Network Working Group
Request for Comments: 3042
Category: Standards Track

M. Allman
NASA GRC/BBN
H. Balakrishnan
MIT
S. Floyd
ACIRI

January 2001

Эффективное восстановление TCP после потерь с использованием ограниченной передачи

Enhancing TCP's Loss Recovery Using Limited Transmit

Статус документа

Документ содержит спецификацию стандартного протокола для сообщества Internet и является приглашением к дискуссии в целях развития и совершенствования протокола. Сведения о стандартизации и состоянии данного протокола можно найти в документе «Internet Official Protocol Standards» (STD 1). Допускается свободное распространение данного документа.

Авторские права

Тезисы

Этот документ предлагает новый механизм TCP[1], который может использоваться для более эффективного восстановления при потере сегментов, когда окно насыщения для соединения достаточно мало или теряется множество сегментов из одного окна передачи. Алгоритм «ограниченной передачи[2]» вызывается для отправки нового сегмента данных в ответ на первый из каждой пары дубликатов подтверждений, полученной отправителем. Передача этих сегментов повышает вероятность того, что TCP сможет выполнить восстановление после потери одного сегмента с помощью алгоритма ускоренного повтора передачи без дорогостоящего ожидания тайм-аута для повтора. Ограниченная передача может использоваться вместе с механизмом селективных подтверждений TCP SACK[3] или независимо от него.

1. Введение

Многие исследователи отмечали, что стратегии восстановления TCP при потере пакетов работают недостаточно хорошо, когда окно насыщения на стороне отправителя TCP мало. Это может происходить, например, по причине передачи ограниченного объема данных, в результате ограничений, вносимых анонсируемым получателем окном, или в силу ограничений, обусловленных механизмом сквозного контроля насыщения при работе через канал с малым значением произведения полоса*задержка [Riz96,Mor97,BPS+98,Bal98,LK98]. Когда TCP детектирует отсутствие сегмента, протокол переходит в фазу восстановления с использованием одного из двух методов. В первом методе если подтверждение (ACK) для данного сегмента не получено в течение некого интервала времени, возникает тайм-аут повтора передачи и сегмент отправляется заново [RFC793,PA00]. Второй метод (алгоритм ускоренного повтора[4]) повторяет передачу сегмента при получении отправителем трех дубликатов ACK [Jac88,RFC2581]. Однако в результате того, что передача дубликатов получателем может быть вызвана также нарушением порядка следования пакетов в Internet, получатель TCP дожидается получения трех дубликатов ACK, чтобы попытаться отличить потерю сегмента от нарушения порядка доставки. В фазе восстановления может использоваться множество методов для повторной передачи потерянных сегментов, включая восстановление на основе процедуры замедленного старта или ускоренного восстановления Fast Recovery [RFC2581], NewReno [RFC2582] или восстановления на базе селективных подтверждений (SACK) [RFC2018,FF96].

Тайм-аут повторной передачи TCP (RTO[5]) определяется на основе измерения времени кругового обхода (RTT[6]) между отправителем и получателем в соответствии со спецификацией [PA00]. Для предотвращения ненужных повторов передачи сегментов, которые были задержаны, но не потеряны, в качестве минимального значения RTO выбрана 1 секунда. Следовательно, это позволяет отправителю TCP детектировать и восстанавливать состояние при потере множества сегментов без продолжительного ожидания в состоянии бездействия. Однако при недостаточном количестве сегментов ACK, принятых от получателя, алгоритм Fast Retransmit не включается - такая ситуация возникает при малом размере окна насыщения или потере большого числа сегментов в одном окне. В качестве примера рассмотрим окно насыщения (cwnd) размером 3 сегмента. Если один сегмент будет отброшен в сети, отправитель будет получать не более двух дубликатов ACK. Поскольку для включения механизма ускоренного повтора требуется три дубликата ACK, повтор передачи отброшенного в сети пакета будет осуществляться по тайм-ауту.

---

[1]Transmission Control Protocol - протокол управления передачей.
[2]Limited Transmit.
[3]Selective acknowledgment.
[4]Fast Retransmit.
[5]Retransmission timeout.
[6]Round-trip time.

В работе [BPS+97] показано, что около 56% повторов передачи от загруженных web-серверов происходит по тайм-ауту RTO и только 44% обрабатывается механизмом Fast Retransmit. Кроме того, лишь 4% повторов по тайм-ауту RTO можно избежать при использовании SACK, который тоже не позволяет четко различать потери и нарушение порядка доставки. Использование метода, описанного в этом документе и работе [Bal98], позволяет избежать 25% повторов по тайм-ауту RTO.

Далее в этом документе описаны незначительные изменения на передающей стороне TCP, которые будут ослаблять зависимость от таймера повтора и, следовательно, повышать производительность TCP для тех случаев, когда механизм ускоренного повтора не включается. Предложенные изменения не оказывают негативного влияния на производительность TCP и на взаимодействие с другими соединениями.

1.1. Уровни требований

Ключевые слова "MUST" (**необходимо**), "MUST NOT" (**недопустимо**), "REQUIRED" (**требуется**), "SHALL" (**следует**), "SHALL NOT" (**не следует**), "SHOULD" (**следует**), "SHOULD NOT" (**не следует**), "RECOMMENDED" (**рекомендуется**), "MAY" (**возможно**) и "OPTIONAL" (**необязательно**) в данном документе трактуются в соответствии с RFC 2119 [1] и указывают уровни требований для протоколов.

2. Алгоритм ограниченной передачи

Когда у отправителя TCP имеются не передававшиеся ранее данные, помещенные в очередь на передачу, отправителю **следует** использовать алгоритм ограниченной передачи, который вызывается для отправки новых данных по прибытии двух первых последовательных дубликатов ACK при выполнении перечисленных ниже условий:

- анонсируемое получателем окно позволяет передать сегмент;
- объем остающихся в сети данных не будет превышать размер окна насыщения + 2 сегмента (иными словами, отправитель может передать не более 2 сегментов сверх размера окна насыщения cwnd).

Окно насыщения (cwnd) **недопустимо** менять при передаче этих новых сегментов. В предположении, что эти новые сегменты и соответствующие сегменты ACK не отбрасываются в сети, эта процедура позволяет отправителю осознать потерю сегментов, используя стандартный порог механизма Fast Retransmit в три дубликата ACK [RFC2581]. Это обеспечивает более высокую устойчивость к нарушению порядка доставки, нежели при повторе передачи старого пакета по прибытии первого или второго дубликата ACK.

Примечание. Если в соединении используется алгоритм селективных подтверждений [RFC2018], отправителю **недопустимо** передавать новые сегменты в ответ на дубликат ACK, не содержащий новой информации SACK, поскольку некорректно работающий получатель может генерировать такие сегменты ACK для инициирования неуместной передачи сегмента. В работе [SCWA99] обсуждаются атаки с использованием некорректно работающих получателей.

Механизм ограниченной передачи позволяет контролировать насыщение по принципу «консервации пакетов» [Jac88]. Каждый первый из пары дубликатов ACK показывает, что сегмент покинул сеть. Более того, отправитель еще не принял решения о том, что сегмент был отброшен и, следовательно, не имеет причин для предположения о некорректности текущего состояния контроля насыщения. Следовательно, передача сегментов не будет отклонением от духа принципов контроля насыщения TCP.

В работе [BPS99] показано, что нарушение порядка доставки в сети не является редким событием. В соответствии с [RFC2581] первые два дубликата ACK, полученные отправителем, не вызывают передачи повтора. Это приводит к всплеску передачи сегментов при получении новых подтверждений вслед за нарушением порядка доставки. При использовании механизма ограниченной передачи пакеты данных будут передаваться в сеть по прибытии сегментов ACK и, следовательно, пиков возникать не будет.

Примечание. Механизм Limited Transmit реализован в имитаторе сети [NS]. Желающие исследовать этот механизм могут воспользоваться для этого имитатором, включив опцию singledup_ для нужного соединения TCP.

3. Связанные работы

Развертывание механизмов явного уведомления о перегрузке (ECN[1]) [Flo94,RFC2481] может принести пользу для соединений с малым размером окна насыщения [SA00]. ECN обеспечивает метод индикации перегрузки конечным хостам без отбрасывания сегментов. Хотя некоторые сегменты все же могут теряться, ECN может повысить эффективность работы TCP при малом размере окна насыщения, поскольку отправитель может избежать многократного использования механизмов ускоренного повтора и повтора по тайм-ауту, которые потребовались бы для детектирования отброшенных сегментов [SA00].

Когда трафик с поддержкой ECN конкурирует с трафиком TCP без ECN, для трафика ECN можно получить рост пропускной способности до 30% по сравнению с трафиком без ECN. Для передачи больших объемов данных относительный рост производительности в результате использования ECN увеличивается если в среднем каждый поток имеет 3-4 остающихся в сети пакета для каждого периода кругового обхода [ZQ00]. Это может служить хорошей оценкой влияния на производительность потока механизма ограниченной передачи, поскольку ECN и Limited Transmit снижают влияние тайм-аута повтора на сигнализацию перегрузки.

Алгоритм контроля насыщения Rate-Halving[2] [MSML99] использует одну из форм ограниченной передачи, отправляя сегмент данных на каждый второй дубликат ACK, полученный отправителем. Алгоритм отделяет решение вопросов «что» и «когда» передавать. Однако, подобно Limited Transmit, этот алгоритм всегда будет передавать новый сегмент данных в ответ на получение отправителей второго дубликата ACK.

[1] Explicit Congestion Notification.
[2] «Ополовинивание» скорости передачи.

4. Вопросы безопасности

Дополнительное влияние на безопасность в результате использования описанных в документе приложений на текущий уровень безопасности TCP является минимальным. Потенциальной проблемой является возможность нарушения работы сквозного контроля насыщения с помощью «ложных» дубликатов ACK, которые на самом деле не подтверждают поступление данных на приемную сторону TCP. Ложные дубликаты ACK могут возникать в результате дублирования подтверждений в сети или некорректного поведения получателей TCP, которые передают ложные сегменты ACK для обхода системы сквозного контроля насыщения [SCWA99,RFC2581].

Если получатель TCP согласен использовать опцию SACK, это обеспечивает отправителю TCP достаточно хорошую защиту от ложных дубликатов ACK. К частности, при использовании SACK дубликат ACK, подтверждающий прибытие новых данных на приемную сторону, сообщает порядковые номера этих новых данных. Таким образом, при использовании SACK отправитель TCP может убедиться, что подтверждаемые дубликатом ACK данные были действительно переданы отправителем и для них еще не было получено подтверждение, прежде, чем передавать новые данные в ответ на полученное подтверждение. Для дополнительной защиты отправитель TCP может сохранять информацию о границах пакетов и признавать корректным не более одного подтверждения для каждого пакета (т. е., первоно подтверждения, которое говорит о приеме всех порядковых номеров из данного пакета).

Можно представить некую ограниченную защиту от ложных дубликатов ACK для соединений TCP без поддержки SACK, когда отправитель TCP сохраняет информацию о переданных пакетах и воспринимает не более одного подтверждения на пакет в качестве триггера предачи новых данных. Однако учет переданных и подтвержденных пакетов требует поддержки дополнительной информации о состоянии соединения и усложняет реализацию TCP на передающей стороне, что представляется неприемлемым.

Наиболее важной защитой от ложных дубликатов ACK является ограниченная возможность нарушать с помощью таких дубликатов сквозной контроль насыщения. Следует различать две ситуации - когда полученное отправителем TCP число дубликатов ACK меньше порогового значения и когда этот порог достигнут. Во втором случае TCP с поддержкой ограниченной передачи будет вести себу, по сути, так же, как TCP без поддержки Limited Transmit в том смысле, что будет снижаться вдвое размер окна насыщения и запускаться процедура восстановления после потерь.

Когда число полученных отправителем TCP дубликатов ACK меньше порогового значения, некорректно работающий получатель может передавать два дубликата ACK после каждого нормального подтверждения. Можно предположить, что отправитель TCP будет передавать втрое быстрее дозволенной скорости. Однако при использовании Limited Transmit отправителю разрешается передавать сверх окна насыщения не более порогового числа сегментов (3), как сказано в параграфе 2, следовательно, в ответ на каждый дубликат новый пакет передаваться не будет.

Благодарности

Bill Fenner, Jamshid Mahdavi и рабочая группа Transport Area предоставили множество полезных откликов на ранние версии этого документа.

Литература

[Bal98] Hari Balakrishnan. Challenges to Reliable Data Transport over Heterogeneous Wireless Networks. Ph.D. Thesis, University of California at Berkeley, August 1998.

[BPS+97] Hari Balakrishnan, Venkata Padmanabhan, Srinivasan Seshan, Mark Stemm, and Randy Katz. TCP Behavior of a Busy Web Server: Analysis and Improvements. Technical Report UCB/CSD-97-966, August 1997. Available from http://nms.lcs.mit.edu/~hari/papers/csd-97-966.ps. (Опубликовано также в Proc. IEEE INFOCOM Conf., San Francisco, CA, March 1998.)

[BPS99] Jon Bennett, Craig Partridge, Nicholas Shectman. Packet Reordering is Not Pathological Network Behavior. IEEE/ACM Transactions on Networking, December 1999.

[FF96] Kevin Fall, Sally Floyd. Simulation-based Comparisons of Tahoe, Reno, and SACK TCP. ACM Computer Communication Review, July 1996.

[Flo94] Sally Floyd. TCP and Explicit Congestion Notification. ACM Computer Communication Review, October 1994.

[Jac88] Van Jacobson. Congestion Avoidance and Control. ACM SIGCOMM 1988.

[LK98] Dong Lin, H.T. Kung. TCP Fast Recovery Strategies: Analysis and Improvements. Proceedings of InfoCom, March 1998.

[MSML99] Matt Mathis, Jeff Semke, Jamshid Mahdavi, Kevin Lahey. The Rate Halving Algorithm, 1999. URL: http://www.psc.edu/networking/rate_halving.html.

[Mor97] Robert Morris. TCP Behavior with Many Flows. Proceedings of the Fifth IEEE International Conference on Network Protocols. October 1997.

[NS] Имитатор сети Ns. URL: http://www.isi.edu/nsnam/.

[PA00] Paxson, V. and M. Allman, "Computing TCP's Retransmission Timer", RFC 2988, November 2000.

[Riz96] Luigi Rizzo. Issues in the Implementation of Selective Acknowledgments for TCP. January, 1996. URL: http://www.iet.unipi.it/~luigi/selack.ps

[SA00] Hadi Salim, J. and U. Ahmed, "Performance Evaluation of Explicit Congestion Notification (ECN) in IP Networks", RFC 2884, July 2000.

[SCWA99] Stefan Savage, Neal Cardwell, David Wetherall, Tom Anderson. TCP Congestion Control with a Misbehaving Receiver. ACM Computer Communications Review, October 1999.

[RFC793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793[1], September 1981.

[1]Перевод этого документа имеется на сайте www.protocols.ru. *Прим. перев.*

[RFC2018] Mathis, M., Mahdavi, J., Floyd, S. and A. Romanow, "TCP Selective Acknowledgement Options", RFC 2018[1], October 1996.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119[1], March 1997.

[RFC2481] Ramakrishnan, K. and S. Floyd, "A Proposal to Add Explicit Congestion Notification (ECN) to IP", RFC 2481[2], January 1999.

[RFC2581] Allman, M., Paxson, V. and W. Stevens, "TCP Congestion Control", RFC 2581[1], April 1999.

[RFC2582] Floyd, S. and T. Henderson, "The NewReno Modification to TCP's Fast Recovery Algorithm", RFC 2582, April 1999.

[ZQ00] Yin Zhang and Lili Qiu, Understanding the End-to-End Performance Impact of RED in a Heterogeneous Environment, Cornell CS Technical Report 2000-1802, July 2000. URL http://www.cs.cornell.edu/yzhang/papers.htm.

Адреса авторов

Mark Allman

NASA Glenn Research Center/BBN Technologies

Lewis Field

21000 Brookpark Rd. MS 54-5

Cleveland, OH 44135

Phone: +1-216-433-6586

Fax: +1-216-433-8705

EMail: mallman@grc.nasa.gov

http://roland.grc.nasa.gov/~mallman

Hari Balakrishnan

Laboratory for Computer Science

545 Technology Square

Massachusetts Institute of Technology

Cambridge, MA 02139

EMail: hari@lcs.mit.edu

http://nms.lcs.mit.edu/~hari/

Sally Floyd

AT&T Center for Internet Research at ICSI (ACIRI)

1947 Center St, Suite 600

Berkeley, CA 94704

Phone: +1-510-666-2989

EMail: floyd@aciri.org

http://www.aciri.org/floyd/

Перевод на русский язык

Николай Малых

nmalykh@protocols.ru

Полное заявление авторских прав

[1]Перевод этого документа имеется на сайте www.protocols.ru. *Прим. перев.*
[2]Этот документ заменен RFC 3168. Переводы обоих документов имеются на сайте www.protocols.ru. *Прим. перев.*

Подтверждение

Финансирование функций RFC Editor обеспечено Internet Society.