which is a business decision mostly orthogonal to the markings themselves.

One expected use of exposed congestion information is to hold the end-to-end transport and the network accountable to each other. The network cannot be relied on to report information to the receiver against its interest, and the same applies for the information the receiver feeds back to the sender, and that the sender reports back to the network. Looking at each in turn:

**The Network** In general it is not in any network's interest to under-declare congestion since this will have potentially negative consequences for all users of that network. It may be in its interest to over-declare congestion if, for instance, it wishes to force traffic to move away to a different network or simply to reduce the amount of traffic it is carrying. Congestion Exposure itself won't significantly alter the incentives for and against honest declaration of congestion by a network, but we can imagine applications of Congestion Exposure that will change these incentives. There is a perception among network operators that their level of congestion is a business secret. Today, congestion is one of the worst-kept secrets a network has, because end-hosts can see congestion better than network operators can. Congestion Exposure will enable network operators to pinpoint whether congestion is on one side or the other of any border. It is conceivable that forwarders with underprovisioned networks may try to obstruct deployment of Congestion Exposure.

**The Receiver** Receivers generally have an incentive to under-declare congestion since they generally wish to receive the data from the sender as rapidly as possible. [Savage] explains how a receiver can significantly improve their throughput by failing to declare congestion. This is a problem with or without Congestion Exposure. [KGao] explains one possible technique to encourage receiver's to be honest in their declaration of congestion.

**The Sender** One proposed mechanism for Congestion Exposure deployment adds a requirement for a sender to advise the network how much congestion it has suffered or caused. Although most senders currently respond to congestion they are informed of, one use of exposed congestion information might be to encourage sources of persistent congestion to back off more aggressively. Then clearly

there may be an incentive for the sender to under-declare congestion. This will be a particular problem with sources of flooding attacks. "Policing" mechanisms have been proposed to deal with this.

In addition there are potential problems from source spoofing. A malicious sender can pretend to be another user by spoofing the source address. Congestion Exposure allows for "Policers" and "Traffic Shapers" so as to be robust against injection of false congestion information into the forward path.

## 8. IANA Considerations

This document does not require actions by IANA.

## 9. Acknowledgments

Bob Briscoe is partly funded by Trilogy, a research project (ICT-216372) supported by the European Community under its Seventh Framework Programme. The views expressed here are those of the author only.

The authors would like to thank Contributing Authors Bernard Aboba, Joao Taveira Araujo, Louise Burness, Alissa Cooper, Philip Eardley, Michael Menth, and Hannes Tschofenig for their inputs to this document. Useful feedback was also provided by Dirk Kutscher.

## 10. References

### 10.1. Normative References

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

### 10.2. Informative References

- |                |   |
|----------------|---|
| [BB-incentive] | MIT Communications Futures Program (CFP) and Cambridge University Communications Research Network, "The Broadband Incentive Problem", September 2005. |
|----------------|---|

- [ConEx-Abstract-Mech]      Briscoe, B., "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-mathis-conex-abstract-mech-00 (work in progress), October 2010.
- [Design-Philosophy]      Clarke, D., "The Design Philosophy of the DARPA Internet Protocols", 1988.
- [Fair-use]      Broadband Choices, "Truth about 'fair usage' broadband", 2009.
- [Fairer-faster]      Briscoe, B., "A Fairer Faster Internet Protocol", IEEE Spectrum Dec 2008 pp38-43, December 2008.
- [I-D.briscoe-tsvwg-re-ecn-tcp-motivation]      Briscoe, B., Jacquet, A., Moncaster, T., and A. Smith, "Re-ECN: A Framework for adding Congestion Accountability to TCP/IP", draft-briscoe-tsvwg-re-ecn-tcp-motivation-01 (work in progress), September 2009.
- [KGao]      Gao, K. and C. Wang, "Incrementally Deployable Prevention to TCP Attack with Misbehaving Receivers", December 2004.
- [Kelly]      Kelly, F., Maulloo, A., and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability", Journal of the Operational Research Society 49(3) 237--252, 1998, <<http://www.statslab.cam.ac.uk/~frank/rate.html>>.

- [LEDBAT] Shalunov, S., "Low Extra Delay Background Transport (LEDBAT)", draft-ietf-ledbat-congestion-01 (work in progress), March 2010.
- [Malice] Briscoe, B., "Using Self Interest to Prevent Malice; Fixing the Denial of Service Flaw of the Internet", WESII - Workshop on the Economics of Securing the Information Infrastructure 2006, 2006, <[http://wesii.econinfosec.org/draft.php?paper\\_id=19](http://wesii.econinfosec.org/draft.php?paper_id=19)>.
- [OfCom] Ofcom: Office of Communications, "UK Broadband Speeds 2008: Research report", January 2009.
- [Padhye] Padhye, J., Firoiu, V., Towsley, D., and J. Kurose, "Modeling TCP Throughput: A Simple Model and its Empirical Validation", ACM SIGCOMM Computer Communications Review Vol 28(4), pages 303-314, May 1998.
- [Policing-freedom] Briscoe, B., Jacquet, A., and T. Moncaster, "Policing Freedom to Use the Internet Resource Pool", RE-Arch 2008 hosted at the 2008 CoNEXT conference , December 2008.
- [QoS-Models] Briscoe, B. and S. Rudkin, "Commercial Models for IP Quality of Service Interconnect", BTTJ

Special Edition on IP  
Quality of Service vol 23  
(2), April 2005.

[RFC2309]

Braden, B., Clark, D.,  
Crowcroft, J., Davie, B.,  
Deering, S., Estrin, D.,  
Floyd, S., Jacobson, V.,  
Minshall, G., Partridge,  
C., Peterson, L.,  
Ramakrishnan, K., Shenker,  
S., Wroclawski, J., and L.  
Zhang, "Recommendations on  
Queue Management and  
Congestion Avoidance in  
the Internet", RFC 2309,  
April 1998.

[Re-Feedback]

Briscoe, B., Jacquet, A.,  
Di Cairano-Gilfedder, C.,  
Salvatori, A., Soppera,  
A., and M. Koyabe,  
"Policing Congestion  
Response in an  
Internetwork Using Re-  
Feedback", ACM SIGCOMM  
CCR 35(4)277--288,  
August 2005, <[http://  
www.acm.org/sigs/sigcomm/  
sigcomm2005/  
techprog.html#session8](http://www.acm.org/sigs/sigcomm/sigcomm2005/techprog.html#session8)>.

[Savage]

Savage, S., Wetherall, D.,  
and T. Anderson, "TCP  
Congestion Control with a  
Misbehaving Receiver", ACM  
SIGCOMM Computer  
Communication Review ,  
1999.

[re-ecn-motive]

Briscoe, B., Jacquet, A.,  
Moncaster, T., and A.  
Smith, "Re-ECN: A  
Framework for adding  
Congestion Accountability  
to TCP/IP", draft-briscoe-  
tsvwg-re-ecn-tcp-  
motivation-01 (work in

progress), September 2009.

## Appendix A. ConEx Architectural Elements

ConEx is a simple concept that has revolutionary implications. It is that rare thing -- a truly disruptive technology, and as such it is hard to imagine the variety of uses it may be put to. Before even thinking what it might be used for we need to address the issue of how it can be used. This section describes four architectural elements that can be placed in the network and which utilise ConEx information to monitor or control traffic flows.

In the following we are assuming the most abstract version of the ConEx mechanism, namely that every packet carries two congestion fields, one for upstream congestion and one for downstream. [ConEx-Abstract-Mech] outlines one possible approach for this.

### A.1. ConEx Monitoring

One of the most useful things ConEx provides is the ability to monitor (and control) the amount of congestion entering or leaving a network. With ConEx, each packet carries sufficient information to work out the Upstream, Downstream and Total Congestion Volume that packet is responsible for. This allows the overall Congestion Volume to be calculated at any point in the network. In effect this gives a measure of how much excess traffic has been sent that was above the instantaneous transmission capacity of the network. A 1 Gbps router that is 0.1% congested implies that there is 1 Mbps of excess traffic at that point in time.

The figure below shows 2 conceptual pieces of network equipment that utilise ConEx information in order to monitor the flow of congestion through the network. The Border Monitor sits at the border between two networks, while the Edge Monitor sits at the Ingress or Egress to the Internet network.

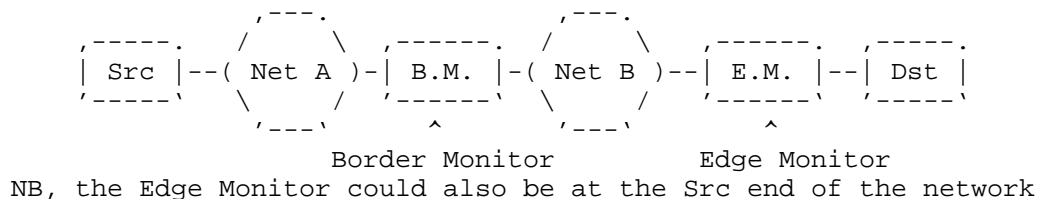


Figure 1: Ingress, egress and border monitors

Note: In the tables below, ConEx-Enabled is as defined in Section 2 and ECN-enabled means any router that fully enables Explicit

Congestion Notification (ECN) as defined in [RFC3168] and any relevant updates to that standard.

#### A.1.1.1. Edge Monitoring

| Network Element    | ECN-Enabled?                                       | ConEx-Enabled?                         | Notes   |
|--------------------|--|--|---|
| Sender             | Yes, if ECN is used as basis for congestion signal | Yes, must be sending ConEx information | Must be receiving congestion feedback                       |
| Sender's Network   | ECN would be beneficial                            | Should understand ConEx markings       | NB, it doesn't have to be fully ConEx-Enabled               |
| Core Network       | ECN would be beneficial                            | Needn't understand ConEx               | ConEx markings must get through the network                 |
| Receiver's Network | ECN would be beneficial                            | Should understand ConEx markings       | Deosn't have to be fully ConEx-Enabled                      |
| Receiver           | Only needed if network is ECN-Enabled              | Should understand ConEx                | Has to feedback the congestion it sees (either ECN or drop) |

Table 1: Requirements for Edge Monitoring

Edge Monitors are ideally positioned to verify the accuracy of ConEx markings. If there is an imbalance between the expected congestion and the actual congestion then this will show up at the egress. Edge Monitors can also be used by an operator to measure the service a given customer is receiving by monitoring how much congestion their traffic is causing. This may allow them to take pre-emptive action if they detect any anomalies.

## A.1.2. Border Monitoring

| Network Element    | ECN-Enabled?                                 | ConEx-Enabled?                         | Notes   |
|--------------------|--|--|---|
| Sender             | Must be ECN-enabled if any of the network is | Yes, must be sending ConEx information | Must receive accurate congestion feedback       |
| Sender's Network   | ECN should be enabled                        | Should understand ConEx markings       | Ideally would be ConEx-Enabled                  |
| Core Network       | ECN should be enabled                        | Should understand ConEx markings       | Ideally would be ConEx-Enabled                  |
| Receiver's Network | ECN should be enabled                        | Should understand ConEx markings       | Ideally would be ConEx-Enabled                  |
| Receiver           | Must be ECN-enabled if any of the network is | Must be ConEx enabled                  | Receiver has to feedback the congestion it sees |

Table 2: Requirements for Border Monitoring

At any border between 2 networks, the operator can see the total Congestion Volume that is being forwarded into its network by the neighbouring network. A Border Monitor is able to measure the bulk congestion markings and establish the flow of Congestion Volume each way across the border. This could be used as the basis for inter-network settlements. It also provides information to target upgrades to where they are actually needed and might help to identify network problems. Border Monitoring really needs the majority of the network to be ECN-Enabled in order to provide the necessary Upstream Congestion signal. Clearly the greatest benefit comes when there is also ConEx deployment in the network. However, as long as the sender is sending accurate ConEx information and the majority of the network is ECN-enabled, border monitoring will work.

## A.2. ConEx Policing

As shown above, ConEx gives an easy method of measuring Congestion Volume. This information can be used as a control metric for making traffic management decisions (such as deciding which traffic to prioritise) or to identify and block sources of persistent and damaging congestion. Simple policer mechanisms, such as those



described in [Policing-freedom] and [re-ecn-motive], can control the overall congestion volume traversing a network. Ingress Policing typically happens at the Ingress Node, Egress Policing typically happens at the Egress Node and Border Policing can happen at any border between two networks. The current charter concentrates on use cases employing Egress Policers.

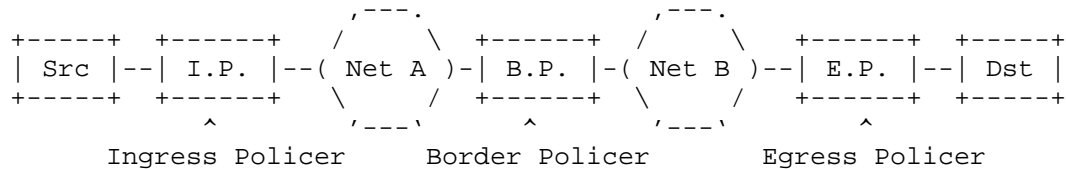


Figure 2: Ingress, egress and border policers

#### A.2.1. Egress Policing

| Network Element    | ECN-Enabled?  | ConEx-Enabled?              | Notes  |
|--------------------|---|-----------------------------|--|
| Sender             | The sender should be ECN-enabled if any of the network is | Must be ConEx-Enabled       | Must be receiving congestion feedback  |
| Sender's Network   | ECN is optional but beneficial                            | ConEx is optional           | ConEx would enable them to do Ingress Policing (see later)   |
| Core Network       | ECN is optional but beneficial                            | Not needed                  | ConEx marks must survive crossing the network  |
| Receiver's Network | ECN is optional but beneficial                            | Must fully understand ConEx | Each receiver needs an Egress Policer  |
| Receiver           | Should be ECN-enabled if any of the network is            | Should understand ConEx     | Must feedback the congestion it sees. ConEx may have a compatibility mode if the receiver is not ConEx-Enabled |

Table 3: Egress Policer Requirements

An Egress Policer allows an ISP to monitor the Congestion Volume a

user's traffic has caused throughout the network, and then use this to prioritise the traffic accordingly. By itself, such a policer cannot tell how much of this congestion was caused in the ISP's own network, but it will identify which users are the "heaviest" in terms of the congestion they have caused. Assuming the ConEx information is accurate then the Egress Policer will be able to see how much congestion exists between it and the final destination (what you might call "last-mile" congestion). There are a number of strategies that could be used to determine how traffic is treated by an Egress Policer. Obviously traffic that is not ConEx enabled needs to receive some form of "default" treatment. Traffic that is ConEx enabled may have under-declared congestion in which case it would be reasonable to give it a low scheduling priority. Traffic that appears to be over-declaring congestion may be simply a result of especially high "last-mile" congestion, in which case the ISP may want to upgrade the access capacity, or may want to try and reduce the volume of traffic. Where the ISP knows what the "last-mile" congestion is (for instance if it is able to measure several users sharing that same capacity) then any remaining over-declared congestion might be seen as a signal that the sender wishes to prioritise this traffic.

#### A.2.2. Ingress Policing

| Network Element    | ECN-Enabled?                                   | ConEx-Enabled?           | Notes  |
|--------------------|--|--------------------------|--|
| Sender             | Should be ECN-enabled                          | Must be ConEx-enabled    | Must be receiving congestion feedback  |
| Sender's Network   | ECN is optional but beneficial                 | Must understand ConEx    |  |
| Core Network       | ECN is optional but beneficial                 | Needn't understand ConEx | ConEx markings must survive crossing the network   |
| Receiver's Network | ECN is optional but beneficial                 | Needn't understand ConEx | ConEx markings must survive crossing the network   |
| Receiver           | Should be ECN-enabled if any of the network is | Should be ConEx-Enabled  | Must feedback the congestion it sees. ConEx may have a compatibility mode if the receiver is not ConEx-Enabled |

Table 4: Ingress Policer Requirements

At the Network Ingress, an ISP can police the amount of congestion a user is causing by limiting the congestion volume they send into the network. One system that achieves this is described in [Policing-freedom]. This uses a modified token bucket to limit the congestion rate being sent rather than the overall rate. Such Ingress policing is relatively simple as it requires no flow state. Furthermore, unlike many mechanisms, it treats all a user's packets equally.

#### A.2.3. Border Policing

| Network Element    | ECN-Enabled?                                   | ConEx-Enabled?          | Notes  |
|--------------------|--|-------------------------|--|
| Sender             | ECN should be enabled                          | Must be ConEx-enabled   | Must receive accurate congestion feedback  |
| Sender's Network   | ECN is optional but beneficial                 | Must be ConEx-enabled   |  |
| Core Network       | ECN is optional but beneficial                 | Should be ConEx-Enabled | Must be ConEx-Enabled if it is doing the policing. At a minimum must pass ConEx markings unaltered             |
| Receiver's Network | ECN is optional but beneficial                 | Should be ConEx-Enabled | At a minimum must pass ConEx markings unaltered  |
| Receiver           | Should be ECN-Enabled if any of the network is | Should be ConEx-Enabled | Must feedback the congestion it sees. ConEx may have a compatibility mode if the receiver is not ConEx-Enabled |

Table 5: Border Policer Requirements

A Border Policer will allow an operator to directly control the congestion that it allows into its network. Normally we would expect the controls to be related to some form of contractual obligation between the two parties. However, such Policing could also be used to mitigate some effects of Distributed Denial of Service (see Section 5.3). In effect a Border Policer encourages the network upstream to take responsibility for congestion it will cause

downstream and could be seen as an incentive for that network to participate in ConEx (e.g. install Ingress Policers)

#### Authors' Addresses

Bob Briscoe

BT

B54/77, Adastral Park

Martlesham Heath

Ipswich IP5 3RE

UK

Phone: +44 1473 645196

EMail: bob.briscoe@bt.com

URI: <http://bobbbriscoe.net/>

Richard Woundy

Comcast

Comcast Cable Communications

27 Industrial Avenue

Chelmsford, MA 01824

US

EMail: richard\_woundy@cable.comcast.com

URI: <http://www.comcast.com>

Toby Moncaster (editor)

Moncaster.com

Dukes

Layer Marney

Colchester CO5 9UZ

UK

EMail: toby@moncaster.com

John Leslie (editor)

JLC.net

10 Souhegan Street

Milford, NH 03055

US

EMail: john@jlc.net

