Network Working Group
Request for Comments: 4786
BCP: 126
Category: Best Current Practice

J. Abley
Afilias Canada
K. Lindqvist
Netnod Internet Exchange

December 2006

Работа anycast-служб

Operation of Anycast Services

Статус документа

В этом документе описывается накопленный опыт (BCP[1]) и содержится запрос к дискуссии в целях дальнейшего развития и совершенствования. Документ может распространяться без ограничений.

Тезисы

Сеть Internet продолжает расти и, по мере того, как корпоративные системы и сетевые службы получают все большее распространение, появляются службы с высокими требованиями по их доступности. Это ведет к росту требований в части надежности инфраструктуры, обеспечивающей работу таких служб.

Были разработаны различные методы повышения уровня доступности служб, развернутых в сети Internet. В данном документе представлены комментарии и рекомендации по распространению услуг с использованием технологии anycast.

Оглавление

[1]Best Current Practices.

1. Введение

Этот документ адресован сетевым операторам, которые заинтересованы в развертывании или уже используют распределенные службы на базе anycast. Документ описывает накопленный опыт, но не дает рекомендаций в части выбора сетевых служб, которые имеет (или не имеет) смысл реализовать с использованием технологии anycast.

Для описания сервиса, использующего anycast, эта услуга сначала ассоциируется со стабильным набором адресов IP и доступность этих адресов анонсируется в систему маршрутизации множеством независимых узлов сервиса. Различные методы реализации услуг на основе anycast рассмотрены в [RFC1546], [ISC-TN-2003-1] и [ISC-TN-2004-1].

Описанные в этом документе методы и соображения применимы как для IPv4, так и для IPv6.

Технология anycast в последние годы стала более популярным решением для резервирования серверов DNS в дополнение к средствам резервирования, обеспечиваемым самой архитектурой DNS. Некоторые операторы корневых серверов DNS рапределяют свои серверы по сети Internet, а преобразователи и полномочные серверы в общем случае распределены по сетям сервис-провайдеров. Технология anycast используется коммерческими операторами полномочных серверов DNS уже в течение нескольких лет. Применение технологии anycast не ограничено DNS, хотя ее использование вносит создает дополнительные ограничения для распространяемого с использованием anycast-адресации сервиса, включая время жизни транзакций, сохранение данных о состоянии транзакций на серверах и возможности синхронизации данных.

Хотя технология anycast проста концептуально, ее реализация создает некоторые сложности в работе служб. Например, усложняется мониторинг доступности сервиса, поскольку наблюдаемая доступность зависит от местоположения клиента в сети, а популяция клиентов, использующих отдельные узлы anycast не является стабильной и надежно детерминированной.

В этом документе описано использование anycast как в локальном масштабе с использованием протокола маршрутизации IGP[1], так и в глобальном масштабе с использованием протокола BGP[2] [RFC4271]. Многие вопросы мониторинга и синхронизации данных являются общими для обоих вариантов, но развертывание сервиса существенно различается.

2. Термины

Service Address — адрес службы (сервиса)

Адрес IP, связанный с определенной службой (например, адрес получателя, используемый преобразователями DNS для доступа к определенному полномочному серверу).

Anycast

Технология, позволяющая сделать определенный адрес службы (Service Address) доступным во множестве дискретных, автономных пунктов, чтобы дейтаграмма, отправленная по anycast-адресу, маршрутизировалась в одно из доступных мест.

Anycast Node — узел anycast

Набор соединенных между собой внутренними связями хостов и маршрутизаторов, совместно обеспечивающх обслуживание anycast-адреса службы. Anycast-узел может быть отдельным хостом, принимающим участие в обмене маршрутными данными со смежными маршрутизаторами, а может представлять собой множество хостов, объединенных тем или иным способом для совместной работы. В любом случае для системы маршрутизации, через которую предоставляются anycast-услуги, каждый узел anycast представляет уникальный путь к адресу службы (Service Address). Сервис anycast включает по крайней мере два отдельных узла anycast.

Catchment — зона охвата

В физической географии — область водосбора реки, которую называют также бассейном реки (drainage basin). По аналогии в данном документе термин используется для обозначения топологической области сети, в пределах которой пакеты, направленные по anycast-адресу, маршрутизируются одному определенному узлу.

Local-Scope Anycast — локальный сервис anycast

Информация о доступности anycast-адреса службы распространяется в системе маршрутизации таким образом, чтобы узел anycast был виден только подмножеству глобальной системы маршрутизации.

Local Node — локальный узел

Узел anycast, обеспечивающий услуги по anycast адресу с локальной значимостью.

Global-Scope Anycast — глобальный сервис anycast

Информация о доступности anycast-адреса службы распространяется в системе маршрутизации таким образом, чтобы узел anycast был виден потенциально во всей глобальной системе маршрутизации.

Global Node — глобальный узел

Узел anycast, обеспечивающий услуги по anycast адресу с глобальной значимостью.

3. Распространение услуг anycast

3.1. Общее описание

Термин anycast используется применительно к практике предоставления одного адреса службы в систему маршрутизации от двух и более anycast-узлов, расположенных в разных местах. Услуги, предоставляемые разными узлами, в общем случае одинаковы, независимо от выбора конкретного узла системой маршрутизации (некоторые

[1]Interior Gateway Protocol — протокол внутреннего шлюза.
[2]Border Gateway Protocol — протокол граничного шлюза.

службы могут обеспечивать определенные преимущества при выборе соответствующего узла для того, чтобы пользователи обращались к более близким сервисным узлам, как описано в параграфе 4.6).

Для услуг, предоставляемых с использованием технологии anycast, не возникает унаследованных требований перенаправления на другие серверы или распространение услуг на сонове имен (round-robin DNS), хотя эти методы могут использоваться совместно с anycast-распространением услуг, если это нужно приложениям. Система маршрутизации решает, какой узел используется для каждого запроса, на основе топологии системы маршрутизации и точек сети, откуда поступают запросы.

На выбор узла anycast для конкретного запроса могут оказывать влияние средства управления трафиком протоколов маршрутизации, используемых в системе маршрутизации. Степень влияния, доступного оператору узла, зависит от масштаба системы маршрутизации, внутри которой используется anycast-адрес службы.

Распределение нагрузки между anycast-узлами в общем случае обеспечить достаточно сложно (в общем случае такое нагрузка между узлами не сбалансирована как по числу запросов, так и по объему трафика). Распределение нагрузки между узлами в целях повышения надежности, и обеспечения масштабируемости популярных служб, тем не менее, может быть достигнуто.

Масштаб системы маршрутизации, через которую предоставляются anycast-услуги, может меняться от небольших систем на основе протокола внутренней маршрутизации, соединяющего небольшое число компонент, до систем на базе протокола BGP [RFC4271] в масштабе Internet, в зависимости от природы распространяемых услуг.

3.2. Цели

Услуги могут предоставляться с использованием anycast по разным причинам, включая нижеперечисленные.

1. Грубое (несбалансированное) распределение нагрузки между узлами, позволяющее масштабировать инфраструктуру по мере роста числа запросов (в том числе, пиковых нагрузок).
2. Снижение угроз от нераспределенных атак на службы за счет локализации негативного воздействия на одном узле anycast.
3. Ограничение распределенных атак на службы и скопления пользователей локальными областями вокруг anycast-узлов. Распространение услуг с использованием технологии anycast обеспечивает возможность обработки трафика вблизи его источника, возможно с использованием высокоскоростных партнерских соединений вместо загрузки дорогостоящих транзитных каналов.
4. Обеспечение дополнительной информации, помогающей определить местоположение источников трафика в случаях атак (или запросов) с использованием подставных адресов отправителей. Эта возможность обысловлена тем, что при распространении услуг по технологии anycast выбор сервисного узла для конкретного запроса может быть топологически связан с источником запроса.
5. Снижения времени отклика на запросы за счет укорочения пути через сеть между клиентом и сервером при получении услуг от локального anycast-узла. Уровень снижения времени отклика на запросы зависит от способа выбора отвечающего узла в системе маршрутизации. Топологическая близость в системе маршрутизации в общем случае не коррелирует со временем кругового обхода через сеть; в некоторых случаях время отклика может не снижаться, а расти.
6. Уменьшения списка серверов до одного распределенного адреса. Например, большое число полномочных серверов имен для зоны может быть развернуто с использованием небольшого числа anycast-адресов служб; такое решение может существенно повысить уровень доступности данных зоны в системе DNS без увеличения времени отклика, связанного с обращением к полномочным серверам родительской зоны.

4. Схема

4.1. Приемлемость протоколов

Когда сервис предоставляется в режиме anycast с использованием двух и более узлов, система маршрутизации принимает решение о выборе узла в интересах клиента. Поскольку обычным требованием является взаимодействие клиента в рамках транзакции с одним сервером, выбор системой маршрутизации узла следует сохранять в течение времени, существенно превышающего предполагаемую продолжительность транзакции, если для сервиса требуется надежность.

В некоторых типах сервиса время транзакции очень мало и транзакции могут просто состоять из одного пакета запроса и одного пакета отклика (например, DNS-транзакции по протоколу UDP). Другие услуги могут использовать более продолжительные транзакции (например, передача больших фалов или потоковое вещание).

Услуги могут предоставляться с использованием anycast в предсказуемых системах маршрутизации, которые могут сохранять стабильность достаточно долго (например, anycast в хорошо управляемых и топологически простых системах IGP, где выбор узла меняется только в случаях отказа на каком-либо канале). Другие варианты развертывания имеют менее предсказуемые характеристики (см. параграф 4.4.7).

Стабильность системы маршрутизации и продолжительность транзакций следует принимать во внимание при решении вопроса о распределении услуг по технологии anycast. В некоторых случаях для новых протоколов может оказаться целесообразным разбиение долгих транзакций на фазу инициализации, обслуживаемую anycast-серверами, и стабильную фазу, которая обслуживается обычными (не-anycast) серверами, возможно выбираемыми в фазе инициализации.

В этом документе преднамеренно не задается никаких правил определения приемлемости тех или иных протоколов для anycast-распространения — такая попытка была бы дерзостью.

Операторам следует соблюдать осторожность, особенно для продолжительных потоков, поскольку отказы в режиме anycast несколько сложней простых отказов «destination unreachable» (адресат недоступен) в режиме unicast.

4.2. Размещение узлов

Выбор места размещения узлов anycast зависит от сферы распространения сервиса. Например:

- Услуги рекурсивного преобразования DNS могут распространяться в сети ISP с использованием одного узла anycast на сайт.
- Корневые серверы DNS могут быть распределены по сети Internet; узлы anycast могут располагаться в регионах со слабой внешней связностью для того, чтобы услуги DNS предоставлялись в регионе даже в случаях отказов на внешних сетях.
- Зеркала сервера FTP могут включать локальные узлы, размещаемые в точках обмена, чтобы подключенные к таким точкам ISP могли позволяли загружать большие файлы без использования дорогих транзитных каналов.

В общем случае при выборе мест размещения узлов следует принимать во внимание требования к трафику, возможное размещение клиентов, стабильность локальной системы маршрутизации и режимы обработки отказов узлов и локальной системы маршрутизации.

4.3. Системы маршрутизации

4.3.1. Anycast в IGP

Существует несколько факторов общего плана, стимулирующих распространение адреса службы в пределах IGP:

1. снижение времени отклика за счет размещения сервиса ближе к потребителям;
2. повышение надежности обслуживания за счет автоматического переключения на резервные узлы;
3. локализация трафика для предотвращения загрузки внешних каналов.

Во всех случаях решение о том, где и как будут предоставляться услуги может приниматься сетевыми инженерами без учета таких аспектов как различия в конфигурации клиентских компьютеров или возможное нарушение когерентности DNS (когда отклики на запросы DNS могут меняться в зависимости от источника запроса).

Когда услуга предоставляется в режиме anycast внутри системы IGP, вся система маршрутизации обычно находится под контролем той же организации, которая предоставляет услуги, и, следовательно, параметры транзакций и стабильности сети известны достаточно хорошо. В таком случае описанная технология применима к большему числу приложений, нежели распространение услуг по технологии anycast в масштабе сети Internet (см. параграф 4.1).

IGP в общем случае в общем случае не имеет унаследованных ограничений на размер анонсируемого префикса. В результате не возникает необходимости создавать покрывающий префикс для конкретного адреса службы — вместо этого в систему маршрутизации могут просто передаваться маршруты к хостам, соответствующие адресам служб. Требования к покрывающим префиксам более подробно рассмотрены в параграфе 4.4.2.

IGP зачастую не использует агрегирования маршрутов или агрегирует их достаточно слабо (отчасти, по причине алгоритмической сложности агрегирования). В большинстве сетей IGP нет важных причин для агрегирования, поскольку объемы маршрутных данных в сети IGP достаточно малы и проблем при масштабировании маршрутизации не возникает. Обсуждение связанных с агрегированием рисков в других системах маршрутизации приведено в параграфе 4.4.8.

За счет сокращения области IGP до обеспечивающих сервис хостов (вместе с одним или несколькими граничными маршрутизаторами) этот метод можно использовать для создания серверных кластеров. Такое применение описано в работе [ISC-TN-2004-1].

4.3.2. Anycast в глобальной сети Internet

Адреса служб могут быть типа anycast в глобальной системе маршрутизации Internet для распространения сервиса в масштабе всей сети. Принципиальные различия между таким применением и распространением сервиса в масштабе IGP, описанном в параграфе 4.3.1, состоят в том, что:

1. система маршрутизации в общем случае контролируется другими людьми;
2. используемый протокол маршрутизации (BGP) и общепринятая практика его использования вносят дополнительные ограничения (см. параграф 4.4).

4.4. Вопросы маршрутизации

4.4.1. Сигнализация доступности услуг

Когда в маршрутную систему представляется информация о доступности адреса службы для отдельного узла, пакеты, направленные по этому адресу, начнут поступать на узел. Поскольку важную роль играет готовность узла в приему пакетов до их прибытия, желательно обеспечить некоторую связь между анонсированием маршрута и доступностью услуг на конкретном узле.

Когда маршрутный анонс от узла соответствует одному адресу службы, связь может быть обеспечения констатацией доступности сервиса по анонсу маршрута и недоступности — по отзыву. Этого можно добиться используя реализатию протокола маршрутизации на том же сервере. Такие реализации обеспечивают поддержку распространяемого сервиса и настраиваются так, чтобы маршрут анонсировался и отзывался в соответствии с доступностью (и состоянием) программ, обеспечивающих обслуживание сервисных запросов. Пример такой реализации для сервиса DNS описан в работе [ISC-TN-2004-1].

Когда маршрутный анонс от узла соответствует двум и более адресам служб, он может оказаться неподходящим способом сигнализации, поскольку может, например, содержать отзыв одного из маршрутов вследствие недоступности одного сервиса. Другим решением для случаев отказа одного anycast-узла является маршрутизация запросов к другому узлу, где сервис нормально работает. Этот вариант обсуждается в параграфе 4.8.

Частые анонсы/отзывы могут вызывать проблемы при работе и узлы следует настраивать на предотвращение таких осцилляций (например, путем реализации минимальной паузы между анонсированием маршрута после его отзыва). Осцилляции маршрутов в BGP рассматриваются в параграфе 4.4.4.

4.4.2. Покрывающий префикс

В некоторых системах маршрутизации (например, в системах BGP глобальной системы маршрутизации Internet) невозможно, в общем случае, распространять через сеть информацию о маршрутах к хосту. Это ограничение связано с политикой маршрутизации, а не с ограничениями протокола.

В таких случаях необходимо распространять маршрут, который покрывает адрес службы и имеет достаточно короткий префикс, который не будет отбрасываться в соответствии с общепринятыми правилами импорта маршрутов. Для адресов IPv4 обычно достаточно сократить префикс до 24 битов, но существуют другие хорошо документированные примеры правил импорта IPv4, которые используют при фильтрации префиксов на границах RIR[1], поэтому могут потребоваться эксперименты по определению подходящего размера префикса. Для префиксов IPv6 также существуют соответствующие правила. Дополнительное рассмотрение адресов служб в формате IPv6 и соответствующих маршрутов anycast приведено в параграфе 4.5.

При распространении одного маршрута на сервис возникают некоторые проблемы масштабирования, описанные в параграфе 4.4.8.

Когда множество адресов служб покрывается одним маршрутом, теряется возможность связать анонсирование маршрута с отдельными службами, связанными с покрываемыми маршрутами к хостам. Связанное с этим влияние на сигнализацию доступности конкретного сервиса рассматривается в параграфах 4.4.1 и 4.8.

4.4.3. Равноценные пути

Некоторые системы маршрутизации поддерживают равноценные пути к получателям. В тех случаях, когда существует несколько равноценных путей, ведущих к различным узлам anycast, возникает риск передачи разных пакетов с запросами одной транзакции нескольким узлам. Услуги, предоставляемые по протоколу TCP [RFC0793] обязательно включают множество пакетов с запросами, поскольку протокол TCP использует процедуру согласования при организации соединений.

Для услуг, распределенных в глобальной сети Internet с использованием протокола BGP, равноценных путей обычно не возникает — алгоритм выбора маршрута BGP обычно завершает работу, возвращая единственный путь к адресату, независимо от числа возможных вариантов. Однако существуют реализации BGP, поддерживающие множество вариантов пути.

Равноценные пути часто встречаются в IGP. Выбора множества узлов для одной транзакции обычно удается избежать при аккуратном учете метрики каналов IGP или за счет использования алгоритма выбора из множества равноценных путей (ECMP[2]), который обеспечивает выбор единственного узла для транзакции, включающей множество пакетов. Например, использование алгоритма ECMP на основе хэширования для распространения anycast-сервиса описано в работе [ISC-TN-2004-1].

Доступны и другие алгоритмы выбора ECMP, включая и те, где не гарантируется маршрутизация всех пакетов одного потока на один адрес. Алгоритмы ECMP, выбирающие маршрут на уровне пакета, а не на уровне потока, обычно называют алгоритма пакетной балансировки нагрузки (PPLB[3]).

В плане распространения услуг с использованием anycast некоторые варианты применения PPLB могут приводить к тому, что пакеты из одной многопакетной транзакции будут доставляться разным anycast-узлам, по сути, делая недоступным anycast-сервис. В любом случае результат воздействия на конкретный сервис anycast будет зависеть от того, как и где узлы anycast развернуты в системе маршрутизации, а также от того, где выполняется PPLB:

1. реализация PPLB на множестве параллельных каналов между парой маршрутизаторов не будет создавать проблем при выборе узла;
2. реализация PPLB на разных путях внутри автономной системы (AS), сходящихся на выходе из AS, будет вызывать проблемы при выборе узла;
3. PPLB на каналах в разные соседние AS, где выбираются разные узлы для одного anycast-сервиса, будет в общем случае приводить к распределению пакетов с запросами между множеством узлов anycast; это приведет к тому, что услуги anycast станут недоступными для клиентов нисходящего направления в сторону маршрутизатора, выполняющего PPLB.

Использование PPLB, способного негативно воздействовать на распространение услуг anycast, может также создавать постоянное изменение порядка доставки пакетов. Путь, на котором постоянно меняется порядок следования сегментов, будет приводить к снижению производительности приложений TCP [Allman2000]. TCP по результатам опубликованных измерений ([McCreary2000], [Fomenkov2004]) широко используется в Internet для транзакций с большими объемами данных. Следовательно, во многих случаях имеет смысл рассматривать сети с таким использованием PPLB, как патологические.

4.4.4. Подавление маршрутов

Частые анонсы и отзывы отдельных префиксов в BGP называют хлопками (flap). Это явление может приводить к перегрузке процессоров (CPU) в маршрутизаторах, достаточно удаленных от источника нестабильности. По этой причине быстрые осцилляции маршрутов зачастую подавляются (демпфируются), как описано в [RFC2439].

Задемпфированный путь будет подавляться маршрутизаторами на интервалы времени, возрастающие с повышением наблюдаемой частоты осцилляций; при этом демпфируемый путь не будет анонсироваться. Следовательно, один маршрутизатор способен предотвратить распространение нестабильного префикса в остальные автономные системы, избавляя другие маршрутизаторы в сети от такой нестабильности.

[1]Regional Internet Registry — региональный регистратор Internet.
[2]equal-cost multi-path.
[3]Per Packet Load Balancing.

Некоторые реализации подавления осцилляций работают на основе наблюдаемых значений AS_PATH, а не NLRI[1] (см. [RFC4271]). По этой причине нестабильность сети, ведущая к осцилляциям маршрута для одного узла anycast, в общем случае не будет вызывать подавления такими реализациями анонсов от других узлов (с иными атрибутами AS_PATH).

Чтобы ограничить в таких реализациях подавление анонсов от других узлов anycast в ответ на осцилляции одного узла, следует сделать атрибуты AS_PATH от разных узлов максимально различающимися. Например, узлам anycast следует использовать в своих анонсах одинаковое значение исходной AS, но они могут указывать разные AS восходящего направления.

Там, где для подавления осцилляций используются в основном реализации иного типа, нестабильность отдельного узла может приводить к недоступности стабильных узлов. Для снижения такого негативного влияния полезны следующие меры:

1. разумное сочетание локальных узлов с достаточно стабильными глобальными узлами (надежные соединения между AS, резервирование оборудования и питания и т. п.) может ограничить область негативного влияния нестабильности локальных узлов;
2. жесткое подавление осцилляций вблизи источника (например, в локальной автономной системе узла anycast или в смежных AS каждого узла anycast) также позволяет снизить негативное влияние осцилляций на удаленные AS.

4.4.5. Проверка обратного пути

Проверка обратного пути (RPF[2]), впервые описанная в [RFC2267], обычно развертывается, как часть входной фильтрации пакетов на маршрутизаторах Internet для отбрасывания пакетов с подставными адресами отправителей (см. также [RFC2827]). Развернутые реализации RPF поддерживают несколько режимов работы (например, loose[3] и strict).

Некоторые режимы RPF могут приводить к отказу от приема пакетов с корректными адресами отправителей, когда эти пакеты приходят от многодомных сайтов, поскольку выбранный путь может представляться не соответствующим входному интерфейсу. Этот вопрос рассматривается в [RFC3704].

Множество anycast-узлов, распределенных по сети Internet, с точки зрения системы маршрутизации трудно отличить от многодомного сайта. Следовательно, существует риск ошибочного отбрасывания anycast-пакетов входными фильтрами маршрутизаторов даже если они отправлены не многодомными узлами. Следует принимать меры, чтобы к каждому узлу anycast фильтры относились, как к многодомной сети, и выполнялись соответствующие рекомендации [RFC3704] в части RPF.

4.4.6. Область распространения

В контексте распространения сервиса anycast в глобальном масштабе Internet глобальными узлами (Global Node) называют те узлы, которые могут предоставлять свои услуги клиентам, находящимся в любом месте сети; информация о доступности такого сервиса распространяется глобально, без ограничений, путем анонсирования маршрутов, покрывающих адреса служб, для глобального транзита одному или множеству провайдеров.

Для одной службы может существовать множество глобальных узлов (на практике это делается достаточно часто для повышения уровня надежности и распределения нагрузки).

Иногда желательно реализовать anycast-узел так, чтобы услуги предоставлялись только в масштабе локальной автономной системы и не были доступными из остальной сети Internet; такие узлы в этом документе называются локальными (Local Node). Примером может служить развертывание локального узла для обслуживания региона с хорошей внутренней связностью, но ненадежными, дорогими и перегруженными соединениями с остальной частью Internet.

Локальные узлы анонсируют покрывающие маршруты для адресов служб так, чтобы распространение этих маршрутов было ограничено. Для ограничения могут использоваться общеизвестные символьные атрибуты групп (community) типа NO_EXPORT [RFC1997] или NOPEER [RFC3765], а также партнерские соглашения по реализации специальной (peering) политики импорта вместо обычной (transit) или некоторым иным мерам ограничения.

Анонсирование доступности адресов служб для локальных узлов в идеальном случае следует выполнять с использованием политики маршрутизации, требующей присутствия явных атрибутов распространения, чтобы не опираться на неявную (принятую по умолчанию) политику. Неадекватное распространение маршрута за пределы запланированного горизонта может приводить к проблемам, связанным с нехваткой ресурсов локального узла, и соответствующему снижению производительности.

4.4.7. Чужие сети

При развертывании сервиса anycast через сети, управляемые другими операторами, информация о доступности сервиса будет зависеть от политики маршрутизации и изменений топологии (планируемых и незапланированных), которые непредсказуемы и не всегда идентифицируются достаточно просто. Поскольку система маршрутизации может включать сети, управляемые множеством не связанных организаций, существует вероятность непреднамеренного воздействия комбинации не связанных между собой факторов.

Стабильность и предсказуемость такой системы маршрутизации следует принимать во внимание при решении вопроса о выборе технологии anycast для конкретных услуг или протоколов (см. также параграф 4.1).

Для упрощения реализации сервиса anycast в таких системах маршрутизации политику маршрутизации следует делать консервативной, а инфраструктуру внутренних и внешних соединений узла следует делать масштабируемой для поддержки уровней нагрузки, существенно превышающих средний уровень. Кроме того, для сервиса следует организовать проактивный мониторинг из множества точек, чтобы избежать неприятных сюрпризов (см. параграф 5.1).

[1] Network Layer Reachability Information — информация о доступности на сетевом уровне.
[2] Reverse Path Forwarding — путь обратной пересылки.
[3] Мягкий, свободный и жесткий, соответственно. *Прим. перев.*

4.4.8. Риски, связанные с агрегированием

Распространение отдельного маршрута для каждого сервиса anycast не обеспечивает требуемого уровня масштабирования для систем маршрутизации, в которых передача маршрутной информации имеет важное значение и возможно использование множества услуг-anycast. Например, автономной системе, поддерживающей доступ в Internet и включающей N адресов служб, обеспечиваемых по технологии anycast, потребуется анонсировать (по крайней мере - *прим. перев.)* N+1 маршрут, если для каждой службы анонсируется отдельный маршрут.

Общепринятая практика использования фильтров минимального размера префиксов в правилах импорта маршрутов в сети Internet (см. параграф 4.4.2) означает, что для маршрута, покрывающего адрес службы, будет полезно анонсировать префикс, размер которого существенно меньше, нежели у простого маршрута к хосту. Однако широкое использование коротких префиксов для отдельных служб ведет к неоправданному расходу адресного пространства.

Обе эти проблемы можно смягчить, если использовать один покрывающий префикс для множества адресов служб, как описано в параграфе 4.8. Однако в этом случае теряется связь между анонсированием маршрута и доступностью сервиса (см. параграф 4.4.1), что ведет к снижению стабильности работы сервиса в целом (см. параграф 4.7).

В общем случае описанная здесь проблема масштабирования делает технологию anycast менее полезной для распространения услуг в глобальном масштабе Internet. Однако эта технология остается весьма полезной для распространения ограниченного числа критически важных услуг, а также для использования в небольших сетях, где не возникает проблем, связанных с агрегированием маршрутов.

4.5. Вопросы адресации

Адреса служб следует делать уникальными в масштабе системы маршрутизации, соединяющей все узлы anycast со всеми возможными клиентами. Адреса служб должны выбираться так, чтобы соответствующие маршруты разрешалось бы распространять внутри данной системы маршрутизации.

Для сервиса IPv4, развернутого в сети Internet, например, адрес можно выбрать из блока, где минимальный размер выделения RIR составляет 24 бита и доступность адреса можно анонсировать с использованием покрывающего префикса размером 24 бита.

Для сервиса IPv4 в частной сети можно использовать специально выделенные для таких сетей адреса [RFC1918] и (32 бита).

Для сервиса IPv6 адреса anycast не рассматриваются отдельно от обычных (unicast) адресов. Поэтому рекомендации по выбору адресов IPv4 подходят и для IPv6. Отметим, давние запреты на распространение услуг по технологии anycast для IPv6 были удалены из спецификации адресов IPv6 [RFC4291].

4.6. Синхронизация данных

Хотя некоторые службы разворачиваются с использованием локализации (клиентам из определенного региона предоставляется информация, относящаяся к этому региону), для многих служб важна согласованность откликов на запросы клиентов, независимо от местоположения источника запроса. Для услуг, распространяемых с использованием anycast, это подразумевает согласованность различных узлов anycast, а в тех случаях, когда согласованность основана на наборе данных, подразумевается синхронизация данных между узлами.

Механизм синхронизации данных зависит от природы этих данных. Примерами могут служить перенос зон для полномочных серверов DNS или операции rsync для архивов FTP. В общем случае синхронизация данных между узлами anycast будет включать транзакции между обычными (не-anycast) адресами.

Синхронизацию данных через публичные сети следует выполнять с использованием средств идентификации и шифрования.

4.7. Автономность узлов

Для развертывания anycast-приложений, обеспечивающих высокую надежность за счет резервирования, важно минимизировать вероятность воздействия одного повреждения на множество (или все) узлов или возникновения каскадных отказов с последовательным отключением улов вплоть до полной остановки сервиса.

Воздействие узлов друг на друга предотвращается за счет того, что каждый узел делается максимально автономным и самодостаточным. Уровень живучести узлов при неполадках в системе зависит от природы предоставляемых услуг, но для служб, которые не требуют постоянного соединения (например, растянутое во времени распространение изменений от ведущего сервера DNS к ведомым), может быть обеспечен высокий уровень автономности.

Вероятность возникновения каскадных аварий в результате перегрузок можно снизить за счет развертывания локальных и глобальных узлов для одного сервиса. Поскольку в общем случае от локального узла к глобальному имеется эффективный, отказоустойчивый путь, трафик, способный привести к отказу одного локального узла обычно не будет оказывать влияния на остальные локальные узлы за исключением очень серьезных аварий.

Вероятность каскадных аварий по причине программных дефектов в операционной системе или сервере во многих случаях может быть снижена за счет развертывания узлов, использующих разные реализации операционных систем, серверных программ, протоколов маршрутизации и т. п., чтобы дефекты одной компоненты не приводили к отказу всей системы.

Следует отметить, что приведенные здесь рекомендации по повышению уровня автономности узлов в той или иной степени противоречат практическим задачам упрощения работы сервиса. Более сложный сервис более подвержен влиянию ошибок операторов, нежели простой сервис. Следовательно, нужно принимать во внимание все обстоятельства и находить разумный компромисс между противоречивыми требованиями.

4.8. Мультисервисные узлы

Для сервисных узлов, распределенных в системе маршрутизации, требующей использования покрывающих префиксов при анонсировании доступности отдельных адресов служб (см. параграф 4.4.2), нужно специально рассматривать случаи, когда на одном наборе узлов требуется развернуть множество разных служб. Такая необходимость

обусловлена требованием поддержки сигнализации доступности отдельных служб в систему маршрутизации, чтобы запросы на обслуживание не передавались узлам, не способным их обработать (см. параграф 4.4.1).

В следующих параграфах описано несколько вариантов решения этой задачи.

4.8.1. Множество покрывающих префиксов

Адреса служб выбираются так, чтобы каждым анонсируемым префиксом покрывался единственный адрес службы. Анонсирование и отзыв одного покрывающего префикса в этом случае будет тесно связано с доступностью одного сервиса.

Это решение является наиболее простым. Однако оно ведет к крайне нерациональному использованию адресов и по этой причине может быть приемлемо лишь для крайне важных инфраструктурных служб типа корневых серверов DNS. Использование этого решения для развертывания служб общего назначения в масштабе Internet не обеспечивает должной масштабируемости.

4.8.2. Общий префикс

Множество адресов служб выбирается таким образом, чтобы все адреса покрывались одним префиксом. Анонсирование и отзыв одного покрывающего префикса будет связано с доступностью всех служб. Если одна из служб становится недоступной, покрывающий префикс отзывается.

Связь между анонсированием покрывающего префикса и доступностью служб осложняется требованием доступности всех адресов служб — управление анонсами должно осуществляться по наличию маршрутов ко всем компонентам, а не просто по одному покрываемому префиксом маршруту.

То, что отказ одной из служб вынуждает отзывать общий префикс и делает недоступными все остальные службы делает этот вариант малопригодным для большинства приложений.

4.8.3. Связность внутри узла

Множество адресов служб выбирается таким образом, чтобы все адреса покрывались одним префиксом. Анонсирование и отзыв префикса связывается с доступностью любого (одного) сервиса. Узлы имеют внутреннюю связность (например, с помощью туннелей). Маршруты к хостам для адресов служб распространяются с использованием IGP, который охватывает маршрутизаторы на всех узлах.

Если сервис перестает быть доступным на одном узле, но все прочие узлы продолжают работать, запрос может маршрутизироваться через внутреннюю сеть в направлении одного из работающих узлов.

В случаях, когда некоторые локальные службы на узле перестают работать и узел теряет связь с другими узлами, анонсирование покрывающего префикса может приводить к тому, что часть запросов уйдет в «черную дыру».

Этот вариант обеспечивает достаточно эффективное использование адресов блока, покрываемого анонсируемым префиксом за счет снижения уровня автономности отдельных узлов. Протокол IGP, в котором участвуют все узлы, является потенциальной точкой отказа.

4.9. Идентификация узла клиентами

Время от времени все использующие сервис клиенты сталкиваются с некими проблемами, которые требуют диагностики. Распределение сервиса с помощью технологии anycast вносит дополнительные сложности в процесс диагностики по сравнению с использованием обычных (unicast) услуг — требуется определить конкретный узел anycast, который обслуживает запросы клиента.

В некоторых случаях может оказаться достаточным использование инструментов общего назначения (например, traceroute) для идентификации используемого клиентом узла. Однако эти инструменты не всегда доступны рядовому пользователю и, кроме того, условия в сети при наличии проблем могут изменяться в процессе использования таких средств контроля.

Поиск неисправностей в anycast-системах можно существенно упростить с помощью механизмов идентификации узлов, встроенных в протокол. Примерами таких механизмов могут служить опция NSID в DNS [NSID] и общепринятое включение имени хоста для сервера SMTP на этапе представления в начале почтовой сессии [RFC2821].

Обеспечение механизмов индикации узла при обычной работе настоятельно рекомендуется для служб, распространяемых с использованием технологии anycast.

5. Управление сервисом

5.1. Мониторинг

Мониторинг распределенного сервиса сложнее, чем мониторинг локализованного сервиса, поскольку точность наблюдений и доступность услуг в общем случае могут зависеть от местоположения клиентов в сети. При обнаружении проблем также не всегда удается легко идентифицировать узел, с которым эти проблемы могут быть связаны.

Для мониторинга распределенных услуг рекомендуется распределять зонды по всей системе маршрутизации и, по возможности, идентифицировать узлы, отвечающие на запросы, сохраняя информацию об узле вместе со статистикой производительности и доступности. Примером такого мониторинга для системы DNS может служить сервис RIPE NCC DNSMON [DNSMON].

Мониторинг системы маршрутизации (из разных точек в тех случаях, когда это уместно) также может давать информацию, полезную при поиске неисправностей в распределенном сервисе. Такой мониторинг можно организовать на основе специализированных зондов или публичных средств контроля маршрутов в Internet типа RIPE NCC Routing Information Service [RIS] или проекта Route Views университета штата Орего [ROUTEVIEWS].

Мониторинг состояния компонент распределенного anycast-сервиса (хостов, маршрутизаторов и т. п.) является достаточно эффективным способом и может быть организован с использованием тех же средств и методов, которые

обычно служат для мониторинга других сетевых инфраструктур. Использование технологии anycast в данном случае не создает дополнительных сложностей.

6. Вопросы безопасности

6.1. Ослабление атак на службы (DoS)

В этом документе описаны механизмы развертывания служб Internet, которые могут снизить влияние ряда атак.

1. Узлы anycast могут «поглощать» вредоносный трафик, происходящих из их сферы влияния, предотвращая воздействие атаки на другие узлы сервиса.
2. Приложения, сталкивающиеся с распределенными атаками на службы, сами по себе распределены между всеми узлами, принимающими участие в сервисе. Поскольку задача сортировки легитимного и вредоносного трафика распределяется по сети, это может обеспечивать более эффективное масштабирование по сравнению с локальной организацией сервиса.

6.2. Компрометация сервиса

Распределение сервиса между несколькими (или многими) автономными узлами усложняет мониторинг и повышает ответственность системных администраторов, что может привести к снижению уровня безопасности хостов и маршрутизаторов.

Потенциальным преимуществом распределенного сервиса является возможность отключения компрометированных серверов без существенного влияния на работу сервиса в целом. Хорошая организация рабочих процедур позволит выполнять такое отключение надежно и быстро.

6.3. Захват сервиса

Возможны случаи, когда не уполномоченные лица будут анонсировать маршруты, соответствующие anycast-адресам служб в сеть и за счет этого захватывать трафик легитимных запросов, обрабатывая его специальным образом с целью компрометации сервиса в целом. Детектирование подставных узлов anycast со стороны клиентов или операторов сервиса может оказаться непростой задачей.

Риск захвата сервиса путем манипуляций в системе маршрутизации существует независимо от использования технологии anycast. Однако присутствие легитимных узлов anycast в системе маршрутизации может осложнять обнаружение подставных узлов.

Многие протоколы, включающие идентификацию и защиту целостности, обеспечивают отказоустойчивость этих функций. Примерами могут служить периодическая повторная идентификация в рамках одного сеанса, контроль целостности на уровне канала (например, [RFC2845], [RFC3207]) или сообщения (например, [RFC4033], [RFC2311]). Менее отказоустойчивые протоколы могут быть более подвержены захвату сеансов, предоставляя возможности недетектируемого захвата anycast-служб. Поэтому для систем anycast рекомендуется использовать протоколы, требующие идентификации и обеспечивающие контроль целостности.

7. Благодарности

Авторы выражают свою признательность участникам рабочей группы GROW и, в частности, Geoff Huston, Pekka Savola, Danny McPherson, Ben Black и Alan Barrett.

Работа поддерживалась US National Science Foundation (исследовательский грант SCI-0427144) и DNS-OARC.

8. Литература

8.1. Нормативные документы

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

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

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

8.2. Дополнительная литература

[Allman2000] Allman, M. and E. Blanton, "On Making TCP More Robust to Packet Reordering", January 2000, <http://www.icir.org/mallman/papers/tcp-reorder-ccr.ps>.

[DNSMON] "RIPE NCC DNS Monitoring Services", <http://dnsmon.ripe.net/>.

[Fomenkov2004] Fomenkov, M., Keys, K., Moore, D., and k. Claffy, "Longitudinal Study of Internet Traffic from 1999-2003", January 2004, <http://www.caida.org/outreach/papers/2003/nlanr/nlanr_overview.pdf>.

[ISC-TN-2003-1] Abley, J., "Hierarchical Anycast for Global Service Distribution", March 2003, <http://www.isc.org/pubs/tn/isc-tn-2003-1.html>.

[ISC-TN-2004-1] Abley, J., "A Software Approach to Distributing Requests for DNS Service using GNU Zebra, ISC BIND 9 and FreeBSD", March 2004, <http://www.isc.org/pubs/tn/isc-tn-2004-1.html>.

[McCreary2000] McCreary, S. and k. claffy, "Trends in Wide Area IP Traffic Patterns: A View from Ames Internet Exchange", September 2000, <http://www.caida.org/outreach/papers/2000/AIX0005/AIX0005.pdf>.

[NSID] Austein, R., "DNS Name Server Identifier Option (NSID)", Work in Progress, June 2006.

[RFC1546] Partridge, C., Mendez, T., and W. Milliken, "Host Anycasting Service", RFC 1546, November 1993.

[RFC2267] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", RFC 2267, January 1998.

[RFC2311] Dusse, S., Hoffman, P., Ramsdell, B., Lundblade, L., and L. Repka, "S/MIME Version 2 Message Specification", RFC 2311, March 1998.

[RFC2821] Klensin, J., "Simple Mail Transfer Protocol", RFC 2821, April 2001.

[RFC2845] Vixie, P., Gudmundsson, O., Eastlake, D., and B. Wellington, "Secret Key Transaction Authentication for DNS (TSIG)", RFC 2845, May 2000.

[RFC3207] Hoffman, P., "SMTP Service Extension for Secure SMTP over Transport Layer Security", RFC 3207, February 2002.

[RFC3765] Huston, G., "NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control", RFC 3765, April 2004.

[RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, March 2005.

[RIS] "RIPE NCC Routing Information Service (RIS)", <http://ris.ripe.net>.

[ROUTEVIEWS] "University of Oregon Route Views Project", <http://www.routeviews.org/>.

Адреса авторов

Joe Abley

Afilias Canada, Corp.

204 - 4141 Yonge Street

Toronto, ON M2P 2A8

Canada

Phone: +1 416 673 4176

EMail: jabley@ca.afilias.info

URI: http://afilias.info/

Kurt Erik Lindqvist

Netnod Internet Exchange

Bellmansgatan 30

118 47 Stockholm

Sweden

EMail: kurtis@kurtis.pp.se

URI: http://www.netnod.se/

Перевод на русский язык

Николай Малых

nmalykh@protocols.ru

Полное заявление авторских прав

Интеллектуальная собственность

Подтверждение

Финансирование функций RFC Editor обеспечено Internet Society.