Congestion Exposure (ConEx) Working Group
Internet-Draft
Intended status: Informational
Expires: September 15, 2011

M. Mathis
Google, Inc
B. Briscoe
BT
March 14, 2011

Congestion Exposure (ConEx) Concepts and Abstract Mechanism
draft-ietf-conex-abstract-mech-01

Abstract

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

Table of Contents

1. Introduction

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

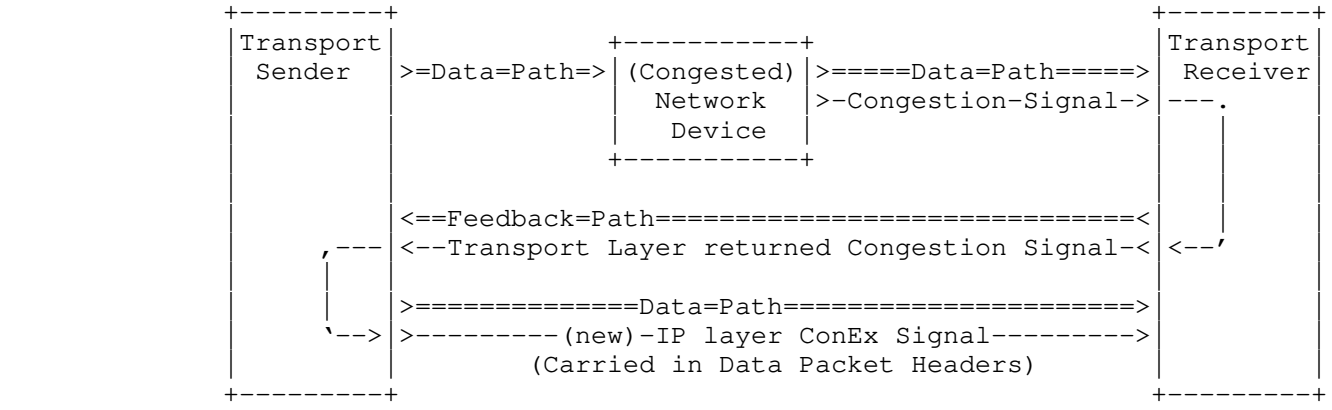
- The most common congestion signal is packet loss. When congested, the network simply discards some packets either as part of an active queue management function [RFC2309] or as the consequence of a queue overflow or other resource starvation. The transport receiver detects that some data is missing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

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

This document proposes to extend the capabilities of the Internet protocol suite with the addition of a ConEx Signal that, to a first approximation, relays the congestion information from the transport sender back through the internetwork layer. That signal is shown in Figure 1. It would be visible to all internetwork layer devices along the forward (data) path and is intended to support a number of

new policy-controlled mechanisms that might be used to manage traffic.

There is no expectation that internetwork layer devices will do fine-grained congestion control using ConEx information. That is still probably best done at the transport sender. Rather, the network will be able to use ConEx information to do better bulk traffic management, which in turn should incentivize end-system transports to be more careful about congesting others [I-D.conex-concepts-uses].



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

Figure 1

1.1. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

ConEx signals in IP packet headers from the sender to the network {ToDo: These are placeholders for whatever words we decide to use}:
Not-ConEx: The transport is not ConEx-capable
ConEx-Capable: The transport is ConEx-Capable. This is the opposite of Not-ConEx and implies one of the following signals
Re-Echo-Loss: (aka Purple) The transport has experienced a loss
Re-Echo-ECN: (aka Black) The transport has experienced an ECN mark

Credit: (aka Green) The transport is building up credit to allow for any future delay in expected ConEx signals (see Section 4.3.1)

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

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

2. Requirements for the ConEx Signal

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

a. The ConEx Signal SHOULD be visible to internetwork layer devices along the entire path from the transport sender to the transport receiver. Equivalently, it SHOULD be present in the IPv4 or IPv6 header, and in the outermost IP header if using IP in IP tunneling. The ConEx Signal SHOULD be immutable once set by the transport sender. A corollary of these requirements is that the chosen ConEx encoding SHOULD pass silently without modification through pre-existing networking gear.

b. The ConEx Signal SHOULD be useful under only partial deployment. A minimal deployment SHOULD only require changes to transport senders. Furthermore, partial deployment SHOULD create incentives for additional deployment, both in terms of enabling ConEx on more devices and adding richer features to existing devices. Nonetheless, ConEx deployment need never be universal, and it is anticipated that some hosts and some transports may never support the ConEx Protocol and some networks may never use the ConEx Signals.

c. The ConEx Signal SHOULD be accurate. In potentially hostile environments such as the public Internet, it SHOULD be possible for techniques to be deployed to audit the Congestion Exposure Signal by comparing it to the actual congestion signals on the forward data path. The auditing mechanism must have a capability for providing sufficient disincentives against misreported congestion, such as by throttling traffic that reports less congestion than it is actually experiencing.

d. The ConEx Signal SHOULD be timely. There will be a delay between the time when an auditing device sees an actual congestion signal and when it sees the subsequent Congestion Exposure Signal from the sender. The minimum delay will be one round trip, but it may be much longer depending on the transport's choice of feedback delay (consider RTCP [RFC3550] for example). It is not practical to expect auditing devices in the network to make allowance for

such feedback delays. Instead, the sender SHOULD be able to send ConEx signals in advance, as 'credit' for any audit function to hold as a balance against the risk of congestion during the feedback delay. This design choice greatly simplifies auditing (see Section 4.3.1).

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

- For congestion signaled by ECN, auditing is most accurate when located near the transport receiver. Within any flow or aggregate of flows, the volume of data tagged with ConEx Signals should never be less than the total volume of ECN marked data seen near the receiver.
- For congestion signaled by loss, totally accurate auditing is not believed to be possible in the general case, because it involves a network node detecting the absence of some packets, when it cannot necessarily see the transport protocol sequence numbers and when the missing packets might simply be taking a different route. But there are common cases where sufficient audit accuracy should be possible:
  - For non-IPsec traffic conforming to standard TCP sequence numbering on a single path, an auditor could detect losses by observing both the original transmission and the retransmission after the loss. Such auditing would be most accurate near the sender.
  - For networks designed so that losses predominantly occur under the management of one IP-aware node on the path, the auditor could be located at this bottleneck. It could simply compare ConEx Signals with actual local losses. This is a good model for most consumer access networks where audit accuracy could well be sufficient even if losses occasionally occur at other nodes in the network, such as border gateways (see Section 4.3 for details).

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

3. Representing Congestion Exposure

Most protocol specifications start with a description of packet formats and codepoints with their associated meanings. This document does not: It is already known that choosing the encoding for the ConEx Signal is likely to entail some engineering compromises that have the potential to reduce the protocol's usefulness in some settings. Rather than making these engineering choices prematurely,

this document side steps the encoding problem by describing an abstract representation of ConEx Signals. All of the elements of the protocol can be defined in terms of this abstract representation. Most important, the preliminary use cases for the protocol are described in terms of the abstract representation in companion documents [I-D.conex-concepts-uses].

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

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

3.1. Strawman Encoding

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

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

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

3.2. ECN Based Encoding

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

The re-ECN specification [I-D.briscoe-tsvwg-re-ecn-tcp] presents an implementation of ConEx that had to be tightly integrated with the

encoding of ECN in order to fit into the IP header. The central theme of the re-ECN work is an audit mechanism that can provide sufficient disincentives against misrepresenting congestion [I-D.briscoe-tsvwg-re-ecn-motiv], which is analyzed extensively in Briscoe's PhD dissertation [Refb-dis].

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

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

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

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

3.2.1. ECN Changes

Although the re-ECN protocol requires no changes to the network part of the ECN protocol, it is important to note that it does propose some relatively minor modifications to the host-to-host aspects of the ECN protocol specified in RFC 3168. They include: redefining the ECT(1) code point (the change is consistent with RFC3168 but requires deprecating the experimental ECN nonce [RFC3540]); modifications to the ECN negotiations carried on the SYN and SYN-ACK; and using a different state machine to carry ECN signals in the transport acknowledgments from a modified Receiver to the Sender. This last change is optional, but it permits the transport protocol to carry multiple congestion signals per round trip. It greatly simplifies accurate auditing, and is likely to be useful in other transports, e.g. DCTCP [DCTCP].

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

3.3. Abstract Encoding

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

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

3.3.1. Independent Bits

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

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

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

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

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

3.3.2. Codepoint Encoding

This encoding involves signaling one of the following five codepoints:

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

Each named codepoint has the same meaning as in the encoding using independent bits (Section 3.3.1). The use of any one codepoint implies the negative of all the others.

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

4. Congestion Exposure Components

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

4.1. Modified Senders

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

4.2. Receivers (Optionally Modified)

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

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

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

4.3. Audit

To audit ConEx Signals against actual losses (as opposed to ECN) an auditor could use one of the following techniques:

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

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

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

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

4.3.1. Using Credit to Simplify Audit

At the audit function,there will be an inherent delay of at least one round trip between a congestion signal and the subsequent ConEx signal it triggers--as it makes the two passes of the feedback loop in Figure 1. However, the audit function cannot be expected to wait for a round trip to check that one signal balances the other, because it is hard for a network device to know the RTT of each transport.

Instead, it considerably simplifies the audit function if the source transport is made responsible for removing the round trip delay in ConEx signals. The transport SHOULD signal sufficient credit in advance to cover any reasonably expected congestion during its feedback delay. Then, the audit function does not need to make allowance for round trip delays--that it cannot quantify. This design choice correctly makes the transport responsible for both minimizing feedback delay and for the risk that packets in flight will cause congestion to others before the source can react.

For example, imagine the audit function keeps a running account of the balance between actual congestion signals (loss or ECN), which it counts as negative, and ConEx signals, which it counts as positive. Having made the transport responsible for round trip delays, it will be expected to have pre-loaded the audit function with some credit at the start. Therefore, if ever the balance does go negative, the audit function can immediately start punishing a flow, without any grace period.

The one-way nature of packet forwarding probably makes per-flow state unavoidable for the audit function. This was a necessary sacrifice

to avoid per-flow state elsewhere in the wider ConEx architecture. Nonetheless, care was taken to ensure that packets could bring soft-state to the audit function, so that it would continue to work if a flow shifted to a different audit device, perhaps after a reroute or an audit device failure. Therefore, although the audit function is likely to need flow state memory, at least it complies with the 'fate-sharing' design principle of the Internet [IntDesPrinciples], and at least per-flow audit is only required at the outer edges of the internetwork, where it is less of a scalability concern.

Note also that ConEx does not intend to embed rules in the network on how individual flows _behave_. The audit function only does per-flow processing to check the integrity of ConEx _information_.

4.3.2. Behaviour Constraints for the Audit Function

There is no intention to standardise how to design or implement the audit function. However, it is necessary to lay down the following normative constraints on audit behaviour so that transport designers will know what to design against and implementers of audit devices will know what pitfalls to avoid:

Minimal False Hits: Audit SHOULD introduce minimal false hits for honest flows;

Minimal False Misses: Audit SHOULD quickly detect and sanction dishonest flows, preferably at the first dishonest packet;

Transport Oblivious: Audit MUST NOT be designed around one particular rate response, such as any particular TCP congestion control algorithm or one particular resource sharing regime such as TCP-friendliness [RFC3448]. An important goal is to give ingress networks the freedom to unilaterally allow different rate responses to congestion and different resource sharing regimes [Evol_cc], without having to coordinate with downstream networks;

Sufficient Sanction: Audit MUST introduce sufficient sanction (e.g. loss in goodput) so that sources cannot understate congestion and play off losses at the audit function against higher allowed throughput at a congestion policer [Salvatori05];

Manage Memory Exhaustion: Audit SHOULD be able to counter state exhaustion attacks. For instance, if the audit function uses flow-state, it should not be possible for sources to exhaust its memory capacity by gratuitously sending numerous packets, each with a different flow ID.

Identifier Accountability: Audit MUST NOT be vulnerable to 'identity whitewashing', where a transport can label a flow with a new ID more cheaply than paying the cost of continuing to use its current ID [CheapPseud];

4.4. Policy Devices

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

4.4.1. Policy Monitoring Devices

Policy devices can typically be decomposed into two functions i) monitoring the ConEx signal to compare it with a policy then ii) acting in some way on the result. Various actions might be invoked against 'out of contract' traffic, such as policing (see next section), re-routing, or downgrading the class of service.

Alternatively a policy device might not act directly on the traffic, but instead report to management systems that are designed to control congestion indirectly. For instance the reports might trigger capacity upgrades, penalty clauses in contracts, levy charges between networks based on congestion, or merely send warnings to clients who are causing excessive congestion.

Nonetheless, whatever action is invoked, the policy monitoring function will always be a necessary part of any policy device.

4.4.2. Congestion Policers

A congestion policer can be implemented in a very similar way to a bit-rate policer, but its effect can be focused solely on traffic causing congestion downstream, which ConEx signals make visible. Without ConEx signals, the only way to mitigate congestion is to blindly limit traffic bit-rate, on the assumption that high bit-rate is more likely to cause congestion.

A congestion policer monitors all ConEx traffic entering a network, or some identifiable subset. Using ConEx signals, it measures the amount of congestion that this traffic is contributing to somewhere downstream. If this exceeds a policy-configured 'congestion-bit-rate' the congestion policer will limit all the monitored ConEx traffic.

A congestion policer can be implemented by a simple token bucket. But unlike a bit-rate policer, it removes a token only when it forwards a packet that is ConEx-Marked, effectively treating Not-ConEx-Marked packets as invisible. Consequently, because tokens give the right to send congested bits, the fill-rate of the token bucket will represent the allowed congestion-bit-rate, which should be

sufficient traffic management without having to additionally constrain the straight bit-rate. See [CongPol] for details.

5. IANA Considerations

This memo includes no request to IANA.

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

6. Security Considerations

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

7. Conclusions

{ToDo:}

8. Acknowledgements

This document was improved by review comments from Toby Moncaster, Nandita Dukkipati, Mirja Kuehlewind and Caitlin Bestler.

9. Comments Solicited

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

10. References

10.1. Normative References

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

10.2. Informative References

[CheapPseud] Friedman, E. and P. Resnick, "The Social Cost of Cheap Pseudonyms", Journal of Economics and Management Strategy 10(2)173--199, 1998.

[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>.

[DCTCP] Alizadeh, M., Greenberg, A., Maltz, D., Padhye, J., Patel, P., Prabhakar, B., Sengupta, S., and M. Sridharan, "Data Center TCP (DCTCP)", ACM SIGCOMM CCR 40(4)63--74, October 2010, <http://portal.acm.org/citation.cfm?id=1851192>.

[Evol_cc] Gibbens, R. and F. Kelly, "Resource pricing and the evolution of congestion control", Automatica 35(12)1969--1985, December 1999, <http://www.statslab.cam.ac.uk/~frank/evol.html>.

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

[I-D.briscoe-tsvwg-re-ecn-motiv] Briscoe, B., Jacquet, A., Moncaster, T., and A. Smith, "Re-ECN: A Framework for adding Congestion Accountability to TCP/IP", draft-briscoe-tsvwg-re-ecn-tcp-motivation-02 (work in progress), October 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.

[I-D.conex-concepts-uses] Briscoe, B., Woundy, R., Moncaster, T., and J. Leslie, "ConEx Concepts

and Use Cases", draft-ietf-conex-concepts-uses-01 (work in progress), March 2011.

[I-D.ietf-ledbat-congestion] Shalunov, S., Hazel, G., and J. Iyengar, "Low Extra Delay Background Transport (LEDBAT)", draft-ietf-ledbat-congestion-03 (work in progress), October 2010.

[I-D.sridharan-tcpm-ctcp] Sridharan, M., Tan, K., Bansal, D., and D. Thaler, "Compound TCP: A New TCP Congestion Control for High-Speed and Long Distance Networks", draft-sridharan-tcpm-ctcp-02 (work in progress), November 2008.

[IntDesPrinciples] Clark, D., "The Design Philosophy of the DARPA Internet Protocols", ACM SIGCOMM CCR 18(4)106--114, August 1988, <http://www.acm.org/sigcomm/ccr/archive/1995/jan95/ccr-9501-clark.pdf>.

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[RFC2309] Braden, B., Clark, D., Crowcroft, J., Davie, B., Deering, S., Estrin, D., Floyd, S., Jacobson, V., Minshall, G., Partridge, C., Peterson, L., Ramakrishnan, K., Shenker, S., Wroclawski, J., and L. Zhang, "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, April 1998.

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

[RFC3448] Handley, M., Floyd, S., Padhye, J., and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 3448, January 2003.

[RFC3514] Bellovin, S., "The Security Flag in the IPv4 Header", RFC 3514, April 1 2003.

[RFC3540] Spring, N., Wetherall, D., and D. Ely, "Robust Explicit Congestion Notification (ECN) Signaling with Nonces", RFC 3540, June 2003.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC5670] Eardley, P., "Metering and Marking Behaviour of PCN-Nodes", RFC 5670, November 2009.

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

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

[Refb-dis] Briscoe, B., "Re-feedback: Freedom with Accountability for Causing Congestion in a Connectionless Internetwork", UCL PhD Dissertation , 2009, <http://bobbriscoe.net/projects/refb/#refb-dis>.

[Salvatori05] Salvatori, A., "Closed Loop Traffic Policing", Politecnico Torino and Institut Eurecom Masters Thesis , September 2005.

[Vegas] Brakmo, L. and L. Peterson, "TCP Vegas: End-to-End Congestion

Avoidance on a Global Internet",
IEEE Journal on Selected Areas in
Communications 13(8)1465--80,
October 1995, <http://
ieeexplore.ieee.org/iel1/49/9740/
00464716.pdf?arnumber=464716>.

Authors' Addresses

Matt Mathis
Google, Inc
1600 Amphitheater Parkway
Mountain View, California 93117
USA

EMail: mattmathis at google.com

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.

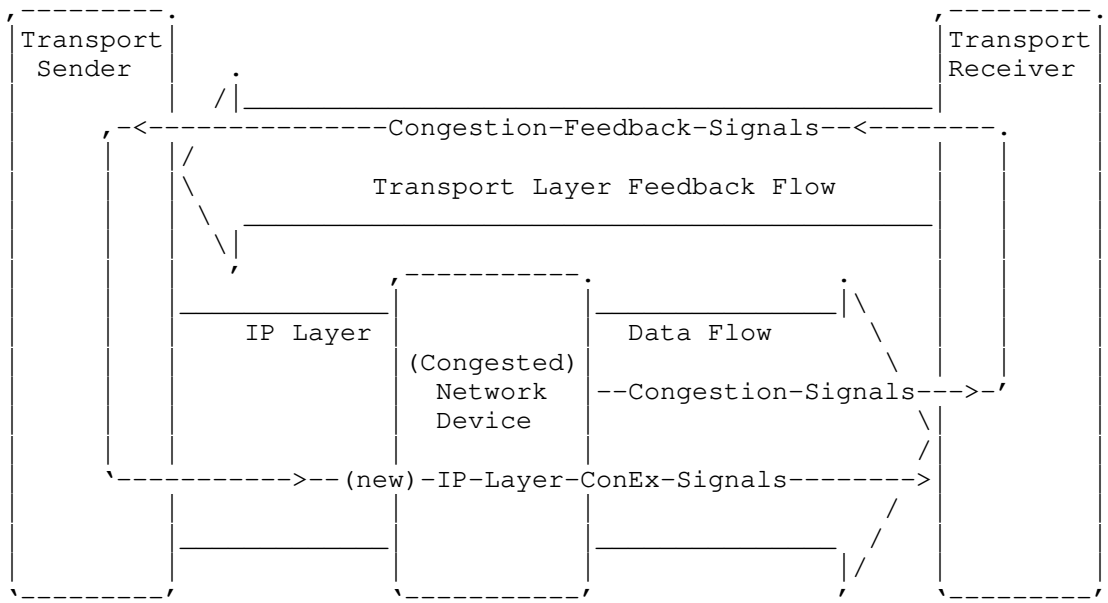


Figure 1: The ConEx Protocol in the Internet Architecture

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

conex Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 15, 2011

S. Krishnan
Ericsson
M. Kuehlewind
IKR University of Stuttgart
C. Ucendo
Telefonica

March 14, 2011

Options for Conex marking in IPv6 packets
draft-krishnan-conex-ipv6-02

Abstract

Conex is a mechanism by which senders inform the network about the congestion encountered by packets earlier in the same flow. This document describes the requirements for conex markings in IPv6 datagrams and describes the various options for performing conex markings in IPv6.

Table of Contents

1. Introduction

Conex is a mechanism by which senders inform the network about the congestion encountered by packets earlier in the same flow. This document describes the requirements for conex markings in IPv6 datagrams and describes the various options for performing conex markings in IPv6.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL","SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Requirements for marking IPv6 packets

R-1: The marking mechanism needs to be visible to all conex-capable nodes on the path.

R-2: The mechanism needs to be able to traverse nodes that do not understand the markings. This is required to ensure that conex can be incrementally deployed over the Internet.

R-3: The presence of the marking mechanism should not significantly alter the processing of the packet. This is required to ensure that conex marked packets do not face any undue delays or drops due to a badly chosen mechanism.

R-4: The markings should be immutable once set by the sender. At the very least, any tampering should be detectable.

4. Possible Solutions

4.1. Hop-by-hop options

The base IPv6 standard [RFC2460] defines hop-by-hop options. These options are processed by every node on the path. Hence they meet R-1. The options have variable semantics based on the 3 MSB of the option code. The state of these bits controls the behavior of nodes to either ignore unknown options or drop packets containing them. It also defines the ICMPv6 error message sending behavior and the mutability of the options en-route. This means that it is possible for hop-by-hop options to satisfy R-2 and R-4. In most commercial router implementations the mere presence of hop-by-hopoptions rResult in the packet being punted to the Slow path instead of being accorded

regular forwarding behavior (Fast Path). This means that R-3 is not satisifed.

4.2. Destination options

The base IPv6 standard [RFC2460] defines the destination options. These options are processed only by the ultimate receiver of the packet (as specified in the Destination Address field) and not by nodes on the path. Hence they do not meet R-1. The options have the same variable semantics based on the 3 MSBs as the hop-by-hop option which means that they can satisfy R-2 and R-4. As intermediate nodes currently do not process destination options R-3 is easily satisifed.

4.3. Header bits

The IPv6 header has no free bits. The only bits in the IPv6 header that are not widely used are the flow label bits [RFC3697]. There are some initiatives to redefine the use of the flow label for other purposes (e.g. Load balancing, nonce). It may be possible (but highly unlikeley) to save a few bits from the flow label for alternate purposes to end up with a shorter flow label. The use of IPv6 header bits can satisfy all the requirements for conex markings but using valuable header bits for experimental purposes (such as conex) may not be acceptable.

4.4. Extension Headers

The base IPv6 standard [RFC2460] defines extension headers as an expansion mechanism to carry optional internet layer information. Extension headers, with the exception of the hop-by-hop options header, are not usually processed on intermediate nodes. This means that R-1 cannot be met. Unknown extension headers cause the packet to be dropped and hence such mechanism is not incrementally deployable. Hence R-3 is not met either.

5. ConEx Encoding

The decision about where to code the ConEx inform might also influence the decision on how to code congestion information itself. Of course, a ConEx capable transport has to inform the network that it is actually ConEx enabled. Thus, as a minimum, every packet has to carry the information that the sender is ConEx enabled and, also whether it is ConEx marked. Moreover, the abstract conex mechanism [CAM] requires that that a distinction between loss or ECN marks as congestion signal is needed in addition to the so-called 'congestion credits'. This implies that a minimum of 4 bits is needed if bit-wise encoding is used , and a minimum of 3 bits is needed if

codepoints are used. Further ideas on additional ConEx information are currently discussed on the mailing list. Moreover, the ConEx information could be represented in a more sophisticated manner than a binary signal (Yes/No), if additional bits are available for use.

6. Acknowledgements

The authors would like to thank Marcelo Bagnulo, Bob Briscoe, Ingemar Johansson, Joel Halpern and John Leslie for the discussions that led to this document.

7. Security Considerations

This document does not bring up any new security issues.

8. IANA Considerations

This document does not require any IANA action.

9. Normative References

[CAM] Briscoe, B., "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-conex-abstract-mech-01 (work in progress), March 2011.

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

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC3697] Rajahalme, J., Conta, A., Carpenter, B., and S. Deering, "IPv6 Flow Label Specification", RFC 3697, March 2004.

Authors' Addresses

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

Email: suresh.krishnan@ericsson.com

Mirja Kuehlewind
IKR University of Stuttgart

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

Carlos Ralli Ucendo
Telefonica

Email: ralli@tid.es

IETF
Internet-Draft
Intended status: Informational
Expires: September 13, 2012

D. Kutscher
F. Mir
R. Winter
NEC
S. Krishnan
Ericsson
C. Cano
Universidad Carlos III de Madrid

March 12, 2012

Mobile Communication Congestion Exposure Scenario
draft-kutscher-conex-mobile-03

Abstract

This memo describes a mobile communications use case for congestion exposure (CONEX) with a particular focus on mobile communication networks such as 3GPP EPS. The draft describes the architecture of these networks (access and core networks), current QoS mechanisms and then discusses how congestion exposure concepts could be applied. Based on this, this memo suggests a set of requirements for CONEX mechanisms that particularly apply to mobile networks.

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 September 13, 2012.

Copyright Notice

Table of Contents

1. Introduction

Mobile data traffic continues to grow rapidly. The challenge wireless operators face is to support more subscribers with higher bandwidth requirements. To meet the bandwidth demand, operators need to seek for new technologies to efficiently utilize the available network resources, in particular, this concerns resource allocation and flow management. Ample statistics for network traffic from cellular networks are available, stating that many short flows exist, but that a few large flows constitute a large part of the overall traffic volume. Measurement studies have shown that a small number of users is responsible for the most part of the traffic in cellular networks. With the highly skewed network access behavior, more expensive resources in cellular networks as compared to other networks and the rapidly increasing network utilization, resource allocation and usage accountability are two important issues for operators to solve in order to achieve a better, accountable network resource utilization. CONEX, as described in [I-D.ietf-conex-concepts-uses], is a technology to base this upon.

The CONEX congestion exposure mechanism is intended as a general technology that could be applied as a key element of congestion management solutions in a variety of use cases. The IETF CONEX WG will however work on a specific use case, where the end hosts and the network that contains the destination end host are CONEX-enabled but other networks need not be.

A specific example of such a use case can be a mobile communication network such as a 3GPP EPS network, where UEs (User Equipment, i.e. mobile end hosts), servers and caches, the access network and possibly an operator's core network can be CONEX-enabled. I.e., hosts support the CONEX mechanisms, and the network provides policing/auditing functions at its edges.

This draft describes the architecture of such networks (access and core networks), current QoS mechanisms and then discusses how congestion exposure concepts can benefit such networks and how they should be applied. Using this use case as a basis, a set of requirements for CONEX mechanisms are described.

2. Overview of 3GPP's Evolved Packet System (EPS)

This section provides an overview of 3GPP's "Evolved Packet System" (EPS [3GPP.36.300]) as a specific example of a mobile communication architecture in order to illustrate congestion exposure applicability in this memo. There are other mobile communication architectures.

The EPS architecture and its standardized interfaces are depicted in Figure 1. The EPS provides IP connectivity to UEs (user equipment, i.e., mobile nodes) and access to operator services, such as global Internet access and voice communications. The EPS comprises the access (evolved UMTS Terrestrial Radio Access Network, E-UTRAN) and the core network (Evolved Packet Core, EPC -- all network elements except the E-UTRAN). QoS is supported through an EPS bearer concept, providing hierarchical bindings within the network.

The evolved NodeB (eNB), the Long Term Evolution (LTE) base station, is part of the access network that provides radio resource management, header compression, security and connectivity to the core network through the S1 interface. In an LTE network, the control plane signaling traffic and the data traffic are handled separately. The eNBs transmit the control traffic and data traffic separately via two logically different interfaces.

The Home Subscriber Server, HSS, is a database that contains user subscriptions and QoS profiles. The Mobility Management Entity, MME, is responsible for user authentication, bearer establishment and modification and maintenance of the UE context.

The Serving gateway, S-GW, is the mobility anchor and manages the user plane data tunnels during the inter-eNB handovers. It tunnels all user data packets and buffers downlink IP packets destined for UEs that happen to be in idle mode.

The Packet Data Network (PDN) Gateway, P-GW, is responsible for IP address allocation to the UE and is a tunnel endpoint for mobility protocols. It is also responsible for charging, packet filtering, and policy-based control of flows. It interconnects the mobile network to external IP networks, e.g. the Internet.

In this architecture, data packets are not sent directly on an IP network between the eNB and the gateways. Instead, every packet is sent in a tunneling protocol - the GPRS Tunneling Protocol (GTP [3GPP.29.060]) over UDP/IP. A GTP path is identified in each node with the IP address and a UDP port number on the eNB/gateways. The GTP protocol carries both the data traffic (GTP-U tunnels) and the control traffic (GTP-C tunnels [3GPP.29.274]). Alternativly Proxy Mobile IP (PMIP) is used on the S8 interface.

The above is very different from an end-to-end path on the Internet where the packet forwarding is performed at the IP level. Importantly, we observe that these tunneling protocols give the operator a large degree of flexibility to control the congestion mechanism incorporated with the GTP/PMIP protocols.

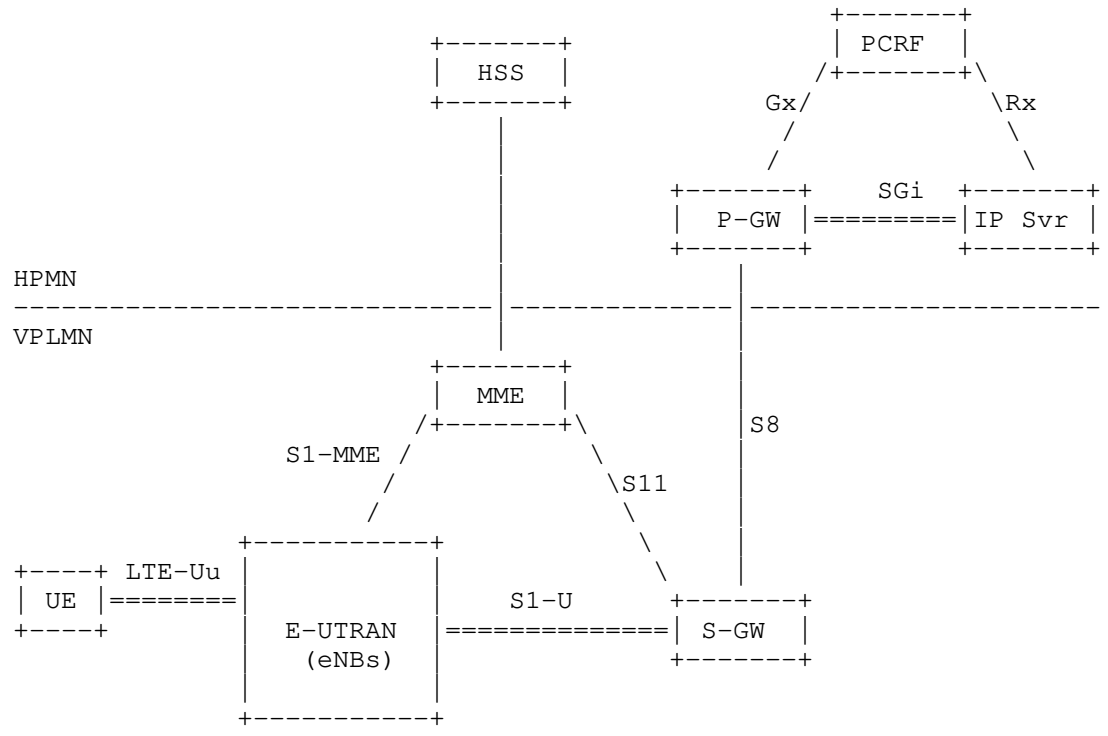


Figure 1: EPS (non-roaming) architecture overview

3. CONEX Use Cases in the Mobile Communication Scenario

In general, quality of service and good network resource utilization are important requirements for mobile communication network operators. Radio access and backhaul networks are considered scarce resources, and bandwidth (and radio resource) demand is difficult to predict precisely due to user mobility, radio propagation effects etc. Hence today's architectures and protocols go to significant extent in order to provide network-controlled quality of service -- for instance by 3GPP's EPS bearer model that enables the network to allocate service data flows (SDFs) to certain EPS bearers with specific quality of service classes (which can be used for fine-granular per-application service differentiation).

In the following, we discuss ways how congestion exposure could be beneficial for supporting resource management in such operated mobile communication networks. [I-D.ietf-conex-concepts-uses] describes fundamental congestion exposure concepts and a set of use cases for applying congestion exposure mechanisms to realize different traffic

management functions, accounting etc. Here, we relate these CONEX use cases to the general mobile communication scenario in order to validate the use cases for this scenario.

3.1. CONEX as a Basis for Traffic Management

Traffic management is a very important function in mobile communication networks. Since wireless resources are considered scarce and since user mobility and shared bandwidth in the wireless access create certain dynamics with respect to available bandwidth, these resources are traditionally managed very tightly (admission control for bearer establishment etc.).

In EPS, the QoS requirements for different applications running on a UE is supported by a bearer concept which is managed by the network. Each bearer has an associated QoS Class identifier (QCI) and an Allocation and Retention Policy (ARP) that has been standardized for uniform traffic handling (across implementations). For the necessary QoS across the mobile network, an EPS bearer is maintained that crosses different interfaces in the network and maps to lower layer bearers for packet forwarding. A radio bearer transports traffic between a UE and eNB whereas S1 bearer transports traffic between the eNB and S-GW. Primarily LTE offers two types of bearer: Guaranteed Bit rate bearer for real time communication, e.g., Voice calls etc. and Non-Guaranteed bit rate, e.g., best effort traffic for web access etc. Packets mapped to the same EPS bearer receive the same bearer level packet forwarding treatment.

In the light of the significant increase of overall data volume in 3G networks, Deep-Packet-Inspection (DPI) is often considered a desirable function to have in the EPC -- on, for example, a PDN (Packet Data Network) gateway, and some operators do in fact deploy DPI today. 3GPP has a current work item on "Service Awareness and Privacy Policies" that is chartered to add DPI-related extensions to the PCC architecture [3GPP.23.203]. The (optional) DPI entity in the EPC is called "Traffic Detection Function" (TDF), and it performs application detection and reporting of detected application and its service data flow description to the Policy Control and Charging Rules Function (PCRF). The TDF and it can perform functions such as traffic blocking, redirection, policing for selected flows.

Congestion exposure can be employed to address these requirements for tight resource management in different ways:

1. It can enhance DPI by providing flow policy-based traffic management. At present, DPI-based resource management is often used to prioritize certain application classes with respect to others in overload situations, so that effectively more users can

be served on the network. In overload situations, operators use DPI to identify dispensable flows and make them yield to other flows (of different application classes) through policing. Such traffic management is thus based on static configuration and some estimation about the future per-flow bandwidth demand. With congestion exposure it would be possible to more accurately and more timely assess the cost that certain flows are causing so that policing can optimize network utilization (better than a pure DPI-based approach can do).

2. It can eventually make DPI obsolete by allowing for a bulk packet traffic management that does not have to consider flows' application classes and individual sessions. Instead traffic management would be based on the current cost (contribution to congestion) incurred by different flows and enable operators to apply policing/accounting depending on their preference. Such traffic management would be simpler and more robust (no real-time flow application type identification required, no static configuration of application classes) and perform better as decisions can be taken based on real-time actual cost contribution.

In summary, it can be said that traffic management in 3GPP EPS and other mobile communication architectures in very important. Previously more static approaches based on admission control and static QoS have been applied, but recently, there has been a perceived need for more dynamic mechanisms such as DPI. Adding CONEX support might thus require revisiting the PCC architecture, depending on the scope and impact of a CONEX-based traffic management approach.

3.2. CONEX to Incentivize Scavenger Transports

As 3G and LTE networks are turning into universal access networks that are shared between mobile (smart) phone users, mobile users with laptop PCs, home users with LTE access etc., it is likely that capacity-sharing among different users and application flows becomes more important in the mobile communication network as a fine-granular differentiation would be too costly.

Most of this traffic is likely to be classified as best-effort traffic, without differentiating (for example) periodic OS updates, application store downloads from web (browser)-based or other more real-time communication. Having said that, the general argument for scavenger transports apply. Especially when wireless and backhaul resources are scarce, incentivizing users to use less-than best effort transport for non-interactive background communication would improve the overall utility of the network. It can be argued that, if this would be done with a CONEX approach, it could be in a more

effective and cost-efficient way compared to the mentioned DPI mechanisms.

This would work best, if the network did not do any traffic class segregation below the IP layer, i.e., if all traffic would be in the same traffic class. With current specifications, that would be possible in principle.

3.3. Accounting for Congestion Volume

3G and LTE networks provide extensive support for accounting and charging already, for example cf. the Policy Charging Control (PCC) architecture. In fact, most operators today account transmitted data volume on a very fine granular basis and either correlate monthly charging to the exact number of packets/bytes transmitted, or employ some form of flat rate (or flexible flat rate), often with a so-called fair-use policy. With such policies, users are typically limited to an administratively configured maximum bandwidth limit, after they have used their data contractual volume budget for the charging period.

Changing this data volume-based accounting to a congestion-based accounting would be possible in principle, especially since there already is an elaborate per-user accounting system available. Also, an operator-provided mobile communication network can be seen as a network domain within such congestion volume accounting would be possible, without requiring any support from the global Internet. Traffic normally leaves/enters the operator's network via well-defined egress/ingress points that would be ideal candidates for policing functions. Moreover, in most commercially operated networks, the user is accounted for both received and sent data, which would facilitate congestion volume accounting as well.

With respect to the current PCC framework, accounting for congestion volume could be added as another feature to the "Usage Monitoring Control" capability that is currently based on data volume. This would not require any new interface (reference points) at all.

3.4. CONEX as a Form of Differential QoS

As mentioned above, 3GPP mobile communication networks provide an elaborate QoS architecture. In LTE, the idea is to map different traffic classes onto different logical channels (bearers) with individual QoS configuration.

It can be argued whether this approach is sufficient in a world where most traffic is on TCP port 80 and whether some more application control would be useful.

With CONEX, accurate downstream path information would be visible to ingress network operators, which can respond to incipient congestion in time. This can be equivalent to offering different levels of QoS, e.g. premium service with zero congestion response.

Again, CONEX could be used in two different ways:

1. as additional information to assist network functions to impose different QoS for different application sessions; and

2. as a tool to let applications decide on their response to congestion notification, while incentivizing them to react (in general) appropriately, e.g., by enforcing overall limits for congestion contribution or by accounting and charging for such congestion contribution.

3.5. Partial vs. Full Deployment

In general CONEX lends itself to partial deployment as the mechanism does not require all routers and hosts to support congestion exposure. Moreover, assuming a policing infrastructure has been put in place, it is not required to modify all hosts. Since CONEX is about senders exposing congestion contribution to the network, senders need to be made supporting CONEX (assuming a congestion notification mechanisms such as ECN is in place).

[I-D.briscoe-conex-initial-deploy] provides specific examples of how CONEX deployment can be initiated, focusing unilaterial deployment by single networks, i.e., by partial deployment.

In mobile communication networks that would for example allow early partial CONEX deployment in the downlink direction only, i.e., servers, gateways and caches would support CONEX but UEs (mobile hosts) would not.

When moving towards full deployment in a specific operator's network, different ways for introducing CONEX support on UEs are feasible. Since mobile communication networks are multi-vendor networks, standardizing CONEX support on UEs (e.g., in 3GPP specifications) appears useful. Still, not all UEs would have to support CONEX, and operators would be free to choose their policing approach in such deployment scenarios. Leveraging existing PCC architectures, 3GPP network operators could for example decide policing/accounting approaches per UE -- i.e., apply fixed volume caps for non-CONEX UEs and more flexible schemes for CONEX-enabled UEs.

Moreover, it should be noted that network support for CONEX is a feature that some operators may implement to deploy if they wish, but

it is not required that all operators (or all other networks) do so.

Depending on the extent of CONEX support, specific aspects such as roaming have to be taken into account. I.e., what happens when a user is roaming in a CONEX-enabled network, but their UE is not CONEX-enabled and vice versa. Although these may not be fundamental problems, they need to be considered. For supporting mobility in general, it can be required to shift users' policing state during hand-over. There is existing work in [raghavan2007] on distributed rate limiting and in [nec.euronf-2011] on specific optimizations for congestion exposure and policing in mobility scenarios.

Another aspect to consider is the addition of Selected IP Traffic Offload (SIPTO) and Local Breakout (LIPA), also see [3GPP.23.829], i.e., the idea that some traffic (e.g., high-volume Internet traffic) is actually not passed through the EPC but is offloaded at a "break-out point" closer to (or in) the access network. On the other hand, CONEX can also enable more dynamic decisions on what traffic to actually offload by considering congestion exposure in bulk traffic aggregates -- thus making traffic offload more effective.

3.6. Summary

In summary, the 3GPP EPS is a system architecture that can benefit from congestion exposure in multiple ways, as we have shown by this brief description of CONEX use cases in this environment. Dynamic traffic and congestion management is an acknowledged important requirement for the EPS, also illustrated by the current DPI work for EPS.

Moreover, we believe that networks such as an EPS mobile communication network would be quite amenable for deploying CONEX as a mechanism, since they represent clearly defined and well separated operational domains, in which local CONEX deployment would be possible. Aside from roaming (which needs to be considered for a specific solution), a single mobile network deployment is under full control of a single operator, which can enable operator-local enhancement without the need to change the complete architecture.

In 3GPP EPS, interfaces between all elements of the architecture are subject to standardization, including UE interfaces and eNodeB interfaces, so that a more general approach, involving more than one single operator's network, can be feasible as well.

4. CONEX in the EPS

The CONEX mechanism is still work in progress in the IETF working

group. Still, we would like to discuss a few options for how such a mechanism (and possibly additional policing functions) could eventually be deployed in 3GPP's EPS. Note that this description of options is not intended as a complete set of possible approaches -- it is merely intended for discussing a few options. More details will be provided in a future revision of this document.

4.1. Possible Deployment Scenarios

There are different possible ways how CONEX functions on hosts and network elements can be used. For example, CONEX could be used for a limited part of the network only -- e.g., for the access network -- congestion exposure and sender adaptation could involve the mobile nodes or not, or, finally, the CONEX feedback loop could extend beyond a single operator's domain or not.

We present three different deployment scenarios for congestion exposure in the figures below:

1. In Figure 2 CONEX is supported by servers for sending data (here: web servers in the Internet and caches in an operator's network) but not by UEs (neither for receiving nor sending). An operator who chooses to run a policing function on the network ingress (e.g., on the P-GW) can still benefit from congestion exposure without requiring any change on UEs.

2. CONEX is universally employed between operators (as depicted in Figure 3), with an end-to-end CONEX feedback loop. Here, operators could still employ local policies, congestion accounting schemes etc., and they could use information about congestion contribution for determining interconnection agreements.

3. Isolated CONEX domains as depicted in Figure 4, CONEX is solely applied locally, in the operator network, and there is no end-to-end congestion exposure. This could be the case when CONEX is only implemented in a few networks, or when operators decide to not expose ECN and account for congestion for inter-domain traffic. Independent of the actual scenario, it is likely that there will be border gateways (as in today's deployments) that are associated with policing and accounting functions.

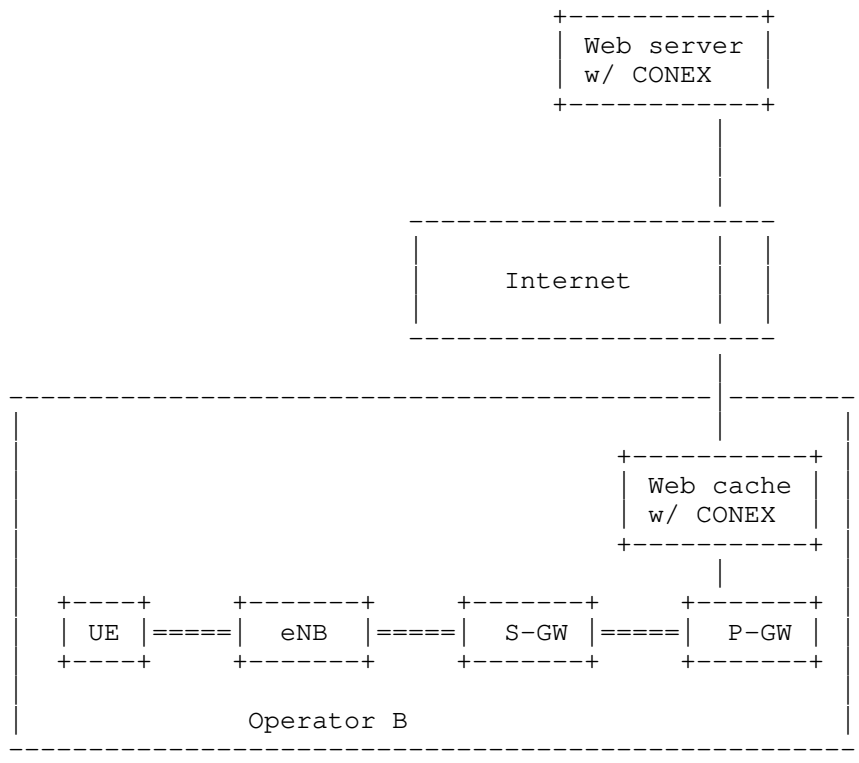


Figure 2: CONEX support on servers and caches

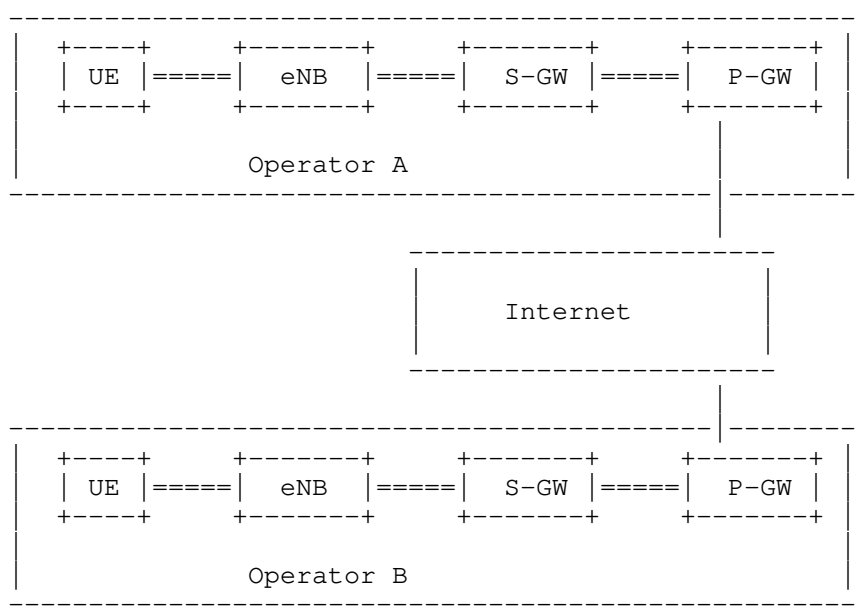


Figure 3: CONEX deployment across operator domains

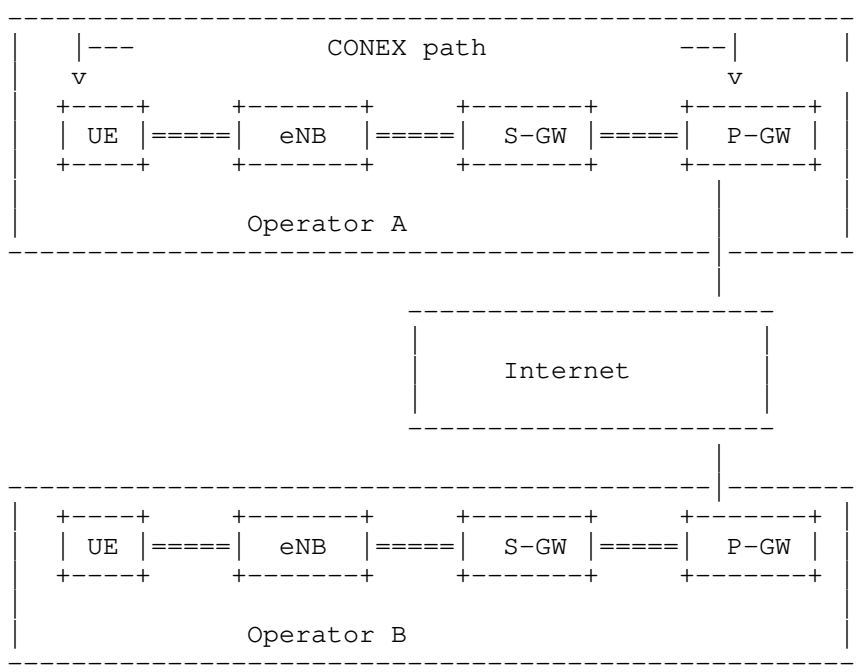


Figure 4: CONEX deployment in a single operator domain

We consider all three scenarios to be relevant and believe that both of them are within the scope of the CONEX WG charter. A more detailed description will be provided in a future version of this document.

4.2. Implementing CONEX Functions in the EPS

We expect a CONEX solution to consist of different functions that should be considered when implementing congestion exposure in 3GPP's EPS. [I-D.ietf-conex-abstract-mech] is describing the following congestion exposure components:

- Modified senders that send congestion exposure information in response to congestion feedback).
- Receivers that generate congestion feedback (leveraging existing behavior or requiring new functions).
- Audit functions that audit CONEX signals against actual congestion, e.g., by monitoring flows or aggregate of flows.

- Policy devices that monitor congestion exposure information and act on the flows according to the operator's policy.

Two aspects are important to consider: 1) how would the CONEX protocol mechanisms be implemented and what modifications to existing networks would be required and 2) where would CONEX functional entities be placed best (to allow for a non-invasive addition). We discuss these two aspects in the following sections.

4.2.1. CONEX Protocol Mechanisms

As described in [I-D.briscoe-conex-initial-deploy], the most important component for introducing CONEX (initially) is adding the congestion exposure functionality to senders. For an initial deployment, no further modification to senders and receivers would be required. Specifically, there is no fundamental dependency on ECN, i.e., CONEX can be introduced without requiring ECN to be implemented.

Congestion exposure information for IPv6 [I-D.ietf-conex-destopt] is represented in a destination option header field, which requires minimal changes at senders and nodes that want to assess path congestion -- and that does not affect non-CONEX nodes in a network.

In 3GPP networks, IP tunneling is used intensively, i.e., using either IP-in-GTP-U or PMIP (i.e., IP-in-IP) tunnels. In general, the CONEX destination option of encapsulated packets should be made available for network nodes on the tunnel path, i.e., a tunnel ingress should copy the CONEX destination option field to the outer header. Details will be provided in a future version of this document.

For an effective and efficient capacity sharing, we envisage the deployment of ECN in conjunction with CONEX so that ECN-enabled receivers and senders get more accurate and more timely information about their flows congestion contribution. ECN is already partially introduced into 3GPP networks: Section 11.6 in [3GPP.36.300] specifies the usage of ECN for congestion notification on the radio link (between eNB and UE), and [3GPP.26.114] specifies how this can be leveraged for voice codec adaptation. A complete, end-to-end support of ECN would require specification of tunneling behaviour, which should be based on [RFC6040] (for IP-in-IP tunnels) and on [I-D.briscoe-tsvwg-ecn-encap-guidelines]. Specifically, a specification for tunneling ECN in GTP-U will be needed.

4.2.2. CONEX Functions in the Mobile Network

In the following, we discuss some possible placement strategies for CONEX functional entities (addressing both policing and auditing functions) in the EPS and for possible optimizations for both the uplink and the downlink.

In general, CONEX information (exposed congestion) is declared by a sender and remains unchanged on the path, hence reading CONEX information (e.g., by policing functions) is placement-agnostic. Auditing CONEX normally requires assessing declared congestion contribution and current actual congestion. If the latter is, for example, done using ECN, such a function would best be placed at the end of the path.

In order to provide a comprehensive CONEX-based capacity management framework for EPS, it would be advantageous to consider user contribution to congestion for both the radio access and the core network. For a non-invasive introduction of CONEX, it can be beneficial to combine CONEX functions with existing logical EPS entities. For example, potential places for CONEX policing and auditing functions would then be eNBs, S-GWs or the P-GWs. Operator deployments may of course still provide additional intermediary CONEX-enabled IP network elements.

For a more specific discussion it will be beneficial to distinguish downlink and uplink traffic directions (also see [nec.globecom2010] for a more detailed discussion). In today's networks and usage models, downlink traffic is dominating (also reflected by the different maximum capacity provided by the air interface). That does however not imply that uplink congestion is not an issue, since the asymmetric maximum bandwidth configuration can create a smaller bottleneck for uplink traffic -- and there are of course backhaul links, gateways etc. that can be overloaded as well.

For managing downlink traffic -- e.g., in scenarios such as the one depicted in Figure 2, operators can have different requirements for policing traffic. Although policing is in principle location-agnostic, it is important to consider requirements related to the EPS architecture (Figure 1) such as tunneling between P-GWs and eNBs. Policing can require access to subscriber information (e.g., congestion contribution quota) or user-specific accounting, which suggests that the CONEX function could be co-located with the P-GW that already has a PCRF interface.

Still, policing can serve different purposes. For example, if the objective is to police bulk traffic induced by peer networks, additional monitoring functions can be placed directly at

corresponding ingress points to monitor traffic and possible drive out-of-band functions such as triggering border contract penalties.

The auditing function which should be placed at the end of the path (at least after/at the last bottleneck) would likely be placed best on the eNB (wireless base station).

For the uplink direction, there are naturally different options for designing monitoring and policy enforcement functions. A likely approach can be to monitor congestion exposure on central gateway nodes (such as P-GWs) that provide the required interfaces to the PCRF, but to perform policing actions in the access network, i.e., in eNBs, e.g., to police traffic at the ingress, before it reaches concentration points in the core network.

Such a setup would enable all the CONEX use cases described in Section 3, without requiring significant changes to the EPS architecture, while enabling operators to re-use existing infrastructure, specifically wireless base stations, PCRF and HSS systems.

For CONEX functions on elements such as the S-GWs and P-GWs, it is important to consider mobility and tunneling protocol requirements. LTE provides two alternative approaches: Proxy-Mobile-IP (PMIP, [3GPP.23.402]) and GPRS Tunneling Protocol (GTP). For the propagation of congestion information (responses) tunneling considerations are therefore very important.

In general, policing will be done based on per-user (per subscriber) information such as congestion quota, current quota usage etc. and network operator policies, e.g., specifying how to react to persistent congestion contribution etc. In the EPS, per-user information is normally part of the user profile (stored in the HSS) that would be accessed by PCC entities such as the PCRF for dynamic updates, enforcement etc.

A more detailed description of the different approaches and their respective advantages will be provided in a future revision of this document.

5. Summary

We have shown how congestion exposure can be useful for efficient resource management in mobile communication networks. The premise for this discussion was the observation that data communication, specifically best-effort bulk data transmission, is becoming a commodity service whereas resources are obviously still limited --

which calls for efficient, scalable, yet effective capacity sharing in such networks.

CONEX can be a mechanism that enables such capacity sharing, while allowing operators to apply these mechanisms in different ways, e.g., for implementing different use cases as described in Section 3. It is important to note that CONEX is fundamentally a mechanism that can be applied in different ways -- to realize different operators policies.

We have described a few possibilities for adding CONEX as a mechanism to 3GPP LTE-based networks and have shown how this could be done incrementally (starting with partial deployment). It is quite feasible that such partial deployments be done on a per-operator-domain basis, without requiring changes to standard 3GPP interfaces. For a network-wide deployment, e.g., with congestion exposure between operators, more considerations might be needed.

We have also identified a few implications/requirements that should be taken into consideration when enabling congestion exposure in such networks:

Performance: In mobile communication networks -- with more expensive resources and more stringent QoS requirements -- the feasibility of applying CONEX as well as its performance and deployment scenarios need to be examined closer. For instance, a mobile communication network may encounter longer delay and higher loss rates, which can impose specific requirements on the timeliness and accuracy of congestion exposure information.

Mobility: One of the unique characteristics in cellular network is the presence of user mobility compared to wired networks. As the user location changes, the same device can be connected to the network via different base stations (eNodeBs) or even go through switching gateways. Thus, the CONEX scheme must to be able to carry latest congestion information per user/flow across multiple network nodes in real time.

Multi-access: In cellular network, multiple access technologies can co-exist. In such cases, a user can use multiple access technologies for multiple applications or even a single application simultaneously. If the congestion policies are set based on each user, then CONEX should have the capability to enable information exchange across multiple access domains.

Tunneling: Both 3G and LTE networks make extensive usage of tunneling. The CONEX mechanism should be designed in a way to support usage with different tunneling protocols such as PMIP and GTP. For ECN-based congestion notification, [RFC6040] specifies how the ECN field of the IP header should be constructed on entry and exit from IP-in-IP tunnels, and [I-D.briscoe-tsvwg-ecn-encap-guidelines] provides guidelines for adding congestion notification to protocols that encapsulate IP.

Roaming: Independent of the specific architecture, mobile communication networks typically differentiate between non-roaming and roaming scenarios. Roaming scenarios are typically more demanding regarding implementing operator policies, charging etc. It can be expected that this would also hold for deploying CONEX. A more detailed analysis of this problem will be provided in a future revision of this document.

It is important to note that CONEX is intended to be used as a supplement and not a replacement to the existing QoS mechanisms in mobile networks. For example, CONEX deployed in 3GPP mobile networks can provide useful input to the existing 3GPP PCC mechanisms by supplying more dynamic network information to supplement the fairly static information used by the PCC. This would enable the mobile network to make better policy control decisions than is possible with only static information.

6. IANA Considerations

No IANA considerations.

7. Security Considerations

Security considerations for applying CONEX to EPS include, but are not limited to, the security considerations that apply to the CONEX protocols.

8. Informative References

[3GPP.23.203]
3GPP, "Policy and charging control architecture", 3GPP TS 23.203 10.5.0, December 2011.

[3GPP.23.402]
3GPP, "Architecture enhancements for non-3GPP accesses", 3GPP TS 23.402 10.6.0, December 2011.

[3GPP.23.829]
3GPP, "Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)", 3GPP TR 23.829 10.0.1, October 2011.

[3GPP.26.114]
3GPP, "IP Multimedia Subsystem (IMS); Multimedia telephony; Media handling and interaction", 3GPP TS 26.114 10.2.0, December 2011.

[3GPP.29.060]
3GPP, "General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface", 3GPP TS 29.060 3.19.0, March 2004.

[3GPP.29.274]
3GPP, "3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control plane (GTPv2-C); Stage 3", 3GPP TS 29.274 8.11.0, December 2011.

[3GPP.36.300]
3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2", 3GPP TS 36.300 8.12.0, April 2010.

[I-D.briscoe-conex-initial-deploy]
Briscoe, B., "Initial Congestion Exposure (ConEx) Deployment Examples", draft-briscoe-conex-initial-deploy-01 (work in progress), November 2011.

[I-D.briscoe-tsvwg-ecn-encap-guidelines]
Briscoe, B., "Guidelines for Adding Congestion Notification to Protocols that Encapsulate IP", draft-briscoe-tsvwg-ecn-encap-guidelines-00 (work in progress), March 2011.

[I-D.ietf-conex-abstract-mech]
Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism", draft-ietf-conex-abstract-mech-03 (work in progress), October 2011.

[I-D.ietf-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.

[I-D.ietf-conex-destopt]
Krishnan, S., Kuehlewind, M., and C. Ucendo, "IPv6 Destination Option for Conex", draft-ietf-conex-destopt-01 (work in progress), October 2011.

[RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, November 2010.

[nec.euronf-2011]
Mir, Kutscher, and Brunner, "Congestion Exposure in Mobility Scenarios", in proceedings of 7th EURO-NF CONFERENCE ON NEXT GENERATION INTERNET, June 2011.

[nec.globecom2010]
Kutscher, Lundqvist, and Mir, "Congestion Exposure in Mobile Wireless Communications", in proceedings of IEEE GLOBECOM 2010, December 2010.

[raghavan2007]
Raghavan, Vishwanath, Ramabhadran, Yocum, and Snoeren, "Cloud Control with Distributed Rate Limiting", in proceedings of ACM SIGCOMM 2007, 2007.

DOI: http://doi.acm.org/10.1145/1282427.1282419

Appendix A. Acknowledgments

We would like to thank Bob Briscoe and Ingemar Johansson for their support in shaping the overall idea and in improving the draft by providing constructive comments.

Authors' Addresses

Dirk Kutscher
NEC
Kurfuersten-Anlage 36
Heidelberg,
Germany

Phone:
Email: kutscher@neclab.eu

Faisal Ghias Mir
NEC
Kurfuersten-Anlage 36
Heidelberg,
Germany

Phone:
Email: faisal.mir@neclab.eu

Rolf Winter
NEC
Kurfuersten-Anlage 36
Heidelberg,
Germany

Phone:
Email: winter@neclab.eu

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

Phone:
Email: suresh.krishnan@ericsson.com

Carlos Jesus Bernardos Cano
Universidad Carlos III de Madrid

Email: cjbc@it.uc3m.es

Conex Group
Internet Draft
Intended Status: Informational
Expires: September 7, 2011

D. McDysan
Verizon

March 7, 2011

Usage/Volume Tier Feedback Use Case for Congestion Exposure

draft-mcdysan-conex-volumetier-usecase-00.txt

Status of this Memo

Copyright Notice

Abstract

As requested in the Beijing meeting, this is an individual draft that expands on the usage tier/volume use case from [McDysan]. The feedback recorded in the Beijing conex meeting minutes was that a number of people were potentially interested in this use case, but that it was out of scope of the current charter, and/or could potentially be built out of the conex abstract mechanism.

Table of Contents

1. Introduction

As requested in the Beijing meeting, this is an individual draft that expands on the usage tier/volume use case from [McDysan]. The feedback recorded in the Beijing conex meeting minutes was that a number of people were interested in this idea, but that it was out of scope of the current charter, and/or could potentially be built out of the conex abstract mechanism.

Section 3 provides some motivational background and a statement of relevant problems involved with congestion pricing.

Section 4 provides text for the usage/volume tier feedback use case. It contains a section that covers a problem statement, objectives for resolving this issue, potential approaches for implementing this use case employing currently defined conex mechanisms, and a description of additional measures and mechanisms that could solve the stated problem and issues.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

2.1. Acronyms

conex = congestion exposure

2.2. Terminology

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

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

3. Motivation and Background

This section provides references to relevant presentations given by experts on congestion pricing from the IETF 78 Technical Plenary in Maastricht. A significant number of ISPs implement some form of usage/volume caps in at least some parts of the world [NewAmerica].

The congestion exposure problems addressed in this document are:

- Volume-based pricing makes it difficult for users to manage costs incurred, [Varian]

- Customers will pay a premium for unmetered use [Varian]

There are three time scales over which congestion pricing can operate [Johari]: short (milliseconds to seconds), medium (minutes to hours to days) and long (months to years). Currently, the short term congestion signal is lost packets or a specific indication of congestion of a particular resource (e.g., a ECN indication for queue congestion), as stated in the conex charter, [UseCases] and [Mechanism]. Usage/Volume tier pricing is typically of a long timescale (months), but users may be able to make effective of shorter time scale feedback (minutes to hours).

4. Usage Tier/ Volume Feedback

Usage/volume caps may be arranged into multiple tiers with different pricing based upon monthly volume. This results in some problems as described in the next section. Next, high-level objectives to address these issues are then proposed. Then potential ways that already defined means [Mechanisms] may be employed are described. Finally, to address some of these issues, other measures and mechanisms that could possibly better meet the objectives are described.

4.1. Problem Statement

Long-term (e.g., monthly) usage/volume based pricing is a widely used incentive, but it creates the following issues;

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

- It does not address the situation where heavy users [Usecases] send at a high rate, but only for a fraction of the usage measurement interval (e.g., only for a few hours or days during a month).

- If usage/volume counting is performed differently dependent upon the degree of congestion experienced, then feedback is more important since in general users will not know when congestion is occurring.

- If a user marks packets as requesting lower effort [LowerEffort], then an incentive could be to not count (or count in a different way) packets marked as lower effort against a usage/volume tier. Some way of ensuring that packets marked as lower effort do not significantly impact packets marked as best (or better) effort.

4.2. Objectives for Addressing this Issue

Provide a way to inform a receiver of the usage/volume incurred to a moment in time. Ideally, this would also include the usage/volume time period (e.g., a month).

Provide a way to inform a receiver of a trend that if usage continues at the same rate then a specific usage/volume tier will be crossed.

Indicate to a receiver whether usage/volume counting is occurring in a different way when congestion measure of a particular form (e.g., loss, ECN marking) is occurring.

Standardize a way to mark packets in a way (e.g., [Lower Effort]) in conjunction with some form of conex signaling that indicates usage/volume counting will not occur (or are counted separately) under the condition that these packets do not create congestion. A means to ensure that these marked packets do not create congestion and do not impact best (and better) effort marked packets is also required.

Enable a means for recharging to occur, where usage/volume counting does not occur for the receiving user since some other party has agreed to incur the cost of usage/volume for that flow.

4.3. Potential Support Using Abstract Mechanism

The conex abstract mechanism [Mechanism] defines implicit signaling of loss and explicit signaling of ECN marking. It also defines re-echoed signal for loss and ECN marking based upon feedback carried by TCP from a receiver back to the sender (in an RFC to be developed by the conex wg as defined in the charter). If this feedback mechanism is designed to be extensible, then a variety of forms of feedback could be developed for use in experiments.

Counting usage/volume differently for congested packets (or congested intervals) based upon [Mechanism] re-echoed congestion experienced signals seems straightforward. This could be a local matter for the IP node which implements usage/volume tier counting.

Also, counting packets marked as Lower Effort differently is a local matter. How to ensure that these packets do not interfere with best effort could be implemented by Diffserv methods locally and at other potential bottlenecks.

Since packets subject to counting in a usage/volume cap may not occur during congestion intervals, reinsertion of such counting information using the re-echoed signals that indicate congestion does not seem possible since the same bits cannot represent usage counting and congestion experienced.

What is missing from the current conex mechanism is a feed forward path operating over a longer timescale that contains sufficient information to meet the objectives.

4.4. Additional Support Using other Measures and Mechanisms

Usage/volume counting has some aspects similar to that of a congestible queue, but on a much longer timescale, as follows:

- Instead of a queue which is typically sized for O(10 ms) at the sending rate, usage/volume counting occurs on a timescale O(month.

- A usage/volume tier is a threshold on a long term usage counter, similar to the way ECN marking can be a threshold on in a queue.

- Queue loss is similar to a usage/volume counter crossing from one tier into the next.

- A usage/volume tier trend warning is similar to a rate estimate for ECN marking based upon queue fill rate, as is described in PCN.

Therefore, in an abstract way a usage/volume counter can be viewed as a congestible resource, but in some ways not the same as a congestible queue. If this information is to be fed forward in a way observable at the IP layer and fed back at the transport layer (e.g.,

TCP), then additional packet and transport fields and/or mechanisms may be better suited to this purpose.

Furthermore, instead of feeding forward information in each IPv6 packet as in [Mechanism], usage/volume congestible resource information can occur much less frequently (e.g., many minutes to hours). The following is an outline of such measures and mechanisms.

The basic idea is based on the fact that the sender and receiver need to be cooperating using the same experimental extensions to TCP, and that if TCP can carry some of the additional information, then the scarcity of IPv6 header bits is avoided. Furthermore, as described previously, "fast path" processing is not required for this use case and the hop-by-options field of the IPv6 header could be used [RFC2460], [IPv6Format] with "slow path" processing used instead. A mechanism to allow the experimental sender to send a "probe" in the IPv6 packet (e.g., using an experimental IPv6 protocol type) could be used by a intermediate IP node(s) to forward the "probe" packet to a special processor (which may be separate from the routers' processor). This special processor could use a polled version of usage/volume count information per user and could also be configured with subscription information (e.g., usage/volume cap tier, cap duration), and threshold settings.

The handling of this "probe" IPv6 packet and associated TCP segment needs to be done within the TCP flow. It could use an Out Of Band mechanism similar to the urgent data capability in TCP. (For example, an experimental usage of the Urgent bit could possibly be employed.) The special processor could insert additional measures and implement some of the proposed mechanisms and then modify/augment the "probe" TCP (urgent-like) segment with the requested information and forwards this modified "probe" packet toward the receiver via the intermediate IP node. Packets from the receiver back to the sender could be sent directly, or could be directed through the special processor at intermediate node(s), depending upon the specifics of the use case involved.

A consequence of the above extension of measures and mechanisms is that the sender and receiver now have much more information which could be used to solve the stated problems and meet the objectives.

The information carried in a "probe" TCP segment could include:

- The service being requested, for example:
    - Request information on the users' usage/volume tier
    - Request statistics on usage
    - Request threshold trend report
    - Request not counting this flow since it is lower effort

- Request recharging

- Information that could be provided by the "special processor" includes:

  - Duration and cap for the usage volume measurement tier (e.g., a month)

  - The absolute count of packets and octets received/sent, and/or fraction of the usage tier already used

  - Count of packets and octets received/sent which experienced congestion

  - Count of packet and octets received/sent that were marked as Lower Effort

  - Estimate of whether the user will exceed the usage tier if the historical usage rate to the reporting instant continues

  - A pointer (e.g., URL) and identification of the authentication method that would enable other queries, and/or implement alternative charging methods (e.g., recharging)

  - Other measures related to the "congestion" of a usage/volume tier use case (or possibly other use cases as well).

One example of a different type of measure is described in [Stanojevic]. In this paper, the Shapley value [Shapley] is used instead of a 95-th percentile measure of hourly usage measurements across a month. The Shapley value has the following desirable intuitive properties [Shapley]: individual fairness, efficiency, symmetry and additivity. Although the 95-th percentile measure is not directly related to the usage/volume tier use case, the authors state that this is a case they plan to address in future research. A mapping, such as an approximation to the Shapley value described in this paper, could be a way to compress the usage/volume tier feed forward/ feedback information into a smaller number of bits that represents the incentives described in the objectives section.

5. Security Considerations

In the proposed mechanisms there are indications that could be spoofed and/or used to game counting and congestion feedback mechanisms, and therefore an authentication mechanism may be needed when this information is handled at the TCP/IPv6 layer in the sender to destination direction or at the TCP layer in the destination to sender direction.

6. IANA Considerations

None

7. Acknowledgements

The idea of not counting lower effort traffic against a usage/volume cap was suggested my Mikael Abrahamsson on the conex mailing list.

8. References

8.1. Normative References

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

8.2. Informative References

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

[Mechanism] M. Mathis, B. Briscoe, "Congestion Exposure (ConEx) Concepts and Abstract Mechanism," draft-mathis-conex-abstract-mech-00, Work in Progress

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

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

[NewAmerica] Li, Losey, "Bandwidth Caps for Residential High- Speed Internet in the U.S. and Japan," August 2009, http://www.newamerica.net/files/Bandwidth%20Caps%20for%20High-Speed%20Internet%20in%20the%20U.S.%20and%20Japan.pdf

[LowerEffort] R. Bless, K. Nichols, K. Wehrle, "A Lower Effort Per-Domain Behavior (PDB) for Differentiated Services," RFC3662, December 2003

[RFC2460] S. Deering, R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification," RFC 2460 December 1998

[IPv6Format] S. Krishnan, M. Kuehlewind, "IPv6 Format," Presentation at conex wg at IETF 79, Beijing

[Stanojevic] Stanojevic, Laoutaris, Rodriguez, Telefonica Research, "On Economic Heavy Hitters: Shapley value analysis of 95th-percentile pricing," IMC'10, November 1-3, 2010, Melbourne, Australia, http://conferences.sigcomm.org/imc/2010/papers/p75.pdf

[Shapley] Wikipedia, "Shapley value," http://en.wikipedia.org/wiki/Shapley_value

9. Acknowledgments

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

Authors' Addresses

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