BGP Part 1

~~*New Hope*~~ *Basics*

by Anatoliy Vedyakin (modemfux@gmail.com)

BGP – история

- Был создан в 1989 году (RFC 1105)
- Актуальная версия протокола – 4, впервые выпущена в 1994 году в RFC 1654
- Актуальный стандарт для BGP-4 – это RFC 4721 от 2006 года
- Существует множество дополнительных RFC, расширяющих и/или исправляющих основной стандарт

BGP – основные данные

- Является дистанционно-векторным протоколом маршрутизации
- De facto единственный протокол типа EGP
- Использует TCP для установки смежных соединений (порт 179)
- Не умеет автоматически находить соседей*
- В современном мире используется для передачи не только маршрутной информации (EVPN, VPLS, Flowspec, RT-Filter, etc)

BGP – Терминология

- AS – Autonomous System (автономная система) – это группа маршрутизаторов, находящая под единым управлением и использующих единую логику принятия маршрутизирующих решений
- EBGP – external BGP, т.е. процесс построения BGP-связности с пиром, находящимся в автономной системе с номером, отличным от номера на устройстве
- IBGP – internal BGP, т.е. процесс построения BGP-связности с пиром, находящимся в той же AS, что и само устройство
- Path Attributes – атрибуты пути, совокупность параметров, характеризующих вектор пути для конкретного префикса и на основе которых происходит выбор наилучшего маршрута

BGP – Терминология

- NLRI – Network Layer Reachability Information – совокупность префикса и атрибутов пути
- RIB – Routing Information Base
- Adj-RIB-In – часть BGP RIB, в которой хранится принятая и необработанная маршрутная информация
- Loc-RIB – часть BGP RIB, в которой хранится маршрутная информация (как сгенерированная локально, так и принятая от пиров после проведенной входной обработки)
- Adj-RIB-Out – часть BGP RIB, в которой хранится маршрутная информация, прошедшая обработку и которая должна быть передана соседям

BGP – Номера автономных систем

- Маршрутизируемые в Интернете номера автономных систем выдаются IANA
- Номера AS можно разделить на два вида по длине номера:
  - 2-байтные (0 - 65 535)
  - 4-байтные (0 - 4 294 967 295)
- 4-байтные ASN могут записываться как в plain виде, так и в dotted-decimal (ASDOT). Например, ASN = 170000:
  - ASPLAIN: 170000
  - ASDOT: 2.38928

BGP – Зарезервированные номера AS

Номер AS	Описание	Reference
0	Зарезервировано	RFC 7607
112	AS112 Project	RFC 7534
23 456	AS_TRANS (используется в качестве номера AS в поле My AS при использовании 32-битной AS)	RFC 6793
64 496 - 64 511	Для использования в документации и примерах	RFC 5398
64 512 - 65 534	Для частного использования	RFC 6996
65 535	Зарезервировано	RFC 7300
65 536 - 65 551	Для использования в документации и примерах	RFC 5398
4 200 000 000 - 4 294 967 294	Для частного использования	RFC 6996
4 294 967 295	Зарезервировано	RFC 7300

BGP – Виды сообщений

- OPEN – отправляется после установки TCP-сессии между устройствами, необходимо для передачи и согласования параметров BGP-сессии. Также в OPEN передается список поддерживаемых опций/функций etc.
- KEEPALIVE – сообщение, служащее для поддержания BGP-сессии в работающем состоянии, также используется как подтверждение получения OPEN-сообщения
- UPDATE – используются для передачи маршрутной информации (NLRI, Path attributes и т.п.). Нужны как для передачи новой маршрутной информации, так и для отзыва неактуальной информации (Withdraw)
- NOTIFICATION – данное сообщение отправляется при обнаружении ошибки в работе. В сообщении указывается код ошибки.

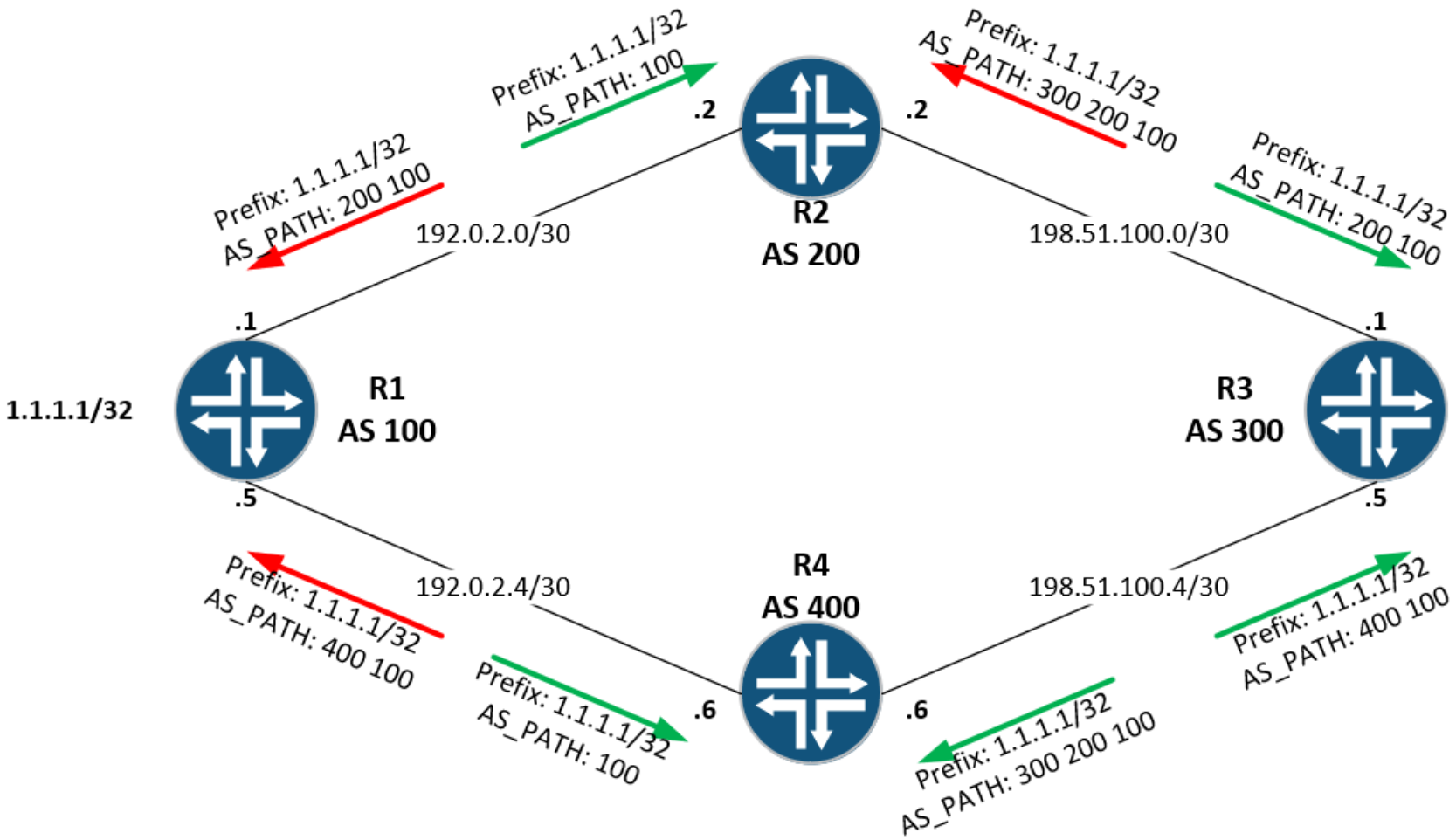
BGP – Установка соседства

- Установление соседства проходит через следующие этапы:
  - Idle – самый первый этап, здесь процесс BGP понимает, что ему необходимо установить TCP-сессию с соседом
  - Connect – процесс BGP пытается установить с соседом TCP-сессию, причем на этом этапе сессию инициирует и управляет ею маршрутизатор с наибольшим IP-адресом (т.е. он отправляет TCP SYN на dst.port = 179). Если TCP-сессия установлена, то процесс переходит на этап OpenSent
  - Active – процесс BGP пытается установить с соседом TCP-сессию, но на этом этапе попыткой установки сессии занимаются вне зависимости от IP-адреса.
  - OpenSent – отправляется и получается сообщение OPEN. Если в полученном сообщении нет никаких ошибок, то происходит переход на этап OpenConfirm.
  - OpenConfirm – на этом этапе процесс BGP ждет получения от соседа сообщения KEEPALIVE или NOTIFICATION. Если был получен KEEPALIVE, то происходит переход на этап Established
  - Established – на этом этапе считается, что соседство установлено и пиры обмениваются друг с другом маршрутной информацией в виде UPDATE-сообщений

BGP – EBGP

- EBGP – защита от петель маршрутизации достигается за счет отслеживания наличия номера собственной автономной системы в AS_PATH получаемых маршрутов
- Обычно EBGP-сессии настраиваются на p2p-адресах
- TTL в BGP-пакетах по умолчанию равен 1
- На оборудовании Cisco и Huawei по умолчанию включается проверка на «прямое подключение» - адрес соседа должен быть в RIB как connected-маршрут (т.е., фактически, подключение должно быть на p2p-стыке.

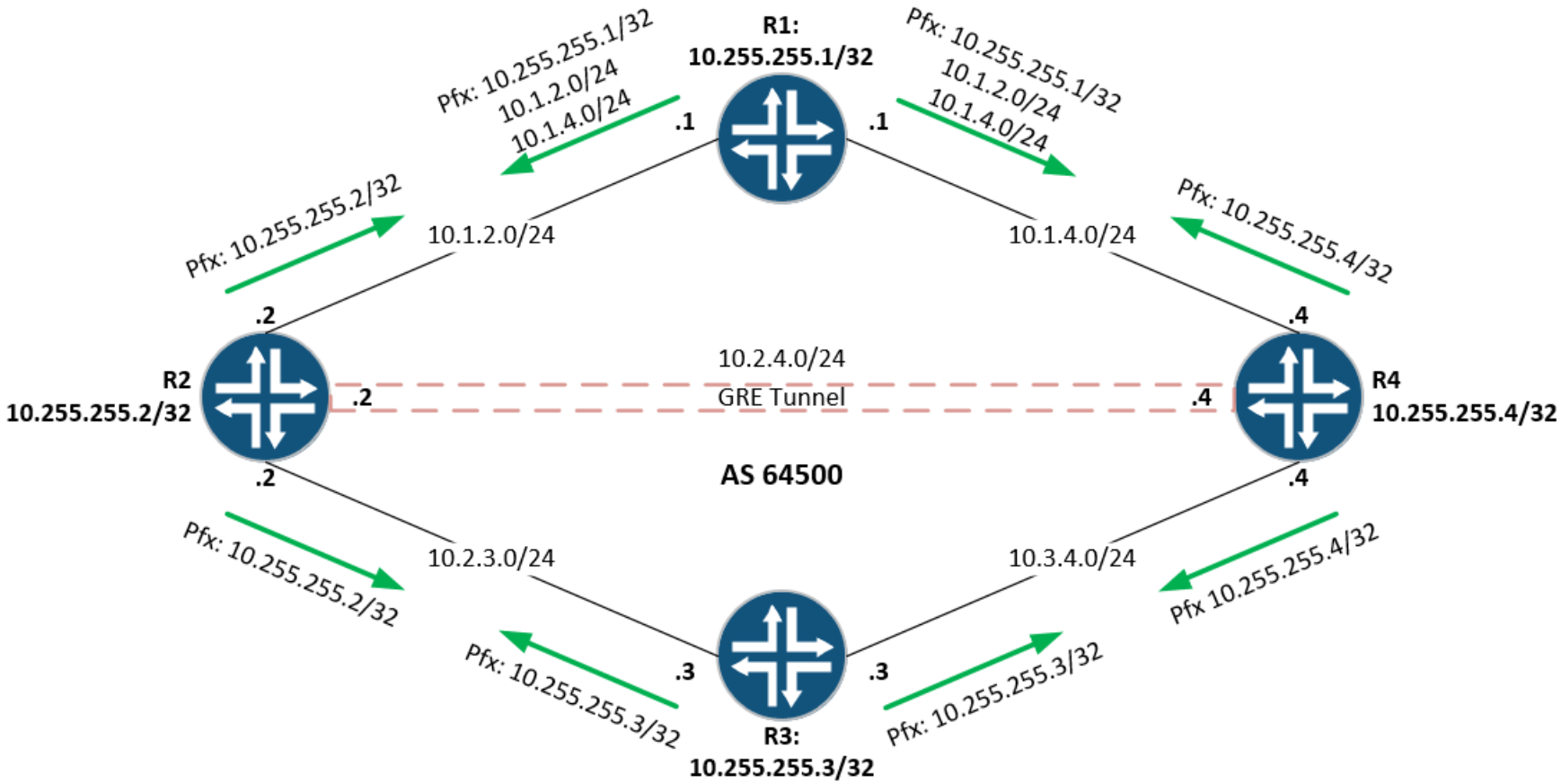
BGP – EBGP


Prefix: 1.1.1.1/32
AS_PATH: 100
Prefix: 1.1.1.1/32
AS_PATH: 300 200 100
Prefix: 1.1.1.1/32
AS_PATH: 200 100
Prefix: 1.1.1.1/32
AS_PATH: 200 100
.2
.2
192.0.2.0/30
198.51.100.0/30
R2
AS 200
.1
.1
1.1.1.1/32
R1
AS 100
R3
AS 300
.5
.5
Prefix: 1.1.1.1/32
AS_PATH: 400 100
Prefix: 1.1.1.1/32
AS_PATH: 100
Prefix: 1.1.1.1/32
AS_PATH: 300 200 100
Prefix: 1.1.1.1/32
AS_PATH: 400 100
192.0.2.4/30
198.51.100.4/30
R4
AS 400
.6
.6

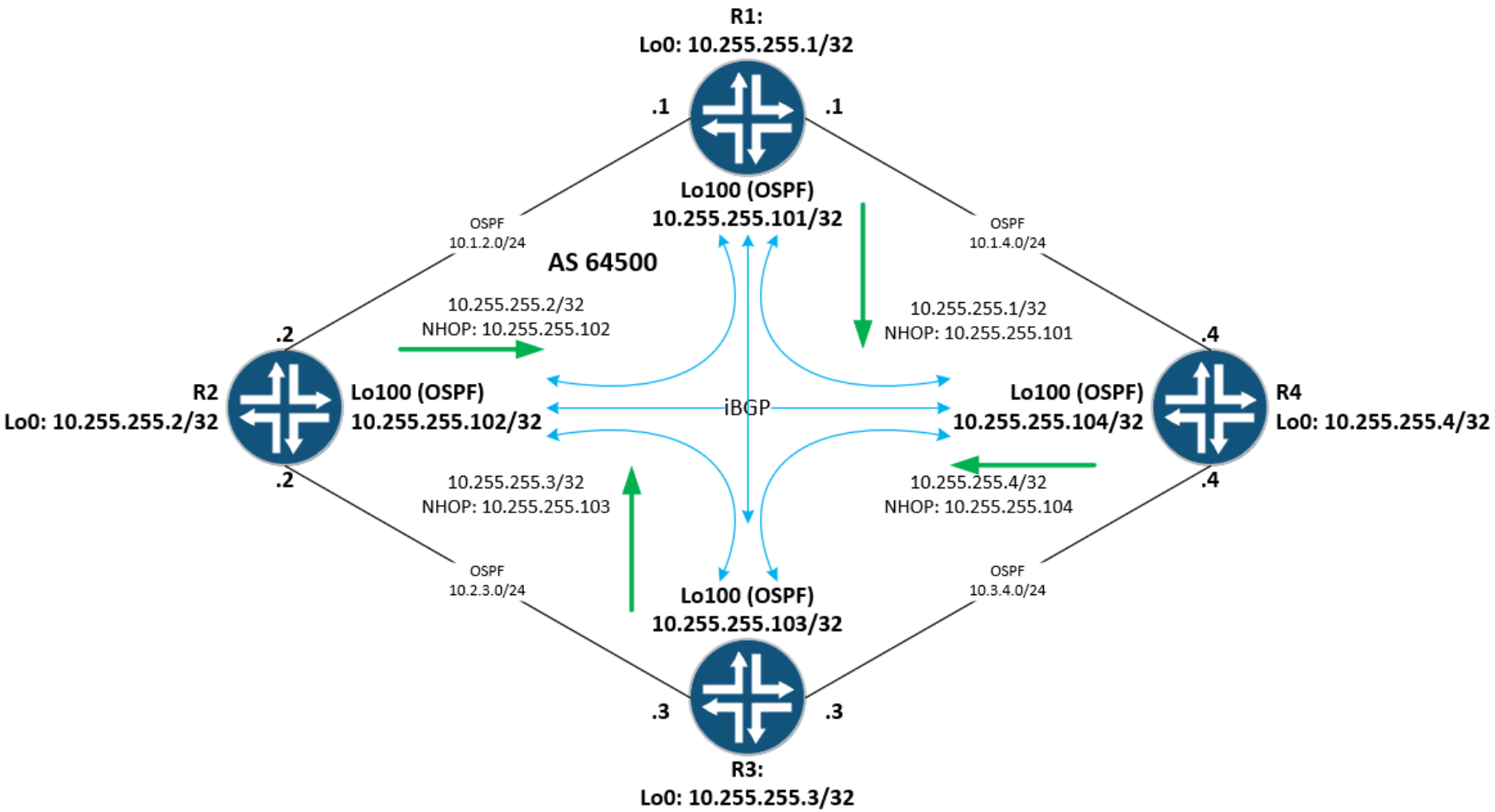
BGP – IBGP

- IBGP – защита от петель маршрутизации достигается за счет отказа от передачи полученных от IBGP-соседа префиксов другим соседям, т.о. приходя к необходимости FullMesh-связности между всеми участниками IBGP-связностей
- TTL в IBGP-пакетах по умолчанию равен 255
- Не производится «connected-check»
- Обычно используется в паре с каким-нибудь IGP и сессии поднимаются между лупбэками

BGP – IBGP


R1:
10.255.255.1/32
Pfx: 10.255.255.1/32
10.1.2.0/24
10.1.4.0/24
Pfx: 10.255.255.1/32
10.1.2.0/24
10.1.4.0/24
.1
.1
Pfx: 10.255.255.2/32
10.1.2.0/24
10.1.4.0/24
Pfx: 10.255.255.4/32
.2
.4
R2
10.255.255.2/32
10.2.4.0/24
GRE Tunnel
.2
.4
R4
10.255.255.4/32
AS 64500
.2
.4
Pfx: 10.255.255.2/32
10.2.3.0/24
10.3.4.0/24
Pfx 10.255.255.4/32
Pfx: 10.255.255.3/32
Pfx: 10.255.255.3/32
.3
.3
R3:
10.255.255.3/32

BGP – IBGP


R1:
Lo0: 10.255.255.1/32
.1
.1
Lo100 (OSPF)
10.255.255.101/32
OSPF
10.1.2.0/24
OSPF
10.1.4.0/24
AS 64500
10.255.255.2/32
NHOP: 10.255.255.102
10.255.255.1/32
NHOP: 10.255.255.101
.2
.4
R2
Lo0: 10.255.255.2/32
Lo100 (OSPF)
10.255.255.102/32
iBGP
Lo100 (OSPF)
10.255.255.104/32
R4
Lo0: 10.255.255.4/32
.2
.4
10.255.255.3/32
NHOP: 10.255.255.103
10.255.255.4/32
NHOP: 10.255.255.104
OSPF
10.2.3.0/24
OSPF
10.3.4.0/24
Lo100 (OSPF)
10.255.255.103/32
.3
.3
R3:
Lo0: 10.255.255.3/32

BGP – Атрибуты

- Атрибуты делятся на 4 типа:
  - Well-known mandatory – обязательно включаются во все BGP UPDATE
  - Well-known discretionary – необязательно включается в BGP UPDATE
  - Optional transitive – при получении атрибут передается дальше, другим пирам (вне зависимости от возможности обработки)
  - Optional non-transitive – при получении атрибут игнорируется и не передается другим пирам.
- Все Well-known атрибуты каждый BGP-маршрутизатор обязан распознавать и обрабатывать при получении, и передавать дальше.
- Optional атрибуты могут и не поддерживаться BGP-маршрутизатором

BGP – Атрибуты

- Список атрибутов, определенных RFC 4271:

Attribute	EBGP	IBGP
ORIGIN	Mandatory	Mandatory
AS_PATH	Mandatory	Mandatory
NEXT_HOP	Mandatory	Mandatory
MULTI_EXIT_DISC	Discretionary	Discretionary
LOCAL_PREF	N/A	Mandatory
ATOMIC_AGGREGATE	Discretionary	Discretionary
AGGREGATOR	Discretionary (Transitive)	Discretionary (Transitive)

BGP – Атрибуты

- ORIGIN – «происхождение» маршрута. Бывает трех* видов:
  - IGP – NLRI сгенерирован внутри AS (фактически, означает, что сеть была анонсирована силами самого процесса BGP через команду network)
  - EGP – NLRI получена от протокола EGP
  - Incomplete – NLRI помещена в BGP RIB любыми другими методами (фактически речь идет о редистрибьюции из других протоколов маршрутизации)
- AS_PATH – список номеров AS, через который «прошел» префикс. Внутри данного атрибута список номеров может принимать одно из двух значений:
  - AS_SET – неупорядоченный список номеров AS
  - AS_SEQUENCE – упорядоченный список номеров AS

BGP – Атрибуты

- MULTI_EXIT_DISC – оно же MED. Используется для управления входящим трафиком на нескольких стыках с одной AS. Чем меньше значение, тем приоритетней путь. Является достаточно «слабым» атрибутом.
  Данный атрибут передается соседу, может быть распространен внутри его AS, но дальше не выходит (это нетранзитивный атрибут)
- LOCAL_PREF – атрибут, определяющий приоритет маршрута. Передается внутри автономной системы, но никогда не отдается наружу*. Даже если данный атрибут был передан в EBGP-сессии, он должен быть проигнорирован принимающей стороной*

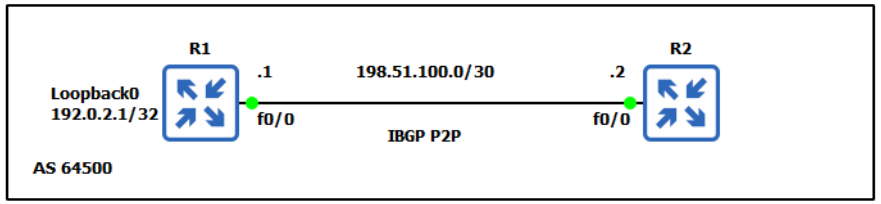
BGP – Атрибуты

- ATOMIC_AGGREGATE – атрибут, который явно объявляет, что маршрут появился в результате суммирования и добавляется в том случае, если из AS_PATH агрегата были удалены какие-либо составляющие (у Cisco, например, по умолчанию в AS_PATH агрегата не добавляется ничего)
- AGGREGATOR – атрибут, который может быть добавлен в информацию об агрегате. Содержит в себе данные об устройстве, сгенерировавшему агрегат: Router-ID устройства и номер его AS.

BGP – Атрибуты – NEXT_HOP

- NEXT_HOP – атрибут, определяющий IP-адрес маршрутизатора, через который доступен тот или иной префикс. В зависимости от типа соединения, может определяться следующим образом:
  - IBGP – маршрут, сгенерированный локально, будет передаваться соседу с NEXT_HOP равным IP-адресом интерфейса, используемым для установления сессии. Маршруты полученные от других пиров передаются с NEXT_HOP без изменения (т.е. NEXT_HOP = IP-адрес source-interface источника маршрута)
  - EBGP – по умолчанию проставляется IP-адрес интерфейса, который используется для установки соседства, однако есть и весьма специфичные варианты, предполагающие, что у пира есть подключенные сети, которые пересекаются с оригинальными next-hop маршрутов

BGP – Атрибуты – NEXT_HOP (IBGP)

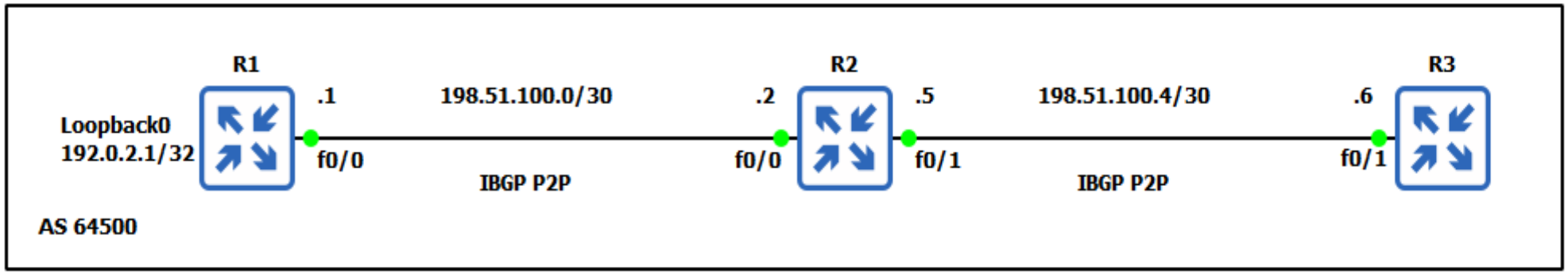



```
R1#show bgp ipv4 unicast | b Network
     Network          Next Hop            Metric LocPrf Weight Path
 *>  192.0.2.1/32     0.0.0.0                  0         32768 i

R2#show bgp ipv4 unicast | b Network
     Network          Next Hop            Metric LocPrf Weight Path
 *>i 192.0.2.1/32     198.51.100.1             0    100      0 i


R2#show ip route bgp | b Gateway
Gateway of last resort is not set

      192.0.2.0/32 is subnetted, 1 subnets
B        192.0.2.1 [200/0] via 198.51.100.1, 00:03:29
```

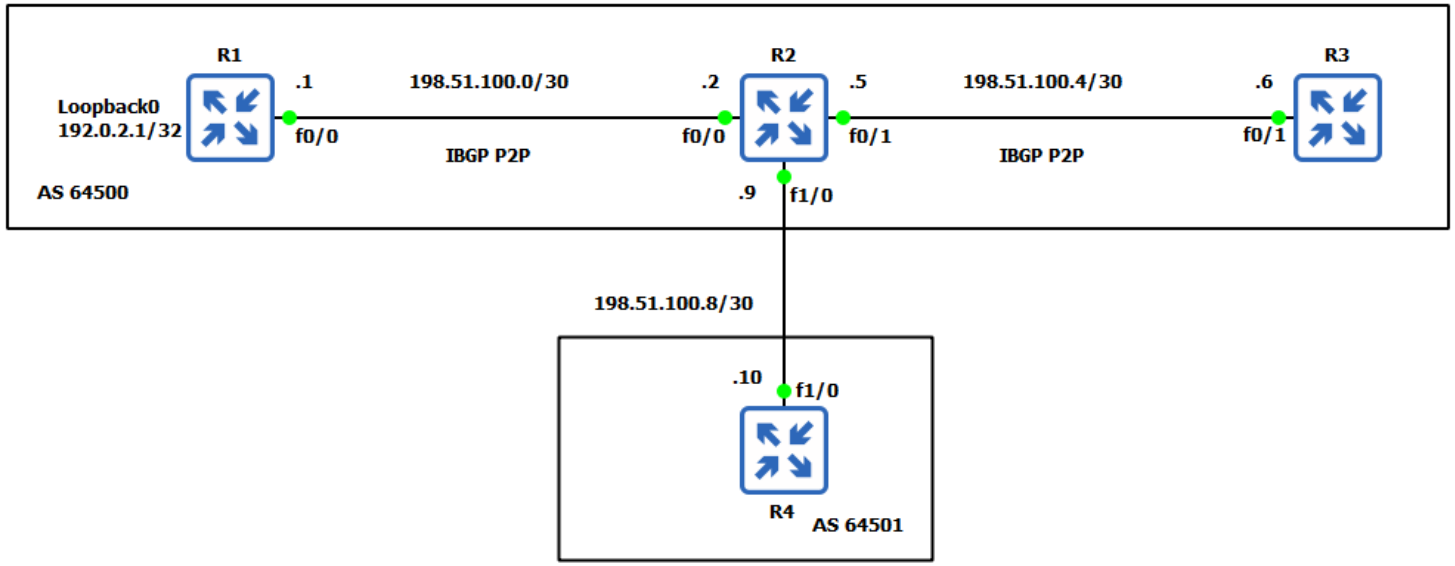



```
R3#show bgp ipv4 unicast | b Network
     Network          Next Hop            Metric LocPrf Weight Path
 * i 192.0.2.1/32     198.51.100.1             0    100      0 i
```

BGP – Атрибуты – NEXT_HOP (EBGP)

```
R2#show bgp ipv4 unicast | b Network
     Network          Next Hop            Metric LocPrf Weight Path
 *>i 192.0.2.1/32     198.51.100.1             0    100      0 i
 *>  203.0.113.4/32   198.51.100.10            0             0 64501 i

R1#show bgp ipv4 unicast | b Network
     Network          Next Hop            Metric LocPrf Weight Path
 *>  192.0.2.1/32     0.0.0.0                  0         32768 i
 * i 203.0.113.4/32   198.51.100.10            0    100      0 64501 i
```




```
R4#show bgp ipv4 unicast | b Network
     Network          Next Hop            Metric LocPrf Weight Path
 *>  192.0.2.1/32     198.51.100.9                           0 64500 i
 *>  203.0.113.4/32   0.0.0.0                  0         32768 i
```

BGP – Атрибуты – Community

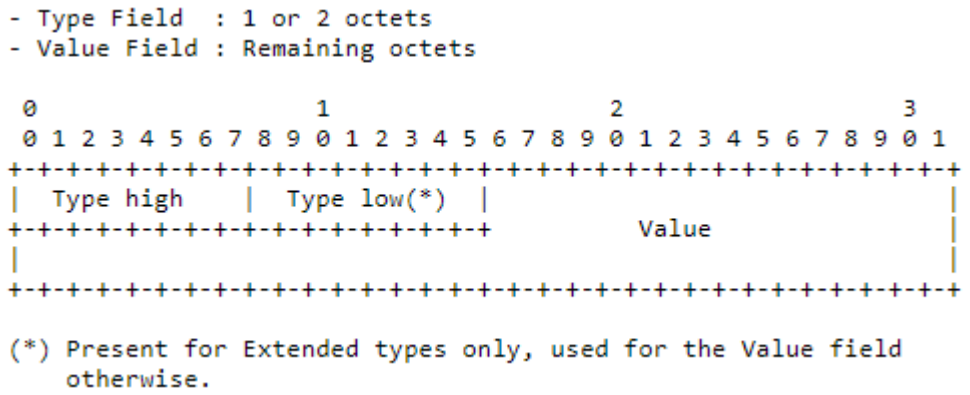
- Community – опциональный транзитивный атрибут. Согласно определения из RFC 1997, является группой маршрутов, имеющих общий признак. На данный момент есть двух видов:
  - Standard – описаны в RFC 1997, представляют собой 32-битное значение.
  - Extended – описаны в RFC 4360, представляют собой 64-битное значение.
- Фактически, используются для управления трафиком в том или ином виде (фильтрация, изменение атрибутов, тегирование и т.п.)
- По умолчанию данный атрибут не передается

BGP – Атрибуты – Standard Community

- Standard community
  - Размер: 32 бита
  - Есть зарезервированные значения:
    - 0x00000000 – 0x0000FFFF (0:0 – 0:65535)
    - 0xFFFF0000 – 0xFFFFFFFF (65535:0 – 65535:65535)
- Well-known communities – группа стандартных community, значения которых и действия, выполняемые процессом BGP, заранее определены в RFC 1997:
  - NO_EXPORT (0xFFFFFF01 – 65535:65281) – все маршруты, отмеченные данным коммьюнити, должны быть исключены из анонсов за пределы автономной системы/конфедерации (т.е. маршруты не анонсируются eBGP-соседям, но анонсируются iBGP-соседям и между субконфедерациями)
  - NO_ADVERTISE (0xFFFFFF02 – 65535:65282) – все маршруты, отмеченные данным коммьюнити, исключаются из анонсов любым соседям
  - NO_EXPORT_SUBCONFED (0xFFFFFF03 – 65535:65283) - все маршруты, отмеченные данным коммьюнити, должны быть исключены из анонсов за пределы автономной системы и/или субконфедерации (т.е. маршруты будут передавать только внутри одной AS)

BGP – Атрибуты – Extended Community

- Extended Community
  - Размер: 64 бита
  - Определены в RFC 4360
  - Появилось поле Type, позволяющее структурировать использование community



- Широко используемые extended community:
  - Route Target – используются для организации MPLS VPN, позволяют определить в какую именно таблицу необходимо поместить маршрут (см. RFC 4364)
  - Route Origin – используются для указания места происхождения маршрута (Site of Origin) в организации MPLS VPN (см. RFC 4364)

BGP – Адресные семейства

- RFC 2858 описал возможность использования MP-BGP (Multi-protocol BGP). Впоследствии этот RFC был заменен RFC 4760.
- Позволяет в рамках одной TCP-сессии передавать маршрутную информацию разных адресных семейств.
- Для разделения маршрутной информации используются атрибуты MP_REACH_NLRI и MP_UNREACH_NLRI, в которых, кроме самих маршрутов, передаются и значения AFI и SAFI
- AFI – Address-family identifier – идентификатор основного адресного семейства (IPv4, IPv6, L2VPN и т.п.)
- SAFI – Subsequent address-family identifier – идентификатор подсемейства основного адресного семейства (н-р, для IPv4 – это unicast, multicast и т.п.)
- Пара AFI/SAFI позволяет однозначно определить адресное семейство, к которому относится тот или иной анонс
- При использовании BGP в vrf используется пара AFI/SAFI соответствующая основному семейству (н-р, для address-family ipv4 vrf TEST_VRF пара AFI/SAFI будет 1/1).
- MP_REACH_NLRI используется для передачи маршрутной информации
- MP_UNREACH_NLRI используется для отзыва маршрутной информации

BGP – Адресные семейства

Address family	AFI	SAFI
IPv4 Unicast	1	1
IPv4 Labeled Unicast	1	4
IPv4 Multicast	1	2
VPNv4 Unicast	1	128
IPv6 Unicast	2	1
IPv6 Labeled Unicast	2	4
IPv6 Multicast	2	2
VPNv6 Unicast	2	128
L2VPN EVPN	25	70
L2VPN VPLS (Kompella)	25	65

MP_REACH_NLRI	
Field Name	**Length (bytes)**
Address Family Identifier	2
Subsequent Address Family Identifier	1
Length of Next Hop Network Address	1
Network Address of Next Hop	Variable
Reserved	1
Network Layer Reachability Information	Variable

MP_UNREACH_NLRI	
Field Name	**Length (bytes)**
Address Family Identifier	2
Subsequent Address Family Identifier	1
Withdrawn Routes	Variable

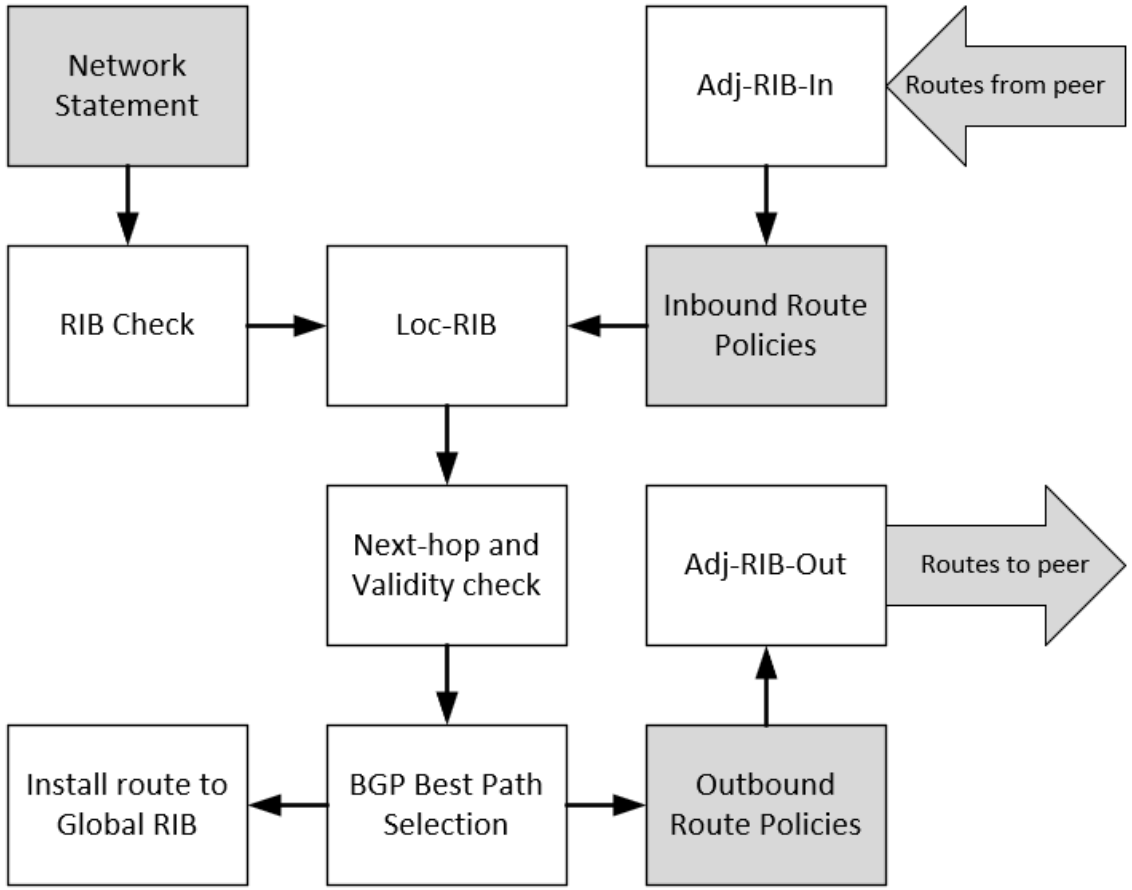
BGP – Обработка маршрутов

- BGP умеет производить те или иные манипуляции с маршрутной информацией, а именно:
  - Фильтрация получаемых префиксов:
  - Фильтрация анонсируемых префиксов
  - Изменение атрибутов и выполнение каких-либо действий

BGP – Обработка маршрутов

- Манипуляции с маршрутами могут производиться на основании:
  1. Содержимое AS_PATH
  2. Содержимое NLRI (за счет применения ACL или prefix-list)
  3. Атрибута Community
  4. Атрибута Next-hop
  5. Прочие атрибуты
- Что можно сделать:
  1. Запретить получение/анонс маршрута
  2. Изменить за счет prepend длину AS_PATH
  3. Добавить/изменить/убрать коммьюнити
  4. Изменить next-hop
  5. Прочие манипуляции с атрибутами

BGP – Обработка маршрутов



BGP – Обработка маршрутов

- Основным местом хранения маршрутной информации является Loc-RIB, в котором хранятся NLRI, как созданные локально, так и полученные от соседей после входной обработки (фильтрации).
- Созданные локально префиксы в Loc-RIB могут следующими путями (предварительно пройдя RIB Check):
  - Явно объявлены через network (origin type = IGP)
  - Через редистрибьюцию из внешнего источника (origin type = incomplete)
- Из Loc-RIB локальные префиксы могут быть переданы соседям по следующему алгоритму:
  - Префикс проходит Validity check (т.е. проверяется корректность NLRI и есть ли проставленный next-hop в RIB)
  - Префикс проходит через фильтрацию и/или применение политик
  - «Обработанный» префикс попадает в Adj-RIB-Out, откуда уже будет передан соседу
- Полученная от соседа информация попадает в исходном виде в Adj-RIB-In
- Из Adj-RIB-In полученные маршруты передаются на обработку входными политиками (на этом этапе происходит фильтрация и/или модификация маршрута)
- «Обработанный» маршрут помещается в Loc-RIB, в котором проходит проверку на валидность (если вдруг проверка не пройдена, то NLRI остается в Loc-RIB, но больше никак не обрабатывается и не передается)
- Если маршрут прошел проверки, то определяется является ли этот маршрут наилучшим
- Если маршрут является наилучшим, то он инсталлируется в RIB и передается на обработку исходящими политиками
- После «выходной» обработки, NLRI попадает в Adj-RIB-Out, откуда уже будет отправлен соседу

BGP – Выбор наилучшего пути

- Для выбора наилучшего пути BGP использует полученные атрибуты, сравнивая их по порядку (не стоит забывать о том, что если маршрут не был признан валидным, то из сравнения он убирается, вне зависимости от наличия альтернативы).
- Список атрибутов, сравниваемых при выборе наилучшего пути, плюс-минус одинаков, за исключением того, что иногда могут применяться какие-то vendor-specific атрибуты (вроде Weight от Cisco)

BGP – Выбор наилучшего пути

1. Local Preference – один из основных атрибутов, работает и передается только внутри одной AS. Принимает значения в диапазоне 0 – 4 294 697 295
2. Local Originated – учитываем источник ориджинации маршрута:
   1) Advertised locally
   2) Aggregated locally
   3) Received by BGP peers
3. AIGP – Accumulated Interior Gateway Protocol – Опциональный нетранзитивный атрибут, применение требует включения поддержки с обеих сторон. Позволяет расчитывать метрику пути. Путь с любой AIGP-метрикой является более предпочтительным, чем путь совсем без AIGP. Путь с наименьшей AIGP-метрикой предпочтительней, чем путь с большей. AIGP обычно берется из IGP

BGP – Выбор наилучшего пути

4. AS_Path – один из основных атрибутов, сравнивается длина AS_PATH, чем меньше, тем лучше.
5. Origin type – один из основных атрибутов, сравнивается тип источника маршрута:
   1) IGP
   2) EGP
   3) Incomplete
6. MED – один из основных атрибутов. Принимает значение в диапазоне 0 – 4 294 697 295. Чем меньше, тем лучше.

BGP – Выбор наилучшего пути

7. “eBGP over IBGP” – еще одно сравнение по источнику маршрута. На этот раз смотрится на тип соседства, в котором был получен маршрут:
    1. eBGP Peers
    2. Confederation member AS Peers
    3. iBGP Peers
8. Lowest IGP Next-hop – проверяется метрика в IGP до адреса next-hop. Очевидно, чем меньше, тем лучше
9. Oldest eBGP Path – предпочитается более старый полученный маршрут
10. Router-ID – сравниваются router-id. У кого меньше, тот и лучше
11. Cluster list length – длина атрибута cluster-list. И опять: меньше – лучше
12. Neighbor Address – актуально для iBGP (как и предыдущий пункт). И как и в предыдущем пункте: меньше – лучше.

Источники

- RFC 4271 - https://datatracker.ietf.org/doc/html/rfc4271
- RFC 4360 - https://datatracker.ietf.org/doc/html/rfc4360
- RFC 4364 - https://datatracker.ietf.org/doc/html/rfc4364
- https://www.iana.org/assignments/iana-as-numbers-special-registry/iana-as-numbers-special-registry.xhtml
- https://www.iana.org/assignments/address-family-numbers/address-family-numbers.xhtml#address-family-numbers-2
- https://www.iana.org/assignments/safi-namespace/safi-namespace.xhtml#safi-namespace-2