ConEx
Internet-Draft
Intended status: Informational
Expires: September 14, 2012

B. Briscoe, Ed.
BT
D. Kutscher
NEC

March 13, 2012

Initial Congestion Exposure (ConEx) Deployment Examples
draft-briscoe-conex-initial-deploy-02

Abstract

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 14, 2012.

Table of Contents

1. Introduction 3
2. Recap: Incremental Deployment Features of the ConEx Protocol 3
3. ConEx Components 4
3.1. Recap of Basic ConEx Components 4
3.2. Per-Network Deployment Concepts 4
4. Example Initial Deployment Arrangements 5
4.1. Single Receiving Network Scenario 5
4.1.1. ConEx Functions in the Single Receiving Network Scenario 7
4.1.2. Incentives to Unilaterally Deploy ConEx in a Receiving Network 8
4.2. Mobile Network Scenario 9
4.2.1. CONEX Functions in a Mobile Network Scenario 12
4.2.2. Incentives to Unilaterally Deploy CONEX in a Mobile Operator Network 13
4.3. Scenario Internal to a Multi-Tenant Data Centre 13
4.3.1. Incremental Deployment of ConEx Scenario in a Multi-Tenant Data Centre 15
5. Security Considerations 15
6. IANA Considerations 15
7. Conclusions 15
8. Acknowledgments 16
9. Informative References 16
Appendix A. Summary of Changes between Drafts 17

1. Introduction

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

2. Recap: Incremental Deployment Features of the ConEx Protocol

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

- The ConEx mechanism is essentially a change to the source, in order to re-insert congestion feedback into the network.
- Source-host-only deployment is possible without any negotiation required, and individual transport protocol implementations within a source host can be updated separately.
- Receiver modification may optionally improve ConEx for some transport protocols with feedback limitations (TCP being the main example), but it is not a necessity
- Proxies for the source and/or receiver are feasible (though not necessarily straightforward)
- Queues and network forwarding do not require any modification for ConEx.
- ECN is not required in the network for ConEx. If some network nodes support ECN, it can be used by ConEx.
- ECN is not required at the receiver for ConEx. The sender should nonetheless attempt to negotiate ECN-usage with the receiver, given some aspects of ConEx work better the more ECN is deployed, particularly auditing and border measurement.
- Given ConEx exposes information for IP-layer policy devices to use, the design does not preclude possible innovative uses of ConEx information by other IP-layer devices, e.g. forwarding itself
- Packets indicate whether or not they support ConEx.

3. ConEx Components

3.1. Recap of Basic ConEx Components

[ConEx-Abstract-Mech] introduces the following components:

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

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

3.2. Per-Network Deployment Concepts

Network deployment-related definitions:

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

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

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

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

Typically, a network will use ConEx information by deploying a policy function at the ingress edge of its network to monitor arriving traffic and to act in some way on the congestion information in those packets that are ConEx-enabled. Actions might include policing,

altering the class of service, or re-routing. Alternatively, less direct actions via a management system might include triggering capacity upgrades, triggering penalty clauses in contracts or levying charges between networks based on ConEx measurements.

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

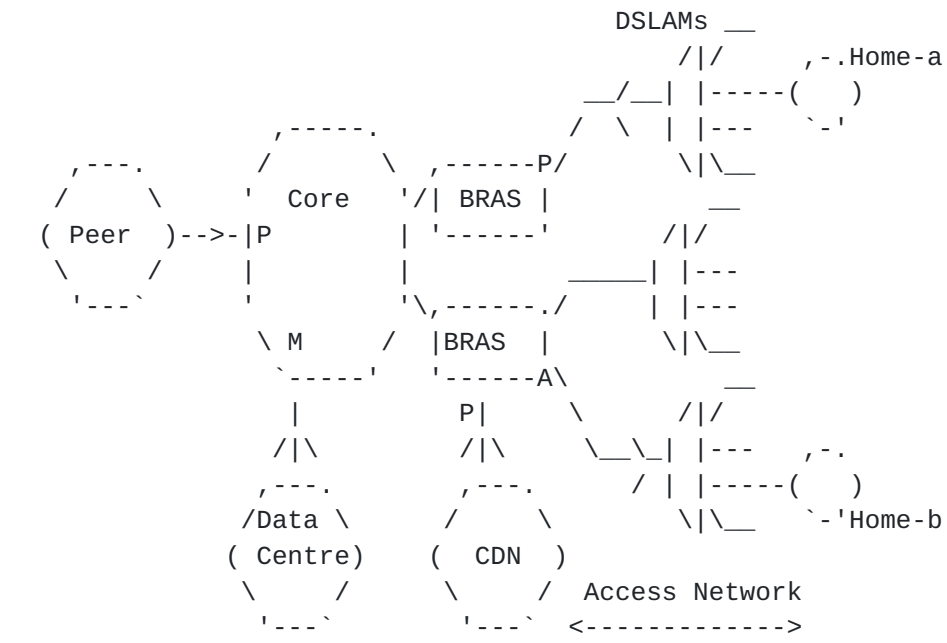
Of course, because each edge device usually serves as both an ingress and an egress, the two functions are both likely to be present in each edge device.

4. Example Initial Deployment Arrangements

In all the deployment scenarios below, we assume that deployment starts with some data sources being modified with ConEx code. The rationale for this is that the developer of a scavenger transport protocol like LEDBAT has a strong incentive to tell the network how little congestion it is causing despite sending large volumes of data. In this case the developer makes the first move expecting it will prompt at least some networks to move in response--so that they use the ConEx information to reward users of the scavenger protocol.

4.1. Single Receiving Network Scenario

The name 'Receiving Network' for this scenario merely emphasises that most data is arriving from connected networks and data centres and being consumed by residential customers on this access network. Some data is of course also travelling in the other direction.



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

Figure 1: Single Receiving Network Scenario

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

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

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

This operator of the data centre, core and access network is the only one in the diagram to have deployed ConEx monitoring and policy devices at the edges of its network. However, it has not enabled ECN on any of its network elements and neither has any other network in the diagram. The operator has deployed a congestion policing function (P) on the provider-edge router where the peer attaches to

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

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

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

4.1.1. ConEx Functions in the Single Receiving Network Scenario

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

- either self-congestion in the queues into the logical ports representing the access lines
- or shared congestion in the shared queue on the BRAS that feeds them.

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

The lower half of the figure shows (downstream congestion) = (whole path) - (upstream congestion). Upstream congestion can only be

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

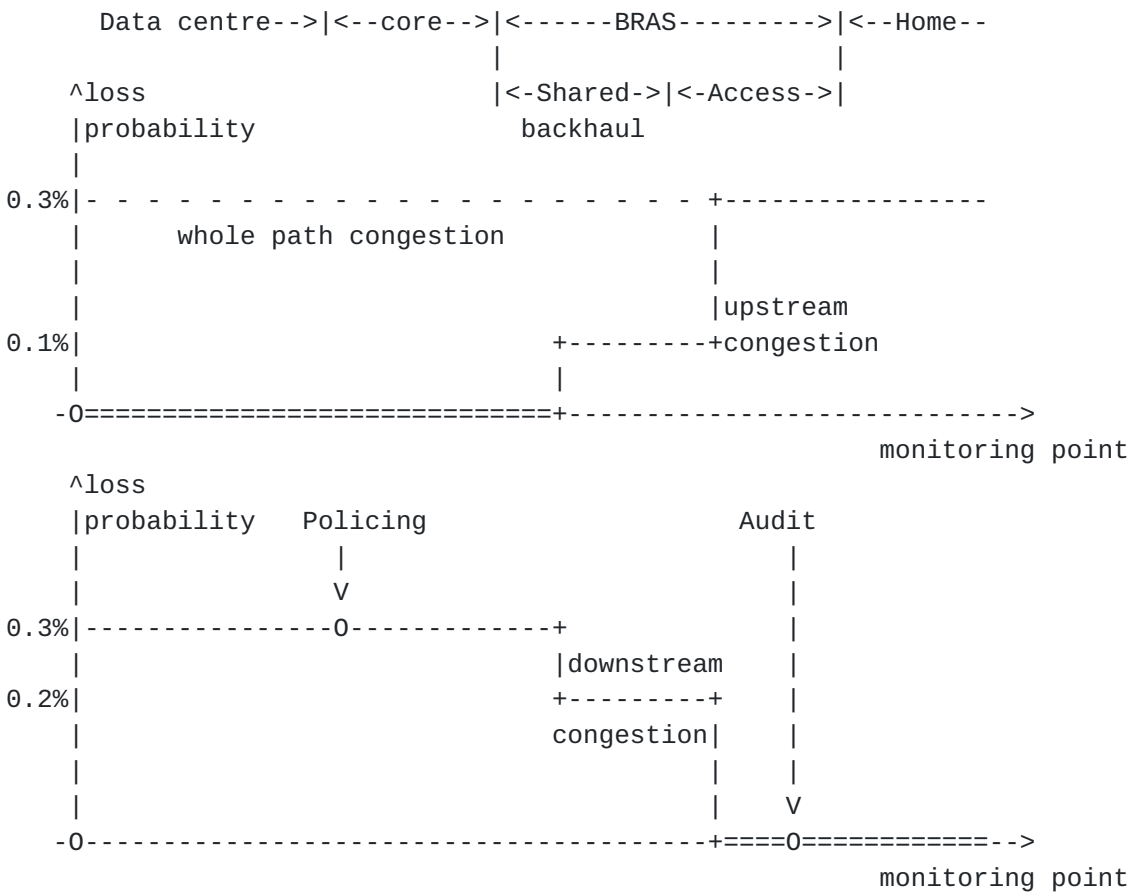


Figure 2: Example plot of loss levels along a path

4.1.2. Incentives to Unilaterally Deploy ConEx in a Receiving Network

Even a sending application that is modified to use ConEx can choose whether to send ConEx or Not-ConEx packets. Nonethelss, ConEx packets bring information to a policer about congestion expected on the rest of the path beyond the policer. Not-ConEx packets bring no such information. Therefore a network that has deployed ConEx policers will tend to rate-limit not-ConEx packets conservatively in order to manage the unknown risk of congestion. In contrast, a network doesn't normally need to rate-limit ConEx-enabled packets

unless they reveal a persistently high contribution to congestion. This natural tendency for networks to favour senders that provide ConEx information encourages senders to choose to use the ConEx protocol whenever they can.

{ToDo: complete this section}

4.2. Mobile Network Scenario

Mobile networks (in general, but we focus on 3GPP EPS here) are another type of network that is generally amenable to initial CONEX deployment because of its need to make congestion visible to the network:

Congestion management is highly important: mobile network operators have traditionally gone to great extent to detect and act upon congestion at different locations in their networks. Capacity investments are high, (especially) wireless resources have been comparatively scarce, and many physical resources (wireless links, backhaul links, core networks) are shared.

Evolving from highly differentiated services to 'best-effort' communication: The conversion to IP-based communication and to ubiquitous Internet access services has rendered traditional models of fine-granular differentiated services too inefficient and complicated. The majority of flows are mapped onto best-effort bearers -- which calls for appropriate resource sharing and accounting models for such flows.

Demand for congestion exposure at different levels: The demand for more appropriate resource sharing in heavy usage scenarios has led to an increased deployment of Deep-Packet Inspection (DPI) -- there is an obvious demand for informing the network about congestion on roundtrip time scales. Moreover, 3GPP mobile network operators require congestion information at different time-scales, specifically on network-management time scales: Identifying hot-spots, analyzing overload situations and assisting network planning is routinely done by "drive tests" -- which could be simplified with a CONEX approach. Congestion and base station load information is also exchanged in Self-Organized Networking (SON) to assist cell capacity optimization and hand-over decisions (at smaller time-scales).

Mobile networks are also amenable to initial CONEX deployment because they already provide many prerequisites:

Elaborate and flexible policy and charging architecture: Mobile networks today employ an elaborate and flexible policy and charging infrastructure that can easily be advanced to account for congestion contribution (instead of data volume as in many current deployments) and that could thus provide incentives for CONEX adoption and sender behavior.

Well integrated overall system: 3GPP specification cover many parts of the overall system, including (for example) ECN usage by mobile terminals. It would thus be quite feasible to introduce CONEX to such networks (without requiring CONEX support in non-3GPP networks) by specifying its detailed usage in the corresponding specifications.

Frequent usage of gateways and proxies It is quite common that actual deployments employ proxy caches, TCP proxies etc., which introduces additional options for an initial deployment (for instance by only modifying proxy TCP senders at a very early phase).

The EPS architecture and its standardized interfaces are depicted in Figure 3. 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. Please see [conex-mobile] for a detailed description of the individual elements.

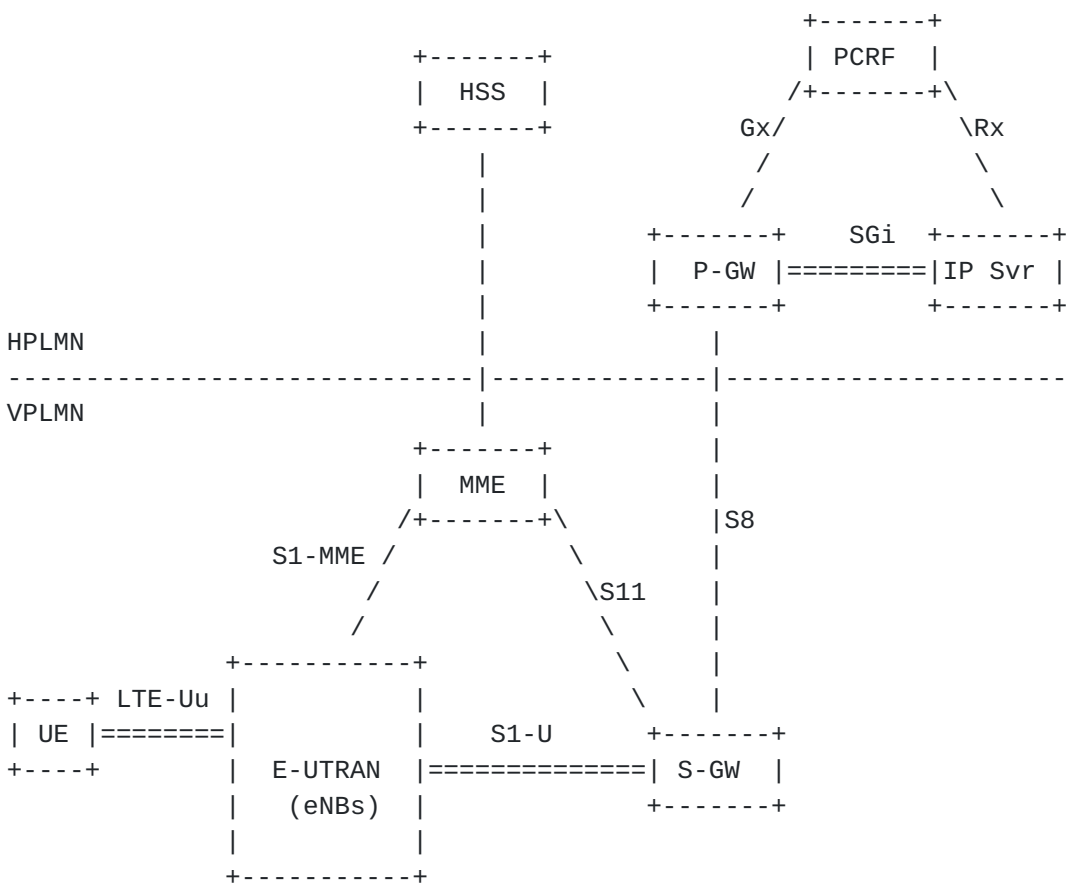


Figure 3: EPS Architecture Overview

Figure 3 does not only depict data path elements but also mobility management, home subscriber servers (HSS) etc, distinguishing home networks and visited networks. Figure 4 depicts a simplified network, focusing on data path elements only.

In Figure 4 depicts a fairly simple deployment scenario, where 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.

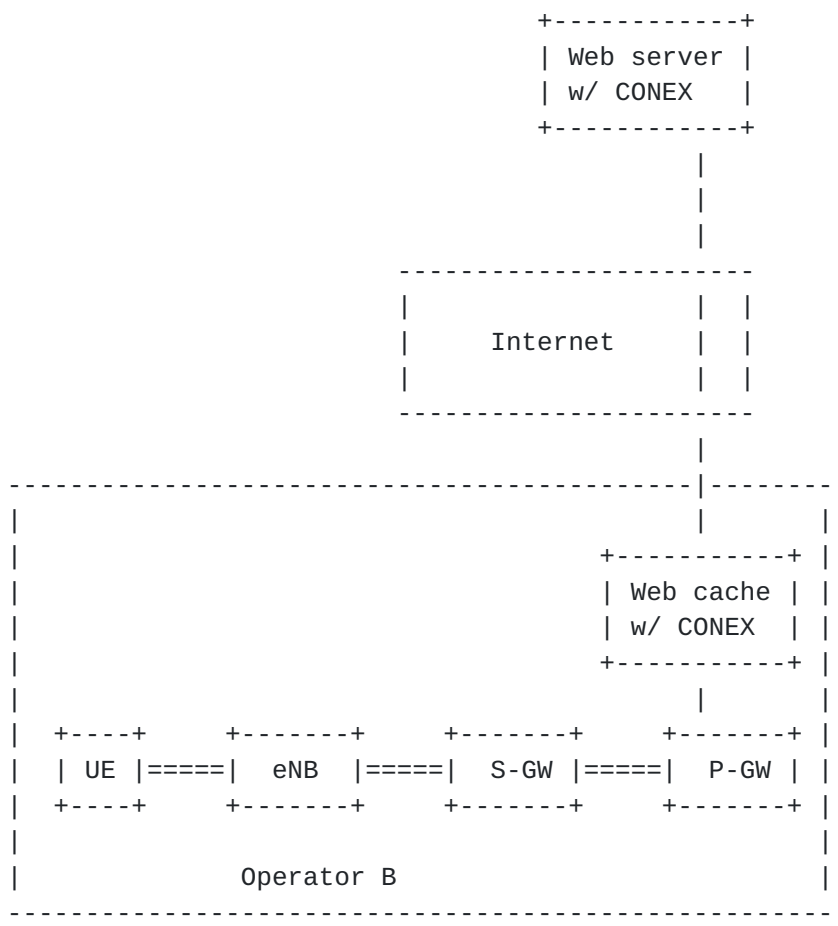


Figure 4: CONEX support on servers and caches

Logical CONEX functions would be mapped to network elements as follows:

CONEX sender: Web cache

(Unmodified) receiver: UE

Policer: P-GW

Audit function: eNB (optional, since operator controls sender)

4.2.1. CONEX Functions in a Mobile Network Scenario

In a mobile network, shared congestion can occur at different places, i.e., in the radio access network, on backhaul links, and in the core network.

In this specific scenario, we assume that not all downlink traffic is CONEX-enabled, but that all (TCP) traffic that originates from the

operators web cache is.

Assuming unmodified receivers (UEs) the main CONEX function that is of interest in this scenario is congestion accountability for web traffic: a policing entity on the P-GW would be able to account for congestion contribution for downlink web traffic per user -- and possibly transfer corresponding information to charging. (Different operator policies are possible -- for instance, it is also possible to police traffic more strictly, after a certain congestion contribution budget has been used in a accounting period.)

Congestion exposure could also be used for traffic offload decisions, for example when downstream entities detect upstream congestion (in the core network).

Moreover, congestion exposure could also be used in longer timer frame network management applications, i.e., downstream nodes in the network access network could report on upstream vs downstream congestion statistics on aggregated flows to assist performance optimizations, network planning etc.

4.2.2. Incentives to Unilaterally Deploy CONEX in a Mobile Operator Network

In mobile networks, both mobile terminals and mobile network equipment are standardised by the 3GPP. This represents a much more centralised standardisation model, where if the 3GPP were to adopt the ConEx protocol, it might mandate ConEx implementation for compliant equipment. Initially 3GPP might mandate ConEx only in user equipment, then each operator could choose (or not) to use ConEx information for traffic management. This would also have the interesting side-effect of making ConEx mode widely available outside cellular networks, given 3GPP user equipment roams elsewhere.

The comparatively non-invasive addition of CONEX support described in the previous section enables operators to add CONEX-based congestion accountability for a considerable fraction of the traffic (all cacheable web traffic). It is independent of other operators and independent of other forms of congestion management (DPI-based for example). But compared to other forms of congestion management, this approach does not require DPI, and it can be extended to other traffic types (in addition to HTTP) in a later deployment phase. The existing policy and charging infrastructure can be leveraged.

4.3. Scenario Internal to a Multi-Tenant Data Centre

A number of companies offer hosting of virtual machines on their data centre infrastructure--so-called infrastructure as a service (IaaS).

A set amount of processing power, memory, storage and network are offered. Although processing power, memory and storage are relatively simple to allocate on the 'pay as you go' basis that has become common, the network is less easy to allocate given it is a naturally distributed system.

The design involves the following elements, all involving changes solely in the hypervisor or operating systems, not network switches:

- A bulk congestion policing function to police all the traffic from a VM into the network (similar to [CongPol]), implemented as a shim in the hypervisor;

  A customer may run virtual machines on multiple physical nodes, in which case the data centre operator would ensure that it deployed a policer in the hypervisor on each node where the customer was running a VM, at the time the each VM was instantiated.The DC operator would arrange for them to collectively enforce the per-customer congestion allowance, as a distributed policer.

- A function to distribute a customer's tokens to the policer associated with each of the customer's VMs. This could be similar to the distributed rate limiting of [DRL]]), or a logically centralised bucket of congestion tokens could be used with simple 1-1 communication between it and the local token bucket in the hypervisor under each VM. Importantly, traditional bit-rate tokens cannot simply be reassigned from one VM to another without implications on the balance of network loading (requiring operator intervention each time), whereas congestion tokens can be freely reassigned between different VMs, because a congestion token is equivalent at any place or time in a network;

- Reinsertion of congestion feedback at the sending side, which may be implemented:

  * either as a shim in both sending and receiving hypervisors using edge-to-edge feedback (as in Seawall [Seawall]).

  * or in the sending operating system using the congestion exposure protocol (ConEx [ConEx-Abstract-Mech]);

  If the Seawall option is used, a feedback proxy will also be required as a shim in the hypervisor at the receiver. This passes congestion feedback that the network operator can trust to the sending hypervisor, by creating a tunnel between the hypervisors. Seawall uses a local variant of the Internet Protocol within the data centre to implement this tunnel.

If the ConEx option is used, a congestion audit function will also be required as a shim in the hypervisor (or container) layer where data leaves the network and enters the receiving host. The ConEx option is only applicable if the guest OS at the sender has been modified to send ConEx markings in IPv6 using [conex-destopt]. In addition, the ConEx options could be encoded in the IPv4 header by hiding them within the packet ID field as described in [intarea-ipv4-id-reuse].

- o Network switches would not need any modification. However, audit would be easier if switches supported ECN. Ideally data centre TCP could be used as well, although not essential. DCTCP is based on ECN and designed for data centres. DCTCP involves a more aggressive AQM in layer 3 switches with a shallow step threshold for ECN marking. DCTCP also requires modified sender and receiver TCP algorithms.

4.3.1. Incremental Deployment of ConEx Scenario in a Multi-Tenant Data Centre

The Seawall option above is a more processing intensive change to the hypervisors, but it can be deployed unilaterally by the data centre operator in al hypervisors (or containers).

The ConEx option above is only applicable if a particular guest OS supports the marking of outgoing packets with ConEx markings.

A simple filter could be installed in each hypervisor to allow ConEx packets through into the data centre network without going through the SeaWall tunnel structure, while non-ConEx packets could be tunnelled as per SeaWall. This would provide an incremental deployment scenario with the best of both worlds: it would work for unmodified guest OSs, but for guest OSs with ConEx support, it would require less processing (therefore being faster) and not require a duplicate feedback channel between hypervisors.

5. Security Considerations

6. IANA Considerations

This document does not require actions by IANA.

7. Conclusions

{ToDo}

8. Acknowledgments

9. Informative References

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

[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>.

[DRL] Raghavan, B., Vishwanath, K., Ramabhadran, S., Yocum, K., and A. Snoeren, "Cloud control with distributed rate limiting", ACM SIGCOMM CCR 37(4)337--348, 2007, <http://doi.acm.org/10.1145/1282427.1282419>.

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

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

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

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

[intarea-ipv4-id-reuse] Briscoe, B., "Reusing the IPv4 Identification Field in Atomic Packets",

draft-briscoe-intarea-ipv4-id-reuse-01 (work in progress), March 2012.

Appendix A. Summary of Changes between Drafts

Detailed changes are available from http://tools.ietf.org/id/draft-briscoe-conex-initial-deploy-00.txt

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

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

*

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

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/

Dirk Kutscher
NEC
Kurfuersten-Anlage 36
Heidelberg,
Germany

Phone:
EMail: kutscher@neclab.eu