

Internet Engineering Task Force (IETF)
Request for Comments: 8511
Category: Experimental
ISSN: 2070-1721

N. Khademi
M. Welzl
University of Oslo
G. Armitage
Netflix
G. Fairhurst
University of Aberdeen
December 2018

TCP Alternative Backoff with ECN (ABE)

Аннотация

Механизмы активного управления очередями (Active Queue Management или AQM) обеспечивают устойчивость к всплескам (пикам) трафика при одновременном сохранении коротких очередей для минимизации времени пребывания пакетов в очередях узких мест сети. Проходящие через узкие места соединения TCP могут терять значительную часть производительности, особенно при наличии небольшого числа потоков или большом значении произведения пропускной способности и задержки (bandwidth-delay product или BDP). Получение маркера перегрузки (Congestion Experienced или CE) явного уведомления о перегрузке (Explicit Congestion Notification или ECN) указывает, что в узком месте применяется механизм AQM и сетевая очередь в этом месте вероятно будет короткой. Эти сигналы позволяют передающей стороне соединения TCP реагировать на ECN и для предотвращения перегрузки сокращать окно перегрузки (Congestion Window или cwnd) на меньшую величину, чем требует механизм контроля перегрузок для предполагаемой потери пакета. Данная спецификация задаёт экспериментальное изменение реакции TCP, заданной в RFC 3168, как это разрешает RFC 8311.

Статус документа

Документ не относится к категории Internet Standards Track и публикуется для проверки, экспериментальной реализации и оценки.

Документ является результатом работы IETF¹ и представляет согласованный взгляд сообщества IETF. Документ прошёл открытое обсуждение и был одобрен для публикации IESG². Не все документы, одобренные IESG, претендуют на статус стандартов Internet, см. раздел 2 в RFC 7841.

Информацию о текущем статусе документа, ошибках и способах обратной связи можно найти по ссылке <https://www.rfc-editor.org/info/rfc8511>.

Авторские права

Авторские права (Copyright (c) 2018) принадлежат IETF Trust и лицам, указанным в качестве авторов документа. Все права защищены.

К документу применимы права и ограничения, указанные в BCP 78 и IETF Trust Legal Provisions и относящиеся к документам IETF (<http://trustee.ietf.org/license-info>), на момент публикации данного документа. Прочтите упомянутые документы внимательно. Фрагменты программного кода, включённые в этот документ, распространяются в соответствии с упрощённой лицензией BSD, как указано в параграфе 4.e документа IETF Trust Legal Provisions, без каких-либо гарантий (как указано в Simplified BSD License).

Оглавление

1. Введение.....	1
2. Уровни требований.....	2
3. Спецификация.....	2
3.1. Выбор коэффициента ABE Multiplier.....	2
4. Обсуждение.....	3
4.1. Обоснования применения ECN для изменения уровня отката.....	3
4.2. Основанный на RTT отклик на указанную перегрузку.....	3
5. Требования к внедрению ABE.....	3
6. Цели экспериментов с ABE.....	4
7. Взаимодействие с IANA.....	4
8. Вопросы безопасности.....	4
9. Литература.....	4
9.1. Нормативные документы.....	4
9.2. Дополнительная литература.....	4
Благодарности.....	5
Адреса авторов.....	5

1. Введение

Явные уведомления о перегрузке (ECN) [RFC3168] позволяют механизму (AQM) информировать о начинающейся перегрузке без необходимости потери пакетов. Это позволяет сети доставлять приложению некоторые пакеты, которые были бы отброшены, если бы приложение или транспорт не поддерживали ECN. Это сокращение потерь пакетов

¹Internet Engineering Task Force - комиссия по решению инженерных задач Internet.

²Internet Engineering Steering Group - комиссия по инженерным разработкам Internet.

является наиболее очевидным преимуществом ECN, но зачастую оно сравнительно мало. Другие преимущества от внедрения ECN документированы в [RFC8087].

Правила для ECN исходно были заданы как очень консервативные и требовали, чтобы алгоритмы контроля перегрузки транспорта с поддержкой ECN (ECN-Capable Transport или ECT) считали индикацию перегрузки через ECN эквивалентом потери пакета [RFC3168]. Исследования показали сокращение сетевых задержек, вызванных контролем перегрузки в TCP на основе потерь и избыточной буферизацией [BUFFERBLOAT]. Это привело к созданию механизмов AQM, таких как улучшенный пропорциональный интегральный контроллер (Proportional Integral Controller Enhanced или PIE) [RFC8033] и контроль задержки в очередях (Controlling Queue Delay или CoDel) [RFC8289], которые предотвращают раздувание очередей, характерное для неуправляемых и слишком больших буферов, развёрнутых в сети Internet [BUFFERBLOAT].

Отмеченные выше механизмы AQM нацелены на поддержку неизменно короткой очереди с допущением кратковременных всплесков пакетов. Однако применяемые в настоящее время механизмы контроля перегрузок на основе потерь не всегда обеспечивают эффективное использование узкого канала (bottleneck) при коротких очередях. Например, отправителю TCP с контролем перегрузок Reno требуется возможность сохранять данные по меньшей мере в объёме BDP в буфере узкого места для обеспечения полной загрузки пути в условиях вызванного потерями сокращения окна перегрузки (cwnd) [RFC5681]. Такой объём буферизации фактически удваивает размер находящихся в сети (in flight) данных и максимальное время кругового обхода (round-trip time или RTT) для отправителя TCP.

Современные механизмы AQM могут использовать ECN для сигнализации о первых признаках надвигающегося заполнения очереди задолго до вынужденного отбрасывания пакетов. Поэтому для алгоритма контроля перегрузок в транспортном протоколе разумно применять более взвешенный отклик на ранние предупреждения о перегрузке после получения удалённой точки пакета с маркером ECN CE. С учётом этих изменений в современной практике AQM строгое требование обработки сигналов ECN CE как предполагаемой потери пакетов было смягчено в [RFC8311]. Этот документ задаёт новый механизм отклика на сигналы контроля перегрузки передающей стороной, названный ABE (Alternative Backoff with ECN). ABE повышает среднюю пропускную способность TCP при использовании в маршрутизаторах буферов с AQM, которые разрешают лишь короткие очереди.

2. Уровни требований

Ключевые слова **необходимо** (MUST), **недопустимо** (MUST NOT), **требуется** (REQUIRED), **нужно** (SHALL), **не следует** (SHALL NOT), **следует** (SHOULD), **не нужно** (SHOULD NOT), **рекомендуется** (RECOMMENDED), **не рекомендуется** (NOT RECOMMENDED), **возможно** (MAY), **необязательно** (OPTIONAL) в данном документе интерпретируются в соответствии с BCP 14 [RFC2119] [RFC8174] тогда и только тогда, когда они выделены шрифтом, как показано здесь.

3. Спецификация

Эта спецификация меняет алгоритм контроля перегрузки поддерживающего ECN транспортного протокола TCP путём смены отклика отправителя TCP обратной связью от получателя TCP, указывающей приём им пакета с маркером CE, т. е. с установленным флагом ECN-Echo (задан в [RFC3168]) в соответствии с процессом, заданным в [RFC8311].

Отклик отправителя TCP в настоящее время задан в параграфе 6.1.2 спецификации ECN [RFC3168] и слегка обновлён в параграфе 4.1 [RFC8311].

Индикацию насыщения следует трактовать просто, как потерю в результате насыщения для TCP без поддержки ECN. Т. е. отправитель TCP снижает вдвое размер окна насыщения cwnd и уменьшает порог замедленного старта ssthresh, если не указано иное в Experimental RFC потока документов IETF.

В соответствии с RFC 8311 этот документ задаёт изменение для передающей стороны TCP, где при получении пакета с флагом ECN-Echo **следует** инициировать у источника TCP установку для порога замедленного старта (slow start threshold или ssthresh) значения $0,8 * \text{FlightSize}$ с нижней границей $2 * \text{SMSS}$ ¹, применяемой к результату. Как и в [RFC5681], отправитель TCP также снижает значение cwnd не более чем до нового значения ssthresh. В параграфе 6.1.2 RFC 3168 даны рекомендации по установке cwnd меньше $2 * \text{SMSS}$.

3.1. Выбор коэффициента ABE Multiplier

ABE отвязывает реакцию отправителя TCP на предполагаемую потерю пакетов от индикации перегрузки сигналом ECN в фазе предотвращения перегрузки. Для этого ABE использует иной коэффициент масштабирования в уравнении 4 из параграфа 3.1 в [RFC5681]. В описании соответственно используются β_{loss} и β_{ecn} для указания факторов мультипликативного снижения, применяемого в ответ на предполагаемую потерю пакета и в ответ на указание получателем перегрузки с помощью ECN. Для соединений TCP без поддержки ECN применяется только β_{loss} . Иными словами, отклик на предполагаемую потерю пакетов имеет вид

$$\text{ssthresh} = \max(\text{FlightSize} * \beta_{\text{loss}}, 2 * \text{SMSS})$$

а отклик на индикацию перегрузки сигналом ECN -

$$\text{ssthresh} = \max(\text{FlightSize} * \beta_{\text{ecn}}, 2 * \text{SMSS})$$

и

$$\text{cwnd} = \text{ssthresh}$$

(для ssthresh == $2 * \text{SMSS}$ в параграфе 6.1.2 RFC 3168 даны рекомендации по установке cwnd меньше $2 * \text{SMSS}$)

FlightSize - размер остающихся в сети данных, ограниченный сверху меньшим из значений cwnd отправителя и анонсированным получателем окном (rwnd) [RFC5681]. Большие значения β_{loss} и β_{ecn} означают менее энергичный отклик на любое отдельное событие отката (backoff).

Подходящий выбор значений β_{loss} и β_{ecn} обеспечивает баланс между загрузкой пути и опустошением очереди в узком месте. Более энергичное поведение (меньшие β_{loss} и β_{ecn}) ведёт к риску недогрузки пути, тогда как менее энергичный откат (большие β_{loss} и β_{ecn}) может приводить к более медленному опустошению очереди в узком месте.

¹Sender Maximum Segment Size - максимальный размер сегмента у отправителя.

В Internet уже несколько лет применяется по меньшей мере два разных значения β_{loss} - стандартное 0,5 [RFC5681] и принятое в реализации Linux CUBIC [RFC8312] значение 0,7, начиная с ядра 2.6.25, выпущенного в 2008 г. ABE не меняет значений β_{loss} , используемых текущими реализациями TCP.

Этот документ рекомендует значение $\beta_{ecn}=0,8$, применяемое только для стандартного контроля перегрузок TCP [RFC5681]. Выбор β_{ecn} позволяет настраивать отклик соединения TCP для снижения порогов маркировки AQM. Значение β_{loss} характеризует отклик алгоритма контроля перегрузки на потерю пакетов, т. е. заполнение буфера (неизвестного размера). Для алгоритмов контроля перегрузок в TCP предлагаются различные значения β_{loss} . Поэтому β_{ecn} , скорей всего, является параметром, связанным с алгоритмом, а не постоянной, кратной значению β_{loss} имеющегося алгоритма. Выбор параметра был изучен в ряде тестов (раздел IV в [ABE2017]) с NewReno и CUBIC через CoDel и PIE в вариантах с незначительным мультиплексированием. Результаты этих тестов показывают, что для соединений CUBIC подходит значение $\beta_{ecn} = 0,85$ (сравн. с $\beta_{loss} = 0,7$), а соединения NewReno показывают улучшения при β_{ecn} от 0,7 до 0,85 (сравн. с $\beta_{loss} = 0,5$).

4. Обсуждение

Большая часть технического обоснования ABE приведена в [ABE2017], где сочетаются эксперименты, теория и моделирование с NewReno [RFC5681] и CUBIC [RFC8312] для оценки производительности. Показано, что ABE обеспечивает значительный рост производительности в сценариях с незначительным мультиплексированием нескольких одновременных потоков без потери преимуществ сокращения задержки в результате внедрения CoDel или PIE. Повышения производительности достигается при реагировании на ECN-Echo для предотвращения перегрузки (когда $ssthresh > cwnd$) путём умножения $cwnd$ и $ssthresh$ на значение из диапазона [0,7, 0,85]. Применение ABE при $cwnd$ не более $ssthresh$ в настоящее время не рекомендуется, но его использование в этом сценарии может выиграть в результате дополнительного внимания, экспериментов и спецификации.

4.1. Обоснования применения ECN для изменения уровня отката

Механизмы AQM, такие как CoDel [RFC8289] и PIE [RFC8033], устанавливают целевую задержку в маршрутизаторах и применяют уведомления о перегрузке для ограничения задержки в очередях, с которыми сталкиваются пакеты, а не в ответ на надвигающееся или фактическое исчерпание буфера в узком месте. С текущими принятыми по умолчанию значениями целевой задержки CoDel и PIE эффективно эмулируют узкое место с короткой очередью (раздел II в [ABE2017]), одновременно допуская в очереди короткие всплески (пики) трафика. Это обеспечивает приемлемую производительность соединений TCP на путях с низким BDP или в сценариях с сильным мультиплексированием (много одновременных потоков). Однако при незначительном мультиплексировании на пути с большим BDP обычный откат (backoff) TCP ведёт к пропускам передачи пакетов и недогрузке пути.

Вместо отбрасывания пакетов механизму AQM разрешено пометить пакеты ECN-Capable маркером ECN CE. Получение отклика с маркером CE не только указывает перегрузку на сетевом пути, но и говорит о наличии в узком месте на пути механизма AQM. Таким образом, появление маркера CE скорей всего объясняется наличием узкого места с контролируемой короткой очередью. Реагирование на сигналы ECN о перегрузке, отличающееся от реакции на предполагаемую потерю пакетов, может обеспечить сокращение отката при коротких очередях. Использование ECN может обеспечивать и другие преимущества [RFC8087]. Идея разного реагирования на предполагаемую потерю пакетов и сигналы о перегрузке от ECN была высказана до этой спецификации, например, в предыдущем исследовании предлагалось использовать отклики ECN CE для смены поведения контроля перегрузок TCP путём увеличения коэффициента сокращения (окна) с сочетанием с уменьшением коэффициента расширения [ICC2002]. Целью этой работы была организация AQM через узкие места (с использованием RED¹) без обязательной настройки для эмуляции короткой очереди (текущее применение RED в качестве Internet AQM ограничено [RFC7567]).

4.2. Основанный на RTT отклик на указанную перегрузку

Эта спецификация применяется к использованию откликов ECN, как задано в [RFC3168], где определён отклик на индикацию перегрузки не чаще 1 раза за интервал кругового обхода. Поскольку ABE реагирует на указанную перегрузку 1 раз за RTT, механизм не реагирует на дополнительные потери в том же интервале RTT, так как отправитель ABE уже сократил окно перегрузки. Если перегрузка продолжается после такого сокращения, ABE продолжает уменьшать окно перегрузки в каждом следующем интервале RTT. Такое последовательное сокращение окна может защитить сеть от продолжительного отсутствия беспристрастности, когда алгоритмы AQM не поддерживают малое значение среднего размера очереди. Механизм не полагается на Accurate ECN [ACC-ECN-FEEDBACK].

Механизмы транспортного протокола, напротив, могут быть созданы для использования более частых и детальных откликов ECN (например, Accurate ECN [ACC-ECN-FEEDBACK]), что позволяет более часто корректировать скорость передачи. Примером такого подхода является Data Center TCP (DCTCP) [RFC8257].

5. Требования к внедрению ABE

Это обновление относится лишь к передающей стороне. Подобно другим изменениям в алгоритмах контроля перегрузки, оно не требует менять что-либо на приёмной стороне TCP или в сетевых устройствах. Обновление не требует вносить связанные с ABE изменения в маршрутизаторы или использовать отклики Accurate ECN [ACC-ECN-FEEDBACK] на приёмной стороне.

Если метод внедряется лишь у части отправителей, те могут получить некоторые преимущества, возможно, за счёт других потоков, не внедривших этот обновлённый метод. Поскольку преимущества относятся лишь к пакетам с маркировкой ECN, а не индикаторам потери пакетов, узкое место с поддержкой ECN все равно будет отбрасывать пакеты, если отправитель TCP, применяющий ABE, будет слишком энергичным. Результат не будет отличаться от использования отправителем TCP лишь традиционного контроля перегрузок на основе потери пакетов.

Эта спецификация не меняет транспортный протокол в узких местах без поддержки ECN.

¹Random Early Detection - случайное раннее обнаружение.

6. Цели экспериментов с ABE

В [RFC3168] сказано, что отклик контроля перегрузки, следующий за индикацией ECN, совпадает с откликом на отбрасывание пакета. [RFC8311] обновляет эту спецификацию, разрешая системам поддерживать разное поведение в случаях индикации ECN и потери пакетов. Данная спецификация определяет такой эксперимент и имеет статус Experimental RFC. Авторы ожидают, что в будущем статус сменится на Standards-Track.

Целью эксперимента в Internet является получение опыта внедрения ABE и подтверждение приемлемой безопасности в сетях, использующих такой обновленный контроль перегрузок в TCP. Для оценки ABE в этом эксперименте нужна поддержка в маршрутизаторах AQM маркировки ECN для пакетов, содержащих код поддерживающего ECN транспорта (ECN-Capable Transport) - ECT(0) [RFC3168].

Результат эксперимента должен включать исследование влияния установки маркера ECN-CE и последующей потери в том же интервале RTT. В конце эксперимента результат будет сообщён рабочей группе TCPM или IESG.

Механизм ABE реализован как исправление (patch) для Linux и FreeBSD, предназначенное для исследований и экспериментов и доступное по ссылке <<https://heim.ifi.uio.no/michawe/research/abe/>>. Код будет применяться для получения результатов теста, представляемых в [ABE2017]. Код FreeBSD был внесён (commit) в основную ветвь ядра (mainline kernel) 9 марта 2018 г. [ABE-REVISION].

7. Взаимодействие с IANA

Документ не требует действий IANA.

8. Вопросы безопасности

Описанный метод вносит изменение в транспорт лишь на стороне отправителя и не меняет обмен протокольными сообщениями, поэтому соображения безопасности для ECN [RFC3168] остаются применимыми.

Это изменение контроля перегрузок в TCP с помощью ECN обычно будет менять доступную производительность для потоков, проходящих через узкое место в сети (bottleneck). Это может приводить к получению некоторыми потоками большей доли пропускной способности. Похожая несправедливость распределения пропускной способности наблюдается и в других механизмах контроля перегрузок, применяемых в Internet долгие годы (например, CUBIC [RFC8312]). Несправедливость может возникать и в результате других факторов, включая интервал кругового обхода для потока. ABE применяется только при получении пакетов с маркировкой ECN, а не в результате потери пакетов. Поэтому использование ABE не может вызывать перегрузочный коллапс (congestion collapse).

9. Литература

9.1. Нормативные документы

- [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>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8257] Bensley, S., Thaler, D., Balasubramanian, P., Eggert, L., and G. Judd, "Data Center TCP (DCTCP): TCP Congestion Control for Data Centers", RFC 8257, DOI 10.17487/RFC8257, October 2017, <<https://www.rfc-editor.org/info/rfc8257>>.
- [RFC8311] Black, D., "Relaxing Restrictions on Explicit Congestion Notification (ECN) Experimentation", [RFC 8311](#), DOI 10.17487/RFC8311, January 2018, <<https://www.rfc-editor.org/info/rfc8311>>.

9.2. Дополнительная литература

- [ABE-REVISION] Stewart, L., "ABE patch review in FreeBSD", Revision 331214, March 2018, <<https://svnweb.freebsd.org/base?view=revision&revision=331214>>.
- [ABE2017] Khademi, N., Armitage, G., Welzl, M., Zander, S., Fairhurst, G., and D. Ros, "Alternative backoff: Achieving low latency and high throughput with ECN and AQM", IFIP Networking Conference and Workshops Stockholm, Sweden, DOI 10.23919/IFIPNetworking.2017.8264863, June 2017.
- [ACC-ECN-FEEDBACK] Briscoe, B., Kuehlewind, M., and R. Scheffenegger, "More Accurate ECN Feedback in TCP", Work in Progress, draft-ietf-tcpm-accurate-ecn-07, July 2018.
- [BUFFERBLOAT] Gettys, J. and K. Nichols, "Bufferbloat: Dark Buffers in the Internet", ACM Queue, Volume 9, Issue 11, DOI 10.1145/2063166.2071893, November 2011, <<https://queue.acm.org/detail.cfm?id=2071893>>.
- [ICC2002] Kwon, M. and S. Fahmy, "TCP increase/decrease behavior with explicit congestion notification (ECN)", 2002 IEEE International Conference on Communications Conference Proceedings, ICC 2002, Cat. No.02CH37333, DOI 10.1109/ICC.2002.997262, May 2002, <<http://dx.doi.org/10.1109/ICC.2002.997262>>.

- [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 ping Explicit Congestion Notification (ECN)", RFC 8087, DOI 10.17487/RFC8087, March 2017, <<https://www.rfc-editor.org/info/rfc8087>>.
- [RFC8289] Nichols, K., Jacobson, V., McGregor, A., Ed., and J. Iyengar, Ed., "Controlled Delay Active Queue Management", RFC 8289, DOI 10.17487/RFC8289, January 2018, <<https://www.rfc-editor.org/info/rfc8289>>.
- [RFC8312] Rhee, I., Xu, L., Ha, S., Zimmermann, A., Eggert, L., and R. Scheffenegger, "CUBIC for Fast Long-Distance Networks", RFC 8312, DOI 10.17487/RFC8312, February 2018, <<https://www.rfc-editor.org/info/rfc8312>>.

Благодарности

Работа N. Khademi, M. Welzl и G. Fairhurst частично финансировалась European Community в рамках Seventh Framework Programme по проекту Reducing Internet Transport Latency (RITE) (ICT-317700). Представленные здесь мнения принадлежат исключительно авторам.

G. Armitage выполнил большую часть работы над этим документом, будучи сотрудником Swinburne University of Technology, Melbourne, Australia.

Авторы благодарны Stuart Cheshire за многочисленные предложения при работе над документом. Спасибо за вклад в [ABE2017] Chamil Kulatunga, David Ros, Stein Gjessing, Sebastian Zander. Спасибо за отклики на документ David Black, Roland Bless, Bob Briscoe, Markku Kojo, John Leslie, Lawrence Stewart (в алфавитном порядке) и рабочей группе TCPM.

Авторы признательны всем членам исследовательской группы IRTF Internet Congestion Control (ICCRG), кто предоставил свои отклики на описанный в этом документе контроль перегрузок.

Адреса авторов

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
Netflix Inc.
Email: garmitage@netflix.com

Godred Fairhurst
University of Aberdeen
School of Engineering, Fraser Noble Building
Aberdeen AB24 3UE
United Kingdom
Email: gorry@erg.abdn.ac.uk

Перевод на русский язык

Николай Малых
nmalykh@protokols.ru