Network Working Group
Request for Comments: 2991
Category: Informational

D. Thaler
Microsoft
C. Hopps
NextHop Technologies

November 2000

Проблема множества путей при выборе Next-Hop для индивидуального и группового трафика

Multipath Issues in Unicast and Multicast Next-Hop Selection

Статус документа

Этот документ содержит информацию для сообщества Internet и не задаёт каких-либо стандартов Internet. Документ можно распространять без ограничений.

Авторские права

Аннотация

Различные протоколы маршрутизации, такие как OSPF[1] и ISIS[2], явно позволяют использовать маршрутизацию через множество равноценных путей (Equal-Cost Multipath или ECMP). Некоторые реализации маршрутизаторов обеспечивают возможность использования равноценных путей в RIP и других протоколах маршрутизации. Эффект такой маршрутизации заключается в том, что пересылающий узел имеет несколько вариантов next-hop (следующий маршрутизатор) для данного получателя и должен использовать тот или иной метод выбора next-hop для пакета.

1. Введение

Различные протоколы маршрутизации, включая OSPF и ISIS, явно позволяют использовать маршрутизацию через множество равноценных путей. Некоторые маршрутизаторы обеспечивают такую возможность для RIP и других протоколов маршрутизации. Использование множества равноценных путей означает, что при наличии у маршрутизатора множеств путей с одной стоимостью к данному адресату, он может обнаруживать и использовать такие пути для распределения нагрузки между ними.

Влияние маршрутизации по множеству путей для пересылающего узла заключается в том, что он имеет множество значений next-hop для данного получателя и должен использовать тот или иной метод выбора next-hop для конкретного пакета данных. В этом документе обобщается опыт, рассматриваются проблемы и предлагаются решения.

2. Постановка задачи

Некоторые реализации маршрутизаторов позволяют пересылать пакету через множество путей. Иногда это выполняется путём простого перебора таких путей по кругу (round-robin), когда пакет, соответствующий маршруту к данному получателя, всякий раз передаётся по следующему варианту next-hop, которые организованы «в кольцо». Это обеспечивает распределение нагрузки, но с круговым или случайным выбором пути связан рад проблем.

Переменное значение Path MTU

Поскольку каждый из путей может иметь своё значение MTU, это означает возможность изменения MTU от пакета к пакету, что снижает пользу определения path MTU.

Переменные задержки

Поскольку задержка на каждом из путей может меняться, использование разных путей может приводить к нарушению доставки пакетов, увеличению задержки при доставке и повышение требований к буферизации.

Нарушение порядка доставки вынуждает TCP предполагать потерю пакетов, когда пакеты с большими номерами приходят раньше. При получении трёх и более пакетов раньше «запоздавшего» TCP активизирует режим ускоренного повтора (fast-retransmit) [6], что ведёт к добавочному расходу пропускной способности (это может вызывать дополнительные потери и снижение пропускной способности). Поэтому нарушение порядка может создавать помехи в работе сети.

Отладка

Обычные средства отладки типа ping и traceroute могут давать менее достоверные результаты при наличии множества путей и даже полностью искажать картину.

При групповой маршрутизации проблема множества путей заключается в том, что протоколы многоадресной маршрутизации предотвращают петли и дублирование пакетов путём построения общего дерева для всех получателей с общим групповым адресом. Реализованные в настоящее время протоколы групповой маршрутизации (DVMRP, PIM-DM, PIM-SM) [2] создают дерево кратчайшего пути с корнем у отправителя или маршрутизатора, называемого ядром (Core) или точкой встречи (Rendezvous Point). Поэтому для предотвращения дубликатов можно использовать лишь одно значение next-hop в направлении корня дерева.

3. Требования

В оставшейся части документа термин «поток» (flow) служит для представления уровня детализации, с которым маршрутизатор сохраняет состояния (если они есть) для классов трафика. Точное определение потока может зависеть от реализации. Например, поток может идентифицироваться триплетом (адрес отправителя, адрес получателя,

[1]Open Shortest Path First - сначала кратчайший путь.
[2]Intermediate System to Intermediate System - протокол маршрутизации между промежуточными системами.

идентификатор протокола). Поэтому поток не обязательно является синонимом «микропотока», определённого в RFC 2474 [7], который учитывает также номера портов. Действительно, включение информации транспортного уровня в процесс выбора next-hop может оказаться проблематичным (например, при фрагментации пакетов информация транспортного уровня может быть недоступна). Кроме того, выбор пути в зависимости от полей транспортного уровня может снизить преимущества, обеспечиваемые кэшированием информации типа MTU, для использования в последующих соединениях между теми же конечными точками.

Все проблемы, упомянутые в предыдущем разделе, возникают в случаях, когда пакеты одного индивидуального или группового «потока» распределяются между разными путями. Естественным решением для таких случаев является сохранение пути для всех пакетов одного потока.

Желательна также поддержка двух указанных ниже функций.

Минимальные нарушения потоков

Использование множества путей означает, что имеется множество маршрутов с пригодными вариантами next-hop и вероятность добавления или удаления путей становится выше, чем при использовании только «лучшего» маршрута (в этом случае изменение метрики дополнительных маршрутов не влияет на пути трафика). Поскольку при наличии множества путей пересылка может выполняться по разным маршрутам, возможно изменение порядка доставки и потеря пакетов в результате флуктуаций маршрутов , возникающее более часто, чем при использовании единственного пути. Поэтому желательно минимизировать число активных потоков, на которые влияет добавление или удаление другого пути (next-hop).

Быстрая реализация

Следует ограничивать объем дополнительных расчётов, требуемых для пересылки пакетов. Например, при использовании кругового перебора путей расчёты могут ограничиваться инкрементированием (по модулю, совпадающему с числом путей) индекса next-hop.

4. Решения

Ниже представлены три возможных варианта повышения производительности при использовании множества путей, а затем рассмотрена их применимость для индивидуальной и групповой пересылки.

Хэш по модулю N

Для выбора варианта next-hop из числа N имеющихся маршрутизатор выполняет хэширование по модулю N для полей заголовка, идентифицирующих поток. Преимуществом метода является высокая скорость, а недостатком нарушение (N-1)/N доли от всех потоков при добавлении или удалении next-hop.

Хэш-порог

Маршрутизатор сначала выбирает ключ с помощью хэширования полей пакета, которые указывают поток. Имеющимся N вариантам next-hop назначаются уникальные области в пространстве значений используемой хэш-функции. Далее путём сопоставления хеш-значения с границами областей маршрутизатор выбирает одно из значений next-hop. Преимуществом этого метода является воздействие только на потоки около границ областей (или порогов) при удалении или добавлении вариантов next-hop. Для ECMP поиск области можно выполнять путём обычного деления hash_value/fixed_region_size. При добавлении или удалении next-hop пути изменяются лишь для части потоков (от 1/4 до 1/2). Анализ этого метода приведён в работе [3].

Максимальный случайный вес (HRW[1])

Маршрутизатор рассчитывает ключ для каждого варианта next-hop путем хеширования идентифицирующих поток полей заголовка и адреса next-hop. После этого маршрутизатор определяет next-hop по весу полученного в результате ключа [4]. Преимуществом этого метода является минимальное влияние добавления или удаления next-hop на потоки (1/N), но приблизительно в N раз большая трудоёмкость расчётов по сравнению с хэшированием по модулю N.

Применимость этих методов зависит (как минимум) от двух факторов - наличие поддержки на устройстве пересылки состояний для каждого потока и возможности загрузки ресурсов CPU.

Некоторые маршрутизаторы поддерживают состояния для каждого потока по причинам, не связанным с наличием множества путей. Например, маршрутизаторы обычно сохраняют состояния групповых потоков для поддержки списка интерфейсов, в которые следует копировать пакеты каждого потока.

Если состояния для потоков поддерживаются в устройстве пересылки с множеством путей, выбор next-hop маршрутизатор может в процессе создания такого состояния. Это не требует дополнительных вычислений при пересылке пакетов по сравнению с обычной пересылкой по единственному пути, поскольку вариант next-hop определяется заранее. В таких случаях можно применять любой метод, включая круговой обход, случайный выбор, modulo-N, hash-threshold или HRW. Хэш-функции типа modulo-N, hash-threshold и HRW будут предпочтительными, если состояние пересылки может быть удалено по той или иной причине в течение срока действия потока, поскольку последующий расчёт next-hop маршрутизатором приведёт к выбору того же пути. Это также улучшает работу с отладочными утилитами типа traceroute. Для максимальной стабильности путей (и эффективности traceroute и т. п.) рекомендуется использовать метод HRW.

Если устройство пересылки не поддерживает состояний для потоков, для использования множества next-hop требуется выбор следующего интервала в момент прибытия пакета. Если экономия ресурсов CPU более важна, чем стабильность путей для потоков, рекомендуется использовать описанный выше метод hash-threshold.

4.1. Индивидуальная пересылка

При пересылке по индивидуальным адресам состояния для потоков могут сохраняться или не сохраняться. При сохранении состояний на устройстве пересылки рекомендуется использовать метод HRW в момент создания состояния (и в момент удаления next-hop) для выбора next-hop, а на устройствах без поддержки состояний использовать метод hash-threshold.

[1] Highest Random Weight.

4.2. Групповая пересылка

Современные машины групповой пересылки используют кэш записей пересылки, индексированный по группам (или групповым префиксам) и отправителям (или префиксам отправителей). Это означает, что такие машины всегда хранят состояния для потоков, хотя в некоторых протоколах групповой маршрутизации «потоки» могут быть достаточно грубыми (например, трафик к одному получателю от всех отправителей). Поскольку устройство пересылки сохраняет состояния потоков, на маршрутизаторе рекомендуется использовать метод HRW для выбора next-hop.

Маршрутизаторам, использующим протоколы с явным присоединением к группам (тип PIM-SM [5]) следует использовать информацию о множестве путей при определении соседа, которому нужно отправить сообщение о вступлении в группу. Например, на данной «точке встречи» (RP) при наличии множества next-hop, по которым нужно отправить сообщение (*,G) Join, рекомендуется использовать метод HRW при выборе next-hop для каждой группы.

5. Применимость

Рассмотренные выше алгоритмы (кроме round-robin) всегда основаны на той или иной хэш-функции. Однородное распределение потоков обеспечивается в случае однородного распределения выхода хэш-функции. Поскольку хэш-функции общего назначения дают равномерное распределение лишь при достаточно большом числе входных данных, эти алгоритмы более применимы для маршрутизаторов с большим числом потоков, нежели для маршрутизаторов небольших компаний.

6. Избыточные параллельные каналы

Похожая проблема возникает при наличии множества параллельных каналов между парой маршрутизаторов. Общим решением является объединение двух каналов в «суперканал» и использовании его для маршрутизации. При групповой маршрутизации эти два канала будут иметь общее значение next-hop (через объединённый канал), которое будет использоваться для предотвращения дубликатов. При постановке индивидуальных или групповых пакетов в очередь объединённого канала нужен тот или иной метод (типа описанных выше) для определения физического канала, через который будет передаваться пакет. Если параллельные каналы идентичны, большинства проблем, рассмотренных в этом документе, на объединённом канале не возникает. Исключением является изменение порядка доставки пакетов, которое может сохраняться для метода round-robin, и оказывать негативное влияние на TCP.

7. Вопросы безопасности

В этом документе рассматриваются вопросы, связанные с разными методами выбора следующего интервала пересылки (next-hop) при наличии множества пригодных путей. Поэтому документ не связан напрямую с вопросами безопасности инфраструктуры и приложений Internet.

Однако следует отметить, что в случае предсказуемого выбора next-hop атакующий может создать трафик, который будет при хэшировании давать один результат, приводящий к выбору одного пути для всех пакетов и таким образом организовать атаку на службы путём перегрузки определённого пути. Поскольку наличие единственного пути является особым (вырожденным) случаем наличия множества next-hop, такую атаку проще всего организовать в тех случаях, когда множество путей не используется. Добавление маршрутизации по многим путям может осложнить организацию таких атак - чем менее предсказуем результат хэширования, тем сложнее будет организовать атаку с перегрузкой любого отдельного канала.

8. Литература

[1] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[2] Maufer, T., "Deploying IP Multicast in the Enterprise", Prentice-Hall, 1998.

[3] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, November 2000.

[4] Thaler, D., and C.V. Ravishankar, "Using Name-Based Mappings to Increase Hit Rates", IEEE/ACM Transactions on Networking, February 1998.

[5] Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., Jacobson, V., Liu, C., Sharma, P. and L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998.

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

[7] Nichols, K., Blake, S., Baker, F. and D. Black., "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

9. Адреса авторов

Dave Thaler
Microsoft
One Microsoft Way
Redmond, WA 98052
Phone: +1 425 703 8835
EMail: dthaler@dthaler.microsoft.com

Christian E. Hopps
NextHop Technologies, Inc.
517 W. William Street
Ann Arbor, MI 48103-4943
U.S.A
Phone: +1 734 936 0291
EMail: chopps@nexthop.com

Перевод на русский язык

Николай Малых

nmalykh@protokols.ru

10. Полное заявление авторских прав

Подтверждение

Финансирование функций RFC Editor обеспечено Internet Society.