CONEX
Internet-Draft
Intended status: Informational
Expires: December 23, 2010

B. Briscoe
BT
R. Woundy
Comcast
T. Moncaster, Ed.
Moncaster.com
J. Leslie, Ed.
JLC.net
June 21, 2010

Congestion Exposure Mechanism Description
draft-conex-mechanism-00

Abstract

Internet Service Providers (ISPs) are facing problems where congestion prevents full utilization of the path between sender and receiver at today's "broadband" speeds. ISPs desire to control the congestion, which often appears to be caused by a small number of users consuming a large amount of bandwidth. Building out more capacity along all of the path to handle this congestion can be expensive; and network operators have sought other ways to manage congestion. The current mechanisms all suffer from difficulty measuring the congestion (as distinguished from the total traffic).

The ConEx Working Group is designing a mechanism to make congestion along any path visible at the Internet Layer. This document discusses this mechanism.

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 December 23, 2010.

Table of Contents

1. Introduction

The growth of "always on" broadband connections, coupled with the steady increase in access speeds [OfCom], has meant network operators are increasingly facing problems with congestion. But congestion results from sharing network capacity with others, not merely from using it. In general, today's "DSL" and cable-internet users cannot "cause" congestion in the absence of competing traffic. (Wireless ISPs and cellular internet have different tradeoffs which we will not discuss here.)

Actual congestion generally results from the interaction of traffic from an ISPs own subscribers with traffic from other users. The tools currently available don't allow an operator to identify the causes of the congestion and so leave them powerless to properly control it.

While building out more capacity to handle increased traffic is always good, the expense and lead-time can be prohibitive, especially for network operators that charge flat-rate feeds to subscribers and are thus unable to charge heavier users more for causing more congestion [BB-incentive]. For an operator facing congestion caused by other operators' networks, building out its own capacity is unlikely to solve the congestion problem. Operators are thus facing increased pressure to find effective solutions to dealing with high-consuming users.

The growth of "scavenger-class" services helps to reduce congestion, but actually make the ISPs problem less tractable. These are services where participating users are not at all interested in paying more, but wish to make good use of the capacity of the path. Thus, users of such services may show very heavy total traffic up until the moment congestion is detected (at the Transport Layer), but immediately back off. ISP monitoring (at the Internet Layer) cannot detect this congestion avoidance if the congestion in question is in a different domain further along the path; and must treat such users

as congestion-causing users.

We propose that Internet Protocol (IP) packets have two "congestion" fields. The exact protocol details of these fields are for another document, but we expect them to provide measures of "congestion so far" and "congestion still expected".

2. Definitions

Since conex expects to build on Explicit Congestion Notification (ECN) [RFC3168], we use the term "congestion" in a manner consistent with ECN, namely that congestion occurs before any packet is dropped.

We define five specific terms carefully:

Congestion: Congestion is a measure of the probability that a given packet will be ECN-marked or dropped as it traverses the network. At any given router it is a function of the queue state at that router. Congestion is added in a combinatorial manner, that is, routers ignore the congestion a packet has already seen when they decide whether to mark it or not.

Upstream Congestion: The congestion that has already been experienced by a packet as it travels along its path. In other words at any point on the path, it is the congestion between the source of the packet and that point.

Downstream Congestion: The congestion that a packet still has to experience on the remainder of its path. In other words at any point it is the congestion still to be experienced as the packet travels between that point and its destination.

Ingress Router: The Ingress Router is the first router a packet traverses that is outside its own network. In a domestic network that will be the first router downstream from the home access equipment. In a commercial network it may be the first router

downstream of the firewall.

Egress Router: The Egress Router is the last router a packet traverses before it enters the destination network.

3. Existing Approaches to Congestion Management

Initial attempts to capture congestion situations have usually focused on the peak hours and aimed at rate limiting heavy users during that time. For example, users who have consumed a certain amount of bandwidth during the last 24 hours got elected as those who get their traffic shaped if the total amount of traffic reaches a congestion situation in certain nodes within the operator's network.

All of the current approaches suffer from some general limitations. First, they introduce 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 connection is being altered or degraded based on how the network operator manages congestion.

Second, none of the approaches is able to make use of what may be the most important factor in managing congestion: the amount that a given endpoint contributes to congestion on the network. This information

simply is not available to network nodes, and neither volume nor rate nor application usage is an adequate proxy for congestion volume, because none of these metrics measures a user or network's actual contribution to congestion on the network.

Finally, none of these solutions accounts for inter-network congestion. Mechanisms may exist that allow an operator to identify and mitigate congestion in their own network, but the design of the Internet means that only the end-hosts have full visibility of congestion information along the whole path. Conex allows this information to be visible to everyone on the path and thus allows operators to make better-informed decisions about controlling traffic.

4. Exposing Congestion

We argue that current traffic-control mechanisms seek to control the wrong quantity. What matters in the network is neither the volume of traffic nor the rate of traffic: it is the contribution to congestion over time -- congestion means that your traffic impacts other users, and conversely that their traffic impacts you. So if there is no congestion there need not be any restriction on the amount a user can send; restrictions only need to apply when others are sending traffic such that there is congestion.

For example, an application intending to transfer large amounts of data could use a congestion control mechanism like [LEDBAT] to reduce its transmission rate before any competing TCP flows do, by detecting an increase in end-to-end delay (as a measure of impending

congestion). However such techniques rely on voluntary, altruistic action by end users and their application providers. ISPs can neither enforce their use nor avoid penalizing them for congestion they avoid.

The Internet was designed so that end-hosts detect and control congestion. We argue that congestion needs to be visible to network nodes as well, not just to the end hosts. More specifically, a network needs to be able to measure how much congestion any particular traffic expects to cause between the monitoring point in the network and the destination ("rest-of-path congestion"). This would be a new capability. Today a network can use Explicit Congestion Notification (ECN) [RFC3168] to detect how much congestion the traffic has suffered between the source and a monitoring point, but not beyond. This new capability would enable an ISP to give incentives for the use of LEDBAT-like applications whilst restricting inappropriate uses of traditional TCP and UDP ones.

So we propose a new approach which we call Congestion Exposure. We propose that congestion information should be made visible at the IP layer, so that any network node can measure the contribution to congestion of an aggregate of traffic as easily as straight volume can be measured today. Once the information is exposed in this way, it is then possible to use it to measure the true impact of any traffic on the network.

In general, congestion exposure gives ISPs a principled way to hold their customers accountable for the impact on others of their network usage and reward them for choosing congestion-sensitive applications.

5. ECN - a Step in the Right Direction

Explicit Congestion Notification [RFC3168] allows routers to explicitly tell end-hosts that they are approaching the point of congestion. ECN builds on Active Queue Mechanisms such as random early discard (RED) [RFC2309] by allowing the router to mark a packet with a Congestion Experienced (CE) codepoint, rather than dropping it. The probability of a packet being marked increases with the

length of the queue and thus the rate of CE marks is a guide to the level of congestion at that queue. This CE codepoint travels forward through the network to the receiver which then informs the sender that it has seen congestion. The sender is then required to respond as if it had experienced a packet loss. Because the CE codepoint is visible in the IP layer, this approach reveals the upstream congestion level for a packet.

Alas, this is not enough - ECN only allows downstream nodes to measure the congestion so far for any flow. This can help hold a receiver accountable for the congestion caused by incoming traffic. But a receiver can only indirectly influence incoming congestion, by politely asking the sender to control it. A receiver cannot make a sender install an adaptive codec, or install LEDBAT instead of TCP congestion-control. And a receiver cannot cause an attacker to stop flooding it with traffic.

What is needed is knowledge of the downstream congestion level, for which you need additional information that is still concealed from the network.

6. The Proposed Congestion Exposure Mechanism

The protocol we propose is based on a concept known as re-feedback [Re-Feedback], and builds on existing active queue management techniques like RED [RFC2309] and ECN [RFC3168] that network elements can already use to measure and expose congestion.

We propose that packets have two "congestion" fields in their IP header:

- A congestion experienced field to record the upstream congestion level along the path. Routers indicate their current congestion level by updating this field in every packet. As the packet traverses the network it builds up a record of the overall congestion along its path in this field. This data is sent back to the sender who uses it to determine its transmission rate.

- A whole-path congestion field that uses re-feedback to record the total congestion expected along the path. The sender does this by re-inserting the current congestion level for the path into this field for every packet it transmits.

Thus at any node downstream of the sender you can see the upstream congestion for the packet (the congestion thus far) and the whole path congestion (with a time lag of one round-trip-time (RTT)) and can calculate the downstream congestion by subtracting one from the other.

So congestion exposure can be achieved by coupling congestion notification from routers with the re-insertion of this information by the sender. This establishes information symmetry between users and network providers.

7. Conex Use Cases

Conex is a simple concept that has revolutionary implications. It is that rare thing -- a truly disruptive technology, and as such it is hard to imagine the variety of uses it may be put to. However there are several obvious use cases that come to mind with a little thought. The authors aren't claiming all of these have equal merit, nor are we claiming conex is the only conceivable solution to achieve these. But these use cases represent a consensus among people that have been working on this approach for some years.

In the following use cases we are assuming the most abstract version of the conex mechanism, namely that every packet carries two congestion fields, one for upstream congestion and one for downstream. At every node that is congested the upstream congestion value will be incremented in some manner and the downstream congestion value will be decremented. Assuming there is accurate feedback in the system then the aim should be for the downstream value to be zero or slightly positive by the time the packet reaches its destination.

If conex information is to be useful it has to be accurate (within the limitations of the available feedback). This raises three issues that need to be addressed:

Distinguishing conex traffic from non-conex traffic: On one level this seems pretty easy -- conex traffic needs to have the downstream congestion field in every packet. However in practise it may not be as simple as this. Re-ECN is one proposed implementation of conex. Here the two congestion fields are unary-encoded into a stream of packets by effectively setting or clearing a single bit. Assuming you are able to identify non-conex (or legacy) traffic, then you need to decide what to do about it. An ISP may reasonably choose to do nothing different with this traffic. Alternatively they might incentivise the conex traffic in order to give it marginally better service.

Over-declaring congestion: Conex relies on the sender accurately declaring the congestion they expect to see. During TCP slow-start a sender is unable to predict the level of congestion they will experience and it is advisable to declare that expect to see some congestion on the first packet. However, if any host or router marks more than a small fraction of total traffic, downstream routers are less likely to trust its congestion markings. We do not initially propose any mechanism to deal with this issue.

Under-declaring congestion: Conex requires the sender to set the downstream congestion field in each packet to their best estimate of what they expect the whole path congestion to be. If this expected congestion level is to be used for traffic management (see use cases) then it benefits the user to under-declare. Mechanisms are needed to prevent this happening.

There are three approaches that may work (individually or in combination):

* An ingress router can monitor a user's feedback to see what their reported congestion level actually is.

* A conex-aware router can drop any packet with a downstream-congestion value of zero or less if that router is even slightly congested.

* An egress router can actively monitor some or all flows to check that they are complying with the requirement that the downstream congestion value should be zero or (slightly positive) when it reaches the egress.

At any point of congestion, it is reasonable to treat conex-marked traffic differently:

- non-conex traffic will mostly be dropped (as now);

- conex-marked traffic which has exhausted its congestion allowance will (all) be dropped;

7.1. Ingress policing for traffic management

Currently many ISPs impose some form of traffic management at peak hours. This is a simple economic necessity -- the only reason the Internet works as a commercial concern is that ISPs are able to rely on statistical multiplexing to share their expensive core network between large numbers of customers. In order to ensure all customers get some chance to access the network, the "heaviest" customers will be subjected to some form of traffic management at peak times

(typically a rate cap for certain types of traffic) [Fair-use]. Often this traffic management is done with expensive flow aware devices such as DPI boxes or flow-aware routers.

Conex enables a new approach that requires simple per-user policing at the ingress. As described above, every packet a user sends should declare the total congestion that the sender expects that packet to encounter on its journey through the network. Congestion volume has been defined [Fairer-faster] as the congestion a packet experiences,

multiplied by the size of that packet. In effect this is a measure of how much traffic was sent that was above the instantaneous transmission capacity of the network. By extension the congestion rate would be the transmission rate multiplied by the congestion level. A 1 Gbps router that is 0.1% congested implies that there is 1 Mbps of excess traffic.

At the Ingress Router an ISP can police the amount of congestion a user is causing by limiting the congestion volume they send into the network. One system that achieves this is described in [Policing-freedom]. This uses a modified token bucket to limit the congestion rate being sent rather than the overall rate. Such ingress policing is relatively simple as it requires no flow state. Furthermore, unlike many mechanisms, it treats all a user's packets equally.

7.2. Conex to incentivise scavenger transports

Recent work proposes a new approach for QoS where traffic is provided with a less than best effort or "scavenger" quality of service. The idea is that low priority but high volume traffic such as OS updates, P2P file transfers and view-later TV programs should be allowed to use any spare network capacity, but should rapidly get out of the way if a higher priority or interactive application starts up. One solution being actively explored is LEDBAT which proposes a new congestion control algorithm that is less aggressive in seeking out bandwidth than TCP.

At present most ISPs assume a strong correlation between the volume of a flow and the impact that flow causes in the network. This assumption has been eroded by the growth of interactive streaming which behaves in an inelastic manner. Assuming the end-user is using

conex marking on all traffic and that LEDBAT leads to the expected low level of congestion and the ingress ISP has deployed a conex-aware ingress policer, then the LEDBAT will not be penalised since it will be causing less congestion. (If LEDBAT is not conex-marking traffic then the ISP will be forced to guess the congestion, probably based on the total volume).

If the ISP has deployed a conex-aware ingress policer then they are able to incentivise the use of LEDBAT because a user will be policed according to the overall congestion volume their traffic generates. If all background file transfers are only generating a low level of congestion then the sender has more "congestion budget" to "spend" on their interactive applications. It can be shown [Kelly] that this approach maximises social welfare -- in other words if you limit the congestion that all users can generate then everyone benefits from a better service.

7.3. Conex to mitigate DDoS

DDoS relies on subverting innocent end users and getting them to send flood traffic to a given destination. This is intended to cause a rapid increase in congestion in the immediate vicinity of that destination. If it fails to do this then it can't be called Denial of Service. If the ingress ISP has deployed conex policers, that ISP will limit how much DDoS traffic enters the 'net. If the compromised user tries to use the 'net during the DDoS attack, they will quickly become aware that something is wrong, and their ISP can show the evidence that their computer has become zombified.

7.4. Conex as a form of differential QoS

Most QoS approaches require the active participation of routers to control the delay and loss characteristics for the traffic. For real-time interactive traffic it is clear that low delay and low jitter are critical and thus these probably always need different treatment at a router. However if low loss is the issue then conex offers an alternative approach. Assuming the ingress ISP has deployed conex-aware ingress policing then the only control on a user's traffic is dependent on the congestion that user has caused. If they want to prioritise some traffic over other traffic then they can allow that traffic to generate more congestion. The price to pay will be to reduce the congestion that their other traffic causes.

7.5. Other issues

make a source believe it has seen more congestion than it has

hijack a user's identity and make it appear they are dishonest at an egress policer

clear or otherwise tamper with the conex markings

...

8. Security Considerations

This document proposes a mechanism tagging onto Explicit Congestion Notification [RFC3168], and inherits the security issues listed therein. The additional issues from Congestion Expected markings relate to the degree of trust each forwarding point places in Congestion Expected markings it receives, which is a business decision mostly orthogonal to the markings themselves.

One expected use of exposed congestion information is to hold the end-to-end transport and the network accountable to each other. The network cannot be relied on to report information to the receiver against its interest, and the same applies for the information the receiver feeds back to the sender, and that the sender reports back to the network. Looking at each in turn:

- The Network. In general it is not in any network's interest to under-declare congestion since this will have potentially negative

consequences for all users of that network. It may be in its interest to over-declare congestion if, for instance, it wishes to force traffic to move away to a different network or simply to reduce the amount of traffic it is carrying. Congestion Exposure itself won't significantly alter the incentives for and against honest declaration of congestion by a network, but we can imagine applications of Congestion Exposure that will change these incentives. There is a perception among network operators that their level of congestion is a business secret. Today, congestion is one of the worst-kept secrets a network has, because end-hosts can see congestion better than network operators can. Congestion Exposure will enable network operators to pinpoint whether congestion is on one side or the other of any border. It is conceivable that forwarders with underprovisioned networks may try to obstruct deployment of Congestion Exposure.

- The Receiver. Receivers generally have an incentive to under-declare congestion since they generally wish to receive the data from the sender as rapidly as possible. [Savage] explains how a receiver can significantly improve their throughput my failing to declare congestion. This is a problem with or without Congestion Exposure. [KGao] explains one possible technique to encourage receiver's to be honest in their declaration of congestion.

- The Sender. One proposed mechanism for Congestion Exposure deployment adds a requirement for a sender to advise the network how much congestion it has suffered or caused. Although most senders currently respond to congestion they are informed of, one use of exposed congestion information might be to encourage sources of excessive congestion to back off more aggressively.

Then clearly there may be an incentive for the sender to under-declare congestion. This will be a particular problem with sources of flooding attacks. "Policing" mechanisms have been proposed to deal with this.

In addition there are potential problems from source spoofing. A malicious sender can pretend to be another user by spoofing the source address. Congestion Exposure allows for "Policers" and "Traffic Shapers" so as to be robust against injection of false congestion information into the forward path.

9. IANA Considerations

This document does not require actions by IANA.

10. Acknowledgments

The authors would like to thank Contributing Authors Bernard Aboba, Joao Taveira Araujo, Louise Burness, Alissa Cooper, Philip Eardley, Michael Menth, and Hannes Tschofenig for their inputs to this document.

11. References

11.1. Normative References

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

11.2. Informative References

[BB-incentive] MIT Communications Futures Program (CFP) and Cambridge University Communications Research Network, "The Broadband Incentive Problem", September 2005.

[Fair-use] Broadband Choices, "Truth about 'fair usage' broadband", 2009.

[Fairer-faster] Briscoe, B., "A Fairer Faster Internet Protocol", IEEE Spectrum Dec 2008 pp38-43, December 2008.

[KGao] Gao, K. and C. Wang, "Incrementally Deployable Prevention to TCP Attack with Misbehaving Receivers", December 2004.

[Kelly] Kelly, F., Maulloo, A., and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability", Journal of the Operational Research Society 49(3) 237--252, 1998, <http://www.statslab.cam.ac.uk/~frank/rate.html>.

[LEDBAT] Shalunov, S., "Low Extra Delay Background Transport (LEDBAT)", draft-ietf-ledbat-congestion-01 (work in progress), March 2010.

[OfCom] Ofcom: Office of Communications, "UK Broadband Speeds 2008: Research report", January 2009.

[Policing-freedom] 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.

[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.

[Re-Feedback] 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>.

[Savage] Savage, S., Wetherall, D., and T. Anderson, "TCP Congestion Control with a Misbehaving Receiver", ACM SIGCOMM Computer Communication Review , 1999.

Authors' Addresses

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/

Richard Woundy
Comcast
Comcast Cable Communications
27 Industrial Avenue
Chelmsford, MA 01824
US

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

Toby Moncaster (editor)

Moncaster.com
Layer Marney
Colchester CO5 9UZ
UK

EMail: toby@moncaster.com

John Leslie (editor)
JLC.net
10 Souhegan Street
Milford, NH 03055
US

EMail: john@jlc.net