Congestion Exposure	T. Moncaster	
Internet-Draft	L. Krug	
Intended status: Informational	BT	
Expires: September 2, 2010	M. Menth	
	University of Wuerzburg	
	J. Araújo	
	UCL	
	S. Blake	
	Extreme Networks	
	R. Woundy, Ed.	
	Comcast	
	March 1, 2010	

TOC

The Need for Congestion Exposure in the Internet
draft-moncaster-conex-problem-00

Abstract

Today's Internet is a product of its history. TCP is the main transport protocol responsible for sharing out bandwidth and preventing a recurrence of congestion collapse while packet drop is the primary signal of congestion at bottlenecks. Since packet drop (and increased delay) impacts all their customers negatively, network operators would like to be able to distinguish between overly aggressive congestion control and a confluence of many low-bandwidth, low-impact flows. But they are unable to see the actual congestion signal and thus, they have to implement bandwidth and/or usage limits based on the only information they can see or measure (the contents of the packet headers and the rate of the traffic). Such measures don't solve the packet-drop problems effectively and are leading to calls for government regulation (which also won't solve the problem).
We propose congestion exposure as a possible solution. This allows packets to carry an accurate prediction of the congestion they expect to cause downstream thus allowing it to be visible to ISPs and network operators. This memo sets out the motivations for congestion exposure and introduces a strawman protocol designed to achieve congestion exposure.

Status of This Memo

This Internet-Draft will expire on September 2, 2010.

Copyright Notice

Table of Contents

---

1. Introduction

TOC

The Internet has grown from humble origins to become a global phenomenon with billions of end-users able to share the network and exchange data and more. One of the key elements in this success has been the use of distributed algorithms such as TCP that share capacity while avoiding congestion collapse. These algorithms rely on the end-systems altruistically reducing their transmission rate in response to any congestion they see.

In recent years ISPs have seen a minority of users taking a larger share of the network by using applications that transfer data continuously for hours or even days at a time and even opening multiple simultaneous TCP connections. This issue became prevalent with the advent of "always on" broadband connections. Frequently peer to peer protocols have been held responsible [RFC5594] (Peterson, J. and A. Cooper, “Report from the IETF Workshop on Peer-to-Peer (P2P) Infrastructure, May 28, 2008,” July 2009.) but streaming video traffic is becoming increasingly significant. In order to improve the network experience for the majority of their customers, many ISPs have chosen to impose controls on how their network's capacity is shared rather than continually buying more capacity. They calculate that most customers will be unwilling to contribute to the cost of extra shared capacity if that will only really benefit a minority of users. Approaches include volume counting or charging, and application rate limiting. Typically these traffic controls, whilst not impacting most customers, set a restriction on a customer's level of network usage, as defined in a "fair usage policy".

We believe that such traffic controls 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. In fact some of the current work at the IETF [LEDBAT] (Shalunov, S., “Low Extra Delay Background Transport (LEDBAT),” October 2009.) and IRTF [CC-open-research] (Welzl, M., Scharf, M., Briscoe, B., and D. Papadimitriou, “Open Research Issues in Internet Congestion Control,” September 2009.) already reflects this thinking. For example, an application intending to transfer large amounts of data could use LEDBAT to try to reduce its transmission rate before any competing TCP flows do, by detecting an increase in end-to-end delay (as a measure of incipient congestion). However these techniques rely on voluntary, altruistic action by end users and their application providers. ISPs cannot enforce their use. This leads to our second point.

The Internet was designed so that end-hosts detect and control congestion. We believe 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 traffic causes 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] (Ramakrishnan, K., Floyd, S., and D. Black, “The Addition of Explicit Congestion Notification (ECN) to IP,” September 2001.) to detect how much congestion traffic has suffered between the source and a monitoring point in the network, but not beyond. Such a capability would enable an ISP to give incentives for the use, without restrictions, of LEDBAT-like applications whilst perhaps restricting excessive use of 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. Lacking the ability to see congestion, some ISPs count the volume each user transfers. On this basis LEDBAT applications would get blamed for hogging the network given the large amount of volume they transfer. However, because they yield rather than hog, they actually contribute very little to congestion. One use of exposed congestion information would be to measure the congestion attributable to a given user, and thereby incentivise the use of protocols such as [LEDBAT] (Shalunov, S., “Low Extra Delay Background Transport (LEDBAT),” October 2009.) which aim to reduce the congestion caused by bulk data transfers.

Creating the incentive to deploy low-congestion protocols such as LEDBAT is just one of many motivations for congestion exposure. 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. It can

measure the impact of an individual consumer, a large enterprise network or the traffic crossing a border from another ISP – anywhere where volume is used today as a (poor) measure of usage. In Section 7 (Use Cases), a range of potential use cases for congestion exposure are given, showing it is possible to imagine a wide range of other ways to use the exposed congestion information.

1.1. Definitions

TOC

Throughout this document we refer to congestion repeatedly. Congestion has a wide range of definitions. For the purposes of this document it is defined using the simplest way that it can be measured - the instantaneous fraction of loss. More precisely, congestion is bits lost divided by bits sent, taken over any brief period. By extension, if explicit congestion notification (ECN) is being used, the fraction of bits marked (rather than lost) gives a useful metric that can be thought of as analagous to congestion. Strictly congestion should measure impairment, whereas ECN aims to avoid any loss or delay impairments due to congestion. But for the purposes of this document, the two will both be called congestion.
We also need to define two specific terms carefully:

> **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 that point and the source of the packet.
>
> **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.

1.2. Changes from previous versions

TOC

> **From -03 to -04 (current version):** Many edits throughout per comments from Bob Briscoe about the intentions of ConEx.
>
> References section updated; reference to Comcast congestion management system added as ISP example.
>
> NOTE: there are still sections needing more work, especially the Use Cases. The whole document also needs trimming in places and checking for repetition or omission.

From -02 to -03:

> Abstract re-written again following comments from John Leslie.
>
> Use Cases Section re-written.
>
> Security Considerations section improved.
>
> This ChangeLog added.

From -01 to -02: Extensive changes throughout the document:

> *Abstract and Introduction re-written.
>
> *The Problem section re-written and extended significantly.
>
> *Why Now? Section re-written and extended.
>
> *Requirements extended.
>
> *Security Considerations expanded.
>
> Other less major changes throughout.

From -00 -01: Significant changes throughout including re-organising the main structure.

> New Abstract and changes to Introduction.

---

2. The Problem

TOC

2.1. Congestion is not the problem

TOC

The problem is not congestion itself. The problem is how best to share available capacity. When too much traffic meets too little capacity, congestion occurs. Then we have to share out what capacity there is. But we should not (and cannot) solve the capacity sharing problem by trying to make it go away – by saying there should somehow be no congestion, slower traffic or more capacity. That misses the whole point of the Internet: to multiplex or share available capacity at maximum bit-rate.

So as we say, the problem is not congestion in itself. Every elastic data transfer should (and usually will) congest a healthy data network. If it doesn't, its transport protocol is broken. There should always be periods approaching 100% utilisation at some link along every data path through the Internet, implying that frequent periods of congestion are a healthy sign. If transport protocols are too weak to congest capacity, they are under-utilising it and hanging around longer than they need to, reducing the capacity available for the next data transfers that might be about to start.

2.2. Increase capacity or manage traffic?

TOC

Some say the problem is that ISPs should invest in more capacity. Certainly increasing capacity should make the congested periods during data transfers shorter and the non-congested gaps between them longer. The argument goes that if capacity were large enough it would make the periods when there is a capacity sharing problem insignificant and not worth solving.

Yet, ISPs are facing a quandary - traffic is growing rapidly and traffic patterns are changing significantly (see Section 4 (Why Now?) and [Cisco-VNI] (Cisco Systems, inc., “Cisco Visual Networking Index: Forecast and Methodology, 2008–2013,” June 2009.)) They know that any increases in capacity will have to be paid for by all their customers but capacity growth will be of most benefit to the heaviest users. Faced with these problems, some ISPs are seeking to reduce what they regard as "heavy usage" in order to improve the service experienced by the majority of their customers.

If done properly, managing traffic should be a valid alternative to increasing capacity. An ISP's customers can vote with their feet if the ISP chooses the wrong balance between managing heavy traffic and charging for too much shared capacity. Current traffic management techniques (Section 3 (Existing Approaches to Traffic Control)) fight against the capacity shares that TCP is aiming for. Ironically, they try to impose something approaching LEDBAT-like behaviour on heavier flows. But as we have seen, they cannot give LEDBAT the credit for doing this itself – the network just sees a LEDBAT flow as a large amount of volume.

Thus the problem for the IETF is to ensure that ISPs and their equipment suppliers have appropriate protocol support – not just to impose good capacity sharing themselves, but to encourage end-to-end protocols to share out capacity in everyone's best interests.

TOC

2.2.1. Making Congestion Visible

Unfortunately ISPs are only able to see limited information about the traffic they forward. As we will see in section 3 they are forced to use the only information they do have available which leads to myopic control that has scant regard for the actual impact of the traffic or the underlying network conditions. All their approaches are unsound because they cannot measure the most useful metric. The volume or rate of a given flow or aggregate doesn't directly affect other users, but the congestion it causes does. This can be seen with a simple illustration. A 5Mbps flow in an otherwise empty 10Mbps bottleneck causes no congestion and so affects no other users. By contrast a 1Mbps flow entering a 10Mbps bottleneck that is already fully occupied causes significant congestion and impacts every other user sharing that bottleneck as well as suffering impairment itself. So the real problem that needs to be addressed is how to close this information gap. How can we expose congestion at the IP layer so that it can be used as the basis for measuring the impact of any traffic on the network as a whole?

2.2.2. ECN - a Step in the Right Directions

TOC

Explicit Congestion Notification [RFC3168] (Ramakrishnan, K., Floyd, S., and D. Black, “The Addition of Explicit Congestion Notification (ECN) to IP,” September 2001.) 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] (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,” April 1998.) 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.
So Is ECN the Solution? Alas not - ECN does allow downstream nodes to measure the upstream congestion for any flow, but this is not enough. This can make a receiver accountable for the congestion caused by incoming traffic. But a receiver can only control incoming congestion indirectly, by politely asking the sender to control it. A receiver

cannot make a sender install an adaptive codec, or install LEDBAT instead of TCP. And a receiver cannot ask 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 - by design.

---

3. Existing Approaches to Traffic Control

TOC

Existing approaches intended to address the problems outlined above can be broadly divided into two groups - those that passively monitor traffic and can thus measure the apparent impact of a given flow of packets and those that can actively discriminate against certain packets, flows, applications or users based on various characteristics or metrics.

3.1. Layer 3 Measurement

TOC

L3 measurement of traffic relies on using the information that can be measured directly or is revealed in the IP header of the packet (or lower layers). Architecturally, L3 measurement is best since it fits with the idea of the hourglass design of the Internet [RFC3439] (Bush, R. and D. Meyer, “Some Internet Architectural Guidelines and Philosophy,” December 2002.). This asserts that "the complexity of the Internet belongs at the edges, and the IP layer of the Internet should remain as simple as possible."

3.1.1. Volume Accounting

TOC

Volume accounting is a technique that is often used to discriminate between heavy and light users. The volume of traffic sent by a given user or network is one of the easiest pieces of information to monitor in a network. Measuring the size of every packet from the header and adding them up is a simple operation. Consequently this has long been a favoured measure used by operators to control their customers.
The precise manner in which this volume information is used may vary. Typically ISPs may impose an overall volume cap on their customers (perhaps 10Gbytes a month). Alternatively they may decide that the heaviest users each month are subjected to some sanction.
Volume is naively thought to indicate the impact that one party's traffic has on others. But the same volume can cause very different impacts on others if it is transferred at slightly different times, or

between slightly different endpoints. Also the impact on others greatly depends on how responsive the transport is to congestion, whether responsive (TCP), very responsive (LEDBAT), aggressive (multiple TCPs) or totally unresponsive.

3.1.2. Rate Measurement

TOC

Rate measurements might be thought indicative of the impact of one aggregate of traffic on others, and rate is often limited to avoid impact on others. However such limits generally constrain everyone much more than they need to, just in case most parties send fast at the same time. And such limits constrain everyone too little at other times, when everyone actually does send fast at the same time.
The problem with measuring rate is that it doesn't say how much the rate is occupying shared capacity over time, and whether the high rate of one user comes at times when others want a high rate.

3.2. Higher Layer Discrimination

TOC

Over recent years a number of traffic management techniques have emerged that explicitly differentiate between different traffic types, applications and even users. This is done because ISPs and operators feel they have a need to use such techniques to better control a new raft of applications that break some of the implicit design assumptions behind TCP (short-lived flows, limited flows per connection, generally between server and client).

3.2.1. Bottleneck Rate Policing

TOC

Bottleneck flow rate policers such as [XCHOKe] (Chhabra, P., Chuig, S., Goel, A., John, A., Kumar, A., Saran, H., and R. Shorey, “XCHOKe: Malicious Source Control for Congestion Avoidance at Internet Gateways,” November 2002.) and [pBox] (Floyd, S. and K. Fall, “Promoting the Use of End-to-End Congestion Control in the Internet,” August 1999.) have been proposed as approaches for rate policing traffic. But they must be deployed at bottlenecks in order to work. Unfortunately, capacity sharing is not only about congestion-responsive behaviour of each flow, but also about how long the flows occupy the capacity and the combined total of multiple flows. Such rate policers also make an assumption about what constitutes acceptable per-flow behaviour. If these bottleneck policers were widely deployed, the

Internet could find itself with one universal rate adaptation policy embedded throughout the network. With TCP's congestion control algorithm approaching its scalability limits as the network bandwidth continues to increase, new algorithms are being developed for high-speed congestion control. Embedding assumptions about acceptable rate adaptation would make evolution to such new algorithms extremely painful.

3.2.2. DPI and Application Rate Policing

TOC

Some operators use deep packet inspection (DPI) and traffic analysis to identify certain applications they believe to have an excessive impact on the network. ISPs generally pick on applications that that they judge as low value to the customer in question and high impact on other customers. A common example is peer-to-peer file-sharing. Having identified a flow as belonging to such an application, the operator uses differential scheduling to limit the impact of that flow on others, which usually limits its throughput as well. This has fuelled the on-going battle between application developers and DPI vendors. When operators first started to limit the throughput of P2P, it soon became common knowledge that turning on encryption could boost your throughput. The DPI vendors then improved their equipment so that it could identify P2P traffic by the pattern of packets it sends. This risks becoming an endless vicious cycle - an arms race that neither side can win. Furthermore such techniques may put the operator in direct conflict with the customers, regulators and content providers.

---

4. Why Now?

TOC

The accountability and capacity sharing problems highlighted so far have always characterised the Internet to some extent. In 1988 Van Jacobson coded capacity sharing into TCP's e2e congestion control algorithms [TCPcc] (Jacobson, V. and M. Karels, “Congestion Avoidance and Control,” August 1988.). But fair queuing algorithms were already being written for network operators to ensure each active user received an equal share of a link and couldn't game the system [RFC0970] (Nagle, J., “On packet switches with infinite storage,” December 1985.). The two approaches have divergent objectives, but they have co-existed ever since.

The main new factor has been the introduction of residential broadband, making 'always-on' available to all, not just campuses and enterprises. Both TCP and approaches like fair queuing don't take account of how much of each user's data is occupying a link over time, which can significantly reduce the capacity available to lighter usage. Therefore

residential ISPs have been introducing new traffic management equipment that can prioritise based on each customer's usage volume, e.g. [Comcast] (Bastian, C., Klieber, T., Livingood, J., Mills, J., and R. Woundy, “Comcast's Protocol-Agnostic Congestion Management System,” February 2010.). Otherwise capacity upgrades get eaten up by transfers of large amounts of data, with little gain for interactive usage [BB-Incentive] (MIT Communications Futures Program (CFP) and Cambridge University Communications Research Network, “The Broadband Incentive Problem,” September 2005.).

In campus networks, capacity upgrades are the easiest way to mitigate the inability of TCP or FQ to take account of activity over time. But capacity upgrades are much more expensive in residential broadband networks that are spread over large geographic areas and customers will only be happy to pay more for their service if the majority can see a significant benefit.

However, these traffic management techniques fight the capacity shares e2e protocols are aiming at, rather than working together in unison. And, the more optimal ISPs try to make their controls, the more they need application knowledge within the network – which isn't how the Internet was designed to work. Congestion exposure hasn't been considered before, because the depth of the problem has only recently been understood. We now understand that both networks and end-systems to focus on contribution to congestion, not volume or rate. Then application knowledge is only needed on the end-system, where it should be. But the reason this isn't happening is because the network cannot see the information it needs (congestion).

As long as ISPs continue to use rate and volume as the key metrics for determining when to control traffic there is no incentive to use LEDBAT or other low-congestion protocols to improve the performance of competing interactive traffic. We believe that congestion exposure gives ISPs the information they need to be able to discriminate in favour of such low-congestion transports. In turn this will give users a direct benefit from using such transports and so encourage their wider use.

---

5. Requirements for a Solution

TOC

This section proposes some requirements for any solution to this problem. We believe that a solution that meets most of these requirements is likely to be better than one that doesn't, but we recognise that if a working group is established in this area, it may have to make tradeoffs.

*Allow both upstream and downstream congestion to be visible at the IP layer -- visibility at the IP layer allows congestion in the heart of the network to be monitored at the edges and without

deploying complicated and intrusive equipment such as DPI boxes. This gives several advantages:

1. It enables bulk policing of traffic based on the congestion it is actually going to cause in the network.

2. It allows the amount of congestion across ISP borders to be monitored.

3. It supports a diversity of intra-domain and inter-domain congestion management practices.

4. It allows the contribution to congestion over time to be counted as easily as volume can be counted today.

5. It supports contractual arrangements for managing traffic (acceptable use policies, SLAs etc) between just the two parties exchanging traffic across their point of attachment, without involving others.

*Avoid making assumptions about the behavior of specific applications (e.g. be agnostic to application and transport behaviour).

*Support the widest possible range of transport protocols for the widest range of data types (elastic, inelastic, real-time, background, etc) -- don't force a "universal rate adaptable policy" such as TCP-friendliness [RFC3448] (Handley, M., Floyd, S., Padhye, J., and J. Widmer, “TCP Friendly Rate Control (TFRC): Protocol Specification,” January 2003.).

*Be responsive to real-time congestion in the network.

*Allow incremental deployment of the solution and ideally design for permanent partial deployment to increase chances of successful deployment.

*Ensure packets supporting congestion exposure are distinguishable from others, so that each transport can control when it chooses to deploy congestion exposure, and ISPs can manage the two types of traffic distinctly.

*Support mechanisms that ensure the integrity of congestion notifications, thus making it hard for a user or network to distort the congestion signal.

*Be robust in the face of DoS attacks, so that congestion information can be used to identify and limit DoS traffic and to protect the hosts and network elements implementing congestion exposure.

Many of these requirements are by no means unique to the problem of congestion exposure. Incremental deployment for instance is a critical requirement for any new protocol that affects something as fundamental as IP. Being robust under attack is also a pre-requisite for any protocol to succeed in the real Internet and this is covered in more detail in Section 9 (Security Considerations).

---

6. A Strawman Congestion Exposure Protocol

TOC

In this section we explore a simple strawman protocol that would solve the congestion exposure problem. This protocol neatly illustrates how a solution might work. A practical implementation of this protocol has been produced and both simulations and real-life testing show that it works. The protocol is based on a concept known as re-feedback [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,” August 2005.) and builds on existing active queue management techniques like RED [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,” April 1998.) and ECN [RFC3168] (Ramakrishnan, K., Floyd, S., and D. Black, “The Addition of Explicit Congestion Notification (ECN) to IP,” September 2001.) that network elements can already use to measure and expose congestion.

Re-feedback, standing for re-inserted feedback, is a system designed to allow end-hosts to reveal to the network information about their network path that they have received via conventional feedback (for instance congestion).

In our strawman protocol we imagine that packets have two "congestion" fields in their IP header:

- The first is 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.

- The other is a whole-path congestion field that uses re-feedback to record the total congestion 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), the whole path congestion (with a time lag of 1RTT) 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. Use Cases

TOC

Once downstream congestion information is revealed in the IP header it can be used for a number of purposes. Precise details of how the information might be used are beyond the scope of this document but this section will give an overview of some possible uses. {ToDo: write up the rest of this section properly. Concentrate on a couple of the most useful potential use cases (traffic management and accountability?) and mention a couple of more arcane uses (traffic engineering and e2e QoS). The key thing is to clarify that Congestion Exposure is a tool that can be used for many other things...}
It allows an ISP to accurately identify which traffic is having the greatest impact on the network and either police directly on that basis or use it to determine which users should be policed. It can form the basis of inter-domain contracts between operators. It could even be used as the basis for inter-domain routing, thus encouraging operators to invest appropriately in improving their infrastructure.
From Rich Woundy: "I would add a section about use cases. The primary use case would seem to be an "incentive environment that ensures optimal sharing of capacity", although that could use a better title. Other use cases may include "DDoS mitigation", "end-to-end QoS", "traffic engineering", and "inter-provider service monitoring". (You can see I am stealing liberally from the motivation draft here. We'll have to see whether the other use cases are "core" to this group, or "freebies" that come along with re-ECN as a particular protocol.)"
My take on this is we need to concentrate on one or two major use cases. The most obvious one is using this to control user-behaviour and encourage the use of "congestion friendly" protocols such as LEDBAT. {Comments from Louise Krug:} simply say that operators MUST turn off any kind of rate limitation for LEDBAT traffic and what they might mean for the amount of bandwidth they see compared to a throttled customer? You could then extend that to say how it leads to better QoS differentiation under the assumption that there is a broad traffic mix any way? Not sure how much detail you want to go into here though? {ToDo: better incorporate this text from Mirja into Michael's text below.} Congestion exposure can enable ISPs to give an incentive to end-systems to response to congestion in a way that leads to a better

share of the available capacity. For example the introduction of a per-user congestion volume might motivate "heavy-user" to back off with their high-bandwidth traffic (when congestion occurs) to save their congestion volume for more time-critical traffic. If every end-system reacts to congestion in such a way that it avoids congestion for non-critical traffic and allow a certain level of congestion for the more important traffic (from the user's point of view), the all-over user experience will be increased. More-over the network might be utilized more equally when less-important traffic is shifted to less congested time slots.

7.1. Improved Policing

TOC

As described earlier in this document, ISPs throttle traffic not because it causes congestion in the network but because users have exceeded their traffic profile or because individual applications or flows are suspected to cause congestion. This is done because it is not possible to police only the traffic that is causing congestion. Congestion exposure allows new possibilities for rate policing.

7.1.1. Per Aggregate Policing

TOC

A straightforward application of congestion exposure is per-flow or per-aggregate congestion policing. Instead of limiting flows or aggregates because they have exceeded certain rate thresholds, they can be throttled if they cause too much congestion in the network. This is throttling on evidence instead of suspicion.

7.1.2. Per customer policing

TOC

The assumption is that every customer has an allowance of congestion per second. If he causes more congestion than this throughout the network, his traffic can be policed or shaped to ensure he stays within his allowance. The nice features of this approach are that it sets incentives for the use of congestion-minimising transport protocols such as LEDBAT and allows tariffs that better reflect the relative impact of customers each other.

> **Incentives for congestion minimising transports:** A user generates foreground and background traffic. Foreground traffic needs to go fast while background traffic can afford to go slow. With per-

customer congestion policing, users can optimise their network experience by using congestion-minimising transport protocols for background traffic and normal TCP-like or even high-speed transport protocols for foreground traffic. Doing so means background traffic only causes minimal congestion so that foreground traffic can go faster than when both were transmitted over the same transport protocols. Hence, per-customer congestion policing sets incentives for selfish users to utilise congestion-minimising transport protocols.

Improved tariff structures: Currently customers are offered tariffs with all manner of differentaitors from peak access rate to volume limit and even specific application rate limits. Congestion-policing offers a better means of distinguishing between tariffs. Heavy users and light users will get equal access in terms of speed and short-term throughput, but customers that cause more congestion and thus have a bigger impact on others will have to pay for the privilege or suffer reduced throuhgput during periods of heavy congestion. However tariffs are a subject best left to the market to determine, not the IETF.

---

8. IANA Considerations

TOC

This document makes no request to IANA.

---

9. Security Considerations

TOC

One intended use of exposed congestion information is to hold the e2e transport and the network accountable to each other. Therefore, any congestion exposure protocol will have to provide the necessary hooks to mechanisms that can assure the integrity of this information. 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 all 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 indeed simply wants to reduce the amonut of traffic it is carrying. Congestion Exposure itself shouldn't significantly

alter the incentives for and against honest declaration of congestion by a network, but it is possible to imagine applications of Congestion Exposure that will change these incentives. There is a general perception among networks that their level of congestion is a business secret. Actually in the Internet architecture congestion is one of the worst-kept secrets a network has, because end-hosts can see congestion better than networks can. Nonetheless, one goal of a congestion exposure protocol is to allow networks to pinpoint whether congestion is in one side or the other of a border. Although this extra transparency should be good for ISPs with low congestion, those with underprovisioned networks may try to obstruct deployment.

*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] (Savage, S., Wetherall, D., and T. Anderson, "TCP Congestion Control with a Misbehaving Receiver," 1999.) explains how a receiver can significantly improve their throughput my failing to declare congestion. This is a problem with or without Congestion Exposure. [KGao] (Gao, K. and C. Wang, "Incrementally Deployable Prevention to TCP Attack with Misbehaving Receivers," December 2004.) explains one possible technique to encourage receiver's to be honest in their declaration of congestion.

*The Sender. One proposed mechanisms for congestion exposure adds a requirement for a sender to let the network know 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 respond more than previously. 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.

In addition there are potential problems from source spoofing. A malicious sender can pretend to be another user by spoofing the source address. A congestion exposure protocol will need to be robust against injection of false congestion information into the forward path that could distort or disrupt the integrity of the congestion signal.

---

10. Conclusions

TOC

Congestion exposure is the idea that traffic itself indicates to all nodes on its path how much congestion it causes on the entire path. It is useful for network operators to police traffic only if it really causes congestion in the Internet instead of doing blind rate capping

independently of the congestion situation. This change would give incentives to users to adopt new transport protocols such as LEDBAT which try to avoid congestion more than TCP does. Requirements for congestion exposure in the IP header were summarized, one technical solution was presented, and additional use cases for congestion exposure were discussed.

---

11. Acknowledgements

TOC

A number of people other than authors have provided text and comments for this memo. The document is being produced in support of a BoF on Congestion Exposure as discussed extensively on the <re-ecn@ietf.org> mailing list.

---

12. Informative References

TOC

[BB-Incentive]	MIT Communications Futures Program (CFP) and Cambridge University Communications Research Network, "The Broadband Incentive Problem," September 2005.
[CC-open-research]	Welzl, M., Scharf, M., Briscoe, B., and D. Papadimitriou, "Open Research Issues in Internet Congestion Control," draft-irtf-iccrg-welzl-congestion-control-open-research-05 (work in progress), September 2009 (TXT).
[Cisco-VNI]	Cisco Systems, inc., "Cisco Visual Networking Index: Forecast and Methodology, 2008–2013," June 2009.
[Comcast]	Bastian, C., Klieber, T., Livingood, J., Mills, J., and R. Woundy, "Comcast's Protocol-Agnostic Congestion Management System," draft-livingood-woundy-congestion-mgmt-03 (work in progress), February 2010 (TXT).
[KGao]	Gao, K. and C. Wang, "Incrementally Deployable Prevention to TCP Attack with Misbehaving Receivers," December 2004.
[LEDBAT]	Shalunov, S., "Low Extra Delay Background Transport (LEDBAT)," draft-ietf-ledbat-congestion-00 (work in progress), October 2009 (TXT).
[RFC0970]	Nagle, J., "On packet switches with infinite storage," RFC 970, December 1985 (TXT).
[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 (TXT, HTML, XML).
[RFC3168]	Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP," RFC 3168, September 2001 (TXT).
[RFC3439]	Bush, R. and D. Meyer, "Some Internet Architectural Guidelines and Philosophy," RFC 3439, December 2002 (TXT).
[RFC3448]	Handley, M., Floyd, S., Padhye, J., and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification," RFC 3448, January 2003 (TXT).
[RFC5594]	Peterson, J. and A. Cooper, "Report from the IETF Workshop on Peer-to-Peer (P2P) Infrastructure, May 28, 2008," RFC 5594, July 2009 (TXT).
[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 (PDF).
[Savage]	Savage, S., Wetherall, D., and T. Anderson, "TCP Congestion Control with a Misbehaving Receiver," ACM SIGCOMM Computer Communication Review , 1999.
[TCPcc]	Jacobson, V. and M. Karels, "Congestion Avoidance and Control," Proc. ACM SIGCOMM'88 Symposium, Computer Communication Review 18(4)314—329, August 1988 (PS, PDF).
[XCHOKe]	Chhabra, P., Chuig, S., Goel, A., John, A., Kumar, A., Saran, H., and R. Shorey, "XCHOKe: Malicious Source Control for Congestion Avoidance at Internet Gateways," Proceedings of IEEE International Conference on Network Protocols (ICNP-02) , November 2002 (PDF).
[pBox]	Floyd, S. and K. Fall, "Promoting the Use of End-to-End Congestion Control in the Internet," IEEE/ACM Transactions on Networking 7(4) 458—472, August 1999 (PDF).

Authors' Addresses

TOC

	Toby Moncaster
	BT
	B54/70, Adastral Park
	Martlesham Heath
	Ipswich IP5 3RE
	UK
Phone:	+44 7918 901170
EMail:	toby.moncaster@bt.com
	Louise Krug
	BT
	B54/77, Adastral Park
	Martlesham Heath
	Ipswich IP5 3RE
	UK
EMail:	louise.burness@bt.com
	Michael Menth
	University of Wuerzburg
	room B206, Institute of Computer Science
	Am Hubland
	Wuerzburg D-97074
	Germany
Phone:	+49 931 888 6644
EMail:	menth@informatik.uni-wuerzburg.de

	João Taveira Araújo
	UCL
	GS206 Department of Electronic and Electrical Engineering
	Torrington Place
	London WC1E 7JE
	UK
EMail:	j.araujo@ee.ucl.ac.uk
	Steven Blake
	Extreme Networks
	Pamlico Building One, Suite 100
	3306/08 E. NC Hwy 54
	RTP, NC 27709
	US
EMail:	sblake@extremenetworks.com
	Richard Woundy (editor)
	Comcast
	Comcast Cable Communications
	27 Industrial Avenue
	Chelmsford, MA 01824
	US
EMail:	richard_woundy@cable.comcast.com
URI:	http://www.comcast.com