Internet Engineering Task Force (IETF)
Request for Comments: 7938
Category: Informational
ISSN: 2070-1721

P. Lapukhov
Facebook
A. Premji
Arista Networks
J. Mitchell, Ed.

August 2016

Использование BGP для маршрутизации в крупных распределенных ЦОД

Use of BGP for Routing in Large-Scale Data Centers

Тезисы

Некоторые сетевые операторы строят и эксплуатируют центры обработки данных (ЦОД) с числом серверов в сотни тысяч. В этом документе такие центры называются крупными (large-scale) в отличие от более мелких инфраструктур. Среды такого масштаба отличаются уникальным набором требований к сети с упором на операционную простоту и стабильность сети. В этом документе обобщен операционный опыт создания и обслуживания крупных ЦОД с использованием BGP в качестве единственного протокола маршрутизации. Цель заключается в описании проверенной и стабильной системы маршрутизации, которая может быть использована в отрасли.

Статус документа

Документ не содержит спецификации Internet Standards Track и публикуется с информационными целями.

Документ является результатом работы IETF[1] и представляет согласованный взгляд сообщества IETF. Документ прошел открытое обсуждение и был одобрен для публикации IESG[2]. Не все документы, одобренные IESG, претендуют на статус стандартов Internet, как указано в разделе 2 в RFC 7841.

Информацию о текущем статусе документа, ошибках и способах обратной связи можно найти по ссылке http://www.rfc-editor.org/info/rfc7938.

Оглавление

[1]Internet Engineering Task Force.
[2]Internet Engineering Steering Group.

1. Введение

Документ описывает практическое решение по маршрутизации в крупных распределенных центрах обработки данных (ЦОД). Такие центры, называемые также hyper-scale или warehouse-scale, могут содержать сотни тысяч серверов. Для организации сетей такого масштаба операторам приходиться пересматривать сетевые решения и используемые платформы.

Представленное здесь решение основано на опыте ЦОД, построенных для поддержки крупномасштабной программной инфраструктуры типа поисковых машин web. Основными требованиями в таких средах являются простота обслуживания и стабильность сети для того, чтобы можно было эффективно поддерживать крупную сеть силами небольшой группы людей.

Эксперименты и широкое тестирование показали, что протокол EBGP[1] [RFC4271] хорошо подходит в качестве автономного протокола маршрутизации для такого типа ЦОД. Это отличается от традиционных проектов ЦОД, которые могут использовать простую древовидную топологию и базироваться на расширении доменов L2[2] через множество сетевых устройств. В документе подробно рассмотрены требования, которые привели к такому выбору, и представлены детали маршрутизации EBGP, а также представлены идеи для дальнейшего развития.

В этом документе впервые представлен обзор требований к устройству сети и проблем крупных ЦОД. Затем традиционные сетевые топологии ЦОД сравниваются с сетями Clos [CLOS1953], которые масштабируются по горизонтали. Далее подробно рассматриваются аргументы для выбора EBGP с топологией Clos как наиболее подходящего протокола маршрутизации, соответствующего требованиям и устройству сети. В заключение рассматриваются некоторые дополнительные соображения и варианты устройства. Предполагается близкое знакомство читателя с протоколом BGP для планирования и реализации описанного здесь решения.

2. Требования к устройству сети

В этом разделе описываются и суммируются требования к устройству сетей для крупных ЦОД.

2.1. Пропускная способность и картина трафика

Основной задачей при построении сети для большого числа серверов являются выполнение требований приложений к пропускной способности и задержкам. До недавнего времени было принято рассматривать основную часть трафика, входящего в ЦОД и выходящего из него, как трафик «север-юг». Традиционной топологии «дерево» было достаточно для размещения таких потоков даже при значительной «переподписке» между уровнями сети. Если требовалась дополнительная пропускная способность, она добавлялась путем «увеличения» элементов сети, например, обновления интерфейсных плат в устройствах или коммутаторах, а также установки устройств с более высокой плотностью портов.

Сегодня многие приложения на хостах крупных ЦОД создают большой объем трафика между серверами, который не выходит из ЦОД и обычно называется трафиком «восток-запад». Примерами могут служить компьютерные кластеры типа Hadoop [HADOOP], репликация данных между приложениями или перенос виртуальных машин. Масштабирование пропускной способности для таких систем на основе традиционных древовидных топологий становится слишком дорогим или просто невозможным физически (например, по причине ограниченной плотности портов в коммутаторах).

2.2. Минимизация капитальных затрат

Капитальные затраты (CAPEX[3]), связанные с сетевой инфраструктурой, обычно составляют 10-15% общей стоимости ЦОД [GREENBERG2009]. Величина в денежном выражении достаточно велика, поэтому постоянно приходится снижать стоимость отдельных элементов сети. Для этого обычно используется два способа:

- унификация элементов и предпочтительным использованием однотипных или даже одинаковых устройств, позволяющая снизить расходы на закупку, поддержку и инвентаризацию;
- снижение расходов за счет использования конкуренции между производителями.

Для диверсификации важно минимизировать требования к программным функциям для элементов сети. Такая стратегия обеспечивает максимальную гибкость при выборе оборудования с сохранением интероперабельности за счет применения открытых стандартов.

[1] External BGP.
[2] Layer 2 — канальный уровень.
[3] Capital Expenditures.

2.3. Минимизация операционных расходов

Эксплуатация крупной инфраструктуры может быть дорогой, поскольку с ростом числа элементов растет и число отказов. Простое устройство и работа с использованием ограниченного набора программных возможностей минимизирует программные отказы.

Важным аспектом минимизации операционных расходов (OPEX[1]) является снижение размера домена отказов в сети. Известно, что сети Ethernet чувствительны к широковещательным и обычным (unicast) «штормам» трафика, которые могут существенно влиять на производительность и доступность сети. Использование полностью маршрутизируемой системы значительно снижает размеро домена отказов на уровне данных, ограничивая его нижним уровнем иерархии сети. Однако при этом возникает проблема отказов в распределенном уровне управления. Это требует более простых и компактных протоколов уровня управления для упрощения протокольных взаимодействий , снижающего вероятность отказов. Минимизация требований к программным функциям, как описано выше для CAPEX, снижает также требования к тестированию и обучению.

2.4. Организация трафика

В любом ЦОД балансировка нагрузки является одной из важных задач, решаемых сетевыми устройствами. Традиционно балансировщики реализуются в виде отдельных устройств на пути пересылки трафика. При росте сети возникает проблема масштабирования балансировщиков. Предпочтительной будет возможность горизонтального масштабирования балансировщиков путем добавления унифицированных узлов и распределения трафика между ними. В таких ситуациях идеальным выбором будет использование самой сетевой инфраструктуры для распределения трафика между этими узлами. Для решения этой задачи может использоваться комбинация anycast-анонсов префиксов [RFC4786] и функциональности ECMP[2]. Для более тонкого распределения нагрузки преимуществом будет возможность организации трафика на уровне этапа маршрутизации (per-hop). Например, преимуществом будет прямое управления набором ECMP next-hop для anycast-префиксов на каждом уровне иерархии сети.

2.5. Суммарные требования

В этом параграфе приведен список требований, очерченных в предыдущих параграфах.

- REQ1. Выбор топологии с возможностью «горизонтального» масштабирования путем добавления каналов и устройств одного типа без обновления самих элементов сети.
- REQ2. Определение небольшого набора программных функций/протоколов, поддерживаемых многими производителями сетевого оборудования.
- REQ3. Выбор протокола маршрутизации, который имеет простую реализацию в части сложности программного кода и операционной поддержки.
- REQ4. Минимизация доменов отказа оборудования и программ.
- REQ5. Поддержка той или иной организации трафика, желательно путем явного управления маршрутизацией с помощью встроенных в протокол механизмов выбора следующего интервала (next hop).

3. Обзор топологий ЦОД

В этом разделе приведен обзор двух базовых вариантов повтроения ЦОД и иерархический (на базе «дерева») и основанный на топологии Clos.

3.1. Традиционная топология ЦОД

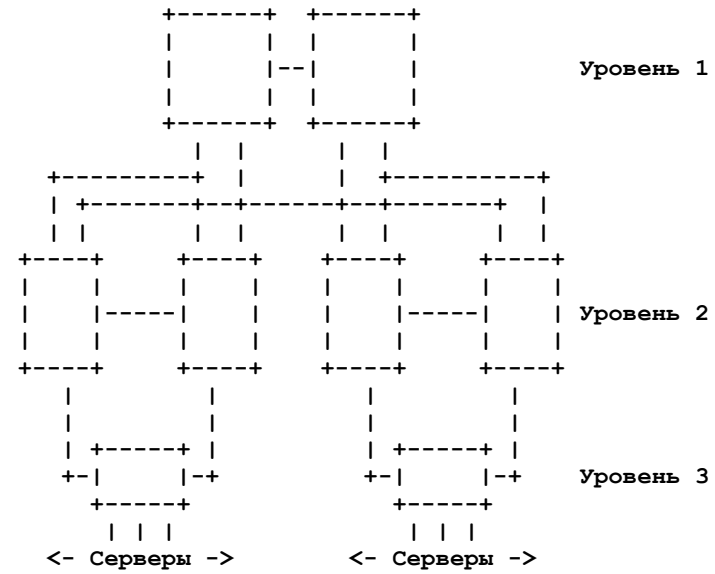

Рисунок 1. Типичная топология сети ЦОД.

В сетевой отрасли распространенное решение для ЦОД обычно выглядит подобно дереву (корнем вверх) с избыточными соединениями и тремя уровнями иерархии — ядро, агрегирование/распределение и доступ (Рисунок 1). Для удовлетворения требований к пропускной способности каждый вышележащий уровень от серверов в направлении выхода из ЦОД или WAN, имеет более высокую плотность портов и пропускную способность , а ядро служит «транком» для древовидной структуры. Для единообразия терминологии в этом документе уровни обозначаются Уровень (Tier) 1, Уровень 2 и Уровень 3 вместо терминов ядро (core), агрегирование (aggregation) и доступ (access).

[1]Operational Expenditure.
[2]Equal Cost Multipath — множество равноценных путей.

К сожалению, как было отмечено выше, древовидную структуру невозможно масштабировать должным образом по причине отсутствия устройств уровня 1 с плотностью портов, достаточной для масштабирования уровня 2. Кроме того, требуется постоянное обновление или замена устройств верхнего уровня по мере роста размера и пропускной способности, что усложняет эксплуатацию. Поэтому требование REQ1 выводит этот вариант из рассмотрения.

3.2. Сетевая топология Clos

В этом разделе описано устройство сети с горизонтально масштабируемой топологией в больших ЦОД в соответствии с REQ1.

3.2.1. Обзор

Общепринятым решение для горизонтально масштабируемой топологии является «сфальцованная» топология Clos, которую называют также fat-tree (например, [INTERCON] и [ALFARES2008]). Эта топология использует нечетное число ступеней (stage), иногда называемых размерностями (dimension) и обычно строится на базе однородных элементов (например, коммутаторов с одинаковым числом портов). Поэтому выбор такой топологии Clos соответствует требованиям REQ1 и REQ2. На рисунке 2 below приведен пример 3-ступенчатой топологии Clos (3 ступени учитывают уровень 2 дважды при трассировке потока пакетов).

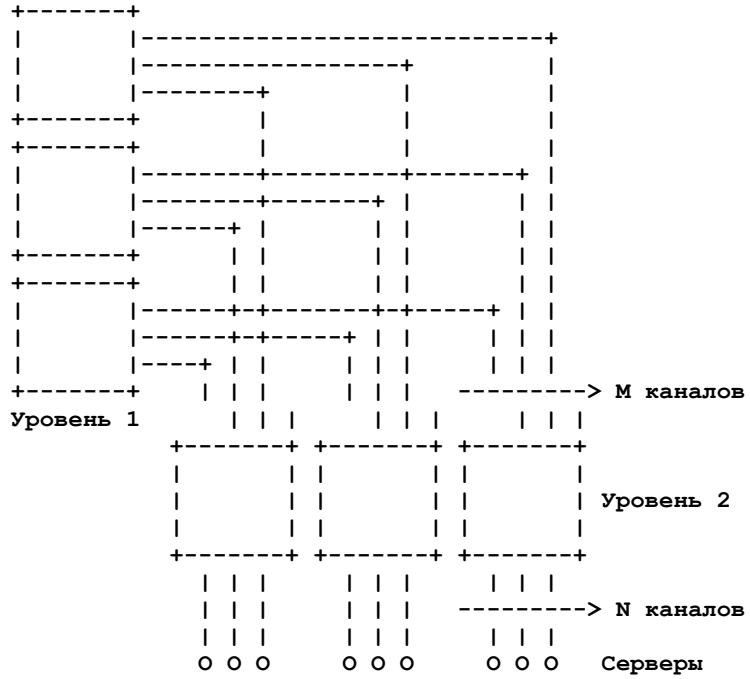

Рисунок 2. 3-ступенчатая «сфальцованная» топология Clos.

Такую топологию часто называют сетью Leaf and Spine (листья и ствол), где термином Spine обозначают среднюю ступень топологии Clos (уровень 1), Leaf — ступени ввода и вывода (уровень 2). Для единообразия в документе для обозначения этих уровней используется обозначение Tier (уровень) n.

3.2.2. Свойства топологии Clos

Ниже перечислены основные свойства топологии Clos.

- Топология является полностью неблокируемой, точнее порты не создают друг другу помех при M >= N, а в остальных случаях работают с переподпиской N/M. Здесь M и N — счетчики восходящего и нисходящего портов, соответственно, для коммутатора уровня 2 на рисунке 2.
- использование этой топологии требует поддержки уровней данных и правления для ECMP с числом вариантов не менее M.
- Коммутаторы уровня 1 имеют в точности один путь к каждому серверу. Это важное свойство делает обобщение маршрутов в данной топологии опасным (см. 8.2. Обобщение маршрутов в топологии Clos).
- Трафик между серверами распределяется (балансируется) по всем возможным путям с использованием ECMP.

3.2.3. Масштабирование топологии Clos

Топологию Clos можно масштабировать путем увеличения плотности портов в устройствах или путем добавления новых «ступеней», например, как в 5-ступенчатой топологии Clos, показанной на рисунке 3.

Небольшой пример топологии на рисунке 3 образован устройствами, имеющими по 4 порта. В этом документе набор напрямую соединенных устройств уровней 2 и 3 вместе с подключенными к ним серверами будет называться кластером. Например, устройства DEV A, B, C, D и серверы, подключенные к DEV A и B, на рисунке 3 образуют кластер. Концепция кластеров может быть полезна также при рассмотрении отдельного блока развертывания или поддержки, который может работать на другой частоте, нежели остальная часть топологии.

На практике уровень 3 в сети, который обычно образован стоечными коммутаторами ToR[1], обычно работает с переподпиской, чтобы обеспечить возможность установки большего числа серверов в ЦОД при выполнении требований разных типов приложений к пропускной способности. Основной причиной ограничения переподписки на одном уровне сети является упрощение разработки приложений, которым в ином случае пришлось бы учитывать множество пулов пропускной способности — в стойке (уровень 3), между стойками (уровень 2) и между кластерами

[1]Top-of-Rack — наверху стойки.

(уровень 1). Поскольку переподписка не связана напрямую с устройством маршрутизации, она больше не рассматривается в этом документе.

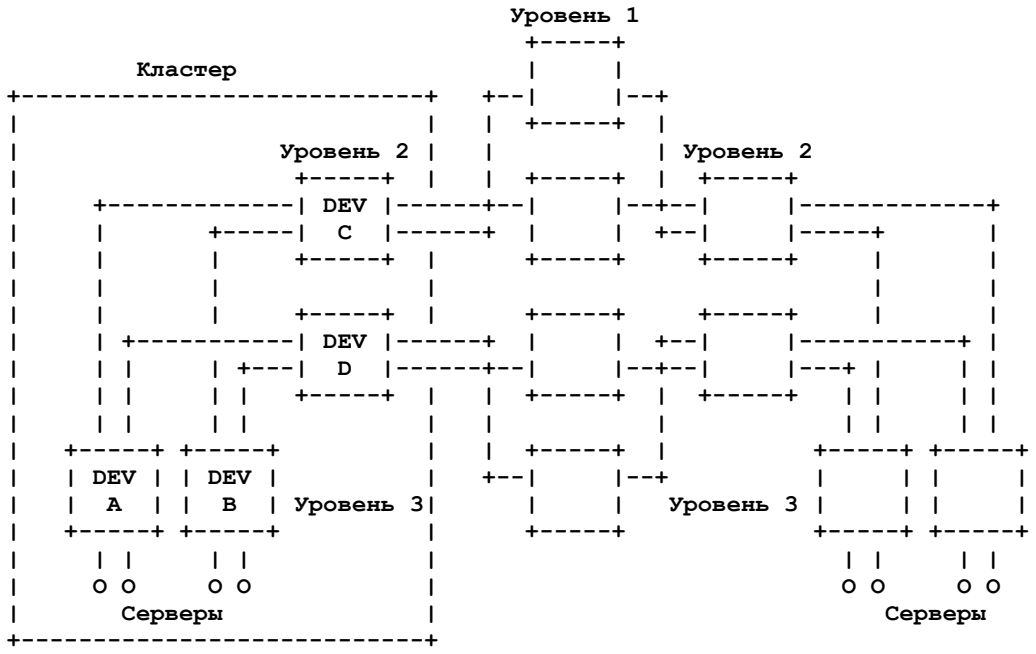

Рисунок 3. 5-ступенчатая топология Clos.

3.2.4. Управление размером уровней топологии Clos

Если размер сети ЦОД невелик, можно вдвое снизить число коммутаторов на уровнях 1 и 2 топологии Clos. Для разъяснения этого рассмотрим уровень 1 в качестве примера. Каждое устройство уровня 2 подключено к одной группе устройств уровня 1. Если половина портов на каждом устройстве уровня 1 не используется, можно уменьшить число устройств уровня 1 просто подключая к одному устройству уровня 1 два восходящих канала от устройства уровня 2, которые раньше были подключены к разным устройствам уровня 1. Этот метод обеспечивает сохранение пропускной способности при снижении числа устройств уровня 1 (и CAPEX). Платой за такую экономию будет снижение вдвое максимального числа серверов в ЦОД.

В этом примере устройства уровня 2 используют два параллельных канала для соединения с каждым устройством уровня 1. При отказе одного из этих каналов другой примет на себя весь трафик, что может привести к значительной перегрузке и снижению качества обслуживания, если процедура определения пути не учитывает пропускную способность, поскольку число восходящих устройств уровня 1 может быть больше 2. Для решения проблемы можно объединять параллельные соединения в группы LAG[1] (например, [IEEE8023AD]), с широко известными настройками, которые будут отключать (down) всю группу при отказе одного канала. Можно также использовать эквивалентный метод «общей судьбы» (fate sharing) на параллельных каналах вместо LAG. В результате из двух (или более) отказавших каналов будет распределен между оставшимися путями по числу устройств уровня 1. В примере для простоты используется 2 канала и наличие большего числа соединений будет снижать влияние отказа на снижение пропускной способности.

4. Обзор маршрутизации в ЦОД

В этом разделе приведен обзор трех основных вариантов организации сетей ЦОД - L2, L3 и гибрид L2/L3.

4.1. Вариант L2

Поначалу большинство ЦОД строились на основе протокола остовного дерева (STP), определенной стандартом [3][IEEE8021D-1990], для создания беспетлевой топологии с использованием одного из традиционных вариантов, описанных в параграфе 3.1. В то время многие коммутаторы ЦОД не поддерживали протоколы маршрутизации L3 или для поддержки этих протоколов требовались дополнительные лицензии, что влияло на выбор решения. Хотя с тех пор было внесено множество усовершенствований типа протокола RSTP[4] в последней версии стандарта [IEEE8021D-2004] и MST[5] из [IEEE8021Q], которые ускорили схождение и улучшили балансировку нагрузки в больших сетях, множество базовых аспектов протокола продолжает ограничивать его применимость в крупных ЦОД. Протокол STP и его новые варианты используют подход «активный-резервный» при выборе пути, а это усложняет развертывание масштабируемых по горизонтали топологий, описанных в параграфе 3.2. Кроме того, операторы накопили большой опыт обработки крупных отказов, связанных с некорректными кабельными соединиениями, ошибками настройки и программными проблемами на отдельном устройстве. Такие отказы обычно влияют на весь домен STP и устранение неполадок сложно из-за самой природы протокола. По этой причине и в связи с тем, что почти весь трафик ЦОД относится к протоколу IP, требующему наличия протокола маршрутизации L3 на границе сети для внешних соединений, организация сетей на основе STP обычно не позволяет выполнить все требования операторов крупных ЦОД. Различные усовершенствования протоколов объединения каналов типа [IEEE8023AD], обычно называемого M-LAG[5], позволили использовать решения L2 с путями «активный-активный» с сохранением протокола STP для предотвращения петель. Основным недостатком такого подхода является отсутствие возможности линейного масштабирования более чем вдвое для большинства реализаций, отсутствие стандартизованных реализаций и возможность синхронного отказа нескольких устройств.

[1]Link aggregation group — группа объединения каналов.
[2]Spanning Tree Protocol.
[3]Rapid Spanning Tree Protocol — ускоренный протокол STP.
[4]Multiple Spanning Tree Protocol — множество экземпляров остовного дерева STP.
[5]Multi-Chassis Link-Aggregation — объединение каналов от нескольких устройств (шасси).

Следует отметить появившуюся недавно возможность построения больших сетей L2 с горизонтальным масштабированием без STP на основе протокола TRILL[1] [RFC6325]. TRILL решает множество проблем STP в сетях крупных ЦОД, однако небольшое число реализаций и необходимость использовать специфическое оборудование с поддержкой этого протокола ограничивают применимость и повышают стоимость таких решений.

Кроме того ни базовая спецификация TRILL, ни подход M-LAG не решают полностью проблему единого домена широковещания, которая возникает для любого решения L2 на базе Ethernet. Были предложены расширения TRILL для решения этой проблемы прежде всего на основе подходов, рассмотренных в [RFC7067], но это еще сильнее ограничивает число доступных интеоперабельных реализаций. Поэтому у решений на базе TRILL возникают сложности с выполнением требований REQ2, REQ3 и REQ4.

4.2. Гибридный вариант L2/L3

Операторы стремились ограничить влияние отказов на уровне данных и строить крупномасштабные топологии на основе протоколов маршрутизации на уровне 1 или 2 с делением доменов L2 на множество более мелких субдоменов. Это позволило масштабировать ЦОД, но добавило сложности, связанные с управлением множеством сетевых протоколов. По перечисленным ниже причинам операторы сохранили L2 на уровне доступа (Tier 3) или на уровнях доступа и агрегирования (Tier 3 и Tier 2).

- Поддержка унаследованных приложений, которым может требоваться прямая смежность L2 или протоколы, не относящиеся к IP.
- Бесшовное перемещение виртуальных машин, которым требуется сохранение адреса IP при переносе на другой коммутатор Tier 3.
- Упрощение адресации IP (меньшее число подсетей IP) для ЦОД.
- Балансировка нагрузки приложений может требовать прямой доступности на уровне L2 для выполнения некоторых функций типа L2 DSR[2] (см. [L3DSR]).
- Сохраняющееся различие CAPEX для коммутаторов L2 и L3.

4.3. Вариант L3

Сетевые проекты, опускающие маршрутизацию IP на уровень (Tier) 3, также обрели популярность. Основным преимуществом таких решений является повышение уровня стабильности и расширяемости сети в результате ограничения размеров доменов широковещания L2. Обычно в таких случаях основным протоколом маршрутизации является тот или иной протокол внутреннего шлюза (IGP[3]) типа OSPF[4] [RFC2328]. ЦОДы растут в размерах и число серверов может достигать десятков тысяч, поэтому полностью маршрутизируемые решения становятся привлекательными.

Выбор варианта L3 существенно упрощает сеть, облегчая выполнение требований REQ1 и REQ2, поэтому такое решение широко распространено в сетях, где наличие больших областей смежности L2 и больших подсетей L3 не столь важно по сравнению со стабильностью и масштабируемостью сети. Поставщики приложений и сетевые операторы продолжают разрабатывать новые решения для удовлетворения требований, которое раньше вынуждали создавать большие домены L2, с помощью различных методов наложения и туннелирования.

5. Устройство протокола маршрутизации

В этом разделе приведена мотивация выбора EBGP в качестве единственного протокола маршрутизации для сетей ЦОД, работающих на основе L3 и топологии Clos. Даны также практические рекомендации по организации сетей на базе EBGP.

5.1. Выбор EBGP в качестве протокола маршрутизации

Выполнение требования REQ2 отдает преимущества выбору единственного протокола маршрутизации в результате упрощения и снижения числа зависимостей. Хотя в таких ситуациях чаще опираются на протоколы IGP, иногда дополняемые EBGP для граничных устройств WAN или IBGP[5] внутри сети, в этом документе предлагается использовать лишь EBGP.

Хотя протокол используется практически повсюду в междоменной маршрутизации Internet и поддерживается многими производителями и сообществами сервис-провайдеров, он по ряду перечисленных ниже причин (некоторые из них связаны между собой) обычно не используется в качестве основного протокола маршрутизации ЦОД.

- BGP считается протоколом для сетей WAN и редко рассматривается для сетей предприятий или ЦОД.
- Считается, что в BGP маршруты сходятся гораздо медленней, чем в IGP.
- Крупномасштабные системы BGP обычно используют IGP для определения BGP next-hop, поскольку не все узлы в топологии IBGP соединены между собой напрямую.
- BGP считается требующим трудоемкой настройки конфигурации и не поддерживающим автоматическое обнаружение соседей.

В этом документе рассматриваются некоторые из этих представлений в связи с предлагаемым решением и указываются преимущества в результате применения протокола.

- BGP менее сложен в части устройства протокола — внутренние структуры данных и машина состояний проще чем для большинства IGP на основе состояния канала типа OSPF. Например, вместо реализации отношений смежности, их поддержки и/или управления потоком данных BGP просто опирается на обеспечиваемые протоколом TCP возможности. Это позволяет выполнить требования REQ2 и REQ3.
- Информационные издержки BGP (лавинные рассылки) меньше по сравнению с IGP по состоянию каналов. Поскольку каждый маршрутизатор BGP рассчитывает и распространяет лишь выбранные лучшие пути, отказ в сети маскируется (обходится), как только узел BGP найдет альтернативный путь, который имеется в топологиях с высокой симметрией (типа Clos), выбранных для решения на основе лишь EBGP. Напротив,

[1]Transparent Interconnection of Lots of Links — прозрачное соединение большого числа каналов.
[2]Direct Server Return.
[3]Interior Gateway Protocol.
[4]Open Shortest Path First — сначала кратчайший путь.
[5]Internal BGP — внутренний BGP.

распространение события в IGP по состоянию канала занимает всю область, независимо от типа отказа. По этой причине BGP лучше подходит для выполнения требований REQ3 и REQ4. Следует также отметить, что все распространенные IGP по состоянию каналов периодически обновляют маршрутную информацию, тогда как BGP не ограничивает срок действия состояния маршрутизации, хотя это редко влияет на современные уровни управления в маршрутизаторах.

- BGP поддерживает сторонние (определяемые рекурсивно) next hop. Это позволяет манипулировать множеством путей без использования ECMP или на основе пересылки по определенным приложениями путям за счет организации партнерской сессии с приложением «контроллер» которое может вносить в систему маршрутную информацию для выполнения требования REQ5. Протокол OSPF обеспечивает похожую функциональность за счет использования концепций типа «адреса пересылки» (Forwarding Address), но это сложнее в реализации и значительно менее управляемо в плане области распространения информации.
- Используя четко определенную схему выделения ASN и стандартное определение петель в AS_PATH, можно управлять «охотой BGP на пути» (BGP path hunting [JAKMA2008]) и игнорировать сложные нежелательные пути. Пример работающей схемы выделения ASN представлен в параграфе 5.2. В IGP по состоянию каналов достижение этой цели требует поддержки множества экземпляров/топологий/процессов, что не всегда возможно в ЦОД и достаточно сложно для настройки и поиска неполадок. Применение традиционной схемы с одним доменом лавинной рассылки, которая используется в большинстве вариантов ЦОД, при некоторых отказах может приводить к нежелательным длинным путям, например, проходящим через множество устройств Tier 2.
- Конфигурация EBGP с минимальной политикой маршрутизации проще в плане поиска и устранения неполадок в сети. В большинстве реализаций просто увидеть содержимое BGP Loc-RIB и сравнить его с базой RIB[1] в маршрутизаторе. Также в большинстве реализаций оператор может видеть структуры Adj-RIB-In и Adj-RIB-Out каждого соседа BGP что позволяет легко сопоставить входящую и исходящую информацию NLRI на обеих сторонах сессии BGP. Таким образом, BGP соответствует требованию REQ3.

5.2. Конфигурация EBGP для топологии Clos

Топологии Clos с числом ступеней более 5 используются весьма редко по причине большого числа требуемых внутренних соединений. Поэтому в приведенных ниже примерах используется 5-ступенчатая топология Clos (без фальцовки).

5.2.1. Рекомендации по настройке EBGP и пример схемы ASN

На рисунке 4 представлен пример схемы выделения ASN. При выделении могут использоваться приведенные ниже рекомендации.

- Одноэтапные (single-hop) сессии EBGP организуются через прямые соединения «точка-точка», связывающие узлы сети, сессии с множеством интервалов (multi-hop) или loopback не применяются даже при наличии множества соединений между парой узлов.
- Используются частные ASN из диапазона 64512-65534 для предотвращения конфликтов ASN.
- Один номер ASN выделяется для всех устройств уровня 1 в топологии Clos.
- Каждому набору устройств уровня 2 в одном кластере выделяется уникальное значение ASN.
- Каждому устройству уровня 3 (например, ToR) выделяется уникальный номер ASN.

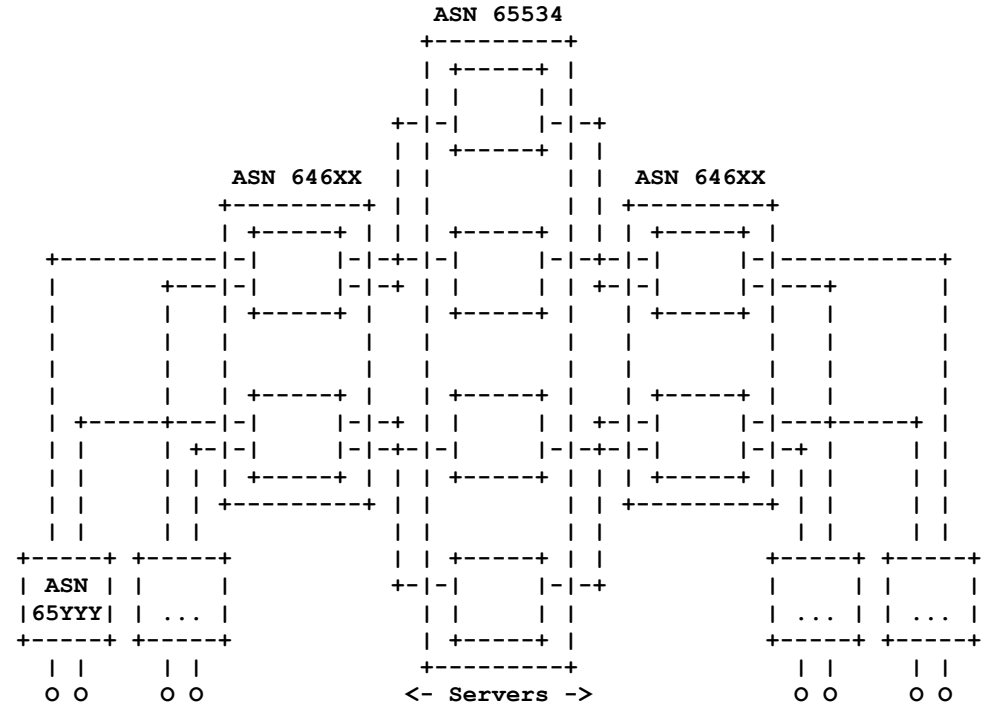

Рисунок 4. Схема BGP ASN для 5-ступенчатой топологии Clos.

5.2.2. Частное применение ASN

Исходный диапазон Private Use ASN [RFC6996] ограничен 1023 уникальными номерами ASN. Поскольку достаточно очевидно, что число устройств в сети превышает это значение, требуется решение проблемы. Одним из вариантов

[1]Routing Information Base — база маршрутной информации.

является повторное использование ASN, выделенных устройствам Tier 3 в разных кластерах. Например, частные ASN 65001, 65002 ... 65032 могут применяться в каждом отдельном кластере и назначаться устройствам Tier 3.

Чтобы предотвратить подавление маршрутов механизмом детектирования петель AS_PATH в BGP, восходящие сессии EBGP на устройствах Tier 3 должны настраиваться с функцией Allowas-in [ALLOWASIN], что позволяет устройству воспринимать в анонсах свой номер ASN. Хотя эта функция не стандартизована, она поддерживается оборудованием многих производителей. Добавление этой функции не повышает вероятность возникновения маршрутных петель, поскольку атрибуты AS_PATH будут добавляться маршрутизаторами на каждом уровне топологии, а размер AS_PATH является ранним способом «отбрасывания лишнего» (early tie breaker) в процессе выбора пути BGP. Дополнительная защита от петель обеспечивается устройствами уровня 1, которые не будут воспринимать маршруты со своим номером ASN. Устройства Tier 2 не имеют прямой связности между собой.

Другим решением этой проблемы является использование 4-октетных ASN [RFC6793], которые обеспечивают значительное расширение пространства Private Use ASN (см. [IANA.AS]). Применение 4-октетных ASN усложняет реализацию BGP и это следует сравнивать со сложностью повторного использования ASN при выборе решения для удовлетворения требований REQ3 и REQ4. Важно понимать, что 4-октетные номера AS поддерживаются не всеми производителями и это может ограничивать выбор оборудования для ЦОД. Если такие номера поддерживаются, следует обеспечить удаление частных ASN на внешних соединениях (параграф 5.2.4), если это требуется.

5.2.3. Анонсирование префиксов

Топология Clos включает множество соединений «точка-точка» и связанных с ними префиксов. Анонсирование всех этих маршрутов в BGP может перегружать таблицы FIB[1] в сетевых устройствах. Анонсирование каналов может также значительно нагружать уровень управления BGP при расчете путей. Ниже описаны два возможных решения задачи.

- Не анонсировать никакие соединения «точка-точка» в BGP. Поскольку в решениях на базе EBGP адрес next-hop меняется на каждом устройстве, удаленные сети автоматически становятся доступными через аносирующих партнеров EBGP и для таких префиксов не требуется доступность. Однако это может осложнять мониторинг, например, популярная утилита traceroute будет указывать такие адреса IP как недоступные.
- Анонсировать соединения «точка-точка» с обобщением на каждом устройстве. Для этого потребуется согласованная схема адресации, например выделение блока последовательных адресов IP на устройство Tier 1 и Tier 2 для использования на интерфейсах «точка-точка» с нижележащими уровнями (адреса для восходящих интерфейсов уровня 2 будут выделяться из блока адресов устройства Tier 1 и т. д.).

Серверные подсети на устройствах Tier 3 должны анонсироваться в BGP без обобщения маршрутов на устройствах уровня 2 и 1. Обобщение подсетей в топологии Clos приводит к возникновению маршрутных «черных дыр» при отказе одного канала (например, между устройствами Tier 2 и Tier 3), поэтому его следует избегать. Использование партнерских соединений внутри одного уровня для решения проблемы «черных дыр» путем создания «обходных путей» нежелательно по причине усложнения $O(N^2)$ соединений между партнерами и неоправданного расхода портов на устройствах. Альтернативой полносвязной сети соединений между партнерами является использование более простой топологии обхода, например, «кольца», описанного в [FB4POST], но это добавляет этапы пересылки и ограничено по пропускной способности. Для обеспечения работы BGP могут потребоваться специальные настройки. В параграфе 8.2 описан менее изощренный метод для ограниченного обобщения маршрутов в сетях Clos и связанные с ним компромиссы.

5.2.4. Внешние соединения

Можно использовать выделенный кластер (или кластеры) в топологии Clos для соединения с краевыми устройствами WAN[2] или маршрутизаторами WAN. Устройства уровня 3 в таком кластере заменяются маршрутизаторами WAN и партнерские отношения EBGP снова будут использоваться, хотя маршрутизаторы WAN будут скорей всего относиться к публичным ASN, если для сети требуется соединение с Internet. Устройства уровня 2 в таком выделенном кластере далее будут называться граничными маршрутизаторами (Border Routers). Эти устройства выполняют некоторые специальные функции, описанные ниже.

- Сокрытие топологической информации при анонсировании путей маршрутизаторам WAN, т. е. удаление частных ASN [RFC6996] из атрибутов AS_PATH. Это обычно делается для предотвращения конфликтов ASN между разными ЦОД, а также для выравнивания размера AS_PATH, передаваемых в WAN для WAN ECMP применительно к anycast-перфиксам, созданным в топологии. Для этого как правило применяется зависящая от реализации функция BGP, которую обычно называют «удаление частных AS» (Remove Private AS). В зависимости от реализации этой функции следует вырезать непрерывные цепочки частных ASN из атрибутов AS_PATH до анонсирования пути соседу. Это предполагает, что все ASN, используемые для внутренней нумерации в ЦОД, относятся к диапазонам частных номеров. Процесс вырезания частных ASN пока не стандартизован (см. [REMOVAL]). Однако большинство реализаций на практике следуют рекомендациям [VENDOR-REMOVE-PRIVATE-AS], что вполне достаточно для целей этого документа.
- Создание принятого по умолчанию маршрута для устройств ЦОД. Это единственное место, где может быть создан такой маршрут, поскольку объединение маршрутов является рискованной задачей для немодифицированной топологии Clos. В дополнение к этому граничные маршрутизаторы могут просто ретранслировать полученный от маршрутизаторов WAN маршрут, принятый по умолчанию. Анонсирование принятого по умолчанию маршрута от граничных маршрутизаторов требует соединения всех этих маршрутизаторов с восходящими маршрутизаторами WAN для обеспечения устойчивости к отказам отдельных каналов, создающим «черные дыры» для трафика. Для предотвращения «черных дыр» в ситуациях, когда все сессии EBGP с маршрутизаторами WAN рвутся на данном устройстве одновременно, желательно заново анонсировать используемый по умолчанию маршрут вместо создания нового маршрута, принятого по умолчанию, с использованием усложненных схем условного расчета маршрута, предлагаемых отдельными реализациями [CONDITIONALROUTE].

5.2.5. Обобщение маршрутов на границе

Зачастую желательно объединить информацию о достижимости сетей до ее анонсирования в WAN по причине большого числа префиксов IP, исходящих из ЦОД при использовании полностью маршрутизируемого решения.

[1]Forwarding Information Base — база информации о пересылке.
[2]Wide Area Network — распределенная (глобальная) сеть.

Например, сеть с 2000 устройств Tier 3 будет давать не менее 2000 анонсов серверных подсетей в BGP вместе с префиксами инфраструктуры. Однако, как указано в параграфе 5.2.3, предлагаемое решение не позволяет обобщать маршруты по причине нехватки партнерских соединений внутри каждого уровня.

Тем не менее, это ограничение можно обойти для граничных маршрутизаторов за счет использования для таких устройств иной модели подключения. Здесь возможны два варианта, описанных ниже.

- Соединить граничные маршрутизаторы между собой с использованием полносвязной сети физических каналов или другой топологии типа кольца онцентратора. BGP на всех граничных маршрутизаторах настраивается на обмен информацией о доступности, например, путем организации полносвязной системы сессий IBGP. Соединения между партнерами должны иметь пропускную способность, достаточную для передачи всего трафика в случае отказа одного устройства или канала в полносвязной сети, соединяющей граничные маршрутизаторы.
- Устройства Tier 1 могут иметь дополнительные физические каналы в направлении граничных маршрутизаторов (которые являются устройствами Tier 2 с точки зрения уровня 1). В частности, если нужна защита от отказа отдельного канала или узла, каждое устройство Tier 1 следует соединить по крайней мере с двумя граничными маршрутизаторами. Это потребует дополнительных портов в устройствах уровня 1 и граничных маршрутизаторах, в результате чего нарушится унификация устройств, поскольку они будут отличаться от других устройств в Clos. Это также уменьшит число портов, доступных для «обычных» коммутаторов Tier 2 и, следовательно, число кластеров, которые можно будет соединить через уровень 1.

При реализации любого из этих вариантов на граничных маршрутизаторах можно выполнить обобщение маршрутов в направлении WAN без риска возникновения «черной дыры» при отказе одного соединения. Оба варианта будут приводить к неоднородности топологии, поскольку некоторые устройства получат дополнительные соединения.

6. Вопросы ECMP

В этом разделе рассматривается функциональность множества равноценных путей (ECMP) для топологии Clos и некоторые специальные требования.

6.1. Базовый механизм ECMP

ECMP представляет собой фундаментальный механизм распределения нагрузки в топологии Clos. Эффективно каждое устройство нижнего уровня будет использовать все подключенные к нему напрямую устройства вышележащего уровня для распределения трафика, направленного в один префикс IP. Число путей ECMP между любыми двумя устройствами Tier 3 в топологии Clos равно числу устройств на средней ступени (Tier 1). Например, рисунок 5 показывает топологию, где устройство A на уровне 3 имеет 4 пути доступа к серверам X и Y через устройства B и C на уровне 2 и устройства 1, 2, 3 и 4 на уровне 1.

Требование ECMP предполагает, что реализация BGP должна поддерживать множество разветвляющихся путей на выходе вплоть до максимального числа устройств, напрямую подключенных к любому порту в восходящем или нисходящем направлении. Обычно в такой топологии это число не превышает половины от числа портов устройства. Например, разветвление ECMP на 32 направления будет требовать создания сети Clos с использованием 64-портовых устройств. Граничным маршрутизаторам могут потребоваться дополнительные разветвления для подключения множества устройств Tier 1, если реализовано описанное в параграфе 5.2.5 обобщение маршрутов на граничных маршрутизаторах. Если оборудование не поддерживает требуемой широты ECMP, можно использовать группировку соединений на канальном уровне (агрегирование каналов L2) для создания «иерархического» ECMP (комбинация L3 ECMP и L2 ECMP) для компенсации ограниченного ветвления. Однако это повышает риск «поляризации» потоков, поскольку снижается энтропия на второй ступени ECMP.

Большинство реализаций BGP объявляют равноценность путей с точки зрения ECMP, если они соответствуют требованиям параграфа 9.1.2.2 [RFC4271] до п. (e) включительно. В предлагаемом решении нет нижележащего IGP, поэтому стоимость IGP принимается равной 0 или в противном случае может применяться одинаковое значение для всех путей и правил, требуемое для выравнивания атрибутов BGP (таких, как MULTI_EXIT_DISC (MED) и код

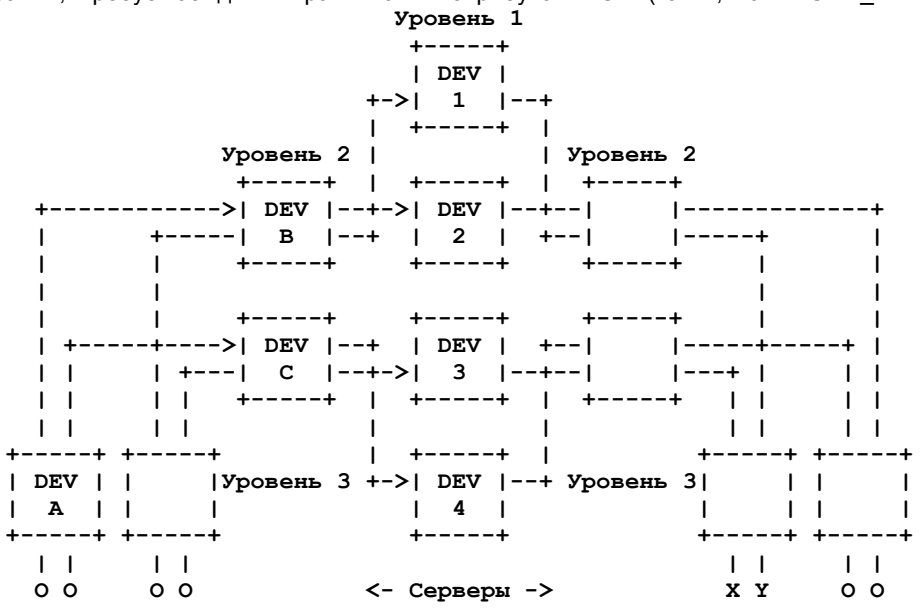

Рисунок 5. Дерево разветвления (Fan-Out Tree) ECMP от A к X и Y.

источника), которые могут меняться у разных производителей. В силу исторических причин лучше не использовать 0 в качестве уравнивающего значения MED (описание этого и другая полезная информация BGP приведены в [RFC4277]). Маршрутные петли маловероятны вследствие использования процесса выбора лучшего пути BGP (который предпочитает наименьший размер AS_PATH), а более длинные пути через устройства Tier 1 (которые не разрешают в пути свой ASN) невозможны.

6.2. BGP ECMP через множество AS

Для распределения нагрузки приложений желательно анонсировать один и тот же префикс от множества устройств Tier 3. С точки зрения других устройств такой префикс будет иметь пути BGP с разными атрибутами AS_PATH одинакового размера. Поэтому реализации BGP должны поддерживать распределение нагрузки между такими путями. Это свойство иногда называют multipath relax или multipath multiple-AS и оно эффективно обеспечивает ECMP через разные соседние ASN, если другие атрибуты совпадают, как описанно в предыдущем параграфе.

6.3. Взвешенный ECMP

Может оказаться желательной реализация устройством «взвешенного» ECMP для того, чтобы передавать больше трафика через отдельные ветви ECMP. Это может быть полезно при компенсации отказов в сети или для передачи большего объема через пути с большей пропускной способностью. Префиксы, которым нужен взвешенный ECMP, будут вноситься с использованием удаленного узла BGP (центральный агент) через multi-hop-сессию, как будет описано в параграфе 8.1. Если реализация поддерживает это, распределение весов между путями BGP может быть передано с использованием метода, описанного в [LINK].

6.4. Согласованное хэширование

Часто желательно, чтобы функция хэширования для ECMP была согласованной (см. [CONS-HASH]) для минимизации воздействия на поток изменения близости next-hop при добавлении или удалении следующего интервала в группе ECMP. Это может применяться для устройств, работающих в качестве балансировщиков, отображающих потоки в направлении множества получателей, - в этом случае потеря или добавление получателя не окажет негативного влияния на существующие потоки. Одна из конкретных рекомендаций по реализации согласованного хэширования представлена в [RFC2992], хотя возможны и другие варианты. Эта функциональность может естественным путем комбинироваться со взвешенным ECMP, при этом влияние смены следующего интервала будет пропорционально весу данного next-hop. Обратной стороной согласованного хэширования является рост использования аппаратных ресурсов, поскольку для ее реализации обычно нужно больше ресурсов (например, пространства TCAM[1]).

7. Схождение маршрутов

В этом разделе рассматриваются свойства схождения маршрутов для предложенного решения. Отмечено, что достижимо схождение за доли секунды, если реализация поддерживает быструю деактивацию сессий EBGP и своевременно обновляет RIB и FIB при отказе связанного канала.

7.1. Время обнаружения отказов

BGP обычно опирается на IGP для обхода отказавших каналов и узлов внутри AS и реализует основанный на опросе или событиях механизм получения обновлений о смене состояний IGP. Предложенный вариант маршрутизации не использует IGP, поэтому для обнаружения отказов остается тайм-аутBGP keep-alive (или иной механизм keep-alive) или триггеры отказа каналов.

Опора исключительно на пакеты BGP keep-alive может привести к значительной задержке схождения до многих секунд (во многих реализациях BGP минимальное значение таймера удержания BGP составляет 3 секунды). Однако многие реализации BGP могут закрывать локальные партнерские сессии EBGP в ответ на событие link down (канал отключен) на исходящем интерфейсе партнерской сессии BGP. Это свойство иногда называют fast fallover (быстрое «падение»). Поскольку каналы в современных ЦОД представляют собой преимущественно оптические соединения «точка-точка», отказ физического интерфейса часто обнаруживается за миллисекунды и инициирует пересчет маршрутов BGP.

Каналы Ethernet могут поддерживать стандарты сигнализации и детектирования отказов типа CFM[2], описанного в [IEEE8021Q], это делает детектирование отказов более надежным. Кроме того, некоторые платформы могут поддерживать BFD[3] [RFC5880], позволяющее обнаруживать отказы каналов и сообщать о них процессу BGP за доли секунды. Однако использование любого из этих методов предъявляет дополнительные требования к фирменным программам и, возможно, оборудованию, а также может противоречить REQ1. До недавнего времени ([RFC7130]) метод BFD также не поддерживал детектирования отказов одного канала в LAG, что ограничивало его применимость.

7.2. Время распространения событий

В предлагаемом решении следует учитывать влияние BGP MinRouteAdvertisementIntervalTimer (таймер MRAI), как указано в параграфе 9.2.1.1 [RFC4271]. По стандарту реализации BGP должны разделять последовательные сообщения BGP UPDATE интервалом не менее MRAI, который часто задается в конфигурации. На начальные сообщения BGP UPDATE после события, связанного с отзывом маршрута, этот таймер обычно не влияет. Таймер MRAI может существенно задерживать схождение когда узел BGP «ждет» получения от партнеров нового пути и не имеет локальной резервной копии информации о путях.

В топологии Clos каждый узел EBGP обычно имеет один путь (устройства Tier 2 не воспринимают пути от других устройств Tier 2 в том же кластере, поскольку у них тот же номер ASN) или N путей для одного префикса, где N имеет достаточно большое значение, например 32 (разветвление ECMP на следующий уровень). Поэтому при отказе канала к другому устройству, от которого получен путь, резервного пути не будет совсем (например, с точки зрения коммутатора Tier 2, теряющего соединение с устройством Tier 3) либо он будет легко доступен в BGP Loc-RIB (например, с точки зрения устройства Tier 2, теряющего соединение с коммутатором Tier 1). В первом случае анонс отзыва BGP будет распространяться без задержки и вызовет пересчет маршрутов на затронутых устройствах. Во втором случае лучший путь будет оцениваться заново и локальная группа ECMP, соответствующая новому значению next-hop, будет изменена. Если путь BGP был ранее лучшим, «неявный отзыв» будет передаваться в сообщении BGP UPDATE, как описано в варианте b параграфа 3.1 [RFC4271] по причине изменения атрибута BGP AS_PATH.

7.3. Влияние разветвлений топологии Clos

Топология Clos имеет большие разветвления, которые могут в некоторых случаях влиять на схождение Up→Down, как описано в этом параграфе. В случае отказа канала между устройствами уровней 3 и 2, устройство Tier 2 будет передавать сообщения BGP UPDATE всем восходящим устройствам 1, отзывая затронутые префиксы. Устройства Tier 1, в свою очередь, будут ретранслировать эти сообщения своим нисходящим устройствам Tier 2 (исключая

[1]Ternary Content-Addressable Memory — троичная ассоциативная память.
[2]Connectivity Fault Management — контроль отказов связности.
[3]Bidirectional Forwarding Detection — двухстороннее детектирование пересылки.

инициатора). Устройствам Tier 2 кроме инициатора UPDATE следует дождаться пока **все** восходящие устройства Tier 1 передадут сообщение UPDATE и только после этого удалять префиксы и передавать соответствующие сообщения UPDATE в нисходящем направлении подключенным устройствам Tier 3. Если исходное устройство Tier 2 или ретранслирующие устройства Tier 1 вносят ту или иную задержку в распространение сообщений UPDATE, в результате может возникнуть «рассеяние» UPDATE, которое может длиться многие секунды. Для предотвращения этого реализации BGP должны поддерживать «группы обновления» (update groups). Группа обновления определяется как набор соседей с общей выходной политикой — локальный узел будет синхронно передавать обновления BGP членам группы.

Влияние такого «рассеяния» возрастает с ростом разветвления топологии и может расти также во время схождения. У некоторых операторов может возникнуть соблазн использовать подавление маршрутных осцилляций (route flap dampening), которое производители включают для снижения воздействия на уровень управления быстрых осцилляций префиксов. Однако в результате известных проблем с ложными срабатываниями в таких реализациях (особенно при таких «рассеянных» событиях), не рекомендуется пользоваться этой функцией в предложенном решении. Дополнительную информацию о подавлении маршрутных осцилляций и связанных с ним изменением в реализациях можно найти в [RFC7196].

7.4. Область влияния отказа

Схождение в результате отказа считается завершенным, когда все устройства в зоне влияния отказа уведомлены о событии, пересчитали свои RIB и обновили FIB. Большоя зона охвата при отказе обычно замедляет схождение, поскольку требуется уведомить больше устройств и в результате стабильность сети снижается. В этом параграфе рассматриваются преимущества BGP по сравнению с протоколами маршрутизации по состояниям каналов в части сужения зоны охвата при отказах в топологии Clos.

BGP ведет себя подобно протоколам distance-vector в том смысле, что соседям передается только лучший путь с точки зрения локального маршрутизатора. Поэтому часть отказов маскируется, если локальный узел незамедлительно нашел резервный путь и больше не передал никаких обновлений. Отметим, что в худшем случае все устройства в ЦОД будут отзывать префикс полностью или обновят группы ECMP в своих FIB. Однако многие отказы не связаны со столь широким охватом. Есть два основных типа отказов, для которых зона влияния может быть сужена.

- Отказ канала между устройствами Tier 2 и Tier 1. В этом случае устройство уровня 2 будет обновлять затронутые группы ECMP, удаляя отказавший канал. Здесь не нужно передавать новую информацию нисходящим устройствам Tier 3, пока путь не был выбран в качестве лучшего процессом BGP — в этом случае требуется передать лишь «неявный отзыв» и это не должно влиять на пересылку. Затронутое устройство Tier 1 потеряет лишь путь к отдельному кластеру и будет отзывать соответствующие префиксы. Этот процесс отзыва префиксов будет влиять лишь на устройства Tier 2, напрямую подключенные к затронутому устройству уровня 1. Устройства Tier 2, получившие сообщение BGP UPDATE с отзывом префиксов, будут просто обновлять свои группы ECMP. Устройства уровня 3 не будут вовлечены в процесс повторного схождения.
- Отказ устройства Tier 1. В этом случае все устройства уровня 2, напрямую подключенные к отказавшему узлу, будут обновлять свои группы ECMP для всех префиксов IP из нелокального кластера. Устройство Tier 3 и в этом случае не будут вовлечены в процесс повторного схождения, но могут получать «неявные отзывы», как указано выше.

Даже при таких отказах, которые требуют перепрограммировать множество IP префиксов в FIB, следует отметить, что эти префиксы будут относиться к одной группе ECMP на устройстве Tier 2. Поэтому для реализации с иерархической базой FIB потребуется единственное изменение в FIB. Иерархической здесь считается FIB, в которой информация next-hop хранится отдельно от таблицы префиксов и в последней содержатся лишь указатели на соответствующие данные пересылки. Описание иерархий FIB и быстрого схождения приведено в [BGP-PIC].

Хотя BGP обеспечивает в некоторых случаях уменьшение зоны влияния отказа, дополнительное уменьшение этой зоны с помощью обобщения не всегда доступно в предложенном решении, поскольку обобщение маршрутов может создавать маршрутные «черные дыры», как было отмечено выше. Поэтому худшим вариантом зоны влияния отказа является уровень управления сети в целом, например в случае отказа на канале между устройствами уровня 2 и 3. Число затронутых отказом префиксов здесь будет много меньше, чем при отказе на верхних уровнях топологии Clos. Столь обширная зона влияния отказа является не результатом выбора решения на основе EBGP, а скорее свойством топологии Clos.

7.5. Микропетли в маршрутизации

Когда устройство нисходящего потока (например, Tier 2) теряет все пути для префикса, оно обычно имеет принятый по умолчанию маршрут, указывающий на восходящее устройство (например, Tier 1). В результате возможны ситуации, когда коммутатор уровня 2 теряет префикс, но коммутатор уровня 1 имеет путь, указывающий на коммутатор Tier 2, что приводит к возникновению временной микропетли, поскольку коммутатор 1 будет передавать пакеты затронутого отказом префикса устройству Tier 2, а оно будет возвращать их по заданному по умолчанию маршруту. Такая микропетля будет существовать, пока восходящее устройство полностью не обновит свои таблицы пересылки.

Для снижения влияния таких микропетель на коммутаторах уровней 2 и 1 можно задать статические маршруты discard или null, которые будут более конкретными по сравнению с принятым по умолчанию маршрутом для утраченного префикса во время схождения. Для коммутаторов Tier 2 отбрасывающий (discard) маршрут должен быть обобщением, включающим все серверные подсети нижележащих устройств Tier 3. Для устройств уровня 1 маршрут отбрасывания следует делать обобщением, включающим подсети серверов, выделенные для всего ЦОД. Эти discard-маршруты будут иметь предпочтение лишь во время схождения, пока устройство ищет более конкретные префиксы через новый путь.

8. Дополнительные варианты

8.1. Вставка стороннего маршрута

BGP позволяет «стороннему» партнеру (т. е. подключенному напрямую узла BGP) внедрять маршруты в любой точке сетевой топологии во исполнение требования REQ5. Это можно реализовать путем организации multi-hop-сессий BGP с некоторыми или даже всеми устройствами в рамках топологии. Кроме того широкое распространение путей BGP [RFC6774] можно использовать для вставки множества BGP next-hop для одного префикса с целью упрощения балансировки трафика или использовать свойство BGP ADD-PATH [RFC7911], если реализация поддерживает его. К

сожалению во многих реализациях обнаружено, что ADD-PATH поддерживает IBGP только в тех случаях, для которых это свойство было изначально оптимизировано. Это ограничивает партнерство third-party только устройствами IBGP.

Для реализации внедрения маршрутов в предлагаемом решении сторонний узел BGP может быть партнером коммутаторов уровня 3 и 1, внедряя один и тот же префикс, но используя специальный набор BGP next-hop для устройств Tier 1. Предполагается что эти next-hop рекурсивно преобразуются через BGP и могут быть, например, IP-адресами устройств Tier 3. Получаемые в результате таблицы пересылки могут обеспечивать желаемое распределение трафика между разными кластерами.

8.2. Обобщение маршрутов в топологии Clos

Как было отмечено выше, обобщение маршрутов в предложенной топологии Clos невозможно, поскольку оно делает сеть уязвимой к маршрутным «черным дырам» при отказах отдельных каналов. Основная проблема заключается в ограниченном числе избыточных путей между элементами сети (например, существует единственный путь между любой парой устройств уровней 1 и 3). Однако некоторые операторы могут счесть агрегирование маршрутов желательным для повышения стабильности уровня управления.

Если планируется какой-либо метод обобщения в такой топологии, следует промоделировать поведение маршрутизации и возможные «черные дыры» не только для отказа одного или нескольких каналов, но и для отказа оптических путей или доменов при включении в топологию физически разнесенных частей. Простое моделирование можно выполнить путем проверки доступности устройств при обобщении маршрутов в случае обрыва канала или пути между множеством устройств в каждом уровне, а также маршрутизаторов WAN при наличии внешних соединений.

Обобщение маршрутов становится возможным при небольшом изменении топологии сети, хотя это приведет к уменьшению общего размера сети, а также перегрузкам при определенных отказах. Этот подход очень похож на описанный выше метод, который позволяет граничным маршрутизаторам обобщать все адресное пространство ЦОД.

8.2.1. «Свертывание» устройств уровня 1

Для добавления путей между устройствами Tier 1 и Tier 3 устройства уровня 2 группируются попарно и пары подключаются к одной группе устройств Tier 1. Это логически эквивалентно «свертыванию» (collapsing) устройств Tier 1 в группу половинного размера с объединением каналов на «свернутых» устройствах. Результат такого свертывания показан на рисунке 6. Например, в этой топологии устройства DEV C и DEV D подключены к одному набору устройств Tier 1 (DEV 1 и DEV 2), тогда как до этого они были подключены к разным группам устройств Tier 1.

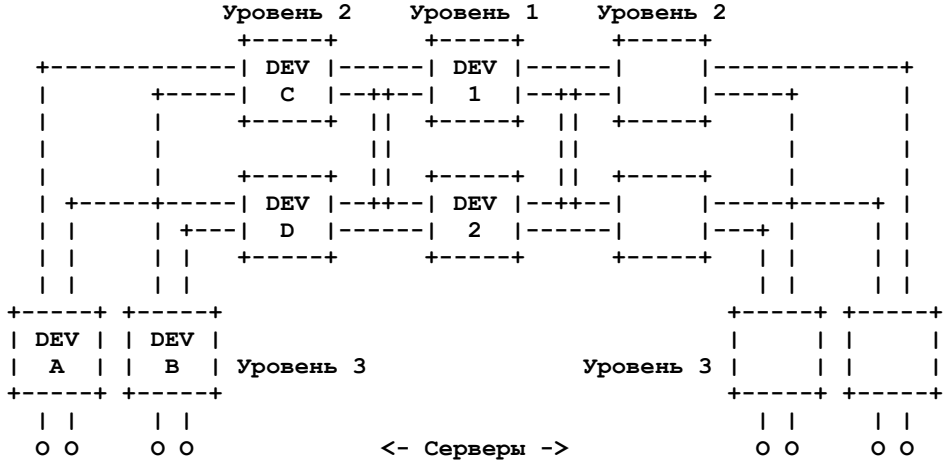

Рисунок 6. 5-ступенчатая топология Clos.

Для такого решения устройства Tier 2 можно настроить на анонсирование устройствам уровня 3 лишь принятого по умолчанию маршрута. При отказе канала между устройствами Tier 2 и Tier 3 трафик будет перемаршрутизирован через второй доступный путь, известный коммутатору Tier 2. Пока еще невозможно анонсировать сводный маршрут, включающий префиксы для одного кластера от устройств Tier 2, поскольку каждое из них имеет лишь один путь вниз к этому префиксу. Для решения задачи потребуются двудомные серверы. Отметим также, что это решение устойчиво лишь к отказам одного канала. При отказе двух каналов, изолирующем устройство Tier 2 от всех путей к конкретному устройству Tier 3, возникнет «черная дыра» в маршрутизации.

Результатом предложенного изменения топологии будет снижение числа портов на устройствах Tier 1. Это ограничит максимальное число подключенных устройств Tier 2 и в результате будет ограничивать общий размер сети ЦОД. Для более крупной сети потребуются устройства Tier 1 с более высокой плотностью портов.

Другой проблемой является перебалансировка трафика при отказах каналов. Поскольку имеется два пути от Tier 1 к Tier 3, отказ канала между коммутаторами уровней 1 и 2 приведе к тому, что весь трафик отказавшего канала будет перенесен на оставшийся путь. Это приведет к удвоению загрузки оставшегося канала.

8.2.2. Простое виртуальное агрегирование

Совершенно иной путь к обобщению маршрутов возможен в том случае, когда основной целью является снижение размера FIB, что позволит уровню управления распространять всю маршрутную информацию. Во-первых, легко заметить, что множество префиксов, часть из которых менее конкретна, использует общий набор next-hop (одна группа ECMP). Например, с точки зрения устройств Tier 3 все маршруты, полученные от восходящих устройств Tier 2 (включая принятый по умолчанию маршрут), будут использовать один набор BGP next-hop при условии отсутствия отказов в сети. Это позволяет использовать методы, похожие на описанный в [RFC6769] и устанавливать в FIB только наименее конкретный маршрут, игнорируя более конкретные, если они используют тот же набор next-hop. Например, при нормальной работе сети в FIB нужно включать лишь принятый по умолчанию маршрут.

Кроме того, если устройства Tier 2 настроены с обобщенными префиксами, включающими префиксы всех подключенных к ним устройств Tier 3, такая же логика применима для устройств Tier 1 путем включения коммутаторов Tier 2/Tier 3 в разные кластеры. Эти обобщенные маршруты позволят отдавать более конкретные префиксы

устройствам Tier 1, что позволит детектировать рассогласование наборов next-hop при отказе отдельного канала и менять набор next-hop для конкретного префикса.

Повторим, что этот метод не снижает число состояний на уровне управления (т. е. размер BGP UPDATEs, BGP Loc-RIB), но позволяет более эффективно использовать FIB за счет нахождения более конкретных префиксов и использования их набора next-hop для менее конкретного префикса.

8.3. Маскирование сообщений ICMP Unreachable

В этом параграфе рассматриваются некоторые эксплуатационные аспекты отказа от анонсирования префиксов каналов «точка-точка» в BGP, отмеченные в параграфе 5.2.3. Влияние этого решения может проявиться при использовании популярной утилиты traceroute. В частности, IP-адреса, отображаемые программой, будут включать и адреса каналов «точка-точка», которые при таком решении недоступны для подключения. Это несколько усложняет поиск неполадок.

Одним из способов решения этой проблемы является использование подсистемы DNS для создания «реверсных» записей для IP-адресов таких каналов «точка-точка» с именами как для адреса петлевого интерфейса (loopback). Связность в этом случает может быть обеспечена путем преобразования «основного» IP-адреса устройства, например, его интерфейса Loopback, который всегда анонсируется в BGP. Однако это создает зависимость от подсистемы DNS, которая может оказаться недоступно во время отказов.

Другим вариантом является маскирование устройством адресов IP, т. е. замена IP-адреса отправителя в соответствующих сообщениях ICMP «основным» адресом устройства. В частности, такая замена требуется в сообщениях ICMP Destination Unreachable Message (тип 3) с кодом 3 (порт недоступен) и ICMP Time Exceeded (тип 11) с кодом 0 для корректной работы программы traceroute. При таком маскировании пробные пакеты traceroute, передаваемые устройствам, всегда будут приводит к отправке откликов с «основного» IP-адреса устройства, что позволит оператору увидеть «доступный» IP-адрес устройства. Недостатком этого подхода является сокрытие «точки входа» в устройство. Если устройство поддерживает [RFC5837], оно сможет предоставить информацию о входном интерфейсе даже при возврате пакетов с «основного» адреса IP.

9. Вопросы безопасности

Рассмотренное здесь решение не создает новых проблем безопасности. Общие вопросы безопасности BGP рассмотрены в [RFC4271] и [RFC4272]. Поскольку домен ЦОД относится к одному оператору, в этом документе предполагается наличие фильтрации для предотвращения атак на сессии BGP извне. Для большинства реализаций это будет более подходящим вариантом, чем использование управления ключами для TCP MD5, описанное в [RFC2385], или реализациями опции аутентификации TCP [RFC5925]. Можно также использовать обобщенный механизм защиты с помощью TTL [RFC5082] для предотвращения подмены сессий BGP.

10. Литература

10.1. Нормативные документы

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC6996] Mitchell, J., "Autonomous System (AS) Reservation for Private Use", BCP 6, RFC 6996, DOI 10.17487/RFC6996, July 2013, <http://www.rfc-editor.org/info/rfc6996>.

10.2. Дополнительная литература

[ALFARES2008] Al-Fares, M., Loukissas, A., and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture", DOI 10.1145/1402958.1402967, August 2008, <http://dl.acm.org/citation.cfm?id=1402967>.

[ALLOWASIN] Cisco Systems, "Allowas-in Feature in BGP Configuration Example", February 2015, <http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/112236-allowas-in-bgp-config-example.html>.

[BGP-PIC] Bashandy, A., Ed., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", Work in Progress, draft-ietf-rtgwg-bgp-pic-02, August 2016.

[CLOS1953] Clos, C., "A Study of Non-Blocking Switching Networks", The Bell System Technical Journal, Vol. 32(2), DOI 10.1002/j.1538-7305.1953.tb01433.x, March 1953.

[CONDITIONALROUTE] Cisco Systems, "Configuring and Verifying the BGP Conditional Advertisement Feature", August 2005, <http://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/16137-cond-adv.html>.

[CONS-HASH] Wikipedia, "Consistent Hashing", July 2016, <https://en.wikipedia.org/w/index.php?title=Consistent_hashing&oldid=728825684>.

[FB4POST] Farrington, N. and A. Andreyev, "Facebook's Data Center Network Architecture", May 2013, <http://nathanfarrington.com/papers/facebook-oic13.pdf>.

[GREENBERG2009] Greenberg, A., Hamilton, J., and D. Maltz, "The Cost of a Cloud: Research Problems in Data Center Networks", DOI 10.1145/1496091.1496103, January 2009, <http://dl.acm.org/citation.cfm?id=1496103>.

[HADOOP] Apache, "Apache Hadoop", April 2016, <https://hadoop.apache.org/>.

[IANA.AS] IANA, "Autonomous System (AS) Numbers", <http://www.iana.org/assignments/as-numbers>.

[IEEE8021D-1990] IEEE, "IEEE Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges", IEEE Std 802.1D, DOI 10.1109/IEEESTD.1991.101050, 1991, <http://ieeexplore.ieee.org/servlet/opac?punumber=2255>.

[IEEE8021D-2004] IEEE, "IEEE Standard for Local and Metropolitan Area Networks: Media Access Control (MAC) Bridges", IEEE Std 802.1D, DOI 10.1109/IEEESTD.2004.94569, June 2004, <http://ieeexplore.ieee.org/servlet/opac?punumber=9155>.

[IEEE8021Q] IEEE, "IEEE Standard for Local and Metropolitan Area Networks: Bridges and Bridged Networks", IEEE Std 802.1Q, DOI 10.1109/IEEESTD.2014.6991462, <http://ieeexplore.ieee.org/servlet/opac?punumber=6991460>.

[IEEE8023AD] IEEE, "Amendment to Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications - Aggregation of Multiple Link Segments", IEEE Std 802.3ad, DOI 10.1109/IEEESTD.2000.91610, October 2000, <http://ieeexplore.ieee.org/servlet/opac?punumber=6867>.

[INTERCON] Dally, W. and B. Towles, "Principles and Practices of Interconnection Networks", ISBN 978-0122007514, January 2004, <http://dl.acm.org/citation.cfm?id=995703>.

[JAKMA2008] Jakma, P., "BGP Path Hunting", 2008, <https://blogs.oracle.com/paulj/entry/bgp_path_hunting>.

[L3DSR] Schaumann, J., "L3DSR - Overcoming Layer 2 Limitations of Direct Server Return Load Balancing", 2011, <https://www.nanog.org/meetings/nanog51/presentations/Monday/NANOG51.Talk45.nanog51-Schaumann.pdf>.

[LINK] Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", Work in Progress, draft-ietf-idr-link-bandwidth-06, January 2013.

[REMOVAL] Mitchell, J., Rao, D., and R. Raszuk, "Private Autonomous System (AS) Removal Requirements", Work in Progress, draft-mitchell-grow-remove-private-as-04, April 2015.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <http://www.rfc-editor.org/info/rfc2328>.

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, DOI 10.17487/RFC2385, August 1998, <http://www.rfc-editor.org/info/rfc2385>.

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <http://www.rfc-editor.org/info/rfc2992>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <http://www.rfc-editor.org/info/rfc4272>.

[RFC4277] McPherson, D. and K. Patel, "Experience with the BGP-4 Protocol", RFC 4277, DOI 10.17487/RFC4277, January 2006, <http://www.rfc-editor.org/info/rfc4277>.

[RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, DOI 10.17487/RFC4786, December 2006, <http://www.rfc-editor.org/info/rfc4786>.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <http://www.rfc-editor.org/info/rfc5082>.

[RFC5837] Atlas, A., Ed., Bonica, R., Ed., Pignataro, C., Ed., Shen, N., and JR. Rivers, "Extending ICMP for Interface and Next-Hop Identification", RFC 5837, DOI 10.17487/RFC5837, April 2010, <http://www.rfc-editor.org/info/rfc5837>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <http://www.rfc-editor.org/info/rfc5880>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <http://www.rfc-editor.org/info/rfc5925>.

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, DOI 10.17487/RFC6325, July 2011, <http://www.rfc-editor.org/info/rfc6325>.

[RFC6769] Raszuk, R., Heitz, J., Lo, A., Zhang, L., and X. Xu, "Simple Virtual Aggregation (S-VA)", RFC 6769, DOI 10.17487/RFC6769, October 2012, <http://www.rfc-editor.org/info/rfc6769>.

[RFC6774] Raszuk, R., Ed., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of Diverse BGP Paths", RFC 6774, DOI 10.17487/RFC6774, November 2012, <http://www.rfc-editor.org/info/rfc6774>.

[RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <http://www.rfc-editor.org/info/rfc6793>.

[RFC7067] Dunbar, L., Eastlake 3rd, D., Perlman, R., and I. Gashinsky, "Directory Assistance Problem and High-Level Design Proposal", RFC 7067, DOI 10.17487/RFC7067, November 2013, <http://www.rfc-editor.org/info/rfc7067>.

[RFC7130] Bhatia, M., Ed., Chen, M., Ed., Boutros, S., Ed., Binderberger, M., Ed., and J. Haas, Ed., "Bidirectional Forwarding Detection (BFD) on Link Aggregation Group (LAG) Interfaces", RFC 7130, DOI 10.17487/RFC7130, February 2014, <http://www.rfc-editor.org/info/rfc7130>.

[RFC7196] Pelsser, C., Bush, R., Patel, K., Mohapatra, P., and O. Maennel, "Making Route Flap Damping Usable", RFC 7196, DOI 10.17487/RFC7196, May 2014, <http://www.rfc-editor.org/info/rfc7196>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <http://www.rfc-editor.org/info/rfc7911>.

[VENDOR-REMOVE-PRIVATE-AS] Cisco Systems, "Removing Private Autonomous System Numbers in BGP", August 2005, <http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080093f27.shtml>.

Благодарности

Эта публикация обобщает работу многих людей, принимавших участие в разработке, тестировании и развертывании предложенного варианта устройства сетей, среди которых были George Chen, Parantap Lahiri, Dave Maltz, Edet Nkposong, Robert Toomey и Lihua Yuan. Авторы благодарны Linda Dunbar, Anoop Ghanwani, Susan Hares, Danny McPherson, Robert Raszuk и Russ White за рецензирование документа и отклики на него, а также Mary Mitchell за предложения по грамматике и стилю.

Адреса авторов

Petr Lapukhov
Facebook
1 Hacker Way
Menlo Park, CA 94025
United States of America
Email: petr@fb.com

Ariff Premji
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
United States of America
Email: ariff@arista.com
URI: http://arista.com/

Jon Mitchell (editor)
Email: jrmitche@puck.nether.net

Перевод на русский язык
Николай Малых
nmalykh@gmail.com