ConEx B. Briscoe
Internet-Draft BT
Intended status: Informational July 16, 2012
Expires: January 17, 2013

Initial Congestion Exposure (ConEx) Deployment Examples
draft-briscoe-conex-initial-deploy-03

Abstract

This document gives examples of how ConEx deployment might get started, focusing on unilateral deployment by a single network.

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 January 17, 2013.

Copyright Notice

Copyright (c) 2012 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

This document gives examples of how ConEx deployment might get started, focusing on unilateral deployment by a single network.

2. Recap: Incremental Deployment Features of the ConEx Protocol

The ConEx mechanism document [conex-abstract-mech] goes to great lengths to design for incremental deployment in all the respects below. It should be referred to for precise details on each of these points:

o The ConEx mechanism is essentially a change to the source, in order to re-insert congestion feedback into the network.

o Source-host-only deployment is possible without any negotiation required, and individual transport protocol implementations within a source host can be updated separately.

o Receiver modification may optionally improve ConEx for some transport protocols with feedback limitations (TCP being the main example), but it is not a necessity

o Proxies for the source and/or receiver are feasible (though not necessarily straightforward)

o Queues and network forwarding do not require any modification for ConEx.

o ECN is not required in the network for ConEx. If some network nodes support ECN, it can be used by ConEx.

o ECN is not required at the receiver for ConEx. The sender should nonetheless attempt to negotiate ECN-usage with the receiver, given some aspects of ConEx work better the more ECN is deployed, particularly auditing and border measurement.

o Given ConEx exposes information for IP-layer policy devices to use, the design does not preclude possible innovative uses of ConEx information by other IP-layer devices, e.g. forwarding itself

o Packets indicate whether or not they support ConEx.

3. ConEx Components

3.1. Recap of Basic ConEx Components

[conex-abstract-mech] introduces the following components:

- o The ConEx Wire Protocol (currently only specified for IPv6 [conex-destopt], although a possible way to fit ConEx into the IPv4 header has been described [intarea-ipv4-id-reuse])

- o Forwarding devices (unmodified)

- o Sender (modified for ConEx)

- o Receiver (optionally modified)

- o Audit

- o Policy Devices:

  * Rest-of-Path Congestion Monitoring Devices (using information from the ConEx wire protocol)

  * Congestion Policers (using rest-of-path congestion monitoring)

[conex-abstract-mech] should be referred to for definitions of each of these components and further explanation.

The goal of all these ConEx elements for this scenario is to expose information about congestion on the whole-path to a congestion-policer. A congestion-policer is nearly identical to a traditional token-bucket-based bit-rate policer except the tokens it fills with arrive at a rate that represents the volume of congestion that the customer is allowed to contribute to over time and tokens drain from the bucket at a rate dependent on the ConEx signals representing rest-of-path congestion. [CongPol] introduces congestion-policing and [conex-concepts-uses] explains the benefits of policing based on congestion-volume compared to methods like weighted round-robin traditionally used in a BRAS.

3.2. Per-Network Deployment Concepts

Network deployment-related definitions:

Internet Ingress: The first IP node a packet traverses that is outside the source's own network. In a residential access network scenario, for traffic from a home this is the first IP-aware node after the home access equipment. For Internet access from an

enterprise network this is the provider edge router.

Internet Egress: The last IP node a packet traverses before reaching the receiver's network.

ConEx-Enabled Network: A network whose edge nodes implement ConEx policy functions.

Each network can unilaterally choose to use any ConEx information given by those sources using ConEx, independently of whether other networks use it.

Typically, a network will use ConEx information by deploying a policy function at the ingress edge of its network to monitor arriving traffic and to act in some way on the congestion information in those packets that are ConEx-enabled. Actions might include policing, altering the class of service, or re-routing. Alternatively, less direct actions via a management system might include triggering capacity upgrades, triggering penalty clauses in contracts or levying charges between networks based on ConEx measurements.

Typically, a network using ConEx info will deploy a ConEx policy function near the ingress edge and a ConEx audit function near the egress edge. The segment of the path between a ConEx policy function and a ConEx audit function can be considered to be a ConEx-protected segment of the path. Assuming a network covers all its ingresses and egresses with policy functions and audit functions respectively, the network within this ring will be a ConEx-protected network.

Of course, because each edge device usually serves as both an ingress and an egress, the two functions are both likely to be present in each edge device.

4. Example Initial Deployment Arrangements

In all the deployment scenarios below, we assume that deployment starts with some data sources being modified with ConEx code. The rationale for this is that the developer of a scavenger transport protocol like LEDBAT has 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.

4.1. Single Receiving Network Scenario

The name 'Receiving Network' for this scenario merely emphasises that most data is arriving from connected networks and data centres and

being consumed by residential customers on this access network. Some data is of course also travelling in the other direction.

```
                                          DSLAMs __
                                              /|/     ,-.Home-a
                                        __/__| |-----(   )
                    ,-----.            /  \  | |---   `-'
       ,---.       /       \  ,------P/       \|\__
      /     \     '  Core   '/| BRAS |          __
     ( Peer  )-->-|P        | '------'       /|/
      \     /     |         |         _____| |---
       '---`      '         '\,------./     | |---
                   \ M     /  |BRAS  |       \|\__
                    `-----'   '------A\          __
                     |          P|     \      /|/
                    /|\         /|\     \__\_| |---   ,-.
                   ,---.        ,---.      / | |-----(   )
                  /Data \      /     \        \|\__   `-'Home-b
                 ( Centre)    (  CDN  )
                  \     /      \     /  Access Network
                   '---`        '---`  <-------------->
```

P=Congestion-Policer; M=Congestion-Monitor; A=Audit function

Figure 1: Single Receiving Network Scenario

Figure Figure 1 is an attempt to show the salient features of a ConEx deployment in a typical broadband access provider's network (within the constraints of ASCII art). Broadband remote access servers (BRASs) control access to the core network from the access network and vice versa. Home networks (and small businesses) connect to the access network, but only two are shown.

In this diagram, all data is travelling towards the access network of Home-b, from the Peer network, the Data centre, the CDN and Home-a. Data actually travels in both directions on all links, but only one direction is shown.

The data centre, core and access network are all run by the same network operator, but each is the responsibility of a different department with internal accounting between them. The content distribution network (CDN) is operated by a third party CDN provider, and of course the peer network is also operated by a third party.

This operator of the data centre, core and access network is the only one in the diagram to have deployed ConEx monitoring and policy devices at the edges of its network. However, it has not enabled ECN on any of its network elements and neither has any other network in

the diagram. The operator has deployed a congestion policing function (P) on the provider-edge router where the peer attaches to its core, on the BRAS where the CDN attaches and on the other BRAS where each of the residential customers like Home-a attach. On the provider-edge router where the data centre attaches it has deployed a congestion monitoring function (M). Each of these policing and monitoring functions handles the aggregate of all traffic traversing it, for all destinations.

The operator has deployed an audit function on each logical output port of the BRAS for each end-customer site like Home-b. The Audit function handles the aggregate of all traffic for that end-customer from all sources. For traffic in the opposite direction (e.g. from Home-b to Home-a, there would be equivalent policing (P) and audit (A) functions in the converse locations to those shown.

Some content sources in the CDN and in the data centre are using the ConEx protocol, but others are not. There is a similar situation for hosts attached to the Peer network and hosts in home networks like Home-a: some are sending ConEx packets at least for bulk data transports, while others are not.

4.1.1. ConEx Functions in the Single Receiving Network Scenario

Within the BRAS there are logical ports that model the rate of each access line from the DSLAM to each home network [TR-059], [TR-101]. They are fed by a shared queue that models the rate of the downstream link from the BRAS to the DSLAM (sometimes called the backhaul network). If there is congestion anywhere in the set of networks in Figure Figure 1 it is nearly always:

- o either self-congestion in the queues into the logical ports representing the access lines

- o or shared congestion in the shared queue on the BRAS that feeds them.

Any ConEx sources sending data through this BRAS will receive feedback about these losses from the destination and re-insert it as ConEx markings into the data. Figure 2 shows an example plot of the loss levels that might be seen at different monitoring points along a path between the data centre and home-b, for instance. The top half of the figure shows the loss probability within the BRAS consists of 0.1% at the shared queue and 0.2% self-congestion in the logical output port that models the access line, making 0.3% in total. This upper diagram also shows whole path congestion as signalled by the ConEx sender, which remains unchanged along the whole path at 0.3%.

The lower half of the figure shows (downstream congestion) = (whole path) - (upstream congestion). Upstream congestion can only be monitored locally where the loss actually happens (within the BRAS output queues). Nonetheless, given there is rarely loss anywhere else but within the BRAS, this limitation is not significant in this scenario. The lower half of the figure also shows the location of the policing and audit functions. Policing anywhere within or upstream ofthe BRAS will be based on the downstream congestion level of 0.3%. While Auditing within the BRAS but after all the queues can check that the whole path congestion signalled by ConEx is no less than the loss levels experienced within the BRAS itself.

```
      Data centre-->|<--core-->|<------BRAS--------->|<--Home--
                               |                      |
    ^loss                      |<-Shared->|<-Access->|
    |probability                 backhaul
    |
0.3%|- - - - - - - - - - - - - - - - - - - - +-----------------
    |      whole path congestion            |
    |                                       |
    |                                       |upstream
0.1%|                             +---------+congestion
    |                             |
   -O=============================+---------------------------->
                                                   monitoring point
    ^loss
    |probability   Policing                   Audit
    |                |                           |
    |                V                           |
0.3%|----------------O-------------+             |
    |                              |downstream   |
0.2%|                              +---------+   |
    |                              congestion|   |
    |                                        |   |
    |                                        |   V
   -O----------------------------------------+====O============-->
                                                   monitoring point
```

Figure 2: Example plot of loss levels along a path

4.1.2. Incentives to Unilaterally Deploy ConEx in a Receiving Network

Even a sending application that is modified to use ConEx can choose whether to send ConEx or Not-ConEx packets. Nonetheless, ConEx packets bring information to a policer about congestion expected on the rest of the path beyond the policer. Not-ConEx packets bring no such information. Therefore a network that has deployed ConEx policers 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 encourages senders to choose to use the ConEx protocol whenever they can.

In particular, high volume sources have the most incentive to deploy ConEx. This is because high volume sources (e.g. video download sites or peer-to-peer file-sharing) can gain by implementing a low 'weight' end-to-end transport (i.e. a less aggressive response to congestion than other transports). Then, although they send a large amount of volume, they need not contribute significantly to congestion. If the ISP currently limits data volume, or offers chargeable tiers based on data volume, such customers stand to gain considerably if they can encourage the ISP to limit usage based on congestion-volume instead of volume.

Figure 3 explains why this is the case. The plots show bit-rate on the vertical axis and time horizontally. A file transfer (e.g. the one labelled from customer 'b') is given a simplified representation as a rectangle, implying it runs at a set rate for a time, then completes. The maximum height of each plot represents the maximum capacity of the shared link across the backhaul network, which is typically the bottleneck in a broadband network. The hatched regions represent unused capacity. 'c' represents the high volume source that we intend to show has an incentive to deploy ConEx.

In the upper half of the figure, customers 'b' & 'c' both use transports with equal weights, which is why they are shown with equal rates when they both compete for the capacity of the line. 'c' sends larger files than 'b', so when 'b' completes each of its file, 'c' can use the full capacity of the line until 'b' starts the next file. In the lower half of the figure, 'c' uses a less aggressive (lower weight) transport, so whenever 'b' sends a file, 'c' yields more of its rate. This allows 'b' to complete its transfer earlier, so that 'c' can take up the full rate earlier. 'b' sends the same volume files (same area in the graph), just faster and therefore they complete sooner (tall & thin instead of shorter and wider). As a result, 'c' hardly finishes any later than in the upper diagram. However, 'c' will have contributed much less to congestion, and 'b' completes the majority of its file transfers much faster. 'b' has also contributed less to congestion.

As we have said, customer 'c' in particular stands to gain if the ISP bases usage-limits (or usage charges) on congestion-volume rather than volume. The ISP also has a strong incentive to reward customers like 'c', because they make the network performance appear far better

than before for customer's like 'b' (e.g. short Web transfers). However, the network cannot make this move until customers like 'c' expose congestion information (ConEx) that the ISP can use in its traffic management or contracts.

```
  ^ bit-rate
  |
  |---------------------------------------------------.--,--.--,-------
  |                                                   |/\|  |\/|
  |                         c                         |\/| b|/\| c
  |------.  ,-----.  ,-----.  ,-----.  ,-----.  ,-----./\|  |\/|  ,----
  | b    |  | b   |  | b   |  | b   |  | b   |  | b   |\/|  |/\|  | b
  |      |  |     |  |     |  |     |  |     |  |     |/\|  |\/|  |
  +------'--'-----'--'-----'--'-----'--'-----'--'-----'--'--'--'--'---->
                                                                    time

  ^ bit-rate
  |
  |---------------------------------------------------.--,--.--,-------
  |---.     ,---.    ,---.    ,---.    ,---.    ,---. |/\|  |\/|  ,---.
  |   |     |   |    |   |  c |   |    |   |    |   | |\/| b|/\| c|   |
  |   |     |   |    |   |    |   |    |   |    |   | |/\|  |\/|  |   |
  | b |     | b |    | b |    | b |    | b |    | b | |\/|  |/\|  | b |
  |   |     |   |    |   |    |   |    |   |    |   | |/\|  |\/|  |   |
  +---'-----'---'----'---'----'---'----'---'----'---'-'--'--'--'--'---'>
                                                                    time
```

Figure 3: Weighted congestion controls with equal weights (upper) and unequal (lower)

Of course, in reality there would be more than two customers. But this would mean that short transfers like 'b' stand to gain even more, as multiple larger files would be yielding at once.

We should point out that not all high-volume customers will be prepared to temporarily shift their usage out of the way as shown -- real-time video for instance would still use a higher weight (more aggressive) so as to ensure timely delivery. However, high volume applications with elastic (non-real-time) requirements are also common (e.g. video streaming, software downloads, etc)

We should also point out that a transport that is less agressive against other customers is similar but not quite the same as LEDBAT [ledbat-congestion]. LEDBAT does indeed yield more to other flows during congestion, but it is designed to only do this if the contention for resources is at a slow link, such as the customer's own home router. If the contention is at a fast link, such as a BRAS, LEDBAT is designed not to yield. This is because ISPs

currently give no reward to a transport that minimises congestion to others -- because they do not have the congestion information to be able to.

5. Security Considerations

6. IANA Considerations

This document does not require actions by IANA.

7. Conclusions

This document has introduced how congestion policing could be deployed at the broadband remote access servers in a typical broadband access network. Congestion policing uses ConEx markings introduced by data sources and packets discarded by the BRAS to determine rest-of-path congestion, and police traffic accordingly.

It has been shown that high-volume elastic data sources have a strong incentive to deploy ConEx speculatively in the expectation that they will be able to encourage their ISPs to account for their usage by congestion-volume, not volume. They can use a less aggressive transport and prove that they are contributing little to congestion despite sending a lot of volume. ISPs also have a strong incentive to use this ConEx information to encourage their elastic high-volume customers to use less agressive transports, given they improve the performance of all the other customers.

Without ConEx information, ISPs can only use volume as a metric of usage, which prevents the above virtuous circle from forming, perversely discouraging high-volume elastic customers from such friendly behaviour.

8. Acknowledgments

9. Comments Solicited

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

10. 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>.

[TR-059] Anschutz, T., Ed., "DSL Forum Technical Report TR-059: Requirements for the Support of QoS-Enabled IP Services", September 2003.

[TR-101] Cohen, A., Ed. and E. Shrum, Ed., "Migration to Ethernet-Based DSL Aggregation", April 2006.

[conex-abstract-mech] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-conex-abstract-mech-05 (work in progress), July 2012.

[conex-concepts-uses] Briscoe, B., Woundy, R., and A. Cooper, "ConEx Concepts and Use Cases", draft-ietf-conex-concepts-uses-04 (work in progress), March 2012.

[conex-destopt] Krishnan, S., Kuehlewind, M., and C. Ucendo, "IPv6 Destination Option for Conex", draft-ietf-conex-destopt-02 (work in progress), March 2012.

[intarea-ipv4-id-reuse] Briscoe, B., "Reusing the IPv4 Identification Field in Atomic Packets", draft-briscoe-intarea-ipv4-id-reuse-01 (work in progress), March 2012.

[ledbat-congestion] Hazel, G., Iyengar, J., Kuehlewind, M., and S. Shalunov, "Low Extra Delay Background Transport (LEDBAT)", draft-ietf-ledbat-congestion-09 (work in progress), October 2011.

Appendix A. Summary of Changes between Drafts

Detailed changes are available from http://tools.ietf.org/html/draft-briscoe-conex-initial-deploy

From draft-briscoe-02 to draft-briscoe-03:

* Removed Mobile and Data Centre scenarios, making this draft solely cover the receiving access network scenario. It then becomes a 'sibling' of the drafts on these two subjects, rather than a 'parent'

* Consequently Dirk Kutscher is no longer a co-author

* Included more comprehensive background information on ConEx

* Completed Incentives section

* Updated refs

From draft-briscoe-01 to draft-briscoe-02:

* Added Mobile Scenario section, and Dirk Kutscher as co-author;

From draft-briscoe-00 to draft-briscoe-01: Re-issued without textual change. Merely re-submitted to correct a processing error causing the whole text of draft-00 to be duplicated within the file.

Author's Address

Bob 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/