Network Working Group
Internet-Draft
Intended status: Experimental
Expires: April 23, 2018

N. Khademi
M. Welzl
University of Oslo
G. Armitage
Swinburne University of Technology
G. Fairhurst
University of Aberdeen
October 20, 2017

TCP Alternative Backoff with ECN (ABE)
draft-ietf-tcpm-alternativebackoff-ecn-02

Abstract

Recent Active Queue Management (AQM) mechanisms instantiate shallow buffers with burst tolerance to minimise the time that packets spend enqueued at a bottleneck. However, shallow buffering can cause noticeable performance degradation when TCP is used over a network path with a large bandwidth-delay-product. Traditional methods rely on detecting network congestion through reported loss of transport packets. Explicit Congestion Notification (ECN) instead allows a router to directly signal incipient congestion. A sending endpoint can distinguish when congestion is signalled via ECN, rather than by packet loss. An ECN signal indicates that an AQM mechanism has done its job, and therefore the bottleneck network queue is likely to be shallow. This document therefore proposes an update to the TCP sender-side ECN reaction in congestion avoidance to reduce the Congestion Window (cwnd) by a smaller amount than the congestion control algorithm’s reaction to loss. This document also recommends this approach to be adopted by any other transport protocol that implements a congestion control reduction to an ECN congestion signal.

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 https://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on April 23, 2018.

Table of Contents

1. Definitions

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

2. Introduction

Explicit Congestion Notification (ECN) [RFC3168] makes it possible for an Active Queue Management (AQM) mechanism to signal the presence of incipient congestion without incurring packet loss. This lets the

network deliver some packets to an application that would have been dropped if the application or transport did not support ECN. This packet loss reduction is the most obvious benefit of ECN, but it is often relatively modest. There are also significant other benefits from deploying ECN [RFC8087], including reduced end-to-end network latency.

The rules for ECN were originally written to be very conservative, and required the congestion control algorithms of ECN-capable transport protocols to treat ECN congestion signals exactly the same as they would treat a packet loss [RFC3168].

Research has demonstrated the benefits of reducing network delays due to excessive buffering [BUFFERBLOAT]; this has led to the creation of new AQM mechanisms like PIE [RFC8033] and CoDel [CODEL2012] [I-D.CoDel], which avoid causing bloated queues that are common with a simple tail-drop behaviour (also known as a First-In First-Out, FIFO, queue).

These AQM mechanisms instantiate short queues that are designed to tolerate packet bursts. However, congestion control mechanisms cannot always utilise a bottleneck link well where there are short queues. For example, to allow a single TCP connection to fully utilise a network path, the queue at the bottleneck link must be able to compensate for TCP halving the "cwnd" and "ssthresh" variables in response to a lost packet [RFC5681]. This requires the bottleneck queue to be able to store at least an end-to-end bandwidth-delay product (BDP) of data, which effectively doubles both the amount of data that can be in flight and the round-trip time (RTT) experience using the network path.

Modern AQM mechanisms can use ECN to signal the early signs of impending queue buildup long before a tail-drop queue would be forced to resort to dropping packets. It is therefore appropriate for the transport protocol congestion control algorithm to have a more measured response when an early-warning signal of congestion is received in the form of an ECN CE-marked packet. Recognizing these changes in modern AQM practices, more recent rules have relaxed the strict requirement that ECN signals be treated identically to packet loss [I-D.ECN-exp]. Following these newer, more flexible rules, this document defines a new sender-side-only congestion control response, called "ABE" (Alternative Backoff with ECN). ABE improves the performance when routers use shallow buffered AQM mechanisms.

3. Specification

This specification describes an update to the congestion control algorithm of an ECN-capable TCP transport protocol. It allows a TCP stack to update the TCP sender response when it receives feedback indicating reception of a CE-marked packet. It RECOMMENDS that a TCP sender multiplies the cwnd by 0.8 and reduces the slow start threshold (ssthresh) in congestion avoidance following reception of a TCP segment that sets the ECN-Echo flag (defined in [RFC3168]). While this specification concerns TCP, other transports also support a per-RTT response to ECN. The method defined in this document is also applicable for such transports.

4. Discussion

Much of the technical background to this congestion control response can be found in a research paper [ABE2017]. This paper used a mix of experiments, theory and simulations with standard NewReno and CUBIC to evaluate the technique. It examined the impact of enabling ECN and letting individual TCP senders back off by a reduced amount in reaction to the receiver that reports ECN CE-marks from AQM-enabled bottlenecks. The technique was shown to present "...significant performance gains in lightly-multiplexed scenarios, without losing the delay-reduction benefits of deploying CoDel or PIE". The performance improvement is achieved when reacting to ECN-Echo in congestion avoidance by multiplying cwnd and ssthresh with a value in the range [0.7..0.85].

4.1. Why Use ECN to Vary the Degree of Backoff?

The classic rule-of-thumb dictates that a network path needs to provide a BDP of bottleneck buffering if a TCP connection wishes to optimise path utilisation. A single TCP bulk transfer running through such a bottleneck will have increased its congestion window (cwnd) up to 2*BDP by the time that packet loss occurs. When packet loss is detected (regarded as a notification of congestion), Standard TCP halves the cwnd and ssthresh [RFC5681], which causes the TCP congestion control to go back to allowing only a BDP of packets in flight -- just sufficient to maintain 100% utilisation of the bottleneck on the network path.

AQM mechanisms such as CoDel [I-D.CoDel] and PIE [RFC8033] set a delay target in routers and use congestion notifications to constrain the queuing delays experienced by packets, rather than in response to impending or actual bottleneck buffer exhaustion. With current default delay targets, CoDel and PIE both effectively emulate a shallow buffered bottleneck (section II, [ABE2017]) while also allowing short traffic bursts into the queue. This provides

acceptable performance for TCP connections over a path with a low BDP, or in highly multiplexed scenarios (many concurrent transport connections). However, it interacts badly for a lightly-multiplexed case (few concurrent connections) over a path with a large BDP. Conventional TCP backoff in such cases leads to gaps in packet transmission and under-utilisation of the path.

Instead of discarding packets, an AQM mechanism is allowed to mark ECN-capable packets with an ECN CE-mark. The reception of a CE-mark not only indicates congestion on the network path, it also indicates that an AQM mechanism exists at the bottleneck along the path, and hence the CE-mark likely came from a bottleneck with a shallow queue. Reacting differently to an ECN CE-mark than to packet loss can then yield the benefit of a reduced back-off, as with CUBIC [I-D.CUBIC], when queues are short, yet it can avoid generating excessive delay when queues are long. Using ECN can also be advantageous for several other reasons [RFC8087].

The idea of reacting differently to loss and detection of an ECN CE-mark pre-dates this document. For example, previous research proposed using ECN CE-marks to modify TCP congestion control behaviour via a larger multiplicative decrease factor in conjunction with a smaller additive increase factor [ICC2002]. The goal of this former work was to operate across AQM bottlenecks using Random Early Detection (RED) that were not necessarily configured to emulate a shallow queue ([RFC7567] notes the current status of RED as an AQM method.)

4.2. Focus on ECN as Defined in RFC3168

Some transport protocol mechanisms rely on ECN semantics that differ from the original ECN definition [RFC3168] -- for example, Congestion Exposure (ConEx) [RFC7713] and Datacenter TCP (DCTCP) [I-D.ietf-tcpm-dctcp] need more accurate ECN information than that offered by the original feedback method. Other mechanisms (e.g., [I-D.ietf-tcpm-accurate-ecn]) allow the sender to adjust the rate more frequently than once each path RTT. Use of these mechanisms is out of the scope of the current document.

4.3. Discussion: Choice of ABE Multiplier

ABE decouples the reaction of a TCP sender to loss and ECN CE-marks when in the congestion avoidance phase by differentiating the scaling factor used in Equation 4 in Section 3.1 of [RFC5681]. The description respectively uses beta_{loss} and beta_{ecn} to refer to the multiplicative decrease factors applied in response to packet loss, and in response to a receiver indicating that an ECN CE-mark was received on an ECN-enabled TCP connection. For non-ECN-enabled

TCP connections, no ECN CE-marks are received and only beta_{loss} applies.

In other words, in response to detected loss:

ssthresh_(t+1) = max (FlightSize_t * beta_{loss}, 2 * SMSS)

and in response to an indication of a received ECN CE-mark:

ssthresh_(t+1) = max (FlightSize_t * beta_{ecn}, 2 * SMSS)

and

cwnd_(t+1) = ssthresh_(t+1)

where FlightSize is the amount of outstanding data in the network, upper-bounded by the sender's cwnd and the receiver's advertised window (rwnd) [RFC5681]. The higher the values of beta_{loss} and beta_{ecn}, the less aggressive the response of any individual backoff event.

The appropriate choice for beta_{loss} and beta_{ecn} values is a balancing act between path utilisation and draining the bottleneck queue. More aggressive backoff (smaller beta_*) risks underutilising the path, while less aggressive backoff (larger beta_*) can result in slower draining of the bottleneck queue.

The Internet has already been running with at least two different beta_{loss} values for several years: the standard value is 0.5 [RFC5681], and the Linux implementation of CUBIC [I-D.CUBIC] has used a multiplier of 0.7 since kernel version 2.6.25 released in 2008. ABE proposes no change to beta_{loss} used by current TCP implementations.

beta_{ecn} depends on how the response of a TCP connection to shallow AQM marking thresholds is optimised. beta_{loss} reflects the preferred response of each congestion control algorithm when faced with exhaustion of buffers (of unknown depth) signalled by packet loss. Consequently, for any given TCP congestion control algorithm the choice of beta_{ecn} is likely to be algorithm-specific, rather than a constant multiple of the algorithm's existing beta_{loss}.

A range of tests (section IV, [ABE2017]) with NewReno and CUBIC over CoDel and PIE in lightly-multiplexed scenarios have explored this choice of parameter. The results of these tests indicate that CUBIC connections benefit from beta_{ecn} of 0.85 (cf. beta_{loss} = 0.7), and NewReno connections see improvements with beta_{ecn} in the range 0.7 to 0.85 (cf. beta_{loss} = 0.5).

5. Status of the Update

This update is a sender-side only change. Like other changes to congestion-control algorithms, it does not require any change to the TCP receiver or to network devices. It does not require any ABE-specific changes in routers or the use of Accurate ECN feedback [I-D.ietf-tcpm-accurate-ecn] by a receiver.

The currently published ECN specification requires that the congestion control response to a CE-marked packet is the same as the response to a dropped packet [RFC3168]. The specification is currently being updated to allow for specifications that do not follow this rule [I-D.ECN-exp]. The present specification defines such an experiment and has thus been assigned an Experimental status before being proposed as a Standards-Track update.

The purpose of the Internet experiment is to collect experience with deployment of ABE, and confirm the safety in deployed networks using this update to TCP congestion control.

When used with bottlenecks that do not support ECN-marking the specification does not modify the transport protocol.

To evaluate the benefit, this experiment therefore requires support in AQM routers (except to enable an ECN-marking mechanism [RFC3168] [RFC7567]) for ECN-marking of packets carrying the ECN Capable Transport, ECT(0), codepoint [RFC3168].

If the method is only deployed by some senders, and not by others, the senders that use this method can gain some advantage, possibly at the expense of other flows that do not use this updated method. Because this advantage applies only to ECN-marked packets and not to loss indications, the new method cannot lead to congestion collapse.

The result of this Internet experiment will be reported by presentation to the TCPM WG (or IESG) or an implementation report at the end of the experiment.

6. Acknowledgements

Authors N. Khademi, M. Welzl and G. Fairhurst were part-funded by the European Community under its Seventh Framework Programme through the Reducing Internet Transport Latency (RITE) project (ICT-317700). The views expressed are solely those of the authors.

The authors would like to thank Stuart Cheshire for many suggestions when revising the draft, and the following people for their contributions to [ABE2017]: Chamil Kulatunga, David Ros, Stein

Gjessing, Sebastian Zander. Thanks also to (in alphabetical order) Bob Briscoe, Markku Kojo, John Leslie, Dave Taht and the TCPM working group for providing valuable feedback on this document.

The authors would finally like to thank everyone who provided feedback on the congestion control behaviour specified in this update received from the IRTF Internet Congestion Control Research Group (ICCRG).

7. IANA Considerations

XX RFC ED - PLEASE REMOVE THIS SECTION XXX

This document includes no request to IANA.

8. Implementation Status

ABE is implemented as a patch for Linux and FreeBSD. It is meant for research and available for download from http://heim.ifi.uio.no/naeemk/research/ABE/ This code was used to produce the test results that are reported in [ABE2017]. An evolved version of the patch for FreeBSD is currently under review for potential inclusion in the mainline kernel [ABE-FreeBSD].

9. Security Considerations

The described method is a sender-side only transport change, and does not change the protocol messages exchanged. The security considerations for ECN [RFC3168] therefore still apply.

This is a change to TCP congestion control with ECN that will typically lead to a change in the capacity achieved when flows share a network bottleneck. This could result in some flows receiving more than their fair share of capacity. Similar unfairness in the way that capacity is shared is also exhibited by other congestion control mechanisms that have been in use in the Internet for many years (e.g., CUBIC [I-D.CUBIC]). Unfairness may also be a result of other factors, including the round trip time experienced by a flow. ABE applies only when ECN-marked packets are received, not when packets are lost, hence use of ABE cannot lead to congestion collapse.

10. Revision Information

XX RFC ED - PLEASE REMOVE THIS SECTION XXX

-02. Corrected the equations in Section 4.3. Updated the affiliations. Lower bound for cwnd is defined. A recommendation for window-based transport protocols is changed to cover all transport

protocols that implements a congestion control reduction to an ECN congestion signal. Added text about ABE's FreeBSD mainline kernel status including a reference to the FreeBSD code review page. References are updated.

-01. Text improved, mainly incorporating comments from Stuart Cheshire. The reference to a technical report has been updated to a published version of the tests [ABE2017]. Used "AQM Mechanism" throughout in place of other alternatives, and more consistent use of technical language and clarification on the intended purpose of the experiments required by EXP status. There was no change to the technical content.

-00. draft-ietf-tcpm-alternativebackoff-ecn-00 replaces draft-khademi-tcpm-alternativebackoff-ecn-01. Text describing the nature of the experiment was added.

Individual draft -01. This I-D now refers to draft-black-tsvwg-ecn-experimentation-02, which replaces draft-khademi-tsvwg-ecn-response-00 to make a broader update to RFC3168 for the sake of allowing experiments. As a result, some of the motivating and discussing text that was moved from draft-khademi-alternativebackoff-ecn-03 to draft-khademi-tsvwg-ecn-response-00 has now been re-inserted here.

Individual draft -00. draft-khademi-tsvwg-ecn-response-00 and draft-khademi-tcpm-alternativebackoff-ecn-00 replace draft-khademi-alternativebackoff-ecn-03, following discussion in the TSVWG and TCPM working groups.

11. References

11.1. Normative References

[I-D.ECN-exp] Black, D., "Explicit Congestion Notification (ECN) Experimentation", Internet-draft, IETF work-in-progress draft-ietf-tsvwg-ecn-experimentation-06, September 2017.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/info/rfc3168>.

[RFC5681] Allman, M., Paxson, V., and E. Blanton, "TCP Congestion Control", RFC 5681, DOI 10.17487/RFC5681, September 2009, <https://www.rfc-editor.org/info/rfc5681>.

[RFC7567] Baker, F., Ed. and G. Fairhurst, Ed., "IETF Recommendations Regarding Active Queue Management", BCP 197, RFC 7567, DOI 10.17487/RFC7567, July 2015, <https://www.rfc-editor.org/info/rfc7567>.

11.2. Informative References

[ABE-FreeBSD] "ABE patch review in FreeBSD", <https://reviews.freebsd.org/D11616>.

[ABE2017] Khademi, N., Armitage, G., Welzl, M., Fairhurst, G., Zander, S., and D. Ros, "Alternative Backoff: Achieving Low Latency and High Throughput with ECN and AQM", IFIP NETWORKING 2017, Stockholm, Sweden, June 2017.

[BUFFERBLOAT] "Bufferbloat project", <https://www.bufferbloat.net/projects/bloat/wiki/Introduction/>.

[CODEL2012] Nichols, K. and V. Jacobson, "Controlling Queue Delay", July 2012, <http://queue.acm.org/detail.cfm?id=2209336>.

[I-D.CoDel] Nichols, K., Jacobson, V., McGregor, V., and J. Iyengar, "Controlled Delay Active Queue Management", Internet-draft, IETF work-in-progress draft-ietf-aqm-codel-09, September 2017.

[I-D.CUBIC] Rhee, I., Xu, L., Ha, S., Zimmermann, A., Eggert, L., and R. Scheffenegger, "CUBIC for Fast Long-Distance Networks", Internet-draft, IETF work-in-progress draft-ietf-tcpm-cubic-06, September 2017.

[I-D.ietf-tcpm-accurate-ecn] Briscoe, B., Kuehlewind, M., and R. Scheffenegger, "More Accurate ECN Feedback in TCP", draft-ietf-tcpm-accurate-ecn-03 (work in progress), May 2017.

[I-D.ietf-tcpm-dctcp]
Bensley, S., Thaler, D., Balasubramanian, P., Eggert, L., and G. Judd, "Datacenter TCP (DCTCP): TCP Congestion Control for Datacenters", draft-ietf-tcpm-dctcp-10 (work in progress), August 2017.

[ICC2002] Kwon, M. and S. Fahmy, "TCP Increase/Decrease Behavior with Explicit Congestion Notification (ECN)", IEEE ICC 2002, New York, New York, USA, May 2002, <http://dx.doi.org/10.1109/ICC.2002.997262>.

[RFC7713] Mathis, M. and B. Briscoe, "Congestion Exposure (ConEx) Concepts, Abstract Mechanism, and Requirements", RFC 7713, DOI 10.17487/RFC7713, December 2015, <https://www.rfc-editor.org/info/rfc7713>.

[RFC8033] Pan, R., Natarajan, P., Baker, F., and G. White, "Proportional Integral Controller Enhanced (PIE): A Lightweight Control Scheme to Address the Bufferbloat Problem", RFC 8033, DOI 10.17487/RFC8033, February 2017, <https://www.rfc-editor.org/info/rfc8033>.

[RFC8087] Fairhurst, G. and M. Welzl, "The Benefits of Using Explicit Congestion Notification (ECN)", RFC 8087, DOI 10.17487/RFC8087, March 2017, <https://www.rfc-editor.org/info/rfc8087>.

Authors' Addresses

Naeem Khademi
University of Oslo
PO Box 1080 Blindern
Oslo N-0316
Norway

Email: naeemk@ifi.uio.no

Michael Welzl
University of Oslo
PO Box 1080 Blindern
Oslo N-0316
Norway

Email: michawe@ifi.uio.no

Grenville Armitage
Internet For Things (I4T) Research Group
Swinburne University of Technology
PO Box 218
John Street, Hawthorn
Victoria 3122
Australia

Email: garmitage@swin.edu.au

Godred Fairhurst
University of Aberdeen
School of Engineering, Fraser Noble Building
Aberdeen AB24 3UE
UK

Email: gorry@erg.abdn.ac.uk