Network Working Group
Request for Comments: 4264
Category: Informational

T. Griffin
University of Cambridge
G. Huston
APNIC

November 2005

BGP Wedgies

Статус документа

Авторские права

Тезисы

Принято считать, что протокол BGP[1] является средством распространения информации о доступности сетей, обеспечивающим создание детерминированных путей пересылки трафика. В этом документе описывается класс конфигураций BGP, для которых существует более одной точки выхода и могут существовать стабильные состояния, отличные от предусмотренных. Кроме того, стабильные состояния BGP могут выбираться протоколом BGP недетерминированными способами. Эти стабильные, но не предусмотренные состояния BGP обозначаются термином BGP Wedgies.

Оглавление

1. Введение

Принято считать, что протокол BGP [RFC1771] является средством распространения информации о доступности сетей, обеспечивающим создание детерминированных путей пересылки трафика. В этом документе (problem statement – констатация проблемы) описывается класс конфигураций BGP, для которых может существовать более одного стабильного состояния пересылки. Одним из таких стабильных является предусмотренное (intended) состояние, а остальные стабильные состояния являются непредусмотренными (unintended). Процесс схождения BGP может приводить к недетерминированному выбору стабильного состояния пересылки.

Такие стабильные, но непреднамеренные состояния BGP обозначаются в этом документе термином BGP Wedgies[2].

2. Описание политики маршрутизации BGP

Политика маршрутизации BGP в общем случае отражает задачи сетевого администратора по оптимизации расходов, производительности и надежности сети.

В части оптимизации расходов принятая по умолчанию локальная политика маршрутизации часто отдает предпочтение маршрутам, полученным от заказчиков по отношению к маршрутам, полученным от центров обмена трафиком. По тем же причинам локальная сеть зачастую настраивается так, чтобы отдавалось предпочтение маршрутам, полученным от заказчиков или партнеров (peer), перед маршрутами, полученными от транзитных upstream-провайдеров. Эти предпочтения могут выражаться через локальные настройки конфигурации, где локальные предпочтения (local preference) имеют более высокий приоритет по сравнению с метрикой AS path length, принятой в BGP.

С точки зрения надежности в общем случае междоменная маршрутизация организована так, что сервис-провайдер имеет каналы к двум или большему количеству вышестоящих (upstream) транзитных провайдеров, передавая маршруты всем этим провайдерам и принимая трафик из всех источников. Если путь к вышестоящему провайдеру рвется, трафик будет передаваться через другие каналы. После восстановления разорванного пути передача трафика по нему возобновляется.

В таких ситуациях для множества upstream-провайдеров также устанавливаются уровни предпочтения так, что одно из соединений является предпочтительным или основным (primary), а остальные рассматриваются как менее предпочтительные или резервные (backup). Смысл этого состоит в том, что резервные соединения используются для передачи трафика только при повреждении основного канала.

[1]Border Gateway Protocol – протокол граничного шлюза.
[2]По-русски это лучше всего будет назвать “расколом BGP”. *Прим. перев.*

Политику “основной-резервный” можно задать с использованием локальных настроек AS path prepending, когда пути (AS path) искусственно удлиняется для резервных провайдеров путем вставки дополнительных значений local AS. Этот алгоритм выбора не является детерминированным, поскольку выбранный в качестве первичного провайдер также может использовать искусственное удлинение пути для своего резервного upstream-провайдера и это может приводить к тому, что в некоторых случаях путь через резервного провайдера может оказаться самым коротким (shortest AS path).

В качестве другого варианта управления политикой маршрутизации используются группы BGP (BGP community) [RFC1997]. В этом случае провайдер публикует набор значений community, который позволяет клиентам выбрать для провайдера локальные параметры предпочтения. Клиент может использовать группу (community) для маркировки как качестве резервного (backup only) маршрута в направлении резервного провайдера и установить primary preferred (предпочтительный) для маршрута в направлении основного провайдера. В этом случае локальные предпочтения будут иметь более высокий приоритет по сравнению с метрикой AS path, поэтому маршрут, помеченный как backup only, будет использоваться только в тех случаях, когда другие маршруты недоступны.

3. BGP Wedgies

Возможности локальной политики маршрутизации, задаваемой с использованием групп, в комбинации с протоколами на основе векторов расстояния типа BGP ведут к возникновению возможности существования нескольких “решений” или стабильных состояний BGP. Пример такой ситуации показан на рисунке 1.

В этом случае AS1 помечает свои анонсы префиксов в AS2 как backup only, а анонсы префиксов в AS4, как primary. AS4 будет анонсировать префиксы AS1 в автономную систему AS3, которая будет видеть анонсы AS4 через партнерское соединение (peering link) и выбирать префиксы AS1 с путем "AS4, AS1". Эти префиксы AS3 будет анонсировать в автономную систему AS2, которая будет видеть два пути к префиксам AS1. Первый путь будет проходить через прямое соединение с AS1, а второй через "AS3, AS4, AS1". AS2 будет выбирать более длинный путь, поскольку маршруты через прямое подключение помечены как резервные, и локальная политика AS2 будет предпочитать анонсы от AS3 перед анонсами от AS1.

Здесь существует преднамеренный результат политики AS1, когда в “нормальном” состоянии трафик совсем не передается из AS2 в AS1 через резервный канал и AS2 обменивается данными с AS1 через путь, включающий AS3 и AS4 и являющийся основным каналом в AS1.

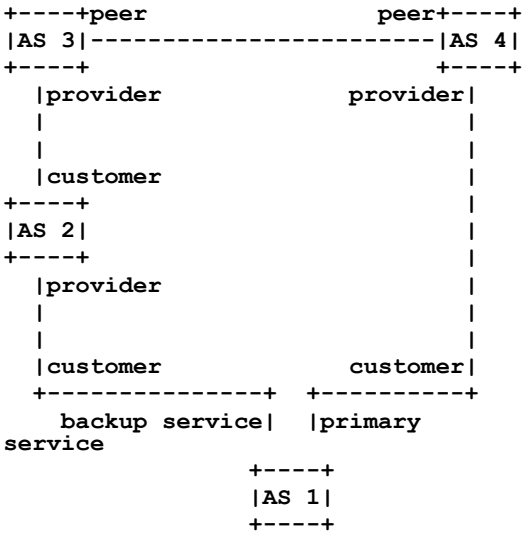

Рисунок 1

Этот преднамеренный результат достигается когда AS1 анонсирует свои маршруты в основной путь через AS4 до анонсирования резервных маршрутов в AS2.

Если путь AS1 - AS4 обрывается, разрывая сессию BGP между AS1 и AS4, автономная система AS4 будет аннулировать свои анонсы маршрутов AS1 в AS3, которая, в свою очередь, будет аннулировать анонсы маршрутов в AS2. В этом случае AS2 будет выбирать резервный путь в AS1. Этот путь AS2 будет анонсировать в AS3, а AS3 будет анонсировать его далее в AS4. Этот процесс является частью преднамеренной политики резервирования каналов и весь трафик в AS1 пойдет по резервному пути.

При восстановлении канала между AS4 и AS1 состояние BGP не вернется к первоначальному. AS4 получит основной путь в AS1 и анонсирует его в AS3, используя "AS4, AS1". Автономная система AS3, используя принятые по умолчанию предпочтения для анонсируемых заказчиками маршрутов по отношению peer-маршрутам, будет по-прежнему выбирать путь "AS2, AS1" и в результате AS3 не будет передавать никаких обновлений в AS2. После восстановления канала между AS4 и AS1 трафик из AS3 и AS2 в AS1 будет передаваться в результате через резервный канал, несмотря на нормальную работу основного пути через AS4.

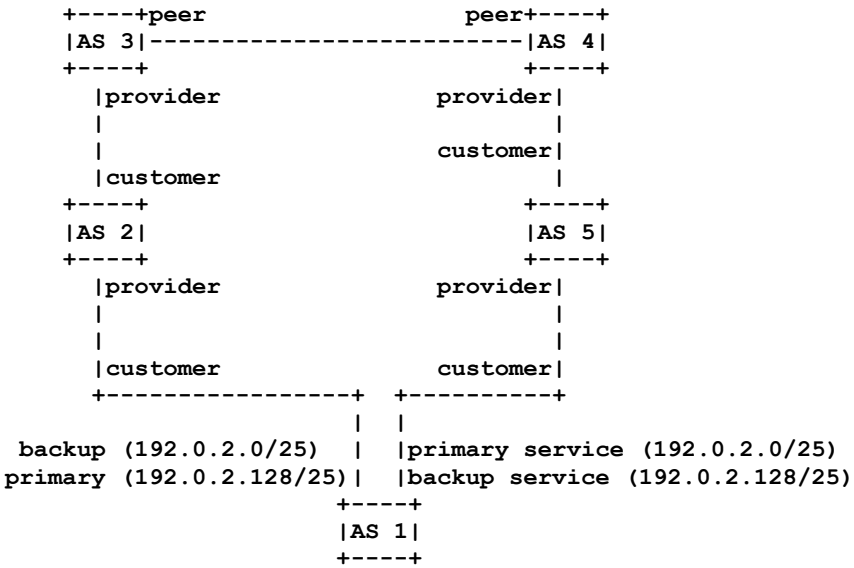

Рисунок 2

Предусмотренное состояние пересылки может быть восстановлено AS1 путем преднамеренного разрыва сессии eBGP с AS2, несмотря на наличие трафика. В результате такого разрыва будет восстановлено предусмотренная конфигурация BGP.

Достаточно часто автономные системы пытаются балансировать входящий трафик между несколькими провайдерами, используя тот же механизм primary-backup. Для некоторых префиксов один канал настраивается в качестве первичного, а остальные служат резервными, тогда как для других префиксов в качестве основного может быть выбран другой канал. Пример такой ситуации показан на рисунке 2.

В преднамеренной конфигурации весь входящий трафик для блока адресов 192.0.2.0/25 приходит по каналу из AS5, а трафик для блока 192.0.2.128/25 – из AS2.

В данном случае при разрыве канала между AS3 и AS4 автономная система AS3 будет получать оба маршрута от AS2, а AS4 – от AS5. Поскольку маршруты от заказчиков являются предпочтительными по сравнению с маршрутами от

партнеров (peer route), при восстановлении связи между AS3 и AS4 ни AS3, ни AS4 не будут менять свое поведение относительно маршрутов AS1. В данном случае описанным выше способом не удастся восстановить преднамеренное состояние, поскольку здесь нет eBGP-партнера для разрыва сессии, способного восстановить состояние. Это один из примеров BGP Wedgie.

Способ восстановления в этом случае заключается в том, что AS1 прекращает анонсировать резервные соединения в оба канала и работает некоторое время без резервирования, после чего восстанавливает анонсы своих префиксов в резервные пути. Продолжительность интервала работы без резервирования невозможно определить заранее и этот интервал должен быть достаточно продолжительным, чтобы AS2 и AS5 узнали дополнительный путь в AS1. После этого можно восстанавливать анонсы маршрутов.

4. Multi-Party BGP Wedgies

Описанная выше ситуация может быть усложнена, если количество транзитных upstream-провайдеров для AS составляет 3 или более. Пример такой ситуации показан на рисунке 3.

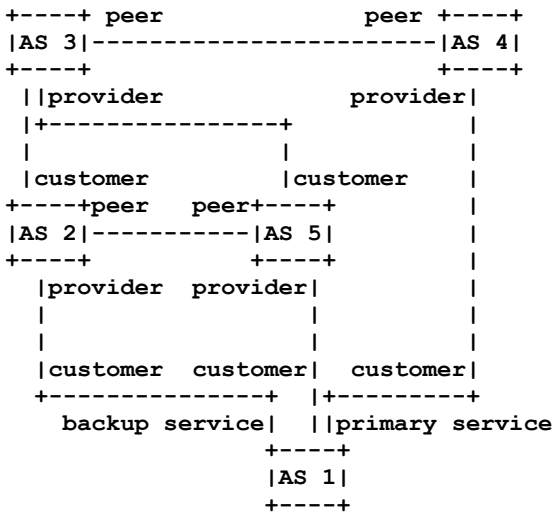

Рисунок 3

В показанном на рисунке примере предусмотренное состояние заключается в том, что AS2 и AS5 являются резервными, а AS4 – основным провайдером для AS1. При разрыве и последующем восстановлении канала между AS1 и AS4 автономная система AS3 будет по-прежнему направлять трафик в AS1 через AS2 или AS5. В такой ситуации однократный разрыв канала между AS2 и AS1 не будет обеспечивать восстановление предусмотренного состояния BGP, поскольку выбранный BGP лучший маршрут в AS1 будет меняться на AS5, а AS2 и AS3 будут узнавать путь в AS1 через AS5.

Предположим, что AS1 получает входящий трафик через резервный канал из AS2. Разрыв этого соединения не будет приводить к восстановлению передачи трафика через основной путь. Вместо этого разрыв лишь приведет к тому, что входящий трафик пойдет через AS5. Для изменения ситуации требуется одновременно разорвать соединения с AS2 и AS5. Это решение может показаться неочевидным, поскольку в любой момент в качестве резервного используется только один канал. Тем не менее требуется разорвать оба сеанса BGP одновременно для того, чтобы восстановить предусмотренное состояние соединений.

5. BGP и детерминированность

Протокол BGP является детерминированным не во всех случаях, следовательно, существуют предусмотренные и непредусмотренные недетерминированные состояния BGP. Например, используемая по умолчанию схема разрыва петель в некоторых реализациях BGP отдает предпочтение наиболее давнему (longest-lived) маршруту. Для получения определенного результата на последнем этапе в таких случаях необходимо использовать оператор сравнения с предсказуемым результатом (например, сравнение идентификаторов в маршрутизаторах). Такой тип недетерминированного поведения называют предусмотренной индетерминированностью[1], и результаты с той или иной достоверностью могут быть предсказаны администратором сети.

BGP может также порождать состояния, которые описываются как непредусмотренная индетерминированность[2] и могут приводить к неожиданным результатам взаимодействия. Такие ситуации нельзя считать некорректностью настройки конфигурации в обычном понимании, поскольку все правила могут локально представляться совершенно разумными, но взаимодействие правил множества объектов маршрутизации может приводить к непредсказуемым результатам и BGP может перейти в недетерминированное и непредусмотренное состояние.

Непредусмотренная индетерминированность в BGP не будет вызывать критических ситуаций, если все стабильные варианты маршрутизации гарантированно совместимы с целями создателя политики. Однако такое наблюдается не во всех ситуациях. Приведенные выше примеры показывают, что BGP может перейти в различные состояния из одного предусмотренного состояния и не все такие состояния могут оказаться совместимыми с принятой политикой. Такие случаи можно рассматривать как форму route pinning, когда маршруты связываются с путями, не являющимися предпочтительными.

Задачей сетевых администраторов является гарантированная поддержка предусмотренного состояния. В некоторых ситуациях такого результата можно добиться только за счет преднамеренного прерывания сервиса, включающего аннулирование маршрутов, используемых для пересылки трафика, и повторное анонсирование маршрутов в определенном порядке для перехода BGP в предусмотренное состояние. Однако для понимания причин, по которым BGP стабилизируется в непредусмотренном состоянии, администратору сети требуется информация о политике конфигурации BGP удаленных сетей. Таким образом информации о локальной политике недостаточно для понимания причин и выработки решения, позволяющего восстановить предусмотренное состояние BGP.

Разумно предположить, что плотность соединений между сетями будет продолжать свой рост и возможности установки предпочтений на основе правил для принятых из других автономных систем и анонсируемых далее маршрутов также будут расширяться. Следовательно, разумно предположить рост числа непредусмотренных, но стабильных состояний BGP, что потребует определения необходимых последовательностей аннулирования и повторного анонсирования маршрутов для восстановления предусмотренных состояний.

Вопрос о том, будет ли это приводить систему маршрутизации BGP в точку, где каждая сеть не сможет согласованно направлять трафик в детерминированной манере, является в данный момент предметом анализа. Примеры BGP Wedgies показывают, что достаточно сложная топология соединений в комбинации с множеством вариантов задания политики могут приводить в возникновению множества стабильных состояний BGP вместо одного предусмотренного состояния. По мере усложнения топологии становится невозможным детерминированное предсказание состояния, в

[1] "Intended" non-determinism
[2] Unintended non-determinism

которое может перейти система маршрутизации BGP. Парадоксально, но потребности управления междоменным трафиком требуют дополнительного расширения возможностей выражения политики в системах с высокой плотностью соединений при сохранении детерминированности. Такое расширение возможностей может приводить к неустойчивости систем маршрутизации на основе BGP.

6. Вопросы безопасности

BGP является трансляционным протоколом, в котором маршрутная информация принимается, обрабатывается и передается дальше. Протокол BGP не включает механизмов предотвращения несанкционированного изменения информации при пересылке, что позволяет изменять или удалять маршрутные данные, или включать в них фальсифицированные маршруты без необходимости принимать во внимание источник маршрутных данных или кого-либо из их получателей.

В этом документе не предлагается изменить протокол BGP или требования к его реализациям и, следовательно, не вносится дополнительных факторов, влияющих на безопасность и целостность междоменной маршрутизации.

Этот документ показывает, что при попытках создания системы выбора пути для входящего трафика на основе правил возможно возникновение конфигураций BGP, имеющих множество стабильных состояний вместо одного предсказуемого состояния. Следовательно, среди множества таких состояний могут возникать такие варианты, при которых выбор BGP перестанет быть детерминированным.

Такие варианты поведения систем могут использоваться сторонними злоумышленниками. Общим для BGP Wedgies является то, что для системы, находящейся в предусмотренном (желательном) состоянии пересылки трафика, разрыв и последующее восстановление партнерских соединений eBGP могут приводить конфигурацию системы пересылки в непредусмотренное и потенциально нежелательное состояние. Усилия администратора, основанные на локальной информации о состоянии и конфигурации BGP, могут оказаться недостаточными для восстановления предусмотренного (желательного) состояния пересылки трафика. Если злоумышленник может по своему усмотрению разрывать соединения BGP, он будет способен перевести систему в нежелательное состояние, которое может оказаться достаточно продолжительным и приводить к потере соединений. Если такое влияние с учетом роста расходов, снижения доступной полосы, увеличения суммарной задержки и снижения надежности обслуживания окажется достаточно серьезным, разрыв BGP может оказаться привлекательным для злоумышленников способом организации атак.

7. Литература

7.1. Нормативные документы

[RFC1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771[1], March 1995.

7.2. Информационные ссылки

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997[2], August 1996.

Адреса авторов

Tim G. Griffin

Computer Laboratory

University of Cambridge

EMail: Timothy.Griffin@cl.cam.ac.uk

Geoff Huston

Asia Pacific Network Information Centre

EMail: gih@apnic.net

Перевод на русский язык

Николай Малых

nmalykh@gmail.com

Полное заявление авторских прав

[1]Этот документ устарел и заменен RFC 4271. Перевод имеется на сайте http://www.protocols.ru. *Прим. перев.*

[2]Перевод этого документа имеется на сайте http://www.protocols.ru. Там же доступен перевод RFC 4360, расширяющего возможности использования групп в BGP. *Прим. перев.*

Интеллектуальная собственность

Подтверждение

Финансирование функций RFC Editor обеспечено Internet Society.