ConEx
Internet-Draft
Intended status: Informational
Expires: April 26, 2012

B. Briscoe
BT

October 24, 2011

Initial Congestion Exposure (ConEx) Deployment Examples
draft-briscoe-conex-initial-deploy-00

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 April 26, 2012.

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:

- The ConEx mechanism is essentially a change to the source, in order to re-insert congestion feedback into the network.
- Source-host-only deployment is possible without any negotiation required, and individual transport protocol implementations within a source host can be updated separately.
- Receiver modification may optionally improve ConEx for some transport protocols with feedback limitations (TCP being the main example), but it is not a necessity
- Proxies for the source and/or receiver are feasible (though not necessarily straightforward)
- Queues and network forwarding do not require any modification for ConEx.
- ECN is not required in the network for ConEx. If some network nodes support ECN, it can be used by ConEx.
- 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.
- 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
- 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:

- The ConEx Wire Protocol
- Forwarding devices (unmodified)
- Sender (modified for ConEx)
- Receiver (optionally modified)
- Audit
- Policy Devices:
  - Rest-of-Path Congestion Monitoring Devices
  - Congestion Policers

[ConEx-Abstract-Mech] should be referred to for definitions of each of these components and further explanation.

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 domestic network that will be the first node downstream from the home access equipment. In 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.

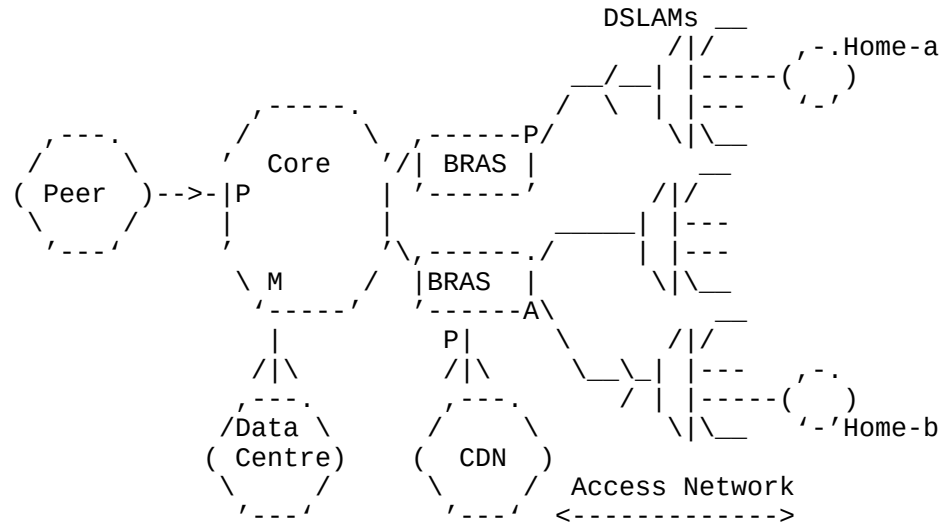
Of course, because each edge device usually serves as both an ingress and an egress, the two functions are both likely to be present in each edge device.

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.



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]. 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:

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

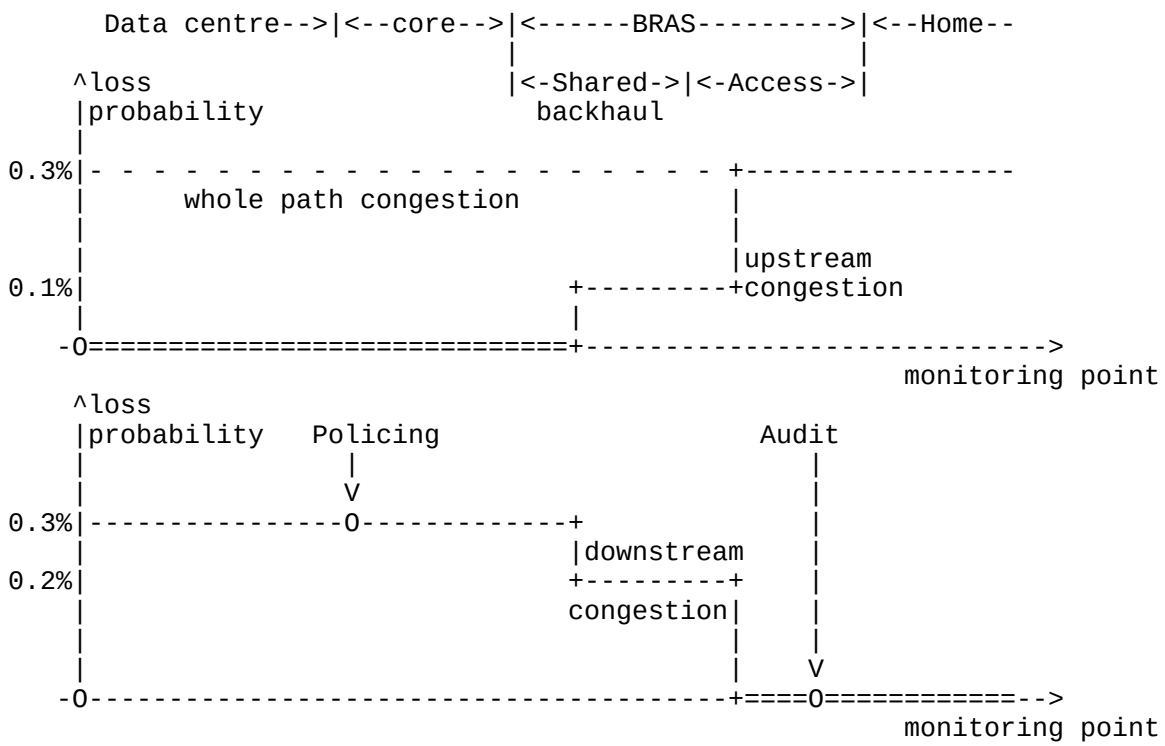


Figure 2: Example plot of loss levels along a path

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. Nonethelss, 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.

{ToDo: complete this section}

4.2. Mobile Network Scenario

Placeholder for summary of the scenario in a mobile network described in [conex-mobile]

In mobile networks, both mobile terminals and mobile network equipment are standardised by the 3GPP. If the 3GPP were to adopt the ConEx protocol, it might mandate ConEx implementation for compliant equipment.

{ToDo: Describe how a central traffic management box can arrange to remotely view upstream congestion as it would be seen from the interface with the mobile terminal.}

4.3. Scenario Internal to a Multi-Tenant Data Centre

A number of companies offer hosting of virtual machines on their data centre infrastructure--so-called infrastructure as a service (IaaS). A set amount of processing power, memory, storage and network are offered. Although processing power, memory and storage are relatively simple to allocate on the 'pay as you go' basis that has become common, the network is less easy to allocate given it is a naturally distributed system.

{ToDo: Complete this section.}

5. Security Considerations

6. IANA Considerations

This document does not require actions by IANA.

7. Conclusions

{ToDo}

8. Acknowledgments

9. Informative References

[ConEx-Abstract-Mech] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-conex-abstract-mech-02 (work in progress), July 2011.

[Seawall] Shieh, A., Kandula, S., Greenberg, A., and C. Kim, "Seawall: Performance Isolation in Cloud Datacenter Networks", Proc 2nd USENIX Workshop on Hot Topics in Cloud Computing , June 2010, <http://research.microsoft.com/en-us/projects/seawall/>.

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

[conex-mobile] Kutscher, D., Mir, F., Winter, R., Krishnan, S., and Y. Zhang, "Mobile Communication Congestion Exposure Scenario", draft-kutscher-conex-mobile-00 (work in progress), March 2011.

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/

ConEx
Internet-Draft
Intended status: Informational
Expires: April 26, 2012

B. Briscoe
BT

October 24, 2011

Initial Congestion Exposure (ConEx) Deployment Examples
draft-briscoe-conex-initial-deploy-00

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 April 26, 2012.

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:

- The ConEx mechanism is essentially a change to the source, in order to re-insert congestion feedback into the network.
- Source-host-only deployment is possible without any negotiation required, and individual transport protocol implementations within a source host can be updated separately.
- Receiver modification may optionally improve ConEx for some transport protocols with feedback limitations (TCP being the main example), but it is not a necessity
- Proxies for the source and/or receiver are feasible (though not necessarily straightforward)
- Queues and network forwarding do not require any modification for ConEx.
- ECN is not required in the network for ConEx. If some network nodes support ECN, it can be used by ConEx.
- 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.
- 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
- 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:

- The ConEx Wire Protocol
- Forwarding devices (unmodified)
- Sender (modified for ConEx)
- Receiver (optionally modified)
- Audit
- Policy Devices:
  - Rest-of-Path Congestion Monitoring Devices
  - Congestion Policers

[ConEx-Abstract-Mech] should be referred to for definitions of each of these components and further explanation.

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 domestic network that will be the first node downstream from the home access equipment. In 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.

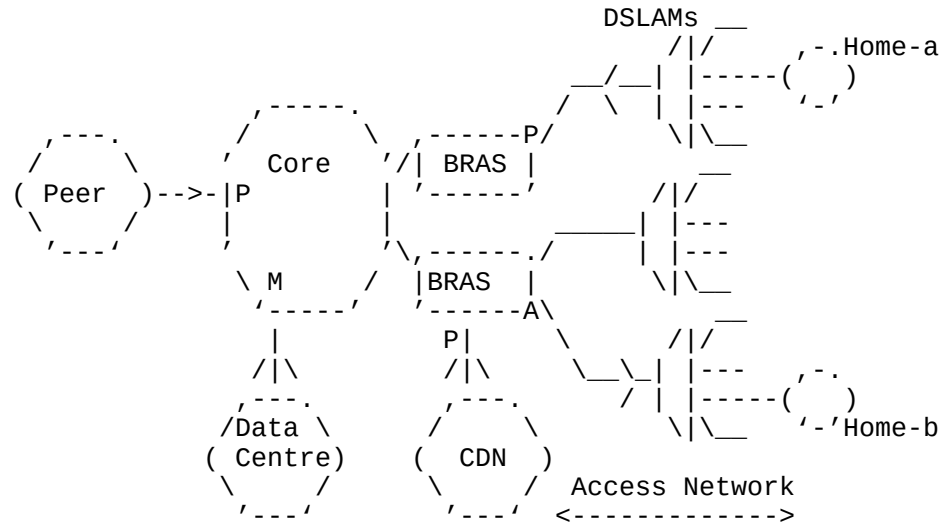
Of course, because each edge device usually serves as both an ingress and an egress, the two functions are both likely to be present in each edge device.

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.



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]. 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:

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

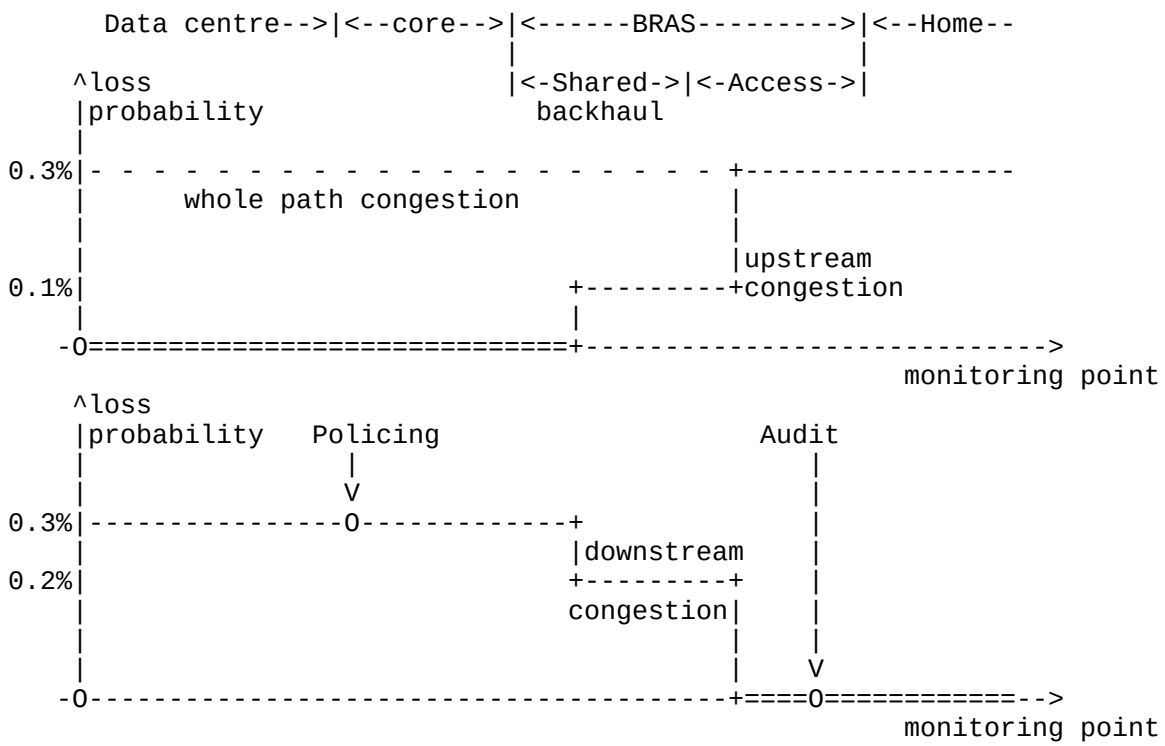


Figure 2: Example plot of loss levels along a path

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. Nonethelss, 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.

{ToDo: complete this section}

4.2. Mobile Network Scenario

Placeholder for summary of the scenario in a mobile network described in [conex-mobile]

In mobile networks, both mobile terminals and mobile network equipment are standardised by the 3GPP. If the 3GPP were to adopt the ConEx protocol, it might mandate ConEx implementation for compliant equipment.

{ToDo: Describe how a central traffic management box can arrange to remotely view upstream congestion as it would be seen from the interface with the mobile terminal.}

4.3. Scenario Internal to a Multi-Tenant Data Centre

A number of companies offer hosting of virtual machines on their data centre infrastructure--so-called infrastructure as a service (IaaS). A set amount of processing power, memory, storage and network are offered. Although processing power, memory and storage are relatively simple to allocate on the 'pay as you go' basis that has become common, the network is less easy to allocate given it is a naturally distributed system.

{ToDo: Complete this section.}

5. Security Considerations

6. IANA Considerations

This document does not require actions by IANA.

7. Conclusions

{ToDo}

8. Acknowledgments

9. Informative References

[ConEx-Abstract-Mech] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-conex-abstract-mech-02 (work in progress), July 2011.

[Seawall] Shieh, A., Kandula, S., Greenberg, A., and C. Kim, "Seawall: Performance Isolation in Cloud Datacenter Networks", Proc 2nd USENIX Workshop on Hot Topics in Cloud Computing , June 2010, <http://research.microsoft.com/en-us/projects/seawall/>.

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

[conex-mobile] Kutscher, D., Mir, F., Winter, R., Krishnan, S., and Y. Zhang, "Mobile Communication Congestion Exposure Scenario", draft-kutscher-conex-mobile-00 (work in progress), March 2011.

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/

ConEx
Internet-Draft
Intended status: Informational
Expires: January 18, 2013

B. Briscoe, Ed.
BT
R. Woundy, Ed.
Comcast
A. Cooper, Ed.
CDT

July 17, 2012

ConEx Concepts and Use Cases
draft-ietf-conex-concepts-uses-05

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 marking 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 by the ConEx marking can serve as the basis for significantly more efficient and effective traffic management than what exists on the Internet today.

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 18, 2013.

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 marking at the IP layer. (A companion document, [I-D.ietf-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.

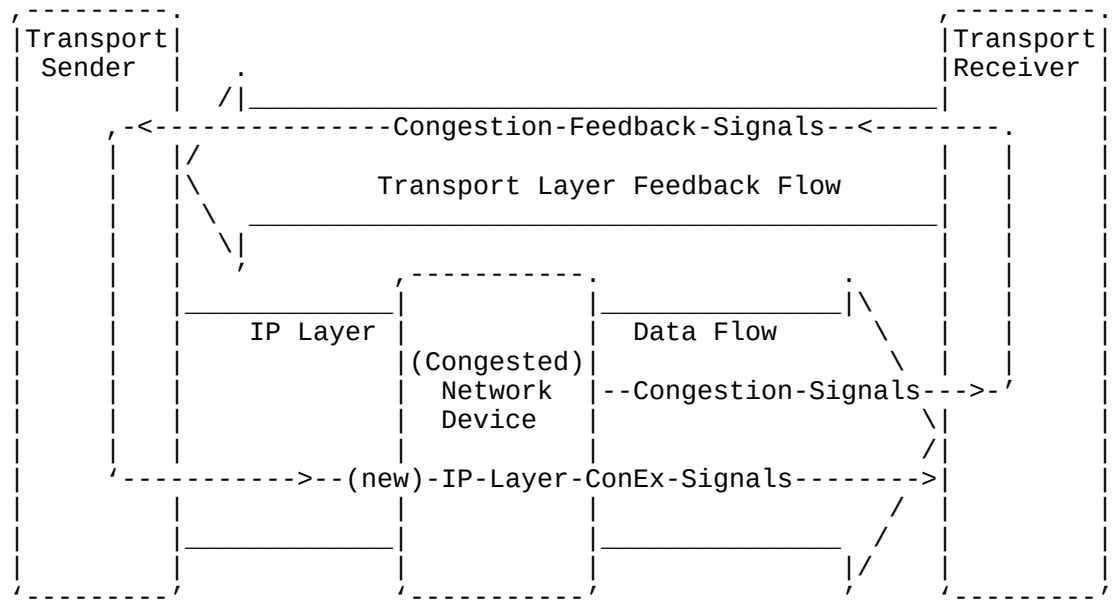


Figure 1: The ConEx Protocol in the Internet Architecture

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. A ConEx-enabled sender signals expected whole path congestion, which is approximately the congestion at least a round trip time earlier as reported by the receiver to the sender (Figure 1). 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 network 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.

The remainder of the core ConEx document suite consists of:

[I-D.ietf-conex-abstract-mech], which provides an abstract encoding of ConEx signals, explains the ConEx audit and security mechanisms, and describes incremental deployment features;

[I-D.ietf-conex-destopt], which specifies the IPv6 destination option encoding for ConEx;

[I-D.ietf-conex-tcp-modifications], which specifies TCP sender modifications for use of ConEx;

and the following documents, which describe some feasible scenarios for deploying ConEx:

[I-D.briscoe-conex-initial-deploy], which describes a scenario around a fixed broadband access network;

[I-D.ietf-conex-mobile], which describes a scenario around a mobile communications provider;

[I-D.briscoe-conex-data-centre], which describes how ConEx could be used for performance isolation between tenants of a data centre.

2. Concepts

ConEx relies on a precise definition of congestion and a number of newer concepts that are introduced in this section. Definitions are summarized in Section 2.4.

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 congestion. Congestion-volume can only be a property of traffic, whereas congestion can be a property of traffic or a property of a link or a path.

To understand congestion-volume, consider a simple example. Imagine Alice sends 1GB of a file 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 another 3GB of the file 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 varying percentage loss does these multiplications and additions inherently.) With ConEx, network operators can count congestion-volume using techniques very similar to those they use for counting volume.

2.3. Rest-of-Path Congestion

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

If traffic is ECN-capable, ECN signals monitored in the middle of a network will indicate the 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 congestion along a whole path from transport source to transport 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). A measurement point can calculate this measurement in the aggregate, across all flows.

A network monitor can usually accurately measure upstream congestion only if the traffic it observes is ECN-capable. [I-D.ietf-conex-abstract-mech] has further discussion of the constraints around the network's ability to measure upstream and rest-of-path congestion in these circumstances. However, there are a number of intial deployment arrangements that benefit from ConEx but work without ECN (see Section 5).

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).

Congestion policer: A logical entity that allows a network operator to monitor each user's congestion-volume and enforce congestion-volume limits (discussed in Section 3.1).

Rest-of-path congestion (or downstream congestion): The 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 congestion the traffic flow has yet to experience as it travels from that point to the receiver. This is usually expressed as a dimensionless percentage.

Upstream congestion: The accumulated 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 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. This is usually expressed as a dimensionless percentage.

Network operators (or providers): 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 operator. 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.

[I-D.ietf-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 network 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 [I-D.ietf-conex-abstract-mech] and elaborated in [CongPol], one way to implement a congestion policer is in a similar way to a bit-rate policer, except that it monitors congestion-volume (based on IP layer ConEx signals) 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 network 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.

By deploying a congestion policer at the BRAS location, the network operator can measure the congestion-volume created by users within the access nodes and police misbehaving users before their traffic

affects others on the access network. 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' apps would be likely to react in the typical way to drops, backing off (assuming at least some use TCP), and thereby lowering the users' congestion-volumes back within the quota limits. If none of a user's apps responds, the policer would continue to increase focused drops and effectively enforce its own congestion control.

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 [I-D.ietf-ledbat-congestion], that provide ways for bulk-transfer applications to rapidly yield when interactive applications require capacity (thereby "scavenging" remaining bandwidth). 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 network operator's traffic management policy than those whose bulk-transfer applications generate the same volume of traffic without being sensitive to congestion. In short, two users who produce similar traffic volumes over the same time interval may produce different congestion-volumes if one of them is using a scavenger transport protocol and the other is not; in that situation the scavenger user's traffic is less likely to be managed by the network operator.

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 reduce the rate of some traffic so that it consumes less congestion-volume "budget", leaving more congestion-volume "budget" for the user to "spend" on making other traffic go faster. 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 transport protocols. They may even penalize users of applications that employ scavenger transports 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. Network 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. It can be used to reduce the performance uncertainty that some users currently experience.

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 ConEx combined with ECN markings 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. This would not be an initial deployment scenario, unless ECN became widely deployed.

Enabling more efficient capacity provisioning: Section 3.2 explained how operators can use ConEx-based traffic management to encourage use of scavenger transport protocols, which significantly improves the performance of interactive applications while still allowing heavy users to transfer high volumes. Here we explain how this can also benefit network operators.

Today, when loss, delay or averaged utilization exceeds a certain threshold, some operators just buy more capacity without attempting to manage the traffic. Other operators prefer to limit a minority of heavy users at peak times, but they still eventually buy more capacity when utilization rises.

With ConEx-based traffic management, a network operator should be able to provision capacity more efficiently. An operator could benefit from this in a variety of ways. For example, the operator could add capacity as it would do without ConEx, but deliver

better quality of service for its users. Or the operator could delay adding capacity while delivering similar quality of service to what it currently provides.

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 stand to benefit from deploying ConEx. In this case the developer makes the first move, expecting it will prompt at least some networks to move in response, using the ConEx information to reward users of the scavenger transport protocol.

On the host side, we have already shown (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.ietf-conex-tcp-modifications] proposes 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.tcpm-accurate-ecn] that improves ConEx precision when using ECN, but ConEx can still use ECN without it. Networks can make use of ConEx even if the implementations of some of the transport protocols on a host do not support ConEx (e.g. the implementation of DNS over UDP might not support ConEx, while perhaps RTP over UDP and TCP will).

On the network side the provider 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 packets that support ConEx or packets that don’t. ConEx-enabled packets bring information to the policer about congestion expected on the rest of the path beyond the policer. Packets that do not support ConEx bring no such information. Therefore the network will tend to conservatively rate-limit non-ConEx-enabled packets 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.

Feasible initial deployment scenarios exist for a broadband access network [I-D.briscoe-conex-initial-deploy], a mobile communications network [I-D.ietf-conex-mobile], and a multi-tenant data centre [I-D.briscoe-conex-data-centre]. The first two of these scenarios are believed to work well without ECN support, while the data center scenario works best with ECN (where it may be more likely for ECN to be deployed in the future).

The above gives only the most salient aspects of ConEx deployment. For further detail, [I-D.ietf-conex-abstract-mech] describes the incremental deployment features of the ConEx protocol and the components that need to be deployed for ConEx to work.

6. Experimental Considerations

ConEx is initially designed as an experimental protocol because it makes an ambitious change at the interoperability (IP) layer, so no amount of careful design can foresee all the potential feature interactions with other uses of IP. This section identifies a number of questions that would be useful to answer through well-designed experiments:

- Are the compromises that were made in order to fit the ConEx encoding into IP (for example, that the initial design was solely for IPv6 and not for IPv4, and that the encoding has limited visibility when tunnelled [I-D.ietf-conex-destopt]) the right ones?

- Is it possible to combine techniques for distinguishing self-congestion from shared congestion with ConEx-based traffic management such that users are not penalized for congestion that does not impact others on the network? Are other techniques needed?

- If ECN deployment remains patchy, are the proposed initial ConEx deployment scenarios (Section 5) still useful enough to kick-start deployment? Is audit effective when based on loss at a primary bottleneck? Can rest-of-path congestion be approximated accurately enough without ECN? Are there other useful deployment scenarios?

- In practice, how does traffic management using ConEx compare with traditional techniques (Section 3.3)? Does it give the benefits claimed in Section 3.1 and Section 3.2?

- Approaches are proposed for congestion policing of ConEx traffic alongside existing management (or lack thereof) of non-ConEx traffic, including UDP traffic [I-D.ietf-conex-abstract-mech]. Are they strategy-proof against users selectively using both? Are there better transition strategies?

- Audit devices have been designed and implemented to assure ConEx signal integrity [I-D.ietf-conex-abstract-mech]. Do they achieve minimal false hits and false misses in a wide range of traffic scenarios? Are there new attacks? Are there better audit designs to defend against these?

ConEx is intended to be a generative technology that might be used for unexpected purposes unforeseen by the designers. Therefore this list of experimental considerations is not intended to be exhaustive.

7. 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 [I-D.ietf-conex-abstract-mech].

8. IANA Considerations

This document does not require actions by IANA.

9. 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, Joao Taveira Araujo, 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.

9.1. 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

10. Informative References

[Bauer09] Bauer, S., Clark, D., and W. Lehr, "The Evolution of Internet Congestion", 2009.

[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.

[I-D.briscoe-conex-data-centre] Briscoe, B. and M. Sridharan, "Network Performance Isolation in Data Centres using Congestion Exposure (ConEx)", draft-briscoe-conex-data-centre-00 (work in progress), July 2012.

[I-D.briscoe-conex-initial-deploy] Briscoe, B., "Initial Congestion Exposure (ConEx) Deployment Examples", draft-briscoe-conex-initial-deploy-02 (work in progress), March 2012.

[I-D.ietf-conex-abstract-mech] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-conex-abstract-mech-04 (work in progress), March 2012.

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

[I-D.ietf-conex-mobile] Kutscher, D., Mir, F., Winter, R., Krishnan, S., Zhang, Y., and C. Bernardos, "Mobile Communication Congestion Exposure Scenario", draft-ietf-conex-mobile-00 (work in progress), July 2012.

[I-D.ietf-conex-tcp-modifications] Kuehlewind, M. and R. Scheffenegger, "TCP modifications for Congestion Exposure", draft-ietf-conex-tcp-modifications-02 (work in progress), May 2012.

[I-D.ietf-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.

[I-D.tcpm-accurate-ecn] Kuehlewind, M. and R. Scheffenegger, "Accurate ECN Feedback Option in TCP", draft-kuehlewind-tcpm-accurate-ecn-option-01 (work in progress), July 2012.

[RFC2475] Blake, S., Black, D., Carlson, M., 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.

[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. Schrum, Ed., "DSL Forum Technical Report TR-101: Migration to Ethernet-Based DSL Aggregation", April 2006.

Authors' Addresses

Bob Briscoe (editor)
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)
Comcast
1701 John F Kennedy Boulevard
Philadelphia, PA 19103
US

EMail: richard_woundy@cable.comcast.com
URI: http://www.comcast.com

Alissa Cooper (editor)
CDT
1634 Eye St. NW, Suite 1100
Washington, DC 20006
US

EMail: acooper@cdt.org

Congestion Exposure (ConEx)
Internet-Draft
Intended status: Experimental
Expires: May 3, 2012

M. Kuehlewind, Ed.
University of Stuttgart
R. Scheffenegger
NetApp, Inc.

October 31, 2011

TCP modifications for Congestion Exposure
draft-kuehlewind-conex-tcp-modifications-01

Abstract

Congestion Exposure (ConEx) is a mechanism by which senders inform the network about the congestion encountered by previous packets on the same flow. This document describes the necessary modifications to use ConEx with the Transmission Control Protocol (TCP).

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 May 3, 2012.

Copyright Notice

Table of Contents

1. Introduction

Congestion Exposure (ConEx) is a mechanism by which senders inform the network about the congestion encountered by previous packets on the same flow. This document describes the necessary modifications to use ConEx with the Transmission Control Protocol (TCP). The ConEx signal is based on loss or ECN marks [RFC3168] as a congestion indication.

With standard TCP without Selective Acknowledgments (SACK) [RFC2018] the actual number of losses is hard to detect, thus we recommend to enable SACK when using ConEx. However, we discuss both cases, with and without SACK support, later on.

Explicit Congestion Notification (ECN) is defined in such a way that only a single congestion signal is guaranteed to be delivered per Round-trip Time (RTT). For ConEx a more accurate feedback signal would be beneficial. Such an extension to ECN is defined in a seperate document [draft-kuehlewind-conex-accurate-ecn], as it can also be useful for other mechanisms, as e.g. [DCTCP] or whenever the congestion control reaction should be proportional to the expirienced congestion.

ConEx is currently/will be defined as an destination option for IPv6. The use of four bits have been defined, namely the X (ConEx-capable), the L (loss experienced), the E (ECN experienced) and C (credit) bit.

1.1. Requirements Language

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].

2. Sender-side Modifications

A ConEx sender MUST negotitate for both SACK and the more accurate ECN feedback in the TCP handshake if these TCP extension are available at the sender. Depending on the capability of the receiver, the following operation modes exist:

- Full-ConEx (SACK and accurate ECN feedback)
- accECN-ConEx (no SACK but accurate ECN feedback)
- ECN-ConEx (no SACK and no accurate ECN feedback but ’classic’ ECN)

- SACK-ECN-ConEx (SACK and 'classic' instead of accurate ECN)
- SACK-ConEx (SACK but no ECN at all)
- Basic-ConEx (neither SACK nor ECN)

A ConEx sender MUST expose congestion to the network according to the congestion information received by ECN or based on loss provided by the TCP feedback loop. A TCP sender MUST account congestion byte-wise (and not packet-wise) and MUST mark the respective number of payload bytes in subsequent packets (after the congestion notification) with the respective ConEx bit in the IP header. The congestion accounting based on different operation modes is described in the next section and the handling of the IPv6 bits itself in the subsequent section afterwards.

3. Accounting congestion

A TCP sender MUST account congestion byte-wise (and not packet-wise) based the congestion information received by ECN or loss detection provided by TCP. For this purpose a TCP sender will maintain two different counters for number outstanding bytes that need to be ConEx marked either with the E bit or the L Bit.

The outstanding bytes accounted based on ECN feedback information are maintained in the congestion exposure gauge (CEG). The accounting of these bytes from the ECN feedback is explained in more detail next.

The outstanding bytes for congestion indications based on loss are maintained in the loss exposure gauge (LEG) and the accounting is explained in subsequent to the CEG accounting.

The subtraction of bytes which have been ConEx marked from both counters is explained in the next section.

Usually all byte of an IP packet must be accounted. If we assume equal sized packets or at least equally distributed packet sizes the sender MAY only account the TCP payload bytes, as the ConEx marked packets as well as the original packets causing the congestion will both contain about the same number of headers. Otherwise the sender MUST take the headers into account. A sender which sends different sized packets with unequally distributed packet sizes should know about reason to do so and thus may be able to reconstruct the exact number of headers based on this information. Otherwise if no additional information is available the worse case number of headers SHOULD be estimated in a conservative way based on a minimum packet size (of all packets sent in the last RTT).

3.1. ECN

A receiver can support the accurate ECN feedback scheme, the 'classic' ECN or neither. In the case ECN is not supported at all, the transport is not ECN-capable and no ECN marks will occur, thus the E bit will never be set. In the other cases a ConEx sender MUST maintain a gauge for the number of outstanding bytes that has to be ConEx marked with the E bit, the congestion exposure gauge (CEG).

The CEG is increased when ECN information is received from an ECN-capable receiver supporting the 'classic' ECN scheme or the accurate ECN feedback scheme. When the ConEx sender receives an ACK indicating one or more segments were received with a CE mark, CEG is increased by the appropriate number of bytes. The two cases, depending on the receiver capability, are discussed in the following sections.

3.1.1. Accurate ECN feedback

With an more accurate ECN feedback scheme either the number of marked packets/received CE marks is know or the number of marked bytes directly. In the later case the CEG can directly be increased by the number of marked bytes. Otherwise when the accurate ECN feedback scheme is supported by the receiver, the receiver will maintain an echo congestion counter (ECC). The ECC will hold the number of CE marks received. A sender that is understanding the accurate ECN feedback will be able to reconstruct this ECC value on the sender side by maintaining a counter ECC.r.

On the arrival of every ACK, the sender calculates the difference D between the local ECC.r counter, and the signaled value of the receiver side ECC counter. The value of ECC.r is increased by D, and D is assumed to be the number of CE marked packets that arrived at the receiver since it sent the previously received ACK.

Whenever the counter ECC.r is increased, the gauge CEG has to be increased by the amount of bytes sent which were marked:

```
CEG += min( SMSS*D, acked_bytes )
```

3.1.2. Classic ECN support

A ConEx sender that communicates with a classic ECN receiver (conforming to [RFC3168] or [RFC5562]) MAY run in one of these modes:

- Full compliance mode:

  The ConEx sender fully conforms to all the semantics of the ECN

signaling as defined by [RFC5562]. In this mode, only a single congestion indication can be signaled by the receiver per RTT. Whenever the ECE flag toggles from "0" to "1", the gauge CEG is increased by the SMSS:

CEG += SMSS

Note that under severe congestion, a session adhering to these semantics may not provide enough ConEx marks. This may cause appropriate sanctions by an audit device in a ConEx enabled network.

- Simple compatibility mode:

  The sender will set the CWR permanently to force the receiver to signal only one ECE per CE mark. Unfortunately, in a high congestion situation where all packets are CE marled over a certain period of time, the use of delayed ACKs, as it is usually done today, will prevent a feedback of every CE mark. With an ACK rate of m, about m-1/m CE indications will not be signaled back by the receiver (e.g. 50% with M=2). Thus, in this mode the ConEx sender MUST increase CEG by a count of M*SMSS for each received ECE signal:

  CEG += M*SMSS

  In case of a congestion event with low congestion (that means when only a very smaller number of packets get marked), the sender might miss the whole congestion event. In average the sender will sent sufficient ConEx marks due to the scheme proposed above but these ConEx marks might be timely shifted. Regarding congestion control it is not a general problem to miss a congestion event as by chance a marking scheme in the network node might also miss a certain flow. Even if then no other flow is reacting, the congestion level will increase and it will get more likely that the congestion feedback is delivered. But to provide a fair share over time, a TCP sender could react more strong when receiving a ECN feedback signal. This of course depends on the congestion control used. A TCP sender using this scheme MUST take the impact on congestion control into account.

- Advanced compatibility mode:

  More sophisticated heuristics, such as a phase locked loop, to set CWR only on those data segments, that will actually trigger an (delayed) ACK, could extract congestion notifications more timely. A ConEx sender MAY choose to implement such a heuristic. In addition, further heuristics SHOULD be implemented, to determine

the value of each ECE notification. E.g. for each consecutive ACK received with the ECE flag set, CEG should be increased by min(M*SSMS, acked_bytes). Else if the predecessor ACK was received with the ECE flag cleared, CEG need only be increase by one SMSS:

```
if previous_marked: CEG += min( M*SSMS, acked_bytes)
else: CEG += SMSS
```

This heuristic is conservative during more serious congestion, and more relaxed at low congestion levels.

3.2. Loss Detection with/without SACK

For all the data segments that are determined by a ConEx sender as lost, an identical number of IP bytes MUST be be sent with the ConEx L bit set. Loss detection typically happens by use of duplicate ACKs, or the firing of the retransmission timer. A ConEx sender MUST maintain a loss exposure gauge (LEG), indicating the number of outstanding bytes that must be sent with the ConEx L bit. When a data segment is retransmitted, LEG will be increased by the size of the TCP payload packet containing the retransmission, assuming equal sized segments such that the retransmitted packet will have the same number of header as the original ones. When sending subsequent segments (including TCP control segments), the ConEx L bit is set as long as LEG is positive, and LEG is decreased by the size of the sent TCP payload with the ConEx L bit set.

Any retransmission may be spurious. To accommodate that, a ConEx sender SHOULD make use of heuristics to detect such spurious retransmissions (e.g. F-RTO [RFC5682], DSACK [RFC3708], and Eifel [RFC3522], [RFC4015]). When such a heuristic has determined, that a certain number of packets were retransmitted erroneously, the ConEx sender should subtract the payload size of these TCP packets from LEG.

Note that the above heuristics delays the ConEx signal by one segment, and also decouples them from the retransmissions themselves, as some control packets (e.g. pure ACKs, window probes, or window updates) may be sent in between data segment retransmissions. A simpler approach would be to set the ConEx signal for each retransmitted data segment. However, it is important to remember, that a ConEx signal and TCP segments do not natively belong together.

4. Setting the ConEx IPv6 Bits

ConEx is currently/will be defined as an destination option for IPv6. The use of four bits have been defined, namely the X (ConEx-capable),

the L (loss experienced), the E (ECN experienced) and C (credit) bit.

By setting the X bit a packet is marked as ConEx-capable. All packets carrying payload MUST be marked with the X bit set including retransmissions. About control packets as pure ACKs which are not carrying any payload no congestion feedback information are available thus these packet should not be take into account when determining ConEx information. These packet MUST carry a ConEx Destination Option with the X bit unset.

4.1. Setting the E and the L Bit

As long as the CEG/LEG is positive, ConEx-capable packets MUST be marked with E or respective L and the CEG/LEG is decreased by the TCP payload bytes carried in this packet. If the CEG/LEG is negative, the CEG/LEG is drained by one byte with every packet sent out, as ConEX information are only meaningful for a certain time:

```
if CEG > 0: CEG -= TCPpayload.length else: CEG--
if LEG > 0: LEG -= TCPpayload.length else: LEG--
```

4.2. Credit Bits

The ConEx abstract mechanism requires that the transport SHOULD signal sufficient credit in advance to cover any reasonably expected congestion during its feedback delay. To be very conservative the number of credits would need to equal the number of packets in flight, as every packet could get lost or congestion marked. With a more moderate view, only an increase in the sending rate should cause congestion.

For TCP sender using the [RFC5681] congestion control algorithm, we recommend to only send credit in Slow Start, as in Congestion Avoidance an increase of one segment per RTT should only cause a minor amount of congestion marks (usually at max one). If a more aggressive congestion control is used, a sufficient amount of credits need to be set.

In TCP Slow Start the sending rate will increase exponentially and that means double every RTT. Thus the number of credits should equal half the number of packets in flight in every RTT. Under the assumption that all marks will not get invalid for the whole Slow Start phase, marks of a previous RTT have to be summed up. Thus the marking of every fourth packet will allow sufficient credits in Slow Start.

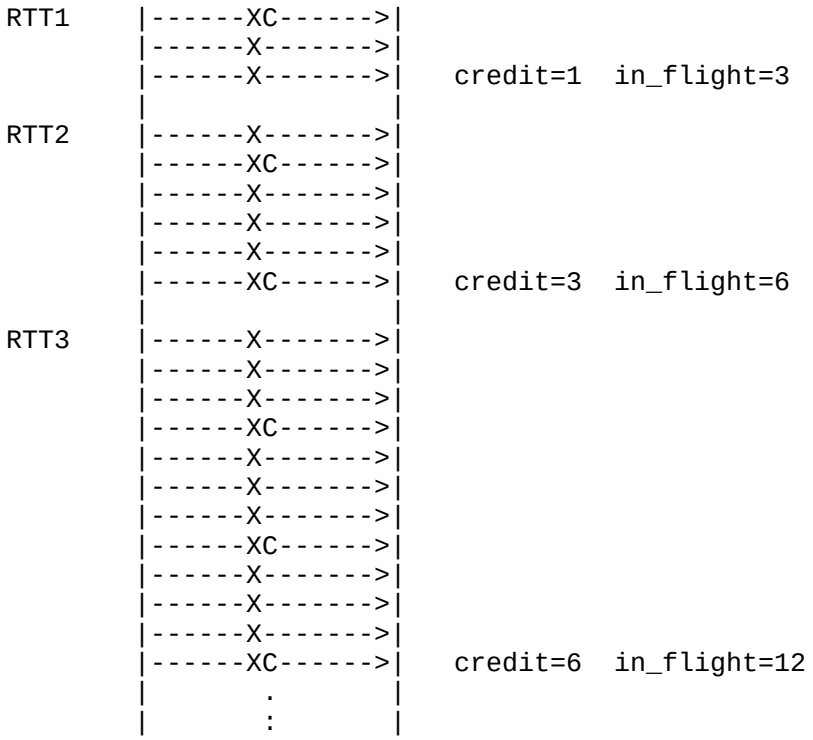


Figure 1: Credits in Slow Start (with an initial window of 3)

If a ConEx sender detects an increasing number of losses even though the sender reduced the sending rate, the sender SHOULD assume that those losses are incorporated by an audit device and thus should send further credits. Up to now its not clear if the credits say valid as long as the connection is established or if an expiration of the credits need to be assumed by the sender.

5. Timeliness of the ConEx Signals

ConEx signals will anyway be evaluated with a slight time delay of about one RTT by a network node. Therefore, it would not be absolutely necessary to immediately signal ConEx bits when they become known (e.g. L and E bits), but a sender SHOULD sent the ConEx signaling with the next available packet. If cases are available where it is preferable to slight delay the ConEx signal, the sender MUST NOT delay the ConEx signal more than one RTT.

Multiple ConEx bits may become available for signaling at the same time, for example when an ACK is received by the sender, that

indicates that at least one segment has been lost, and that one or more ECN marks were received at the same time. This may happen during excessive congestion, where buffer queues overflow and some packets are marked, while others have to be dropped nevertheless. Another possibility when this may happen are lost ACKs, so that a subsequent ACK carries summary information not previously available to the sender.

It is important to remember, that ConEx bits and TCP retransmissions do not interact with each other. However, a retransmission should be accompanied by one ConEx L bit in close proximity nevertheless. This does not mean, that TCP retransmissions may never contain ConEx marks. In a typical scenario using SACK, the first retransmission would not carry a ConEx L bit, while subsequent retransmissions in the same recovery episode, would be marked with the ConEx L bit. Spreading the ConEx bits over a small number of segments increases the likelihood that most devices along the path will see some ConEx marks even during heavy congestion.

6. Acknowledgements

7. IANA Considerations

8. Security Considerations

9. References

9.1. Normative References

[RFC2018] Mathis, M., Mahdavi, J., Floyd, S., and A. Romanow, "TCP Selective Acknowledgment Options", RFC 2018, October 1996.

[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.

[RFC5562] Kuzmanovic, A., Mondal, A., Floyd, S., and K. Ramakrishnan, "Adding Explicit Congestion Notification (ECN) Capability to TCP's SYN/ACK Packets", RFC 5562, June 2009.

9.2. Informative References

[DCTCP] Alizadeh, M., Greenberg, A., Maltz, D., Padhye, J., Patel, P., Prabhakar, B., Sengupta, S., and M. Sridharan, "DCTCP: Efficient Packet Transport for the Commoditized Data Center", Jan 2010.

[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-09 (work in progress), October 2010.

[RFC3522] Ludwig, R. and M. Meyer, "The Eifel Detection Algorithm for TCP", RFC 3522, April 2003.

[RFC3708] Blanton, E. and M. Allman, "Using TCP Duplicate Selective Acknowledgement (DSACKs) and Stream Control Transmission Protocol (SCTP) Duplicate Transmission Sequence Numbers (TSNs) to Detect Spurious Retransmissions", RFC 3708, February 2004.

[RFC4015] Ludwig, R. and A. Gurtov, "The Eifel Response Algorithm for TCP", RFC 4015, February 2005.

[RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, September 2009.

[RFC5682] Sarolahti, P., Kojo, M., Yamamoto, K., and M. Hata, "Forward RTO-Recovery (F-RTO): An Algorithm for Detecting Spurious Retransmission Timeouts with TCP", RFC 5682, September 2009.

[draft-kuehlewind-conex-accurate-ecn] Kuehlewind, M. and R. Scheffenegger, "Accurate ECN Feedback in TCP", draft-kuehlewind-conex-accurate-ecn-00 (work in progress), Jun 2011.

Authors' Addresses

Mirja Kuehlewind (editor)
University of Stuttgart
Pfaffenwaldring 47
Stuttgart 70569
Germany

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

Richard Scheffenegger
NetApp, Inc.
Am Euro Platz 2
Vienna, 1120
Austria

Phone: +43 1 3676811 3146
Email: rs@netapp.com