Internet Engineering Task Force (IETF)
Request for Comments: 8277
Obsoletes: 3107
Category: Standards Track
ISSN: 2070-1721

E. Rosen
Juniper Networks, Inc.

October 2017

Using BGP to Bind MPLS Labels to Address Prefixes

Использование BGP для привязки меток MPLS к адресным префиксам

Тезисы

В этом документе задан набор процедур использования протокола BGP для анонсирования фактов привязки конкретным маршрутизатором указанной метки MPLS (или последовательности меток, являющейся непрерывной частью стека) к заданному префиксу адресов. Это может быть сделано путем передачи сообщения BGP UPDATE, в котором поле NLRI[1] содержит префикс и метку (метки) MPLS, а поле Next Hop указывает узел, который заявил привязку метки (меток) к префиксу. Документ отменяет RFC 3107.

Статус документа

Документ относится к категории Internet Standards Track.

Документ является результатом работы IETF[2] и представляет согласованный взгляд сообщества IETF. Документ прошел открытое обсуждение и был одобрен для публикации IESG[3]. Дополнительную информацию о стандартах Internet можно найти в разделе 2 в RFC 7841.

Информацию о текущем статусе документа, ошибках и способах обратной связи можно найти по ссылке https://www.rfc-editor.org/info/rfc8277.

Оглавление

1. Введение

В [RFC3107] задано кодирование и процедуры для использования BGP с целью индикации того, что определенный маршрутизатор связал метку MPLS (или последовательность меток, являющуюся непрерывной частью стека меток, как определено в [RFC3031] и [RFC3032]) с адресным префиксом. Это делается путем передачи сообщения BGP UPDATE, поле NLRI в котором содержит префикс и метку (метки) MPLS, а поле Next Hop указывает узел, который заявил о

[1]Network Layer Reachability Information - информация о доступности на сетевом уровне.
[2]Internet Engineering Task Force.
[3]Internet Engineering Steering Group.

привязке метки или меток. Сообщение UPDATE также анонсирует путь к конкретному префиксу через указанный следующий маршрутизатор (next hop).

Несмотря на наличие множества реализаций и обширное применение [RFC3107], имеется ряд проблем, которые препятствуют или могут воспрепятствовать взаимодействию. Проблемы эти перечислены ниже.

- Хотя заданное в [RFC3107] кодирование позволяет связать с адресным префиксом последовательность меток MPLS (а не просто одиночную метку), документ не определяет семантику связывания последовательности меток с префиксом.
- Многие реализации [RFC3107] предполагают связывание с префиксом единственной мети и не могут корректно обработать сообщения BGP UPDATE с привязкой префикса к последовательности меток. В результате попытка реализации использовать такую возможность с высокой вероятностью приведет к проблемам при взаимодействии с другими реализациями.
- Процедуры отзыва привязки метки или последовательности меток к адресному префиксу не заданы в [RFC3107] четко и корректно.
- [RFC3107] задает необязательную функцию анонсирования множества маршрутов к получателю (Advertising Multiple Routes to a Destination), которая по сведениям автора никогда не было реализована в соответствии со спецификацией. Эту функцию можно реализовать другим путем с использованием процедур [RFC7911], которые не были доступны на момент написания [RFC3107]. В [RFC3107] для управления этой функцией служил код BGP Capability, который не был реализован и в настоящее время отменен (см. раздел 6).
- Узел BGP может получить для адресного префикса два сообщения BGP UPDATE, одно из которых будет связывать с префиксом метку (или последовательность меток), другое - нет. В [RFC3107] ничего не сказано о влиянии двух таких сообщений UPDATE на процесс решения BGP и не сказано явно, какое из этих двух сообщений UPDATE следует распространять. Это привело к тому, что разные реализации по-разному обрабатывают такие события.
- Большая часть [RFC3107] применима к семействам адресов VPN-IPv4 ([RFC4364]) и VPN-IPv6 ([RFC4659]), но эти семейства не упомянуты в документе.

Данный документ заменяет и отменяет [RFC3107]. Здесь определяется новая возможность BGP (BGP Capability) для использования в случаях привязки префикса к последовательности меток. Использование этой возможности позволяет предотвратить упомянутую выше проблему взаимодействия. Документ также отменяет не реализованную функцию «Advertising Multiple Routes to a Destination» (см. раздел 4 в [RFC3107]) и задает использование [RFC7911] для решения тех же задач. Документ решает вопрос взаимоотношений сообщений UPDATE для одного префикса, одно из которых анонсирует привязку к метке, другое не включает такой привязки. Однако для совместимости с имеющимися реализациями указано, что большая часть таких взаимодействий определяется локальной политикой.

Места, где данная спецификация отличается от [RFC3107], указаны в тексте. Предполагается, что реализации, соответствующие данному документу, будут совместимы с развернутыми реализациями [RFC3107].

Ключевые слова **необходимо** (MUST), **недопустимо** (MUST NOT), **требуется** (REQUIRED), **нужно** (SHALL), **не нужно** (SHALL NOT), **следует** (SHOULD), **не следует** (SHOULD NOT), **рекомендуется** (RECOMMENDED), **не рекомендуется** (NOT RECOMMENDED), **возможно** (MAY), **необязательно** (OPTIONAL) в данном документе должны интерпретироваться в соответствии с BCP 14 [RFC2119] [RFC8174] тогда и только тогда, когда они набраны заглавными буквами, как показано здесь.

2. Использование BGP для привязки адресного префикса к MPLS

Протокол BGP может служить для анонсирования информации о привязке определенным узлом (назовем его N) конкретной метки MPLS (или определенной последовательности меток, составляющей непрерывную часть стека) к конкретному адресному префиксу. Это выполняется путем передачи сообщений Multiprotocol BGP UPDATE, т. е. сообщений UPDATE с атрибутом MP_REACH_NLRI, соответствующим [RFC4760]. Сетевой адрес в поле Next Hop этого атрибута указывает IP-адрес узла N. Метка (метки) и префикс кодируются в поле NLRI атрибута MP_REACH_NLRI, как описано в параграфах 2.2 и 2.3.

Если префикс относится к IPv4 или VPN-IPv4 ([RFC4364]), поле AFI[1] атрибута MP_REACH_NLRI имеет значение 1, а для префиксов IPv6 или VPN-IPv6 ([RFC4659]) указывается AFI = 2.

Если префикс относится к IPv4 или IPv6, в поле SAFI[2] устанавливается значение 4, в для префикса VPN-IPv4 или VPN-IPv6 - SAFI = 128.

Использование SAFI = 4 или SAFI = 128 при AFI отличном от 1 и 2 выходит за рамки этого документа.

Данный документ не задает формат адреса в поле Next Hop атрибута MP_REACH_NLRI. Формат поля Next Hop зависит от множества факторов и рассматривается во многих RFC (см. [RFC4364], [RFC4659], [RFC4798] и [RFC5549]).

Существует множество приложений с другими методами использования BGP для анонсирования привязки меток MPLS, например [RFC7432], [RFC6514], [TUNNEL-ENCAPS]. Описанный здесь метод не объявляется единственным вариантом анонсирования привязки меток MPLS с помощью BGP. Обсуждение выбора метода для конкретных приложений выходит за рамки документа.

Далее в документе SAFI-x UPDATE будет обозначать сообщение BGP UPDATE с атрибутом MP_REACH_NLRI или MP_UNREACH_NLRI ([RFC4760]), в котором поле SAFI содержит значение x.

Этот документ определяет параметр BGP Optional Capabilities ([RFC5492]), известный как Multiple Labels Capability.

[1]Address Family Identifier - идентификатор семейства адресов.
[2]Subsequent Address Family Identifier - идентификатор последующего семейтсва адресов.

- Пока эта возможность не указана в данной сессии BGP обоими узлами BGP, эти узлы **должны** передавать в данной сессии сообщения UPDATE для SAFI-4 или SAFI-128 с привязкой единственной метки и **должны** использовать кодирование, описанное в параграфе 2.2.

- Если эта возможность указана обоими узлами в данной сессии BGP, оба узла **должны** передавать в данной сессии сообщения UPDATE с кодированием, описанным в параграфе 2.3 и **могут** связывать префикс с последовательностью меток.

Представление Multiple Labels Capability описано в параграфе 2.1.

Процедуры для явного отзыва привязки меток описаны в параграфе 2.4. Процедуры изменения привязки метки (меток) к данному префиксу на данном узле описаны в параграфе 2.5.

Процедуры распространения сообщений UPDATE для SAFI-4 и SAFI-128 описаны в разделе 3.

Когда узел BGP устанавливает и распространяет обновление (UPDATE) для SAFI-4 или SAFI-128 или меняет адрес в поле Next Hop, он должен соответствующим образом программировать свой уровень данных. Это описано в разделе 4.

2.1. Поддержка множества меток

[RFC5492] определяет Capabilities Optional Parameter. Узел BGP может включить Capabilities Optional Parameter в свое сообщение BGP OPEN. Параметр представляет собой триплет, включающий 1 октет Capability Code, 1 октет размера и значение переменной длины.

Данный документ определяет код возможности (Capability Code), известный как Multiple Labels Capability. Агентство IANA назначило для этого кода значение 8 (данный код введен этим документом и отсутствует в [RFC3107].)

Если узел BGP не передает Multiple Labels Capability в своем сообщении BGP OPEN для конкретной сессии BGP или не получает Multiple Labels Capability в сообщении BGP OPEN от партнера в данной сессии BGP, ему **недопустимо** передавать в этой сессии какие-либо сообщений UPDATE, связывающие с префиксом более одной метки MPLS. При анонсировании привязки единственной метки к префиксу узел BGP **должен** использовать кодирование, описанное в параграфе 2.2.

Значение поля Multiple Labels Capability (рисунок 1) состоит из одного или множества триплетов, размером 4 октета. Два первых октета задают значение AFI, третий - SAFI, а четвертый - Count. Поле Count в триплете <AFI, SAFI, Count> указывает максимальное число меток, которые узел BGP, указавший поддержку множества меток, способен обработать в сообщении UPDATE для указанных AFI и SAFI. Значение 255 показывает отсутствие ограничения на число меток в сообщении UPDATE для указанных AFI и SAFI.

Любая реализация, передающая Multiple Labels Capability **должна** быть способна обрабатывать не менее двух меток в NLRI. Однако в некоторых вариантах развертывания может требоваться большее число меток.

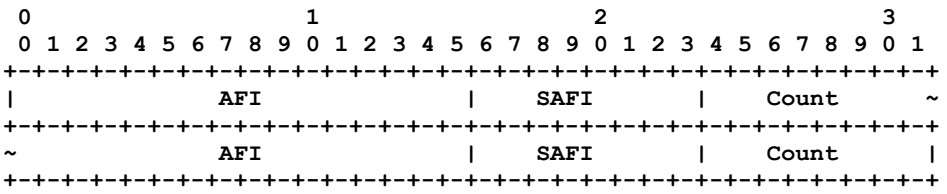

Рисунок 1. Формат Multiple Labels Capability.

Если указано более 1 триплета с данной парой AFI и SAFI, все триплеты, кроме первого, **должны** игнорироваться.

Триплеты вида <AFI=x, SAFI=y, Count=0> или <AFI=x, SAFI=y, Count=1> передавать **недопустимо**, а при получении такие триплеты **должны** игнорироваться.

Multiple Labels Capability с размером, не кратным 4, **должны** считаться некорректно сформированными.

Механизм «Graceful Restart Mechanism for BGP» [RFC4724] описывает процедуру, которая позволяет поддерживать маршруты, полученные через данную сессию BGP, после отказа и последующего перезапуска сессии. Эта процедура требует передачи RIB целиком при перезапуске сессии. Если в отказавшей сессии был выполнен обмен Multiple Labels Capability для данного AFI/SAFI, но этого обмена не произошло в восстановленной сессии, все префиксы, анонсированные в этом AFI/SAFI с множеством меток, **должны** быть явно отозваны. Точно так же при снижении максимального числа меток (указанного в Capability для данного AFI/SAFI) все префиксы, анонсированные с большим числом меток, **должны** быть явно отозваны.

В документе «Accelerated Routing Convergence for BGP Graceful Restart» [Enhanced-GR] описана другая процедура, позволяющая поддерживать полученные в сессии BGP маршруты при отказе и перезапуске этой сессии. Эти процедуры **недопустимо** применять при наличии любого из приведенных ниже условий.

- Обмен Multiple Labels Capability для данного AFI/SAFI был выполнен до перезапуска сессии, но не был повторен после перезапуска.

- Обмен Multiple Labels Capability для данного AFI/SAFI до перезапуска был выполнен со значением Count, а после перезапуска повторен с меньшим значением.

При выполнении любого из условий **должен** быть выполнен обмен полным набором маршрутов для данных AFI/SAFI.

Если сообщение BGP OPEN содержит множество копий Multiple Labels Capability, значение будет иметь лишь первая, а последующие копии **должны** игнорироваться.

Если (a) узел BGP передал Multiple Labels Capability в своем сообщении BGP OPEN для определенной сессии BGP, (b) получил Multiple Labels Capability в сообщении BGP OPEN от партнера в этой сессии и (c) обе возможности указывают AFI/SAFI x/y, тогда при использовании UPDATE для AFI x и SAFI y для анонсирования привязки метки или последовательности меток к данному префиксу узел BGP **должен** применять кодирование, описанное в параграфе 2.3. Это кодирование **должно** применяться даже при связывании с префиксом лишь одной метки.

Если оба партнера в сессии BGP передали Multiple Labels Capability, но AFI/SAFI x/y не было задано обеими сторонами, сообщения UPDATE для AFI/SAFI x/y в этой сессии **должны** использовать кодирование параграфа 2.2, а такие сообщения UPDATE могут привязывать к префиксу лишь одну метку.

Узлу BGP **не следует** передавать сообщений UPDATE, которые привязывают к префиксу больше меток, нежели партнер способен воспринять в соответствии с переданной им возможностью Multiple Labels Capability. Если узел BGP получает сообщение UPDATE, привязывающее к префиксу больше меток, чем этот узел готов получить (как он указал в Multiple Labels Capability), этот узел **должен** применять к данному сообщения UPDATE стратегию treat-as-withdraw[1] из [RFC7606].

Несмотря на заявленное узлом BGP число меток, которые он способен принять, его партнеру **недопустимо** пытаться передать больше меток, нежели можно корректно поместить в поле NLRI атрибута MP_REACH_NLRI. Следует подчеркнуть, что пространство для меток в поле NLRI ограничено.

- В соответствии с [RFC4760] размер поля ограничен 255 битами (не октетами), включая число битов префикса.
- В SAFI-128 UPDATE префикс имеет размер не менее 64 битов и может достигать 192 битов (например, в маршруте к хосту VPN-IPv6).

2.2. Кодирование NLRI без поддержки множества меток

Если возможность Multiple Labels Capability не была передана и получена в данной сессии BGP, в сообщениях BGP UPDATE с атрибутом MP_REACH_NLRI, содержащим одну комбинацию AFI/SAFI, заданную в разделе 2, поле NLRI кодируется, как показано на рисунке 2.

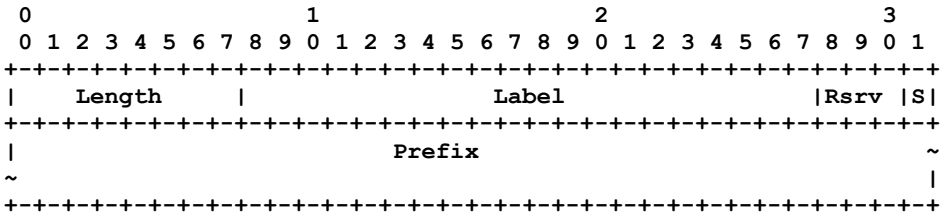

Рисунок 2. NLRI с одной меткой.

Length

Однооктетное поле Length указывает число битов в оставшейся части поля NLRI.

Отметим, что размер всегда будет составлять 20 (число битов поле Label) + 3 (число битов в поле Rsrv) + 1 (число битов в поле S) + размер префикса в битах.

В атрибуте MP_REACH_NLRI с AFI/SAFI = 1/4 размер префикса составляет не более 32 битов, при AFI/SAFI = 2/4 - не более 128 битов. В атрибуте MP_REACH_NLRI с SAFI = 128 размер префикса будет не более 96 битов, если AFI = 1 и не более 192 битов для AFI = 2.

Как указано в [RFC4760], действительный размер поля NLRI будет равен числу битов, указанному в поле Length, с округлением до ближайшего большего целого числа октетов.

Label

20-битовое поле со значением метки MPLS (см. [RFC3032]).

Rsrv

В этом 3-битовом поле при передаче **следует** устанавливать 0, а на приеме оно **должно** игнорироваться.

S

Это 1-битовое поле **должно** быть установлено (1) при передаче и **должно** игнорироваться при получении.

Отметим, что сообщение UPDATE анонсирует не только привязку метки и префикса, но и путь к префиксу через узел, указанный в Network Address поля Next Hop атрибута MP_REACH_NLRI.

[RFC3107] требует установки бита S, если с префиксом связана лишь одна метка. Если бит S не установлен, [RFC3107] указывает присутствие других меток в NLRI. Однако некоторые реализации считают, что в NLRI бывает лишь одна метка и не проверяют бит S. Процедуры, заданные в этом документе, обеспечат взаимодействие с такими реализациями. Пока возможность Multiple Labels Capability не передана и не принята обеими сторонами сессии BGP, этот документ **требует** указывать в NLRI лишь одну метку, что ведет к установке бита S на передающей стороне и игнорированию на приемной.

Если применяются процедуры [RFC7911], 4-октетный идентификатор пути (определен в разделе 3 [RFC7911]) является частью NLRI и предшествует полю Length.

2.3. Кодирование NLRI при поддержке множества меток

Если возможность Multiple Labels Capability передана и получена обеими сторона данной сессии BGP, в сообщениях BGP UPDATE этой сессии, где атрибут MP_REACH_NLRI содержит одну из комбинаций AFI/SAFI, заданных в разделе 2, поле NLRI кодируется, как показано на рисунке 3.

Length

Однооктетное поле Length указывает число битов в оставшейся части поля NLRI.

Отметим, что каждая метка увеличивает размер на 24 бита (20 битов поля Label, 3 бита Rsrv и бит S).

В атрибуте MP_REACH_NLRI с AFI/SAFI = 1/4 размер префикса составляет не более 32 битов, при AFI/SAFI = 2/4 - не более 128 битов. В атрибуте MP_REACH_NLRI с SAFI = 128 размер префикса будет не более 96 битов, если AFI = 1 и не более 192 битов для AFI = 2.

Как указано в [RFC4760], действительный размер поля NLRI будет равен числу битов, указанному в поле Length, с округлением до ближайшего большего целого числа октетов.

Label

20-битовое поле со значением метки MPLS (см. [RFC3032]).

[1]Рассматривать как отзыв.

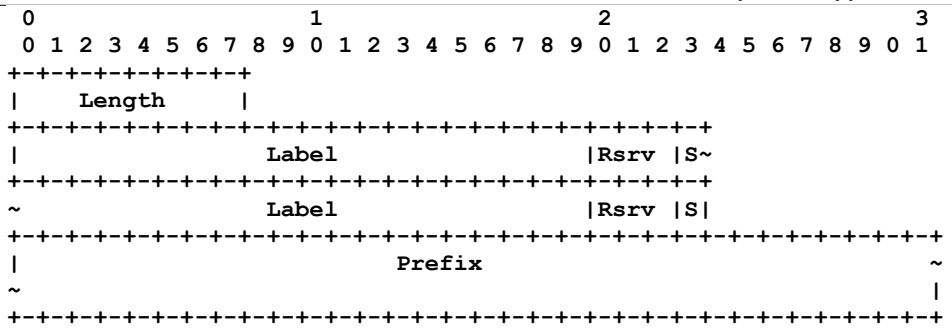

Рисунок 3. NLRI с последовательностью меток.

Rsrv

В этом 3-битовом поле при передаче **следует** устанавливать 0, а на приеме оно **должно** игнорироваться.

S

Во всех метках, кроме последней (т. е. непосредственно предшествующей префиксу) бит S **должен** быть сброшен (0), а в последней метке **должен** быть установлен (1).

Отметим, что отсутствие установленного бита S в последней метке делает невозможным корректный разбор NLRI. Обсуждение обработки ошибок при отказе в обработке NLRI приведено в разделе 3, п. j [RFC7606].

Отметим, что сообщение UPDATE анонсирует не только привязку метки и префикса, но и путь к префиксу через узел, указанный в Network Address поля Next Hop атрибута MP_REACH_NLRI.

Если применяются процедуры [RFC7911], 4-октетный идентификатор пути (определен в разделе 3 [RFC7911]) является частью NLRI и предшествует полю Length.

2.4. Явный отзыв привязки метки к префиксу

Предположим, что узел BGP анонсировал в данной сессии привязку метки или последовательности меток к данному префиксу, а сейчас желает отозвать эту привязку. Для этого узел может передать сообщение BGP UPDATE с атрибутом MP_UNREACH_NLRI, поле NLRI которого показано на рисунке 4.

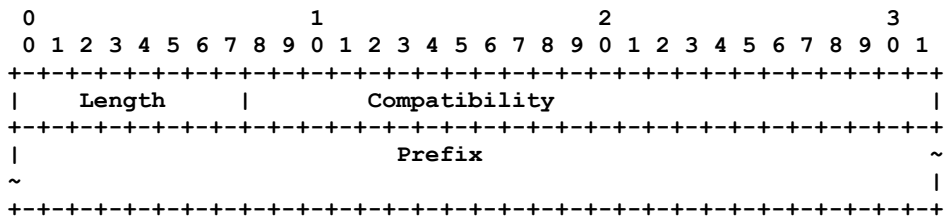

Рисунок 4. NLRI для отзыва.

При передаче в поле Compatibility **следует** установить 0x800000, а при получении это поле **должно** игнорировтаься.

Это кодирование применяется для явного отзыва (в данной сессии BGP) привязки указанного префикса к любой метке или последовательности меток, которая ранее была выполнена процедурами этого документа для данного префикса. Кодирование не зависит от передачи и приема возможности Multiple Labels Capability для этой сессии. Отметим, что привязки, которые не были анонсированы в этой сессии, не могут быть отозваны таким методом. Однако привязки, анонсированные в предыдущей сессии с тем же партнером могут быть отозваны этим методом, если текущая сессия является результатом «аккуратного перезапуска» (graceful restart [RFC4724]) предыдущей сессии.

При использовании атрибута MP_UNREACH_NLRI для отзыва маршрута, в котором поле NLRI было ранее задано атрибутом MP_REACH_NLRI, размеры и значения соответствующих префиксов и AFI/SAFI должны совпадать. При использовании процедур [RFC7911] должны также совпадать соответствующие значения полей идентификаторов пути. Отметим, что размер префикса - это не размер NLRI и для определения размера префикса в MP_UNREACH_NLRI из размера поля Compatibility должно вычитаться значение размера NLRI.

Явный отзыв в сообщении SAFI-x UPDATE для данной сессии BGP отзывает не только привязки префиксов к меткам, но и пути к этим префиксам, анонсированные ранее в SAFI-x UPDATE для этой сессии.

[RFC3107] делает возможным указание конкретного значения метки в поле Compatibility. Однако, функциональность, требующая присутствия конкретного значения метки (или последовательности значений меток), никогда не была реализована и не присутствует в данном документе. Следовательно, значение этого поля не имеет смысла и нет никаких причин включать в него значение метки или последовательность значений меток.

[RFC3107] также делает возможным отзыв без явного указания метки путем установки в поле Compatibility значения 0x800000. Однако некоторые реализации устанавливают в этом поле 0x000000. Для совместимости с предыдущими версиями этот документ **рекомендует** устанавливать Compatibility = 0x800000 и **требует** игнорировать поле при получении.

2.5. Смена привязанной к префиксу метки

Предположим, что узел BGP S1 в данной сессии BGP получил сообщение SAFI-4 или SAFI-128 UPDATE U1, которое указывает метку (или последовательность меток) L1, префикс P и следующий маршрутизатор N1. Как указано выше, это показывает, что метка (последовательность меток) L1 связана с префиксом P на узле N1. Предположим, что S1 затем получает в той же сессии сообщение UPDATE U2 с тем же AFI/SAFI, которое указывает метку (последовательность меток) L2, префикс P и тот же следующий интервал N1.

- Если [RFC7911] не используется, сообщение UPDATE U2 **должно** трактоваться как привязка L2 к префиксу P на узле N1 и отмена имеющейся привязки L1 к префксу P на N1. Т. е. UPDATE U1 неявно отзывается и заменяется UPDATE U2.

- Предположим, что используется [RFC7911], сообщение UPDATE U1 имеет Path Identifier I1, а UPDATE U2 - Path Identifier I2.
    - Если I1 совпадает с I2, сообщение UPDATE U2 **должно** должно интерпретироваться как привязка L2 к префиксу P на узле N1 и отмена привязки L1 к P на узле N1. Сообщение UPDATE U1 неявно отзывается.
    - Если I1 и I2 различаются, сообщение U2 **должно** интерпретироваться как привязка метки L2 к префиксу P на узле N1, но U2 **недопустимо** трактовать как отмену привязки L1 к P на узле N1. При некоторых условиях (их указание выходит за рамки документа) S1 может выбрать балансировку трафика между путями, представленными U1 и U2. Для отправки трафика по пути, представленному U1, S1 использует метку (метки), анонсированную в U1, а для отправки по пути, представленному U2, - метку (метки), анонсированную в U2 (хотя эти пути ведут к одному маршрутизатору, можно предположить, что далее они разойдутся).

Предположим, что узел BGP S1 получил в данной сессии BGP сообщение SAFI-4 или SAFI-128 UPDATE, задающее метку L1, префикс P и next hop N1. Затем предположим, что S1 получил в другой сессии BGP сообщение UPDATE с тем же AFI/SAFI, которое задает метку L2, префикс P и тот же next hop N1. Узлу BGP S1 **следует** трактовать это как наличие у N1 не менее двух путей к P и S1 **может** использовать это для распределения трафика, передаваемого в P.

Отметим, что здесь рассмотрены лишь случаи, где два сообщения UPDATE имеют одинаковый следующий интервал -- next hop. Процедуры для разных next hop в двух сообщениях UPDATE рассмотрены в [RFC4271].

3. Установка и распространение маршрутов SAFI-4 или SAFI-128

3.1. Совместимость маршрутов

Предположим, что узел BGP получил два сообщения SAFI-4 UPDATE с одним значением Prefix:

- в разных сессиях BGP;
- в одной сессии, использующей добавление пути, и NLRI в сообщениях имеют разные идентификаторы путей.

Эти два маршрута **должны** считаться сравнимыми, даже если они указывают разные метки. Таким образом, применяется процедура выбора лучшего пути BGP (параграф [RFC4271]) для выбора одного из путей. Если процедуры [RFC7911] не применяются в данной сессии BGP, в такой сессии будет распространяться только лучший путь. При использовании процедур [RFC7911] в сессии BGP для этой сессии могут распространяться оба пути с разными идентификаторами.

То же самое применимо и для маршрутов SAFI-128.

3.2. Изменение поля меток в процессе распространения

3.2.1. Поле Next Hop не меняется

Если при распространении маршрутов SAFI-4 или SAFI-128 сетевой адрес (Network Address) в поле Next Hop не меняется, поле (поля) Label также **должно** оставаться неизменным.

Отметим, что данный маршрут **недопустимо** распространять партнеру, если NLRI в маршруте имеет множество меток, но возможность Multiple Labels Capability не согласована с этим партнером. Точно так же данный маршрут **недопустимо** распространять партнеру, если NLRI в маршруте имеет больше меток, чем может обработать (указал в Multiple Labels Capability) этот партнер. В любом случае, если предыдущий маршрут с теми же AFI, SAFI и префиксом (но с меньшим числом меток) уже был распространен партнеру, этот маршрут **должен** быть отозван с использованием процедуры, описанной в параграфе 2.4.

3.2.2. Поле Next Hop меняется

Если Network Address в поле Next Hop меняется перед распространением маршрута SAFI-4 или SAFI-128, поле (поля) Label в распространяемом маршруте **должно** содержать метку (метки), привязанную к префиксу в новом next hop.

Предположим, что узел BGP S1 получил сообщение UPDATE, связывающее последовательность из одной или нескольких меток с определенным префиксом. Если S1 принимает решение о распространении этого маршрута после изменения next hop, S1 может изменить метку одним из перечисленных ниже способов, в зависимости от локальной политики.

- Одна метка может быть заменена на другую с таким же или иным значением.
- Последовательность меток может быть заменена одной меткой.
- Одна метка может быть заменена последовательностью меток.
- Последовательность меток может быть заменена другой с тем же или иным числом меток.

При решении вопроса о распространении сообщений UPDATE с привязкой последовательности из нескольких меток узел BGP должен принимать во внимание информацию Multiple Labels Capability (параграф 2.1). Узлу BGP **недопустимо** передавать множество меток партнерам, с которыми не было обмена Multiple Labels Capability, а также **недопустимо** передавать больше меток, чем партнер может обработать (указал в Multiple Labels Capability).

Возможно, что локальная политика узла BGP задает кодирование N меток в NLRI данного маршрута перед его распространением, но один из партнеров BGP не способен обработать N в NLRI. В этом случае возможны 2 варианта.

- Можно распространить маршрут с меньшим числом меток. Если это имеет смысл и делается, выбор меток определяется локальной политикой.
- Можно отказаться от распространения маршрута данному партнеру. В этом случае распространенный ранее данному партнеру маршрут с теми же AFI, SAFI и префиксом (но меньшим числом меток) **должен** быть отозван по процедуре параграфа 2.4.

4. Уровень данных

Далее в документе фраза «узел S туннелирует пакет P узлу N» предполагает, что P является пакетом MPLS. Это означает, что узел S инкапсулирует пакет P и организует доставку P на узел N так, что стек меток пакета P до инкапсуляции придет узлу N неизменным, но не будет виден другими узлами между S и N (если они имеются).

Если туннель является LSP[1], инкапсуляция может заключаться в простом вталкивании дополнительной метки в стек MPLS. Если узлы N и S смежны на канальном уровне, может оказаться достаточно инкапсуляции L2. Могут применяться и другие типы туннелей (например, IP, GRE, UDP) в зависимости от конкретного типа развертывания.

Предположим, что узел BGP S1 получает сообщение SAFI-4 или SAFI-128 BGP UPDATE с MP_REACH_NLRI, указывающим метку L1, префикс P и next hop N1, а также предположим, что S1 устанавливает этот маршрут в качестве лучшего (или одного из лучших) пути к P. Затем предположим, что S1 распространяет этот маршрут, заменив next hop на себя и метку - на L2. Если S1 получит пакет данных MPLS и в процессе его обработки для пересылки увидит, что метка L2 поднялась на вершину стека, S1 должен будет заменить метку L2 на L1 в качестве верхней метки стека и туннелировать этот пакет узлу N1.

Предположим, что полученный S1 маршрут содержит не одну метку, а последовательность из k меток <L11, L12, ..., L1k>, где L11 - первая метка в NLRI, а L1k - последняя. Снова предположим, что S1 распространяет этот маршрут, указав себя в качестве next hop и заменив поле Label на единственную метку L2. Если S1 получает пакет данных MPLS и в процессе его обработки для пересылки видит метку L2 на вершине стека, вместо простой замены L2 на L1 он будет удалять L2 и вталкивать в стек метки с L1k по L11 так, чтобы метка L11 оказалась в стеке верхней. После этого узел S1 должен туннелировать пакет узлу N1 (отметим, что L1k не будет нижней меткой стека и у нее не будет установлен флаг bottom of stack, если метка L2 не была нижней в стеке).

В предыдущих абзацах предполагается, что S1 распространяет маршрут SAFI-4 или SAFI-128 после указания себя в качестве next hop и меняет метку или последовательность меток в NLRI этого маршрута на одну метку. Однако локальная политика может разрешать узлу BGP задавать множество меток в распространяемом маршруте SAFI-4 или SAFI-128 после установки себя в качестве next hop.

Предположим, например, что S1 поддерживает метки контекста ([RFC5331]). Пусть L21 будет меткой контекста, поддерживаемой S1, а L22 - меткой в пространстве, указанном (для S1) меткой L21. Предположим, что S1 получает сообщение SAFI-4 или SAFI-128 UPDATE для префикса P, в котором поле Label имеет значение <L11, L12, ..., L1k>, а next hop - N1. Перед распространением UPDATE узел S1 может указать себя в качестве next hop (заменив N1 на S1) и заменить стек меток <L11, L12, ..., L1k> на пару меток <L21, L22>.

Если в таком случае узел S1 получает пакет данных MPLS с верхней меткой L21 и второй меткой L22, он будет удалять из стека обе метки, заменяя их последовательностью <L11, L12, ..., L1k>. Отметим, что контекстный характер метки L21 известен лишь S1, а другие узлы BGP не знают, как узел S1 будет интерпретировать L21 (или L22).

Способность заменять одну или множество меток другой меткой (или множеством) может обеспечить высокий уровень гибкости, но делать это нужно с осторожностью. Предположим снова, что S1 получает сообщение UPDATE с префиксом P, стеком меток <L11, L12, ..., L1k> и next hop N1. Пусть S1 распространяет UPDATE узлу BGP S2, устанавливая себя в качестве next hop и заменяя поле Label последовательностью <L21, L22, ..., L2k>. Далее предположим, что S1 программирует свой уровень данных так, что при обработке пересылаемого пакета MPLS с верхней меткой L21 эта метка заменяется последовательностью <L11, L12, ..., L1k>, а затем пакет туннелируется N1.

В этом случае узел BGP S2 будет получать маршрут с префиксом P, полем Label <L21, L22, ..., L2k> и next hop S1. Если S2 решит переслать пакет IP по этому маршруту, он будет вталкивать метки <L21, L22, ..., L2k> в стек пакета и туннелировать пакет узлу S1, который заменит L21 на <L11, L12, ..., L1k> и туннелирует пакет узлу N1. Узел N1 получит пакет со стеком меток <L11, L12, ..., L1k, L22, ..., L2k>. В некоторых случаях это будет полезно, а в других может приводить к неожиданным результатам.

Процедуры выбора, организации, поддержки и проверки живучести конкретного туннеля или типа туннелей выходят за рамки документа.

При вталкивании меток в стек пакета поля TTL[2] ([RFC3032], [RFC3443]) и TC[3] ([RFC3032], [RFC5462]) должны быть установлены для каждой метки. Документ не задает правил установки полей, оставляя это локальной политике.

Этот документ не задает никаких новых правил обработки стека меток во входящих пакетах данных.

Решение вопроса о возможности применения маршрутов SAFI-4 для базовой пересылки пакетов IP или только для пересылки пакетов MPLS определяется локальной политикой. Если узел BGP S1 пересылает пакеты в соответствии с маршрутами SAFI-4, тогда при пересылке пакета IP с адресом получателя D маршрут, в котором префикс P является самым длинным префиксом, совпадающим с D, рассматривается наряду с другими маршрутами, применяемыми для пересылки пакетов IP. Предположим, что пакет пересылается в соответствии с маршрутом SAFI-4, имеющим префикс P, next hop N1 и последовательность меток L1. Для пересылки пакета по этому маршруту S1 должен создать стек меток для пакета, втолкнуть в него последовательность меток L1 и туннелировать пакет узлу N1.

5. Связь между маршрутами SAFI-4 и SAFI-1

Узел BGP может получить маршруты SAFI-1 и SAFI-4 для префикса P. Разные реализации по разному трактуют это.

Например, некоторые реализации могут считать маршруты SAFI-1 и SAFI-4 совершенно независимыми и трактовать их по принципу «корабль в ночи» (ships in the night). В таком случае выбор лучшего пути для двух SAFI будет независимым и будет выбран лучший маршрут SAFI-1 к префиксу P и лучший маршрут SAFI-4 к тому же префиксу. Выбор маршрута для реальной пересылки пакетов по этому префиксу будет зависеть от локальной политики.

Другие реализации могут считать маршруты SAFI-1 и SAFI-4 для одного префикса сравнимыми и при выборе маршрута к префиксу P останется маршрут SAFI-1 или SAFI-4, но не оба. В таких реализациях при распределении нагрузки

[1]Label Switched Path - путь с коммутацией по меткам.
[2]Time-to-Live - время жизни.
[3]Traffic Class - класс трафика.

между равноценными маршрутами некоторые из таких маршрутов могут оказаться SAFI-1, а другие - SAFI-4. Какой из маршрутов будет применяться, определяет локальная политика.

Некоторые реализации могут позволять передачу в одной сессии BGP сообщений UPDATE для SAFI-1 и SAFI-4, а другие могут это запрещать. Некоторые реализации, разрешающие оба SAFI в одной сессии, могут считать получение маршрута SAFI-1 для префикса P в данной сессии неявным отзывом предыдущего маршрута SAFI-4 для префикса P в этой сессии и наоборот. Поведение других реализаций может быть иным.

Узел BGP может получить маршрут SAFI-4 через данную сессию BGP, имея другие сессии BGP, где SAFI-4 не разрешаются. В таких случаях узел BGP **может** преобразовать маршрут SAFI-4 в маршрут SAFI-1 и распространять результат через другие сессии, где SAFI-4 не поддерживается. Это определяется локальной политикой.

Различия в поведении реализаций могут приводить к неожиданностям или проблемам при взаимодействии. В некоторых случаях может оказаться сложно или невозможно добиться желаемого поведения от некоторых реализаций или комбинаций разных реализаций.

6. Взаимодействие с IANA

Агентство IANA выделило значение 8 для Multiple Labels Capability в реестре BGP Capability Codes со ссылкой на данный документ.

Реестр BGP Capability Codes был изменен с указанием кода возможности 4 (множество маршрутов к адресату) как отмененного со ссылкой на данный документ.

Агентство IANA изменило ссылку для SAFI 4 в реестре Subsequent Address Family Identifiers (SAFI) Parameters указанием на данный документ.

Этот документ также добавлен в качестве ссылки для SAFI 128 в том же реестре.

7. Вопросы безопасности

Вопросы безопасности BGP, рассмотренные в [RFC4271], применимы в данном случае.

Если реализация BGP, не соответствующая данному документу, кодирует множество меток в NLRI, но возможность Multiple Labels Capability не была передана и принята, соответствующая данному документу реализация BGP будет вероятно сбрасывать сессию BGP.

Этот документ указывает, что некоторые пакеты данных будут «туннелироваться» от одного узла BGP к другому. Это требует инкапсуляции пакетов «на лету». Данный документ не задает применяемой инкапсуляции, однако при использовании той или иной инкапсуляции применимы и соответствующие вопросы безопасности.

Если туннельная инкапсуляция не обеспечивает контроля целостности и подлинности, пакеты данных со стеком меток могут быть изменены в результате ошибок или злонамеренно, пока они передаются через сеть. Это может приводить к нарушению доставки пакетов. Следует также отметить, что туннельная инкапсуляция (MPLS), наиболее часто применяемая в реализациях данного документа, не обеспечивает контроля целостности и подлинности, равно как и другие типы инкапсуляции, упомянутые в разделе 4.

Имеются различные методы для ограничения области распространения сообщений BGP UPDATE. Если BGP UPDATE анонсирует привязку метки или последовательности меток к адресному префиксу, эти методы могут применяться для контроля набора узлов BGP, которые будут знать о такой привязке. Однако другие маршрутизатору также смогут узнать о привязке, если сессии BGP не используют защиты конфиденциальности.

Когда узел BGP получает пакет данных MPLS, верхнюю метку которого он анонсировал, нет гарантии, что эта метка была помещена в пакет маршрутизатором, которому следует знать о данной привязке. Если узел BGP использует процедуры, описанные в этом документе, ему может быть полезно отличать свои «внутренние» интерфейсы от «внешних» и избегать анонсирования одинаковых меток партнерам BGP через внутренние и внешние интерфейсы. Тогда пакет данных можно отбросить, если верхняя метка не анонсировалась через интерфейс, на котором был принят пакет. Это снижает вероятность пересылки пакетов, в которых метки были подделаны недоверенными источниками.

8. Литература

8.1. Нормативные документы

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <https://www.rfc-editor.org/info/rfc3032>.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <https://www.rfc-editor.org/info/rfc3107>.

[RFC3443] Agarwal, P. and B. Akyol, "Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks", RFC 3443, DOI 10.17487/RFC3443, January 2003, <https://www.rfc-editor.org/info/rfc3443>.

[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, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, DOI 10.17487/RFC4659, September 2006, <https://www.rfc-editor.org/info/rfc4659>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, DOI 10.17487/RFC4798, February 2007, <https://www.rfc-editor.org/info/rfc4798>.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, DOI 10.17487/RFC5462, February 2009, <https://www.rfc-editor.org/info/rfc5462>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[RFC5549] Le Faucheur, F. and E. Rosen, "Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop", RFC 5549, DOI 10.17487/RFC5549, May 2009, <https://www.rfc-editor.org/info/rfc5549>.

[RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <https://www.rfc-editor.org/info/rfc7606>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

8.2. Дополнительная литература

[Enhanced-GR] Patel, K., Chen, E., Fernando, R., and J. Scudder, "Accelerated Routing Convergence for BGP Graceful Restart", Work in Progress, draft-ietf-idr-enhanced-gr-06, June 2016.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <https://www.rfc-editor.org/info/rfc4724>.

[RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, DOI 10.17487/RFC5331, August 2008, <https://www.rfc-editor.org/info/rfc5331>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <https://www.rfc-editor.org/info/rfc6514>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

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

[TUNNEL-ENCAPS] Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", Work in Progress, draft-ietf-idr-tunnel-encaps-07[1], July 2017.

Благодарности

Этот документ отменяет RFC 3107. Спасибо Yakov Rekhter - соавтору RFC 3107 за его работу над документом. Спасибо также Ravi Chandra, Enke Chen, Srihari R. Sangli, Eric Gray и Liam Casey за их рецензии и замечания.

Спасибо Alexander Okonnikov и David Lamparter за найденные в RFC 3107 ошибки.

Спасибо Lili Wang и Kaliraj Vairavakkalai за помощью и советы при подготовке документа.

Спасибо Mach Chen, Bruno Decraene, Jie Dong, Adrian Farrel, Jeff Haas, Jonathan Hardwick, Jakob Heitz, Alexander Okonnikov, Keyur Patel, Kevin Wang и Lucy Yong за их рецензии и замечания к данному документу.

Адрес автора

Eric C. Rosen

Juniper Networks, Inc.

10 Technology Park Drive

Westford, Massachusetts 01886

United States of America

Email: erosen@juniper.net

Перевод на русский язык

Николай Малых

nmalykh@protocols.ru

[1] В https://tools.ietf.org/html/draft-ietf-idr-tunnel-encaps-10 имеется более свежий вариант документа. *Прим. перев.*