ConEx	B. Briscoe, Ed.
Internet-Draft	BT
Intended status: Informational	R. Woundy, Ed.
Expires: April 27, 2012	Comcast
	A. Cooper, Ed.
	CDT
	October 25, 2011

ConEx Concepts and Use Cases
draft-ietf-conex-concepts-uses-03

Abstract

This document provides the entry point to the set of documentation about the Congestion Exposure (ConEx) protocol. It explains the motivation for including a ConEx field at the IP layer: to expose information about congestion to network nodes. Although such information may have a number of uses, this document focuses on how the information communicated in the ConEx field can serve as the basis for significantly more efficient and effective traffic management than what exists on the Internet today.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

The power of Internet technology comes from multiplexing shared capacity with packets rather than circuits. Network operators aim to provide sufficient shared capacity, but when too much packet load meets too little shared capacity, congestion results. Congestion appears as either increased delay, dropped packets or packets explicitly marked with Explicit Congestion Notification (ECN) markings [RFC3168]. As described in Figure 1, congestion control currently relies on the transport receiver detecting these 'Congestion Signals' and informing the transport sender in 'Congestion Feedback Signals.' The sender is then expected to reduce its rate in response.

This document provides the entry point to the set of documentation about the Congestion Exposure (ConEx) protocol. It focuses on the motivation for including a ConEx field at the IP layer. (A companion document, [ConEx-Abstract-Mech], focuses on the mechanics of the protocol.) Briefly, the idea is for the sender to continually signal expected congestion in the headers of any data it sends. To a first approximation, the sender does this by relaying the 'Congestion Feedback Signals' back into the IP layer. They then travel unchanged across the network to the receiver (shown as 'IP-Layer-ConEx-Signals' in Figure 1). This enables IP layer devices on the path to see information about the whole path congestion.

```
,---------.                                               ,---------.
|Transport|                                               |Transport|
| Sender  |   .                                           |Receiver |
|         |  /|___________________________________________|         |
|     ,-<---------------Congestion-Feedback-Signals--<--------.     |
|     |   |/                                              |   |     |
|     |   |\           Transport Layer Feedback Flow      |   |     |
|     |   | \  ___________________________________________|   |     |
|     |   |  \|                                           |   |     |
|     |   |   '         ,-----------.               .     |   |     |
|     |   |_____________|           |_______________|\    |   |     |
|     |   |    IP Layer |           |  Data Flow      \   |   |     |
|     |   |             |(Congested)|                  \  |   |     |
|     |   |             |  Network  |--Congestion-Signals--->-'     |
|     |   |             |  Device   |                    \|         |
|     |   |             |           |                    /|         |
|     `----------->--(new)-IP-Layer-ConEx-Signals-------->|         |
|         |             |           |                  /  |         |
|         |_____________|           |_______________  /   |         |
|         |             |           |               |/    |         |
`---------'             `-----------'               '     `---------'
```

One of the key benefits of exposing this congestion information at the IP layer is that it makes the information available to network operators for use as input into their traffic management procedures. As shown in Figure 1, a ConEx-enabled sender signals whole path congestion, which is (approximately) the congestion one round trip time earlier as reported by the receiver to the sender. The ConEx signal is a mark in the IP header that is easy for any IP device to read. Therefore a node performing traffic management can count congestion as easily as it might count data volume today by simply counting the volume of packets with ConEx markings.

ConEx-based traffic management can make highly efficient use of capacity. In times of no congestion, all traffic management restraints can be removed, leaving the network's full capacity available to all its users. If some users on the network cause disproportionate congestion, the traffic management function can learn about this and directly limit those users' traffic in order to protect the service of other users sharing the same capacity. ConEx-based traffic management thus presents a step change in terms of the options available to operators for managing traffic on their networks.

The remainder of this document explains the concepts behind ConEx and how exposing congestion can significantly improve Internet traffic management, among other benefits. Section 2 introduces a number of concepts that are fundamental to understanding how ConEx-based traffic management works. Section 3 shows how ConEx can be used for traffic management, discusses additional benefits from such usage, and compares ConEx-based traffic management to existing traffic management approaches. Section 4 discusses other related use cases. Section 5 briefly discusses deployment arrangements. The final sections are standard RFC back matter.

2. Concepts

ConEx relies on a precise definition of congestion and a number of newer concepts that are introduced and defined in this section.

2.1. Congestion

Despite its central role in network control and management, congestion is a remarkably difficult concept to define. Experts in different disciplines and with different perspectives define congestion in a variety of ways [Bauer09].

The definition used for the purposes of ConEx is expressed as the probability of packet loss (or the probability of packet marking if ECN is in use). This definition focuses on how congestion is measured, rather than describing congestion as a condition or state.

2.2. Congestion-Volume

The metric that ConEx exposes is congestion-volume: the volume of bytes dropped or ECN-marked in a given period of time. Counting congestion-volume allows each user to be held responsible for his or her contribution to causing congestion. Congestion-volume is a property of traffic, whereas congestion is a property of a link or a path.

To understand congestion-volume, consider a simple example. Imagine Alice sends 1GB while the loss-probability is a constant 0.2%. Her contribution to congestion -- her congestion-volume -- is 1GB x 0.2% = 2MB. If she then sends 3GB while the loss-probability is 0.1%, this adds 3MB to her congestion-volume. Her total contribution to congestion is then 2MB+3MB = 5MB.

Fortunately, measuring Alice's congestion-volume on a real network does not require the kind of arithmetic shown above because congestion-volume can be directly measured by counting the total volume of Alice's traffic that gets discarded or ECN-marked. (A queue with a percentage loss involves multiplication inherently.)

2.3. Rest-of-Path Congestion

At a particular measurement point within a network, "rest-of-path congestion" (also known as "downstream congestion") measures the level of congestion that a traffic flow is expected to experience between the measurement point and its final destination. "Upstream congestion" measures the level of congestion experienced up to the measurement point.

Measurement points that only observe ECN marks are capable of measuring upstream congestion, whereas measurement points that observe ConEx marks in addition to ECN marks can use both kinds of marks to calculate rest-of-path congestion. When ECN signals are monitored in the middle of a network, they indicate the level of congestion experienced so far on the path (upstream congestion). In contrast, the ConEx signals inserted into IP headers as shown in Figure 1 indicate the level of congestion along a whole path from source to destination. Therefore if a measurement point detects both of these signals, it can subtract the level of ECN (upstream congestion) from the level of ConEx (whole path) to derive a measure of the congestion that packets are likely to experience between the monitoring point and their destination (rest-of-path congestion).

[ConEx-Abstract-Mech] has further discussion of the constraints around the network's ability to measure rest-of-path congestion.

2.4. Definitions

Congestion: In general, congestion occurs when any user's traffic suffers loss, ECN marking, or increased delay as a result of one or more network resources becoming overloaded. For the purposes of ConEx, congestion is measured using the concrete signals provided by loss and ECN markings (delay is not considered). Congestion is measured as the probability of loss or the probability of ECN marking, usually expressed as a dimensionless percentage.

Congestion-volume: For any granularity of traffic (packet, flow, aggregate, link, etc.), the volume of bytes dropped or ECN-marked in a given period of time. Conceptually, data volume multiplied by the congestion each packet of the volume experienced. Usually expressed in bytes (or MB or GB).

Rest-of-path congestion (or downstream congestion): The level of congestion a flow of traffic is expected to experience on the remainder of its path. In other words, at a measurement point in the network the rest-of-path congestion is the level of congestion the traffic flow has yet to experience as it travels from that point to the receiver.

Upstream congestion: The accumulated level of congestion experienced by a traffic flow thus far, relative to a point along its path. In other words, at a measurement point in the network 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.

Network provider (or operator): Operator of a residential, commercial, enterprise, campus or other network.

User: The contractual entity that represents an individual, household, business, or institution that uses the service of a network provider. There is no implication that the contract has to be commercial; for instance, the users of a university or enterprise network service could be students or employees who do not pay for access but may be required to comply with some form of contract or acceptable use policy. There is also no implication that every user is an end user. Where two networks form a customer-provider relationship, the term user applies to the customer network.

[ConEx-Abstract-Mech] gives further definitions for aspects of ConEx related to protocol mechanisms.

3. Core Use Case: Informing Traffic Management

This section explains how ConEx could be used as the basis for traffic management, highlights additional benefits derived from having ConEx-aware nodes on the network, and compares ConEx-based traffic management to existing approaches.

3.1. Use Case Description

One of the key benefits that ConEx can deliver is in helping network operators to improve how they manage traffic on their networks. Consider the common case of a commercial broadband network where a relatively small number of users place disproportionate demand on network resources, at times resulting in congestion. The network operator seeks a way to manage traffic such that the traffic that contributes more to congestion bears more of the brunt of the management.
Assuming ConEx signals are visible at the IP layer, the operator can accomplish this by placing a congestion policer at an enforcement point within the network and configuring it with a traffic management policy that monitors each user's contribution to congestion. As described in [ConEx-Abstract-Mech] and elaborated in [CongPol], a congestion policer can be implemented in a similar way to a bit-rate policer, except that it monitors and polices congestion-volume rather than bit-rate. When implemented as a token bucket, the tokens provide users with the right

to cause bits of congestion-volume, rather than to send bits of data volume. The fill rate represents each user's congestion-volume quota. The congestion policer monitors the ConEx signals of the traffic entering the network. As long as the network remains uncongested and users stay within their quotas, no action is taken. When the network becomes congested and a user exhausts his quota, some action is taken against the traffic that breached the quota in accordance with the operator's traffic management policy. For example, the traffic may be dropped, delayed, or marked with a lower QoS class. In this way, traffic is managed according to its contribution to congestion -- not some application- or flow-specific policy -- and is not managed at all during times of no congestion.

As an example of how a network operator might employ a ConEx-based traffic management system, consider a typical DSL network architecture (as elaborated in [TR-059] and [TR-101]). Traffic is routed from regional and global IP networks to an operator-controlled IP node, the Broadband Remote Access Server (BRAS). From the BRAS, traffic is delivered to access nodes. The BRAS carries enhanced functionality including IP QoS and traffic management capabilities.

Based on typical network designs and current traffic patterns, the BRAS is located at a point in the network where congestion may be most likely to occur. As a consequence, the BRAS is a logical choice of location for deploying traffic management functionality. By deploying a congestion policer at the BRAS location, the operator can measure the congestion-volume created by users within the access nodes. The policer would be provisioned with a traffic management policy, perhaps directing the BRAS to drop packets from users that exceed their congestion-volume quotas during times of congestion. Those users would be expected to react in the typical way to drops, backing off (assuming use of standard TCP), and thereby lowering their congestion-volumes back within the quota limits.

3.2. Additional Benefits

The ConEx-based approach to traffic management has a number of benefits in addition to efficient management of traffic. It provides incentives for users to make use of scavenger transport protocols, such as [LEDBAT], that provide ways for bulk-transfer applications to rapidly yield when interactive applications require capacity. With a congestion policer in place as described in Section 3.1, users of these protocols will be less likely to run afoul of the operator's traffic management policy than those whose bulk-transfer applications generate the same volume of traffic without being sensitive to congestion.

ConEx-based traffic management also makes it possible for a user to control the relative performance among its own traffic flows. If a user wants some flows to have more bandwidth than others, it can allow the higher bandwidth traffic to generate more congestion signals, leaving less congestion "budget" for the user to "spend" on other traffic. This approach is most relevant if congestion is signalled by ECN, because no impairment due to loss is involved and delay can remain low.

3.3. Comparison with Existing Approaches

A variety of approaches already exist for network operators to manage congestion, traffic, and the disproportionate usage of scarce capacity by a small number of users. Common approaches can be categorized as rate-based, volume-based, or application-based.

Rate-based approaches constrain the traffic rate per user or per network. A user's peak and average (or "committed") rate may be limited. These approaches have the potential to either over- or under-constrain the network, suppressing rates even when the network is uncongested or not suppressing them enough during heavy usage periods. Round-robin scheduling and fair queuing were developed to address these problems. They equalize relative rates between active users (or flows) at a known bottleneck. The bit-rate allocated to any one user depends on the number of active users at each instant. The drawback of these approaches is that they favor heavy users over light users over time, because they do not have any memory of usage. Heavy users will be active at every instant whereas light users will only occupy their share of the link occassionally, but bit-rate is shared instant by instant.

Volume-based approaches measure the overall volume of traffic a user sends (and/or receives) over time. Users may be subject to an absolute volume cap (for example, 10GB per month) or the "heaviest" users may be sanctioned in some other manner. Many providers use monthly volume limits and count volume regardless of whether the network is congested or not, creating the potential for over- or under-constraining problems, as with the original rate-based approaches.

ConEx-based approaches, by comparison, only react during times of congestion and in proportion to each user's congestion contribution, making more efficient use of capacity and more proportionate management decisions.

Unlike ConEx-based approaches, neither rate-based nor volume-based approaches provide incentives for applications to use scavenger transports. They may even penalize users of applications that employ scavenger services for the large amount of volume they send, rather than rewarding them for carefully avoiding congestion while sending it. While the volume-based approach described in Comcast's Protocol-Agnostic Congestion Management System [RFC6057] aims to overcome the over/under-constraining problem by only measuring volume and triggering traffic management action during periods of high utilization, it still does not provide incentives to use scavenger transports because congestion-causing volume cannot be distinguished from volume overall. ConEx provides this ability.

Application-based approaches use deep packet inspection or other techniques 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, in some cases only during peak hours. These approaches suffer from being at odds with IPSec and some application-layer encryption, and they may raise additional policy concerns. In contrast, ConEx offers an application-agnostic metric to serve as the basis for traffic management decisions.

The existing types of approaches share a further limitation that ConEx can help to overcome: 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 connections is being altered or degraded based on how the network operator is attempting to manage congestion. By exposing congestion information at the IP layer, ConEx instead provides a metric that can serve as an open, transparent basis for traffic management policies that both providers and their customers can measure and verify.

4. Other Use Cases

ConEx information can be put to a number of uses other than informing traffic management. These include:

Informing inter-operator contracts: ConEx information is made visible to every IP node, including border nodes between networks. Network operators can use this information to measure how much traffic from each network contributes to congestion in the other. As such, congestion-volume could be included as a metric in inter-operator contracts, just as volume or bit-rate are included today.

Enabling more efficient capacity provisioning: Operators currently provision capacity based on observations of a number of network characteristics, including averaged utilization and congestion. Without ConEx, a user may have little incentive to back off during times of congestion, even if the reduction in performance resulting from backing off certain applications (bulk transfer, for example) would go largely unnoticed by the user. Using ConEx to ration congestion-volume directly creates incentives where appropriate for users and applications to switch to scavenger transports, resulting in traffic demand that more accurately reflects the actual capacity needed for the mix of applications on the network to perform well. This enables capacity to be provisioned more efficiently because traffic more closely tracks users' real capacity needs.

5. Deployment Arrangements

ConEx is designed so that it can be incrementally deployed in the Internet and still be valuable for early adopters. As long as some senders are ConEx-enabled, a network on the path can unilaterally use

ConEx-aware policy devices for traffic management; no changes to network forwarding elements are needed and ConEx still works if there are other networks on the path that are unaware of ConEx marks.

The above two steps seem to represent a stand-off where neither step is useful until the other has made the first move: i) some sending hosts must be modifed to give information to the network and ii) a network must deploy policy devices to monitor this information and act on it. Nonetheless, the developer of a scavenger transport protocol like LEDBAT does have a strong incentive to tell the network how little congestion it is causing despite sending large volumes of data. In this case the developer makes the first move expecting it will prompt at least some networks to move in response—so that they use the ConEx information to reward users of the scavenger protocol.

On the host side, we have already shown (Figure Figure 1) how the sender piggy-backs ConEx signals on normal data packets to re-insert feedback about packet drops (and/or ECN) back into the IP layer. In the case of TCP, [I-D.conex-tcp-mods] specifies the required sender modifications. ConEx works with any TCP receiver as long as it uses SACK, which most do. There is a receiver optimisation [I-D.conex-accurate-ecn] that improves ConEx precision when using ECN, but ConEx can still use ECN without it.

On the network side the operator solely needs to place ConEx congestion policers at each ingress to its network, in a similar arrangement to the edge-policed architecture of Diffserv [RFC2475].

A sender can choose whether to send ConEx or Not-ConEx packets. ConEx packets bring information to the policer about congestion expected on the rest of the path beyond the policer. Not-ConEx packets bring no such information. Therefore the network will tend to rate-limit not-ConEx packets conservatively in order to manage the unknown risk of congestion. In contrast, a network doesn't normally need to rate-limit ConEx-enabled packets unless they reveal a persistently high contribution to congestion. This natural tendency for networks to favour senders that provide ConEx information reinforces ConEx deployment.

The above gives only the most salient aspects of ConEx deployment. For further detail, [ConEx-Abstract-Mech] describes the incremental deployment features of the ConEx protocol and the components that need to be deployed for ConEx to work. Then [I-D.conex-init-deploy] gives concrete examples of feasible initial deployment scenarios.

6. Security Considerations

This document does not specify a mechanism, it merely motivates congestion exposure at the IP layer. Therefore security considerations are described in the companion document that gives an abstract description of the ConEx protocol and the components that would use it [ConEx-Abstract-Mech].

7. IANA Considerations

This document does not require actions by IANA.

8. Acknowledgments

Bob Briscoe was 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 the many people that have commented on this document: Bernard Aboba, Mikael Abrahamsson, João Taveira Araújo, Marcelo Bagnulo Braun, Steve Bauer, Caitlin Bestler, Steven Blake, Louise Burness, Ken Carlberg, Nandita Dukkipati, Dave McDysan, Wes Eddy, Matthew Ford, Ingemar Johansson, Georgios Karagiannis, Mirja Kuehlewind, Dirk Kutscher, Zhu Lei, Kevin Mason, Matt Mathis, Michael Menth, Chris Morrow, Tim Shepard, Hannes Tschofenig and Stuart Venters. Please accept our apologies if your name has been missed off this list.

8.1. Contributors

Philip Eardley and Andrea Soppera made helpful text contributions to this document.
The following co-edited this document through most of its life:

Toby Moncaster
Computer Laboratory
William Gates Building
JJ Thomson Avenue
Cambridge, CB3 0FD
UK
EMail: toby.moncaster@cl.cam.ac.uk

John Leslie
JLC.net
10 Souhegan Street
Milford, NH 03055
US
EMail: john@jlc.net

9. References

[RFC2475]	Blake, S., Black, D.L., Carlson, M.A., Davies, E., Wang, Z. and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
[RFC3168]	Ramakrishnan, K., Floyd, S. and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.
[RFC6057]	

	Bastian, C., Klieber, T., Livingood, J., Mills, J. and R. Woundy, "Comcast's Protocol-Agnostic Congestion Management System", RFC 6057, December 2010.
[LEDBAT]	Shalunov, S, Hazel, G, Iyengar, J and M Kuehlewind, "Low Extra Delay Background Transport (LEDBAT)", Internet-Draft draft-ietf-ledbat-congestion-08, May 2011.
[CongPol]	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.
[ConEx-Abstract-Mech]	Mathis, M and B Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", Internet-Draft draft-ietf-conex-abstract-mech-02, July 2011.
[I-D.conex-tcp-mods]	Kuehlewind, M and R Scheffenegger, "TCP modifications for Congestion Exposure", Internet-Draft draft-kuehlewind-conex-tcp-modifications-00, July 2011.
[I-D.conex-accurate-ecn]	Kuehlewind, M and R Scheffenegger, "Accurate ECN Feedback in TCP", Internet-Draft draft-kuehlewind-conex-accurate-ecn-00, July 2011.
[I-D.conex-init-deploy]	Briscoe, B, "Initial Congestion Exposure (ConEx) Deployment Examples", Internet-Draft draft-briscoe-conex-initial-deploy-00, October 2011.
[Bauer09]	Bauer, S, Clark, D and W Lehr, "The Evolution of Internet Congestion", 2009.
[TR-059]	Anschutz, T., "DSL Forum Technical Report TR-059: Requirements for the Support of QoS-Enabled IP Services", September 2003.
[TR-101]	Cohen, A. and E. Schrum, "DSL Forum Technical Report TR-101: Migration to Ethernet-Based DSL Aggregation", April 2006.

Appendix A. Changes from previous drafts (to be removed by the RFC Editor)

From draft-ietf-conex-concepts-uses-02 to -03: Reorganization and re-write of most sections.

From draft-ietf-conex-concepts-uses-01 to -02: New Abstract & Introduction. Concepts and Misconceptions sections added around definitions. Minor clarifications to Existing Traffic Management and

Use-Cases sections, with Other use Cases Added. Deployment Arrangements Section added.

From draft-ietf-conex-concepts-uses-00 to -01: Added section on timescales: Section 6

Revised introduction to clarify congestion definitions

Changed source for congestion definition in Section 2

Other minor changes

From draft-moncaster-conex-concepts-uses-02 to draft-ietf-conex-concepts-uses-00 (per decisions of working group):
Removed section on DDoS mitigation use case.

Removed appendix on ConEx Architectural Elements. PLEASE NOTE: Alignment of terminology with the Abstract Mechanism draft has been deferred to the next version.

From draft-moncaster-conex-concepts-uses-01 to draft-moncaster-conex-concepts-uses-02:
Updated document to take account of the new Abstract Mechanism draft [ConEx-Abstract-Mech].

Updated the definitions section.

Removed sections on Requirements and Mechanism.

Moved section on ConEx Architectural Elements to appendix.

Minor changes throughout.

From draft-moncaster-conex-concepts-uses-00 to draft-moncaster-conex-concepts-uses-01:
Changed end of Abstract to better reflect new title

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

Extensive re-write of use cases 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 5.5)

Added a discussion on ConEx as a Business Secret

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 capitalization 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 use cases section.

Authors' Addresses

Bob Briscoe editor Briscoe BT
B54/77, Adastral Park Martlesham Heath Ipswich, IP5 3RE UK Phone: +44 1473 645196 EMail: bob.briscoe@bt.com URI: http://bobbriscoe.net/

Richard Woundy editor Woundy Comcast 1701 John F Kennedy Boulevard Philadelphia, PA 19103 US EMail: richard_woundy@cable.comcast.com URI: http://www.comcast.com

Alissa Cooper editor Cooper CDT 1634 Eye St. NW, Suite 1100 Washington, DC 20006 US EMail: acooper@cdt.org