BGP路由优选

前言

- BGP是一个应用非常广泛的边界网关路由协议，在全球范围内被大量部署。BGP定义了多种路径属性，并且拥有丰富的路由策略工具，这使得BGP在路由操控和路径决策上变得非常灵活。
- 针对BGP路由的各种属性的操作都可能影响路由的优选，从而对网络的流量产生影响，因此掌握BGP路由的优选规则十分重要。
- 本章节将会详细学习BGP路由的优选规则。

目标

- 学习完本课程后，你将能够：
  - 描述BGP路由优选规则
  - 实现BGP路由控制

目录

BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。

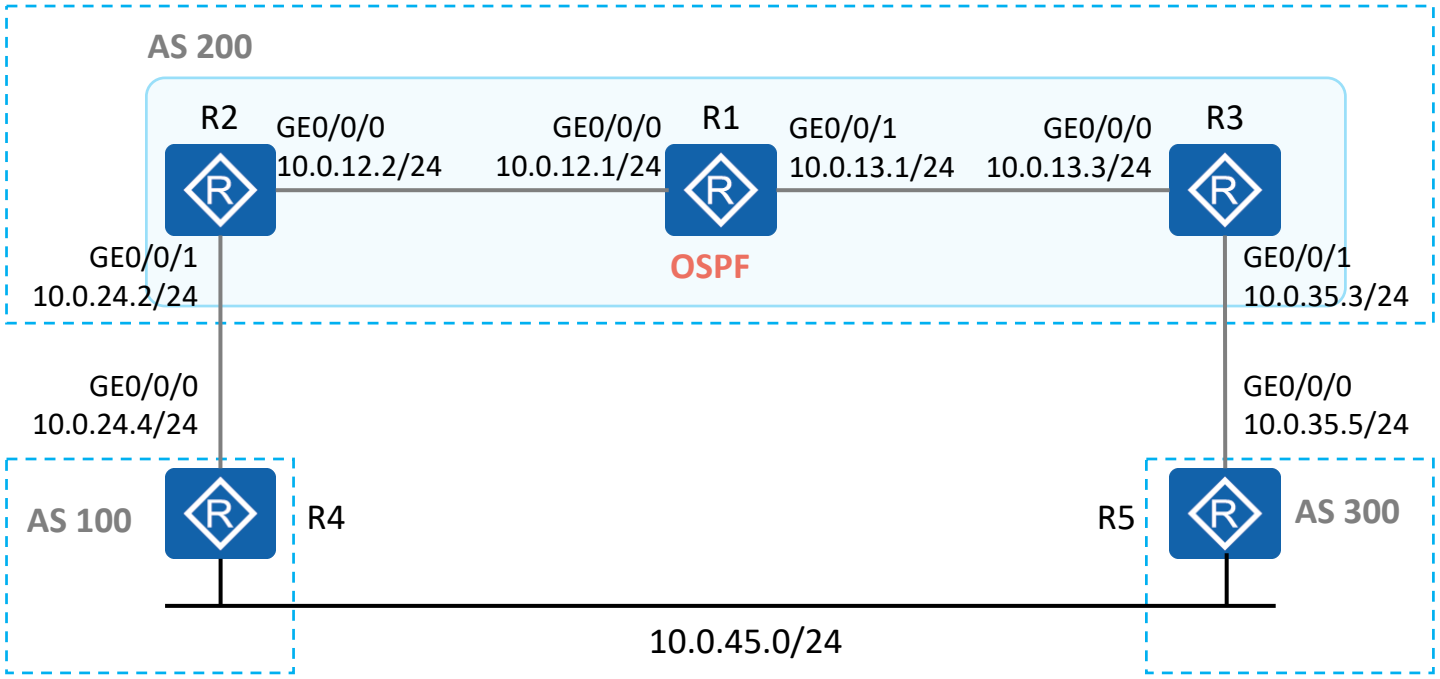
↑ 取值越大越优

3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

↓ 取值越小越优

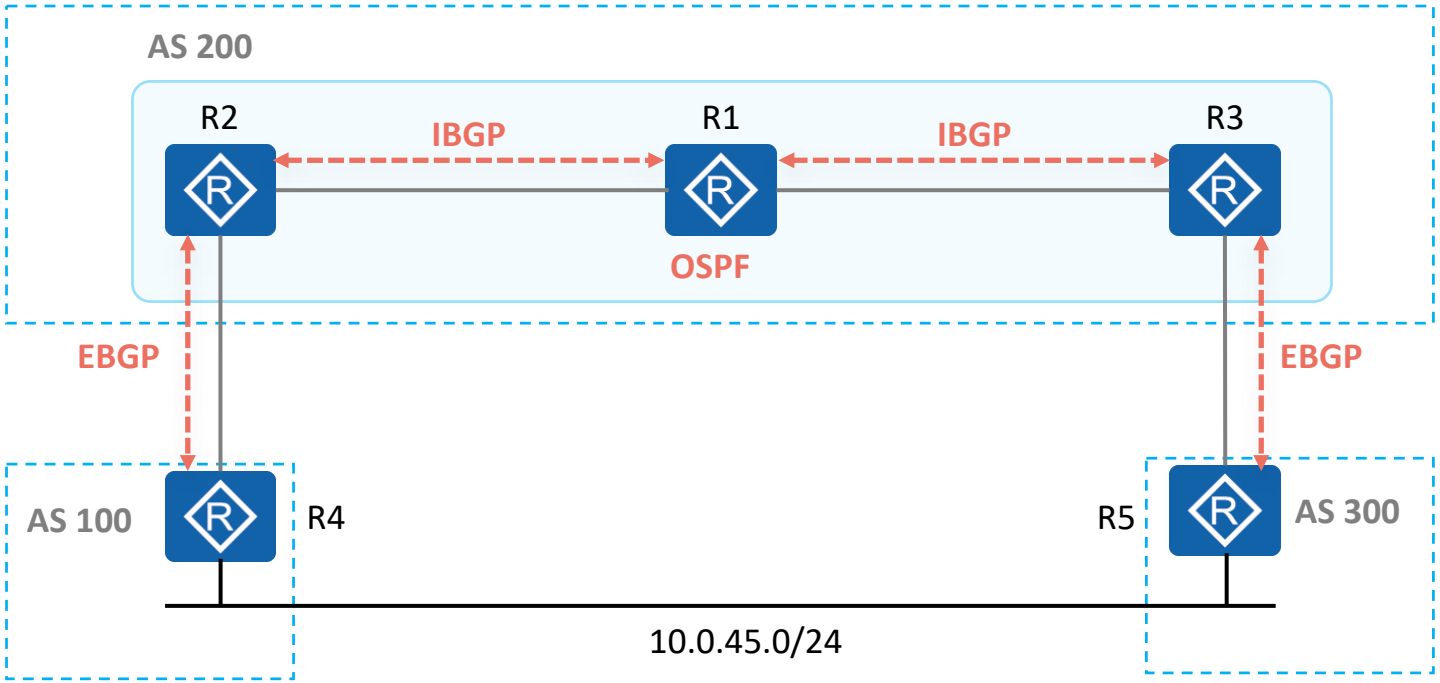
当前8条属性全部相同时可以形成路由负载分担

拓扑说明 (1)



- AS、设备互联地址如图所示，所有设备均创建Loopback0接口，IP地址为10.0.x.x（x为设备编号），所有设备使用环回口地址作为Router ID。
- AS200内运行OSPF，在内部互联接口（不包含连接外部AS的接口）、Loopback接口上激活OSPF。

拓扑说明 (2)



- AS内部基于Loopback0接口建立IBGP对等体关系，AS之间基于直连接口建立EBGP对等体关系。
- R4、R5上存在相同的网段：10.0.45.0/24，通过import-route命令将该网段的直连路由注入到BGP，用于验证BGP路由优选规则。

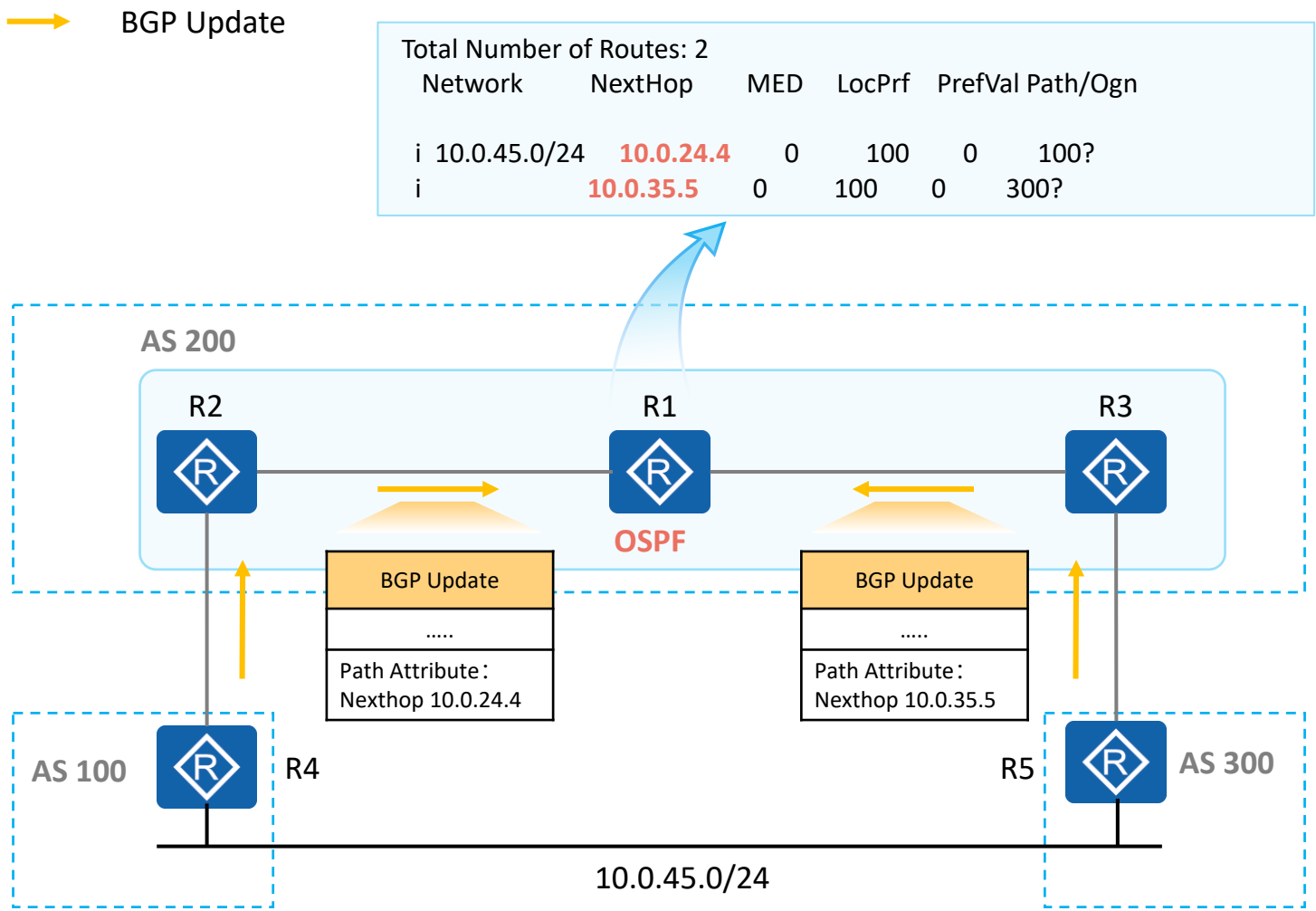
BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

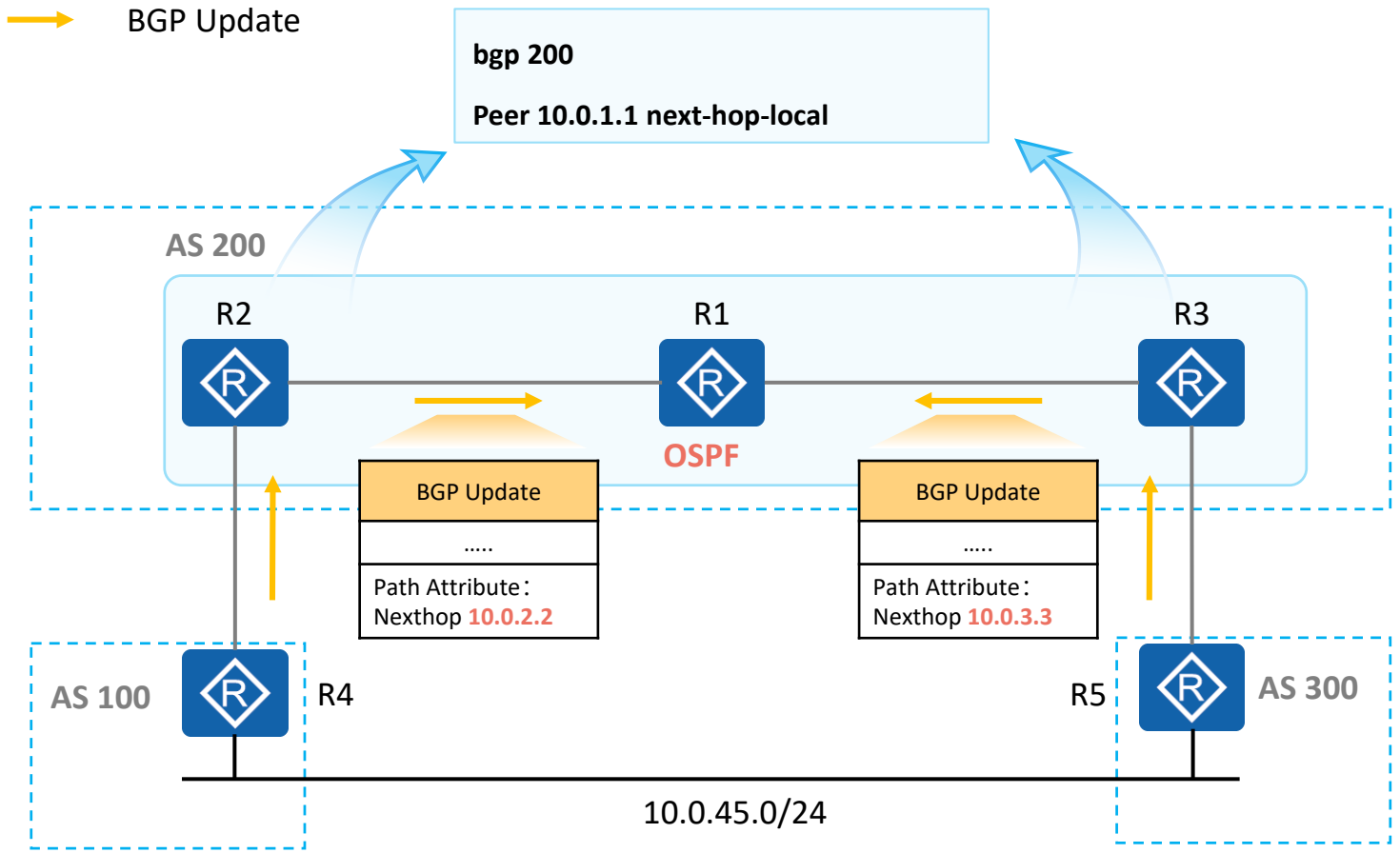
1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

丢弃下一跳不可达的路由 (1)



- R4、R5将BGP路由10.0.45.0/24通告给AS200时Next_Hop属性值为10.0.24.4、10.0.34.5。
- R2、R3将路由通告给R1时不修改Next_Hop属性值，R1学习到的两条BGP路由10.0.45.0/24下一跳为10.0.24.4、10.0.35.5。
- R1进行BGP路由下一跳迭代查询时，由于R2、R3未在连接外部AS的接口上激活OSPF，导致路由迭代失败，R1上的BGP路由10.0.45.0/24下一跳不可达。
- 在R1上通过**display bgp routing**查看BGP路由表，此时BGP路由10.0.45.0/24为非有效路由条目。

丢弃下一跳不可达的路由 (2)



- 在R2、R3上通过**next-hop-local**命令修改Next_Hop属性值为本地更新源地址。
- R2、R3向R1通告BGP路由时Next_Hop属性值将会变为：10.0.2.2、10.0.3.3。
- 这两个下一跳地址在R1上能够成功进行路由迭代，BGP路由的下一跳地址将会变成可达。

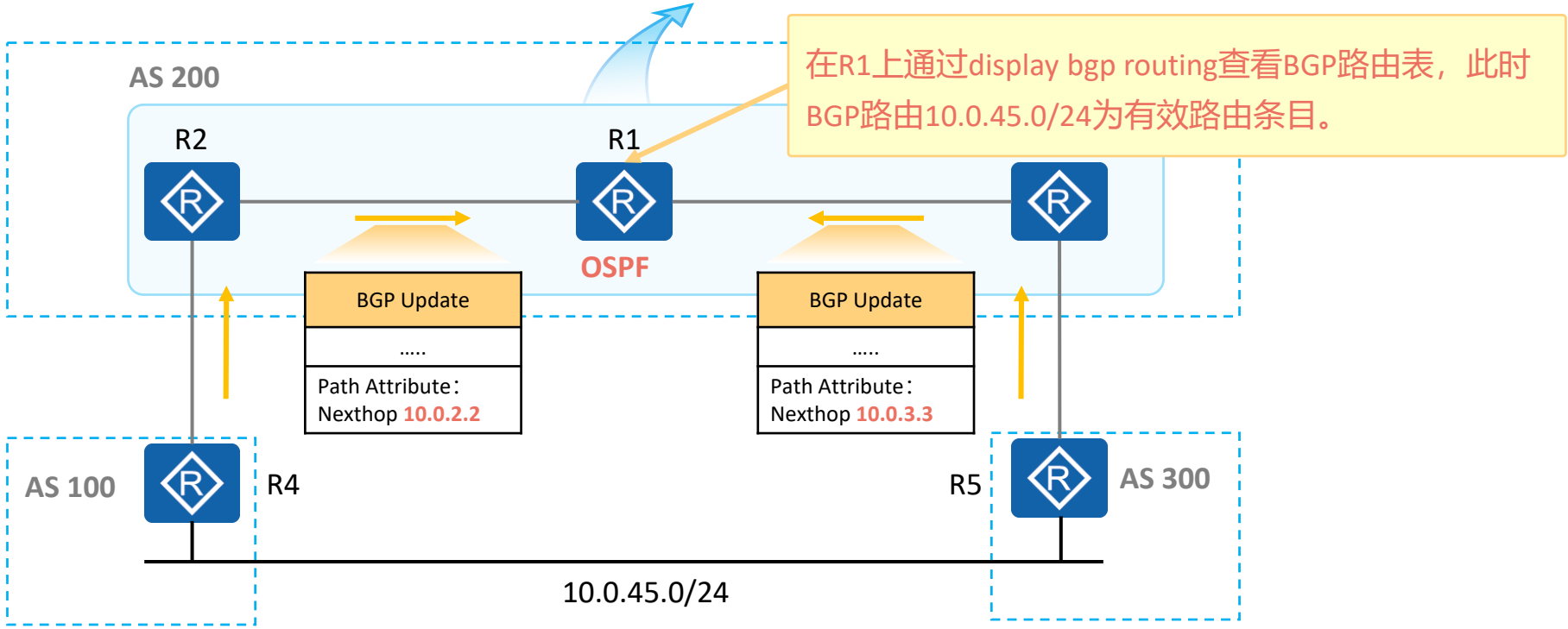
如无特殊说明，后续所有案例的初始配置都为基础配置加R2、R3开启了next-hop-local

丢弃下一跳不可达的路由 (3)

BGP Update

```
Total Number of Routes: 2
     Network         NextHop     MED    LocPrf    PrefVal Path/Ogn

 *>i 10.0.45.0/24    10.0.2.2      0       100       0     100?
 * i                 10.0.3.3      0       100       0     300?
```



两条BGP路由下一跳都可达的情况下，为什么下一跳为10.0.2.2的BGP路由为最优?

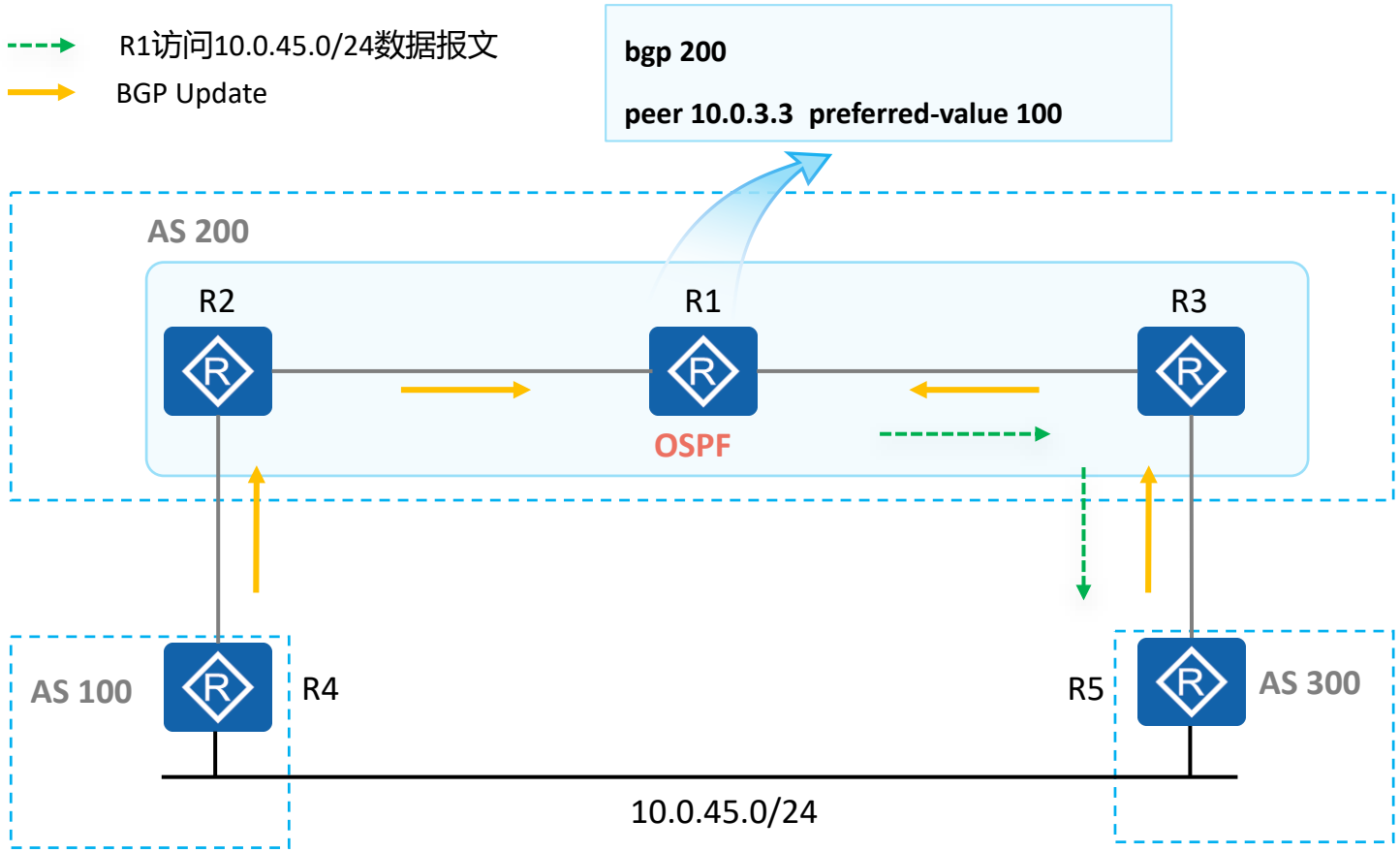
BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

1. **优选Preferred-Value属性值最大的路由。**
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

修改Preferred-Value



使用**preferred-value**命令修改R3通告的BGP路由其Preferred-Value为100，优于R2通告BGP路由的默认Preferred-Value，R1将会优选R3通告的BGP路由10.0.45.0/24。

查看R1 BGP路由表

```
[R1] display bgp routing-table
BGP Local router ID is 10.0.1.1
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete

 Total Number of Routes: 4
     Network            NextHop        MED        LocPrf     PrefVal  Path/Ogn
*>i  10.0.45.0/24       10.0.3.3       0                     100      300 i
* i                     10.0.2.2       0                     0        100 i
```

R3（10.0.3.3）通告的BGP路由拥有更高的Preferred-Value（100），因此R1将会优选R3通告的BGP路由10.0.45.0/24。

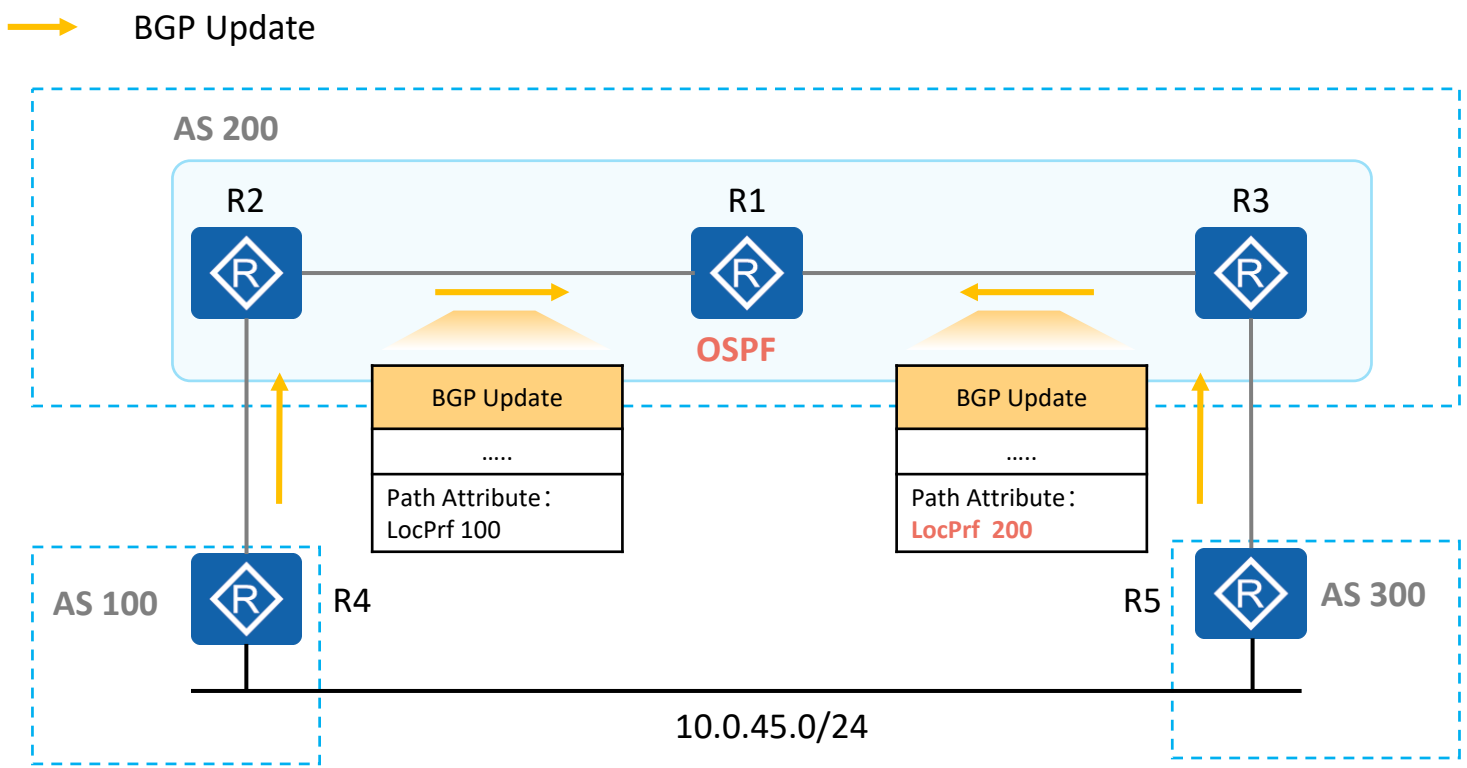
BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. **优选Local_Preference属性值最大的路由。**
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

修改Local_Preference (1)



R3上执行如下操作：

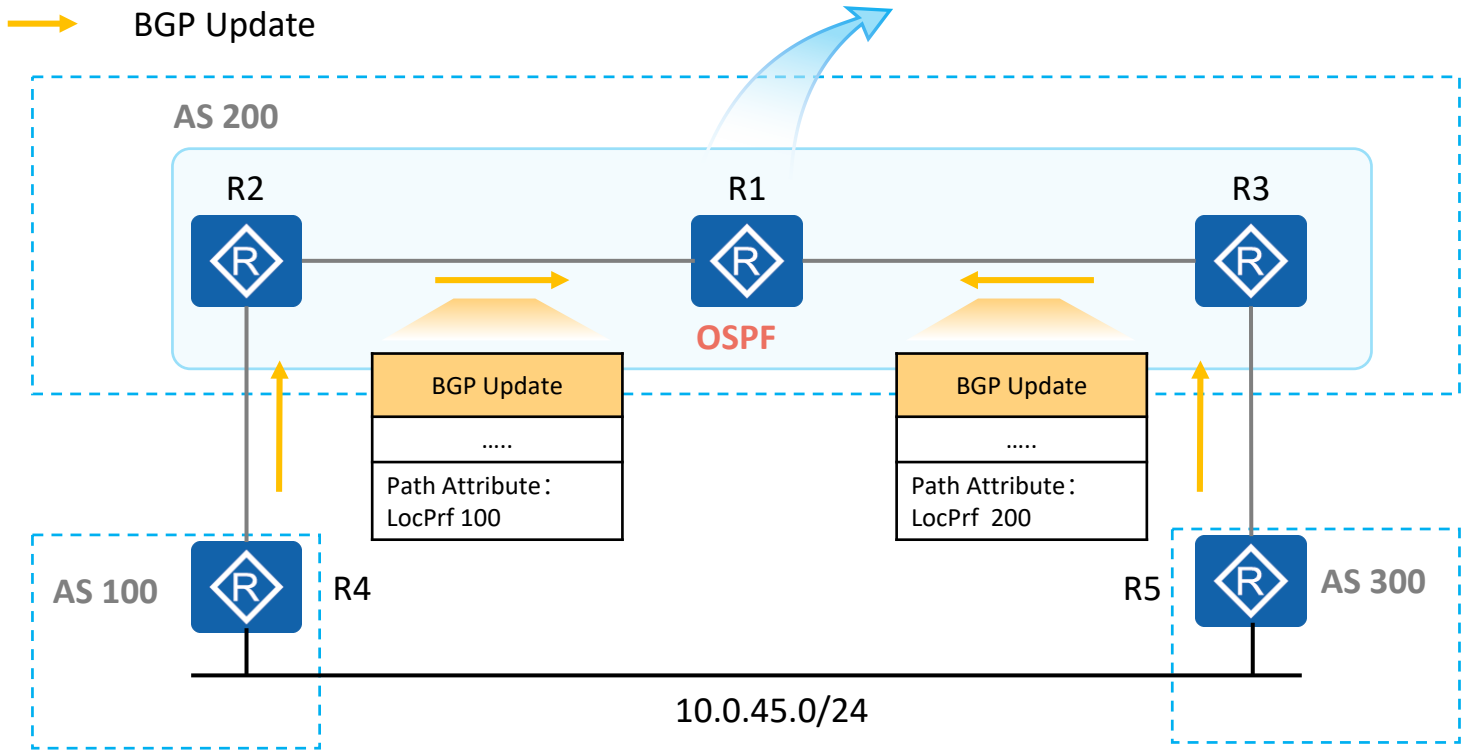
```
ip ip-prefix local_pref index 10 permit 10.0.45.0 24
#
route-policy local_pref permit node 10
 if-match ip-prefix local_pref
 apply local-preference 200
route-policy local_pref permit node 20
#
bgp 200
peer 10.0.1.1 route-policy local_pref export
```

R3上通过路由策略修改通告给R1的BGP路由10.0.45.0/24其Local_Preference属性值。

修改Local_Preference (2)

```
Total Number of Routes: 2
     Network         NextHop   MED    LocPrf    PrefVal Path/Ogn

*>i  10.0.45.0/24    10.0.3.3        0       200       0     300?
* i                  10.0.2.2        0       100       0     100?
```



下一跳可达、相同Preferred-Value的情况下将会比较Local_Preference，R3通告的BGP路由Local_Preference值为200，高于R2通告的BGP路由，R1将会优选R3通告的BGP路由。

BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

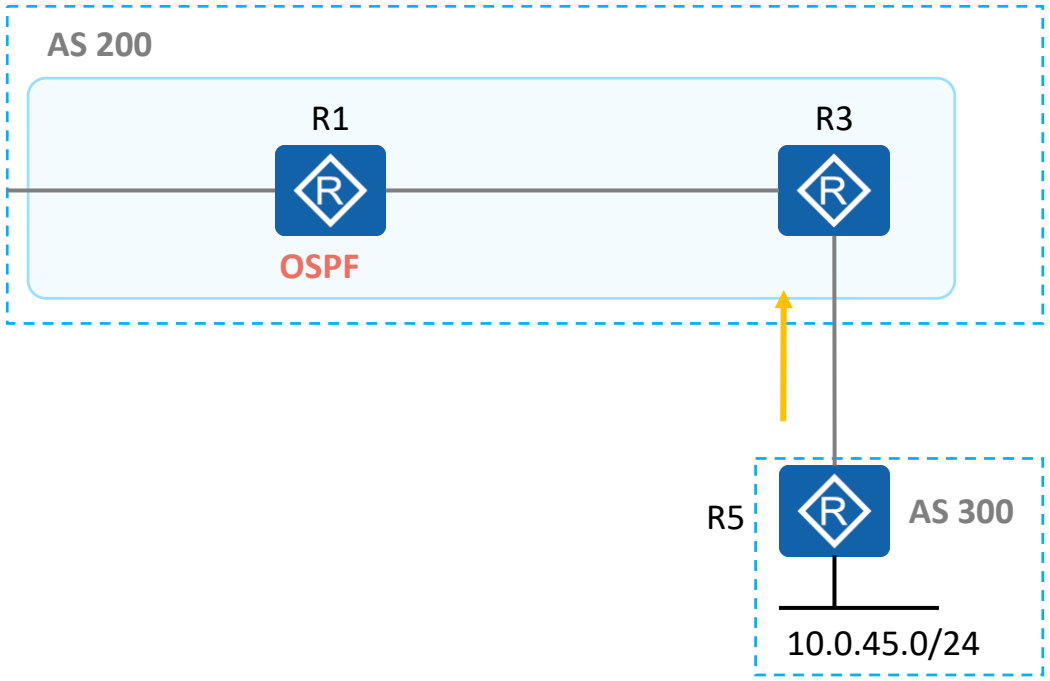
丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. **本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。**
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

本地优先

- 本条规则可以概括为在相同条件下，优选本地生成的路由，从对等体学习到的路由条目为次优。
- 同时本地生成的路由也可能存在多种途径，当本地存在多种途径学习到相同路由时，从高到低优先级如下：
  - 手动聚合：手动通过aggregate命令在BGP视图内聚合生成的聚合路由
  - 自动聚合：Summary automatic命令生成的自动聚合路由
  - Network方式注入的路由
  - Import-route方式注入的路由

手动聚合 (1)



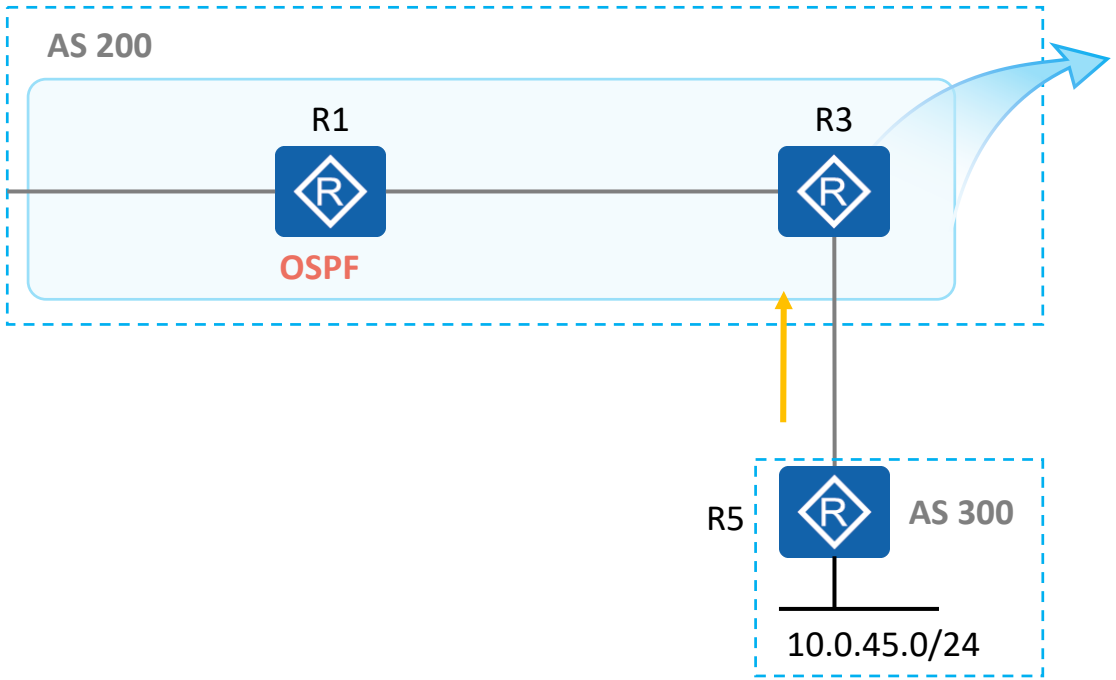
为了在R3上进行手动聚合，在R3上配置两条指向null0的静态路由，用于注入到BGP。

- R3上执行如下操作：

```
ip route-static 10.0.45.0 255.255.255.128 null0
ip route-static 10.0.45.128 255.255.255.128 null0
bgp 200
 aggregate 10.0.45.0 255.255.255.0 detail-suppressed
 import-route static
```

- R3上配置两条静态路由，将静态路由通过**import-route**注入到BGP，并通过**aggregate**命令进行手动聚合，同时增加关键字**detail-suppressed**抑制明细路由的对外通告。

手动聚合 (2)



	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.0.45.0/24	127.0.0.1			0	?
*		10.0.35.5	0		0	300?
s>	10.0.45.0/25	0.0.0.0	0		0	?
s>	10.0.45.128/25	0.0.0.0	0		0	?

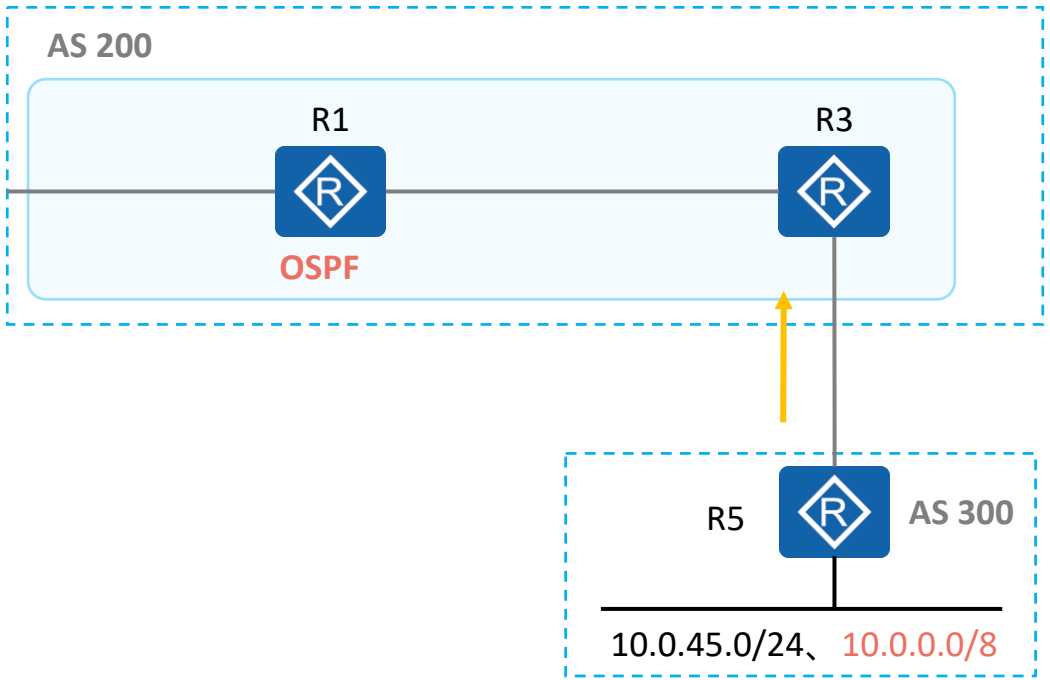
- R3上查看BGP路由表存在两条BGP路由10.0.45.0/24：
  - 本地产生的：静态路由注入到BGP中，由手动聚合产生
  - 对等体通告：由对等体R5（10.0.35.5）通告
- 在R3上这两条路由都不存在local_preference、Preferred-Value值，此时比较路由来源：手动聚合最优，R3将会优选本地手动聚合产生的BGP路由。

手动聚合 (3)

```
[R3]display bgp routing-table 10.0.45.0 24
 BGP local router ID : 10.0.3.3
 Local AS number : 200
 Paths:   2 available, 1 best, 1 select
 BGP routing table entry information of 10.0.45.0/24:
 Aggregated route.
 Route Duration: 00h00m14s
 Direct Out-interface: NULL0
 Original nexthop: 127.0.0.1
 Qos information : 0x0
 AS-path Nil, origin incomplete, pref-val 0, valid, local, best, select, active,
pre 255
 Aggregator: AS 200, Aggregator ID 10.0.3.3, Atomic-aggregate
 Advertised to such 2 peers:
   10.0.35.5
   10.0.1.1
```

- R3上通过**display bgp routing-table 10.0.45.0 24**查看BGP路由10.0.45.0/24的详细信息，存在两条有效路由，其中最优的为手动聚合产生的路由。
- 在本案例中我们验证了本地产生的BGP路由优于从对等体学习的BGP路由。

自动聚合 (1)



此时R1、R3、R5上的配置和手动聚合案例中已执行的配置无关。

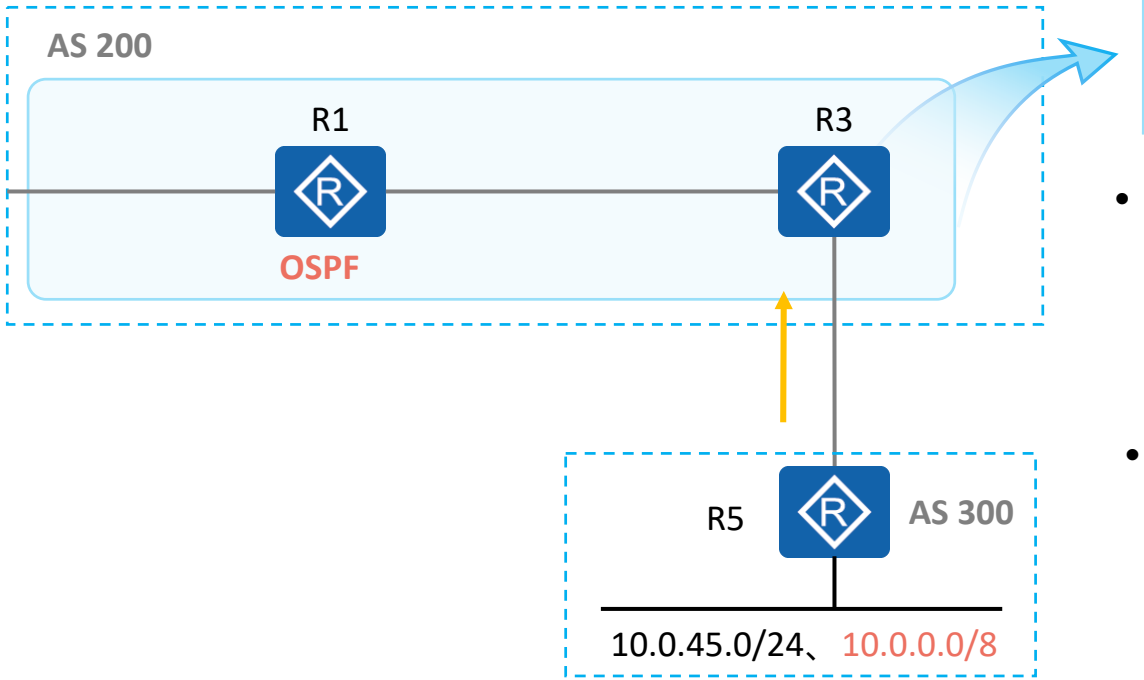
- R3上执行如下操作：

```
ip route-static 10.0.45.0 255.255.255.128 null0
ip route-static 10.0.45.128 255.255.255.128 null0

bgp 200
summary automatic
import-route static
```

- R3上配置两条静态路由，将静态路由通过**import-route**注入到BGP，并开启自动聚合，BGP将按照自然网段聚合路由（例如非自然网段A类地址10.1.1.1/24和10.2.1.1/24将聚合为自然网段A类地址10.0.0.0/8），并且BGP只向对等体通告聚合后的路由。
- 在R3上将会看到路由被聚合为10.0.0.0/8。
- R5上又注入了路由10.0.0.0/8，并通告给了R3。

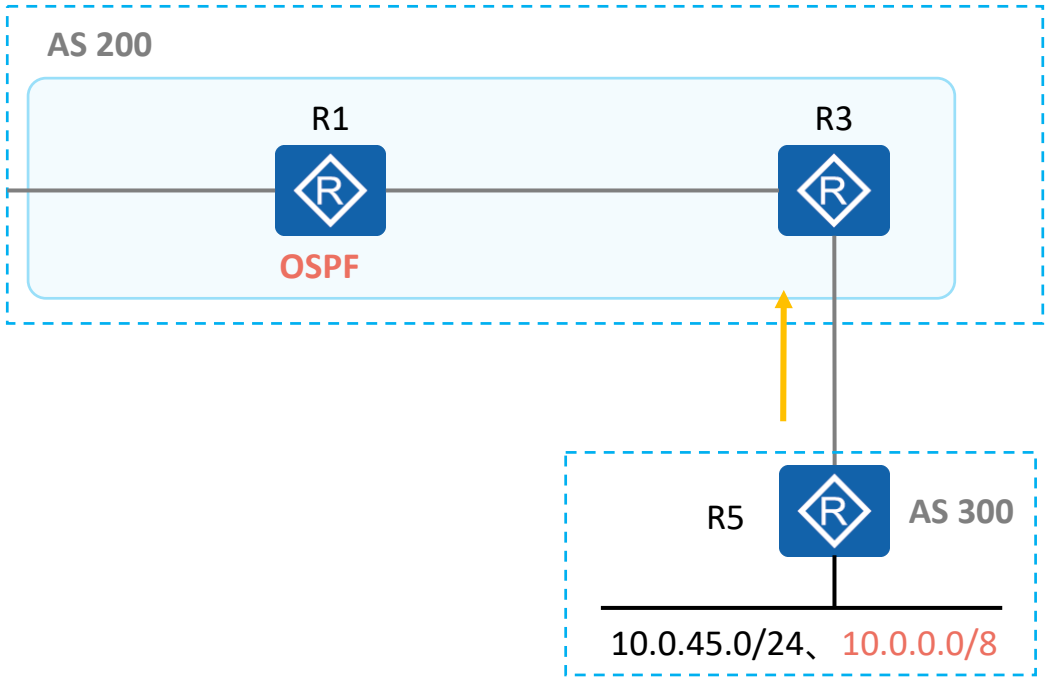
自动聚合 (2)



	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.0.0.0	127.0.0.1			0	?
*		10.0.35.5	0		0	300?

- R3上查看BGP路由表存在两条BGP路由10.0.0.0：
  - 本地产生：静态路由注入到BGP中，自动聚合产生
  - 对等体通告：由对等体R5（10.0.35.5）通告
- 在R3上这两条路由都不存在local_preference、Preferred-Value值，此时比较路由来源：本地产生优于从对等体学习到的，R3将会优选本地自动聚合产生的BGP路由。

自动聚合 (3)



- 在R3上执行手动聚合：

```
bgp 200
 aggregate 10.0.0.0 255.0.0.0 detail-suppressed
```

- 查看R3的BGP路由表

	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.0.0.0	127.0.0.1			0	?
*		127.0.0.1			0	?
*		10.0.35.5	0		0	300?

- 优选的依旧是本地产生的BGP路由，但是可以看到本地产生的BGP路由有两条，从该表项无法判断出优选的为手动聚合还是自动聚合产生的BGP路由。

自动聚合 (4)

BGP local router ID : 10.0.3.3

Local AS number : 200

Paths: 3 available, 1 best, 1 select

BGP routing table entry information of 10.0.0.0/8:

Aggregated route.

Route Duration: 00h08m17s

Direct Out-interface: NULL0

Original nexthop: 127.0.0.1

Qos information : 0x0

AS-path Nil, origin incomplete, pref-val 0, valid, local, best, select, active, pre 255

Aggregator: AS 200, Aggregator ID 10.0.3.3, Atomic-aggregate

Advertised to such 2 peers:

10.0.35.5

10.0.1.1

- R3上通过**display bgp routing-table 10.0.0.0** 查看BGP路由10.0.0.0/8的详细信息，存在三条有效路由，其中最优的条目由聚合产生，并且存在Atomic-aggregate属性，由此可以看出该聚合条目为手动聚合产生的条目。
- R3上相同的BGP聚合路由：手动聚合 > 自动聚合。
- 在该案例中我们验证了手动聚合产生的BGP路由优于自动聚合产生的BGP路由。

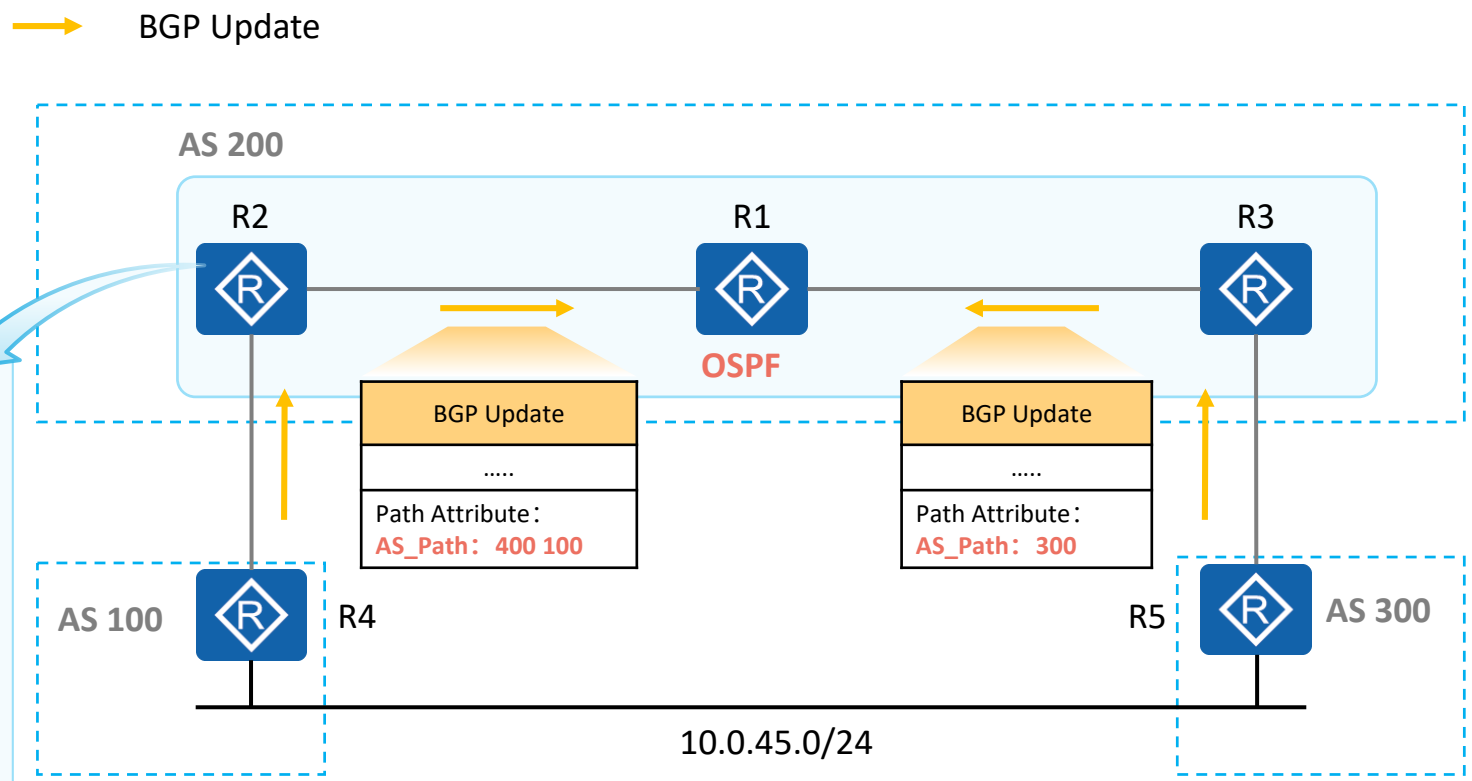
BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. **优选AS_Path属性值最短的路由。**
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

优选AS_Path最短 (1)




```
ip ip-prefix as_path index 10 permit 10.0.45.0 24
#
route-policy as_path permit node 10
 if-match ip-prefix as_path
 apply as-path 400 additive
route-policy as_path permit node 20
#
bgp 200
peer 10.0.1.1 route-policy local_pref export
```

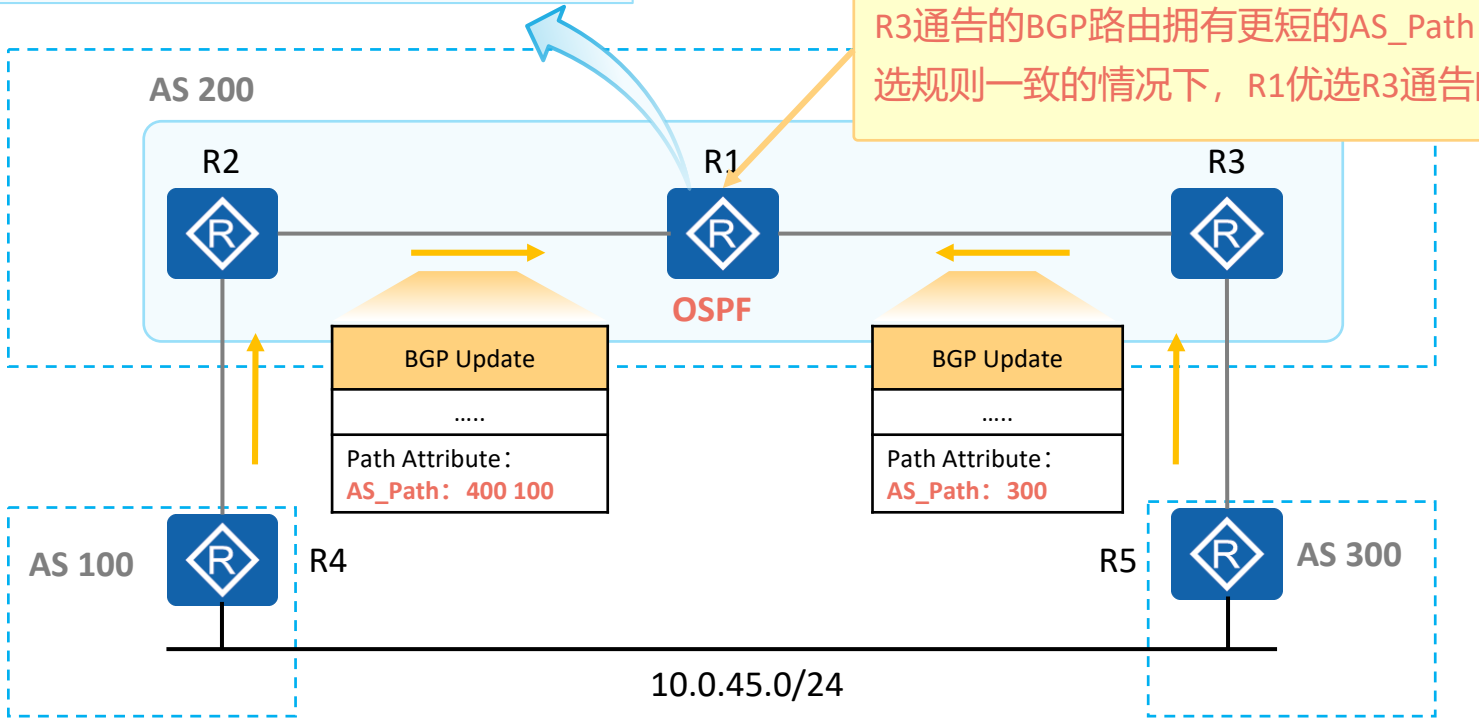
R2上通过路由策略修改通告给R1的BGP路由其AS_Path属性值。

优选AS_Path最短 (2)

```
Total Number of Routes: 2
     Network          NextHop   MED    LocPrf    PrefVal Path/Ogn

*>i  10.0.45.0/24     10.0.3.3        0         100       0     300?
* i                   10.0.2.2        0         100       0     400 100?
```

R3通告的BGP路由拥有更短的AS_Path，在前几条优选规则一致的情况下，R1优选R3通告的BGP路由。



BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

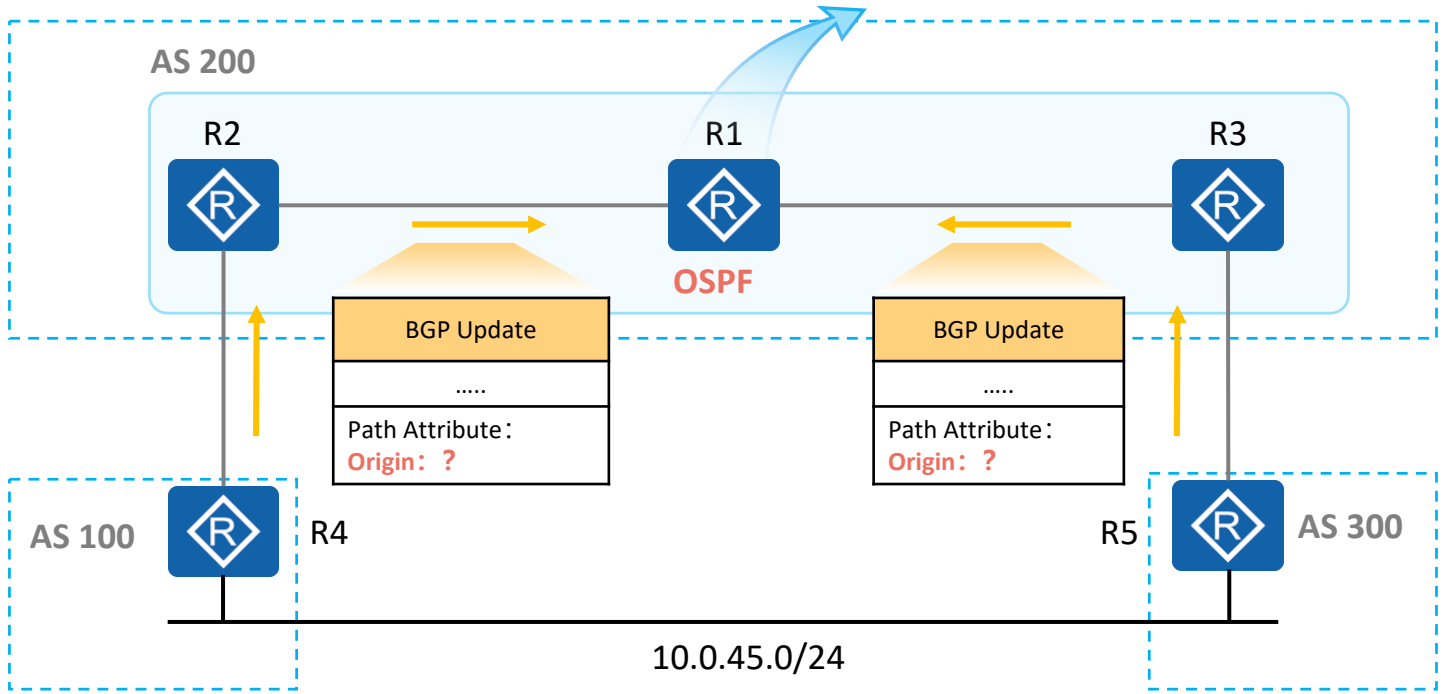
丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. **优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。**
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

Origin属性验证 (1)

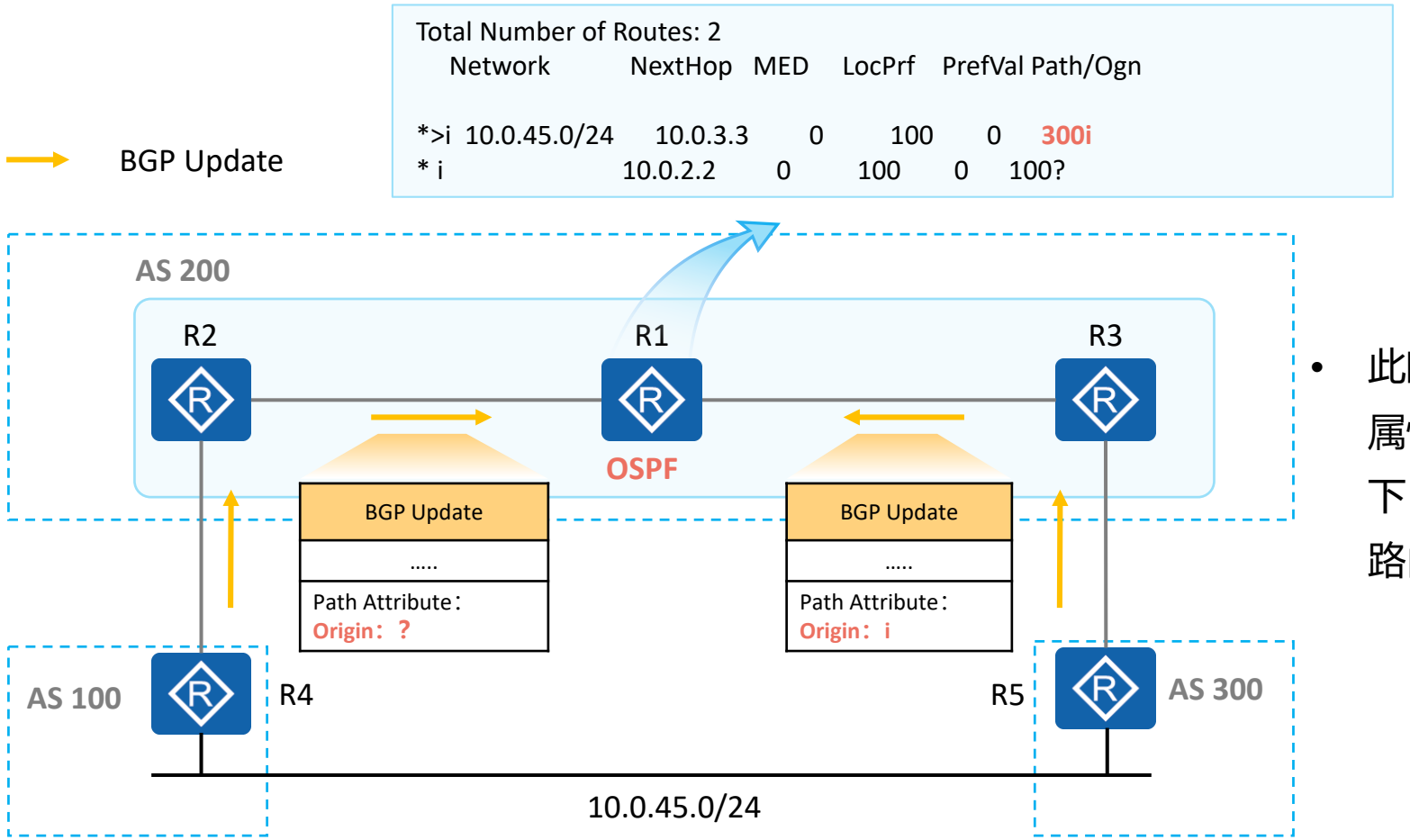
```
Total Number of Routes: 2
     Network         NextHop   MED    LocPrf    PrefVal Path/Ogn

*>i  10.0.45.0/24    10.0.2.2       0        100       0     100?
* i                  10.0.3.3       0        100       0     300?
```



- R4、R5上默认采用import-route方式将路由10.0.45.0/24注入到BGP，R1的BGP路由表中两条BGP路由10.0.45.0/24其Origin属性都是“？”，此时R1优选R4注入的BGP路由。
- 在R5上修改注入路由的方式为network
- 之后在R1上再次查看BGP路由表。

Origin属性验证 (2)



- 此时R5注入的BGP路由10.0.45.0/24其Origin属性为“i”，在前几条优选规则相同情况下，起源类型为“i”的BGP路由成为优选路由。

BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

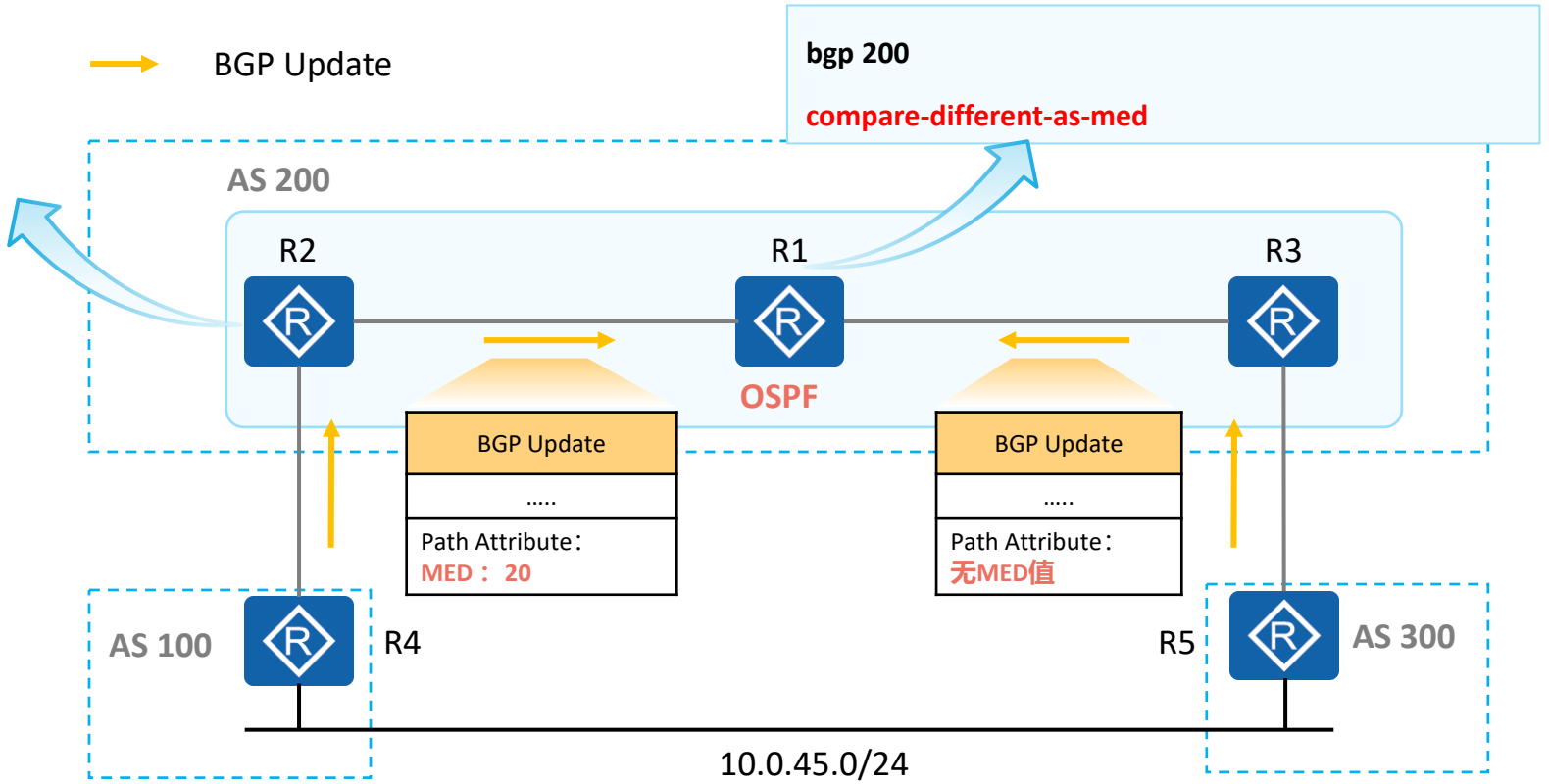
丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. **优选MED属性值最小的路由。**
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

优选MED最小 (1)

默认情况下BGP只会对来自同一个AS的相同路由比较MED值，可以通过命令开启来自不同AS的相同路由也比较MED值。

```
ip ip-prefix med index 10 permit 10.0.45.0 24
#
route-policy med permit node 10
 if-match ip-prefix med
 apply cost 20
route-policy med permit node 20
#
bgp 200
peer 10.0.1.1 route-policy med export
```

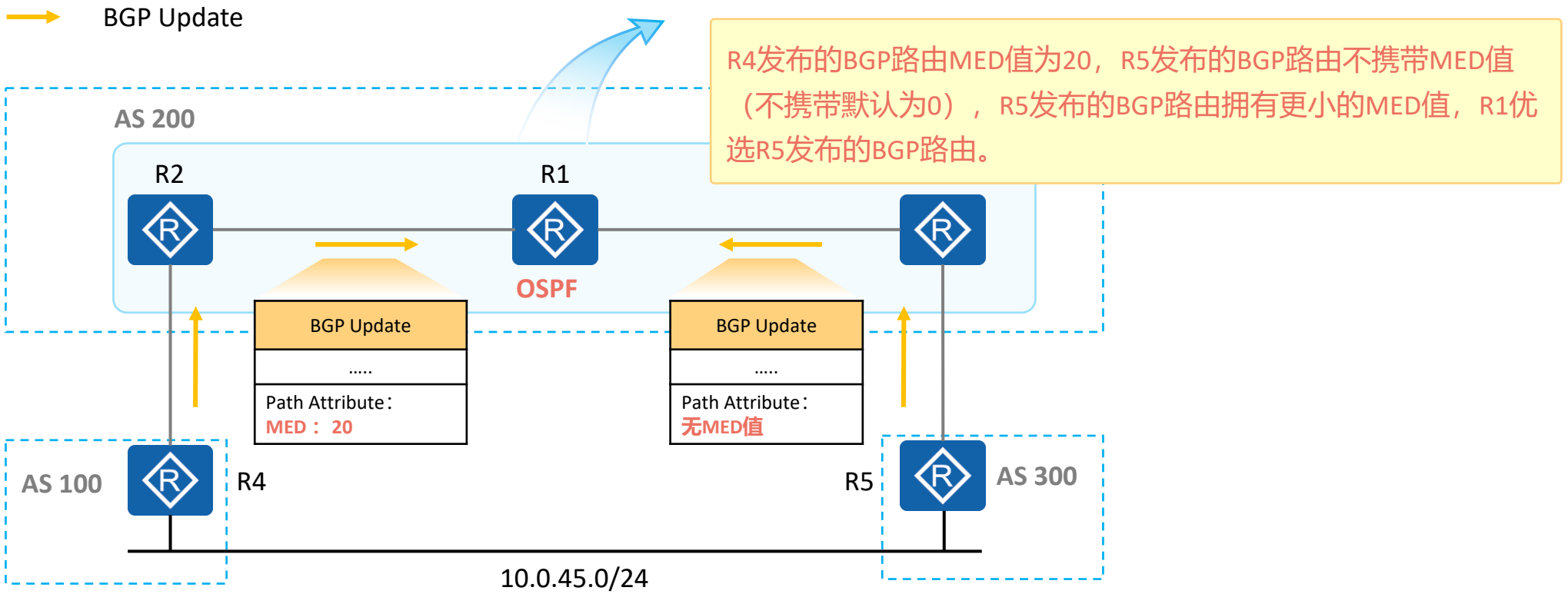


R2上通过路由策略修改通告给R1的BGP路由其MED属性值。

优选MED最小 (2)

```
Total Number of Routes: 2
     Network         NextHop   MED    LocPrf    PrefVal Path/Ogn

*>i  10.0.45.0/24    10.0.3.3       0        100       0     300?
* i                   10.0.2.2      20        100       0     100?
```



R4发布的BGP路由MED值为20，R5发布的BGP路由不携带MED值（不携带默认为0），R5发布的BGP路由拥有更小的MED值，R1优选R5发布的BGP路由。

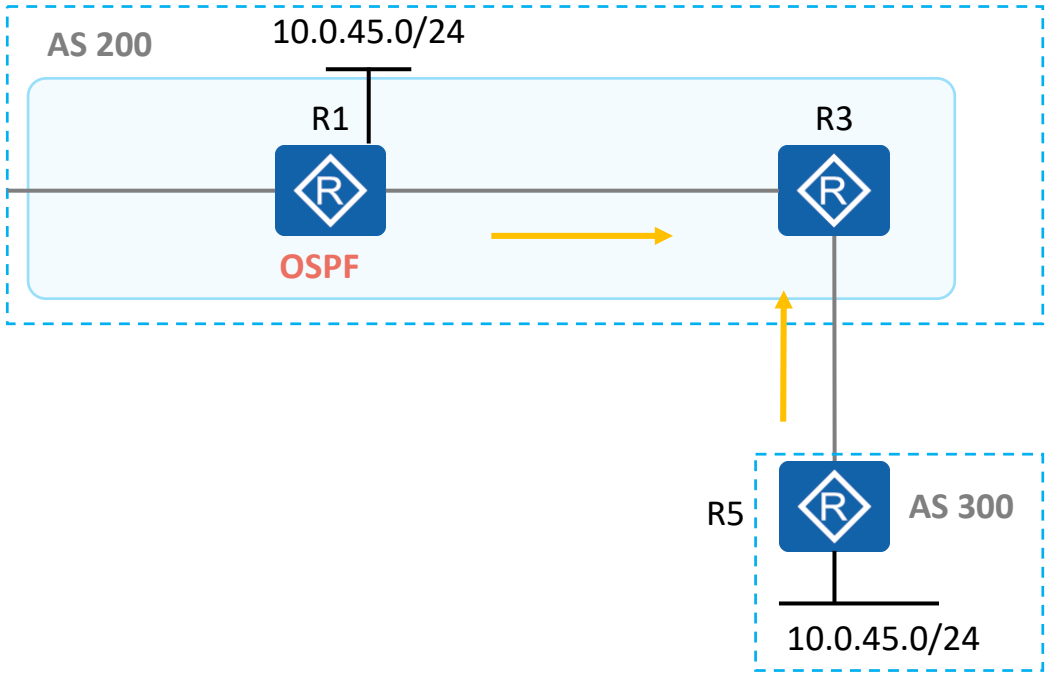
BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. **优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。**
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

优选从EBGP对等体学来的路由 (1)



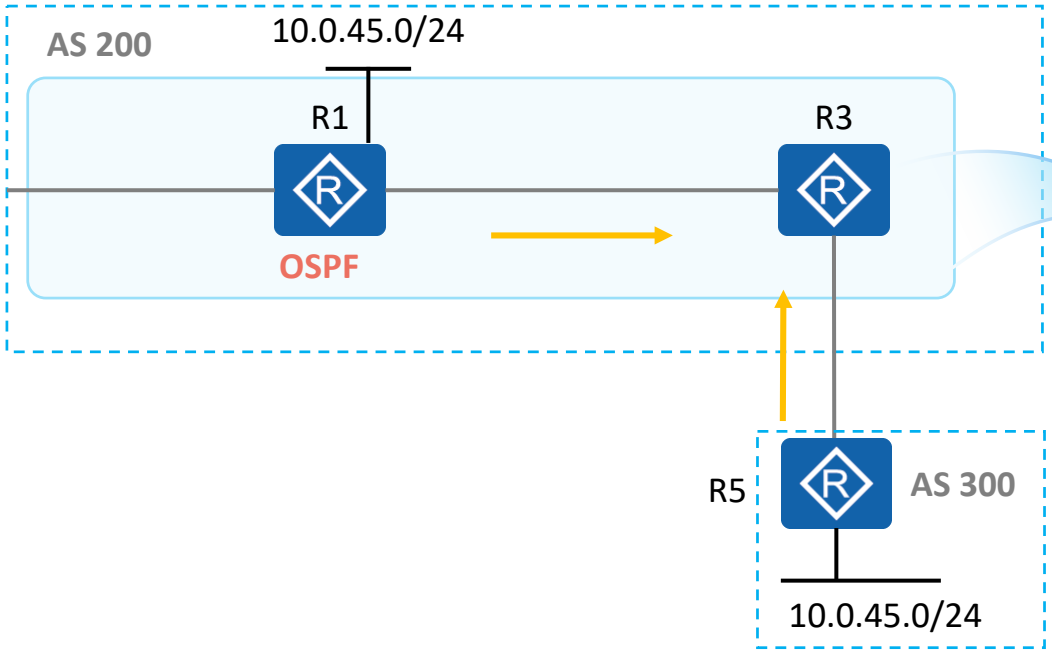
在R1上创建一条10.0.45.0/24的静态路由（指向null0），将该条路由发布到BGP，同时为了保证R1、R5通告给R3的BGP路由AS_Path长度相同，使用路由策略为R1通告给R3的路由加上AS_Path属性，其值为：500。

- R1上执行如下操作：

```
ip route-static 10.0.45.0 255.255.255.0 null0
ip ip-prefix ebgp index 10 permit 10.0.45.0 24
#
route-policy ebgp permit node 10
 if-match ip-prefix ebgp
 apply as-path 500 additive
route-policy ebgp permit node 20
#
bgp 200
  import-route static
  peer 10.0.3.3 route-policy ebgp export
```

- R3上将会同时收到R1、R5通告的BGP路由10.0.45.0/24，并且前面的优选规则无法比较出优选路由。

优选从EBGP对等体学来的路由 (2)



	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	10.0.45.0/24	10.0.35.5	0		0	300?
* i		10.0.1.1	0	100	0	500?

- 此时比较通告路由的对等体类型，R5为EBGP对等体，R1为IBGP对等体，EBGP对等体通告的BGP路由优于IBGP对等体通告的BGP路由，R3优选R5通告的BGP路由。

优选从EBGP对等体学来的路由 (3)

BGP routing table entry information of 10.0.45.0/24:

From: 10.0.1.1 (10.0.1.1)

Route Duration: 00h06m43s

Relay IP Nexthop: 10.0.13.1

Relay IP Out-Interface: GigabitEthernet0/0/0

Original nexthop: 10.0.1.1

Qos information : 0x0

AS-path 500, origin incomplete, MED 0, localpref 100, pref-val 0, valid, internal, pre 255, IGP cost 1, **not preferred for peer type**

Not advertised to any peer yet

- R3上通过**display bgp routing-table 10.0.45.0 24**查看BGP路由的详细信息，可以看到如下内容：

  not preferred for peer type

  表明该路由因为对等体类型没有被优选。

BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. **优选到Next_Hop的IGP度量值最小的路由。**
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

IGP Cost

BGP local router ID : 10.0.1.1

Local AS number : 200

Paths: 2 available, 1 best, 1 select

BGP routing table entry information of 10.0.45.0/24:

From: 10.0.3.3 (10.0.3.3)

Route Duration: 00h22m35s

Relay IP Nexthop: 10.0.13.3

Relay IP Out-Interface: GigabitEthernet0/0/1

Original nexthop: 10.0.3.3

Qos information : 0x0

AS-path 300, origin incomplete, MED 0, localpref 100, pref-val 0, valid, internal, best, select, active, pre 255, **IGP cost 1**

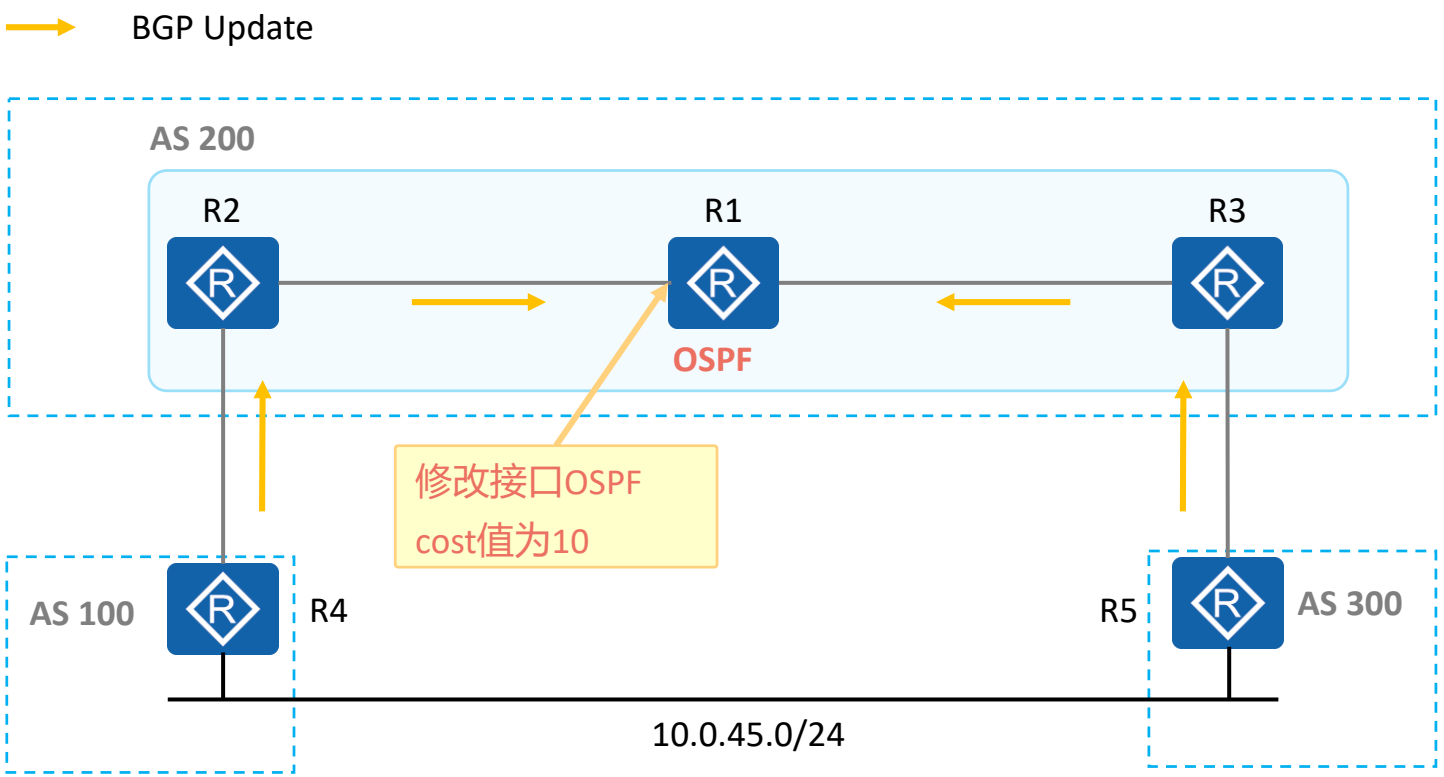
Not advertised to any peer yet

- 在BGP路由详细信息中存在IGP cost值这一内容，该值为本地IP路由表中去往Original nexthop地址的路由Cost值。

Destination/Mask	Proto	Pre	Cost	NextHop	Interface
10.0.3.3/32	OSPF	10	1	10.0.13.3	GigabitEthernet0/0/1

- 当前7条优选规则无法比较出优选BGP路由时将会比较前往下一跳地址的IGP cost值。

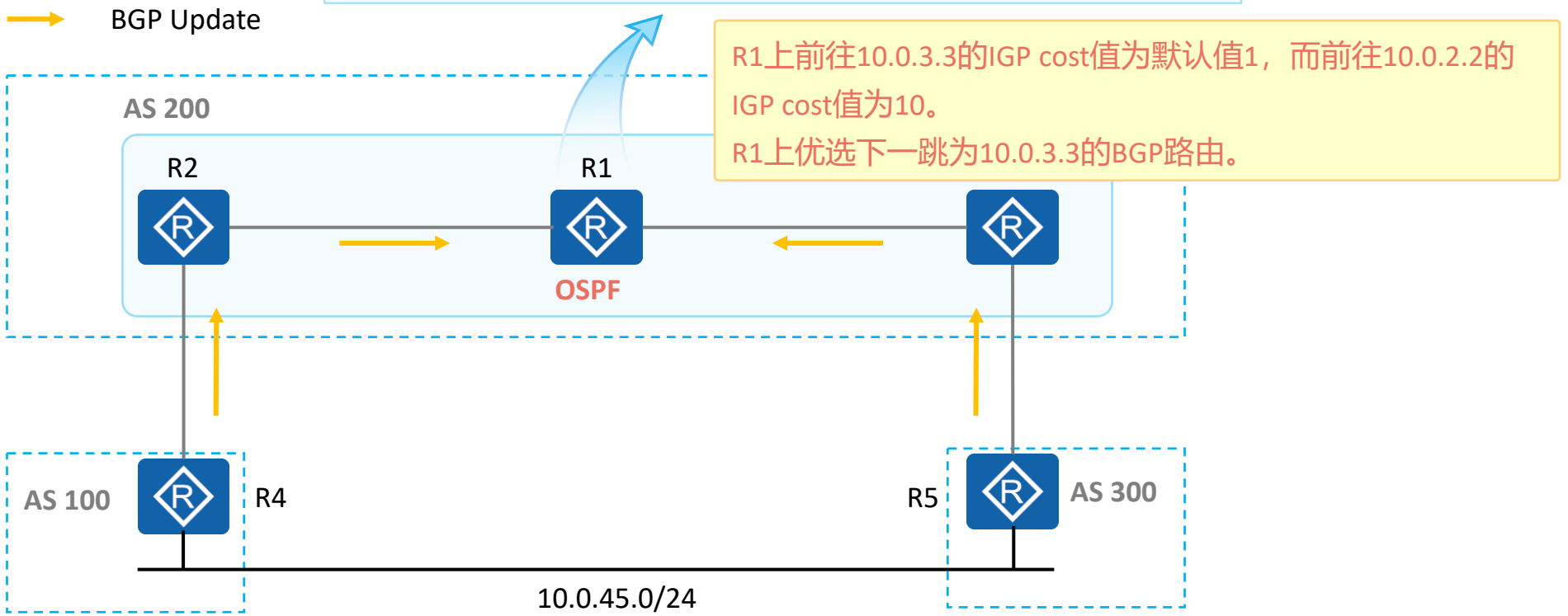
优选IGP Cost值最小 (1)



优选IGP Cost值最小 (2)

```
Total Number of Routes: 2
     Network         NextHop     MED    LocPrf    PrefVal Path/Ogn

 *>i 10.0.45.0/24    10.0.3.3      0       100       0     300?
 * i                 10.0.2.2      0       100       0     100?
```



优选IGP Cost值最小 (3)

BGP routing table entry information of 10.0.45.0/24:

From: 10.0.2.2 (10.0.2.2)

Route Duration: 00h24m07s

Relay IP Nexthop: 10.0.12.2

Relay IP Out-Interface: GigabitEthernet0/0/0

Original nexthop: 10.0.2.2

Qos information : 0x0

AS-path 100, origin incomplete, MED 0, localpref 100, pref-val 0, valid, internal, pre 255, **IGP cost 10, not preferred for IGP cost**

Not advertised to any peer yet

- R1上通过**display bgp routing-table 10.0.45.0 24** 查看BGP路由的详细信息，下一跳10.0.2.2的BGP路由其IGP cost值变为了10，而下一跳为10.0.3.3的BGP路由其IGP cost为默认值1，所以R1优选下一跳为10.0.3.3的路由。
- 在R1的路由详细信息中可以看到如下内容：

  not preferred for IGP cost

  表明该路由因为IGP cost未被优选。

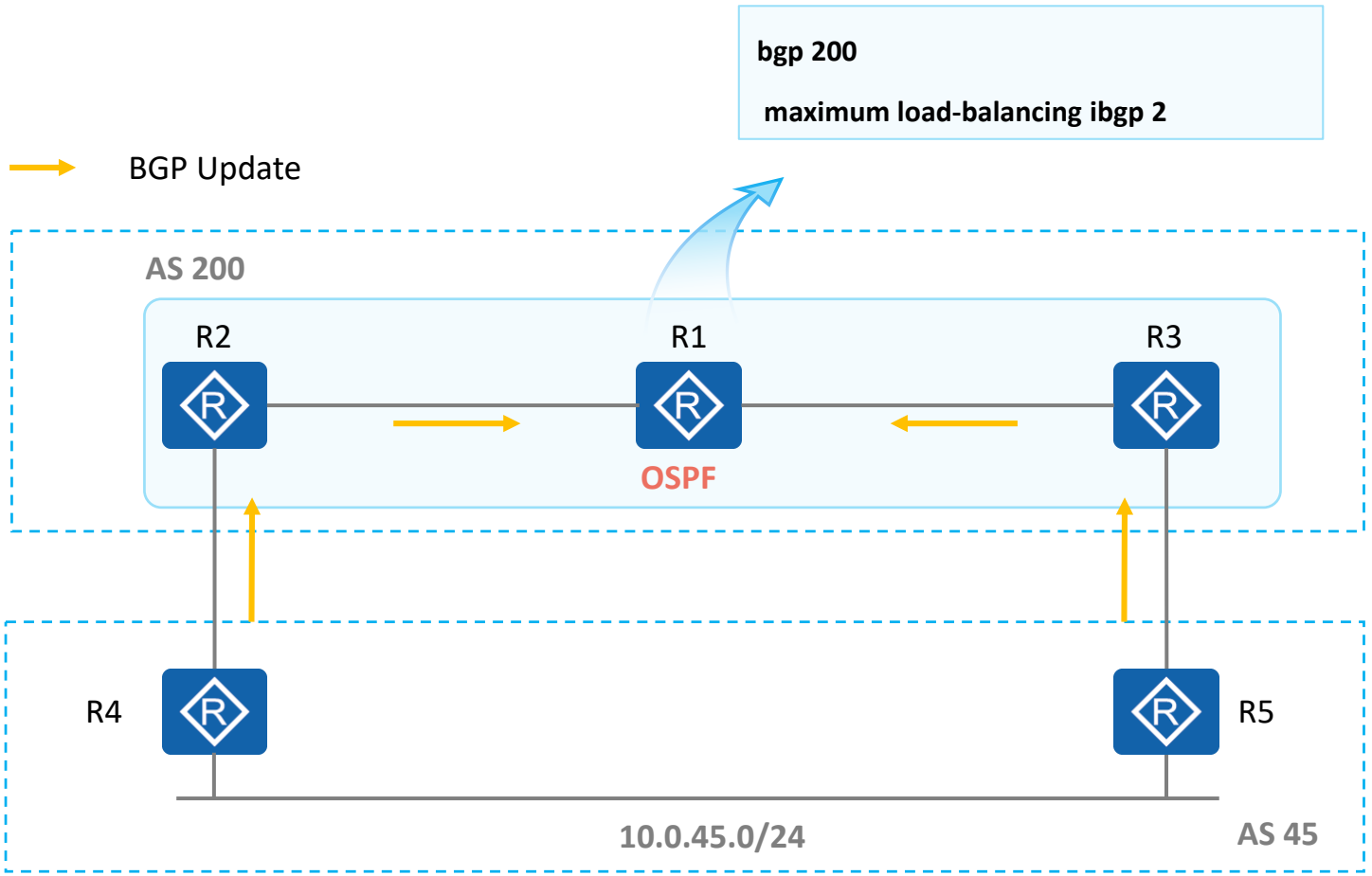
BGP路由等价负载分担

- 在大型网络中，到达同一目的地通常会存在多条有效BGP路由，设备只会优选一条最优的BGP路由，将该路由加载到路由表中使用，这一特点往往会造成很多流量负载不均衡的情况。
- 通过配置BGP负载分担，可以使得设备同时将多条等代价的BGP路由加载到路由表，实现流量负载均衡，减少网络拥塞。
- 值得注意的是，尽管配置了BGP负载分担，设备依然只会在多条到达同一目的地的BGP路由中优选一条路由，并只将这条路由通告给其他对等体。
- 在设备上使能BGP负载分担功能后，只有满足条件的多条BGP路由才会成为等价路由，进行负载分担。

形成BGP路由等价负载分担的条件

- Preferred-Value属性值相同。
- Local_Preference属性值相同。
- 都是聚合路由或者非聚合路由。
- AS_Path属性长度相同。
- Origin类型（IGP、EGP、Incomplete）相同。
- MED属性值相同。
- 都是EBGP路由或都是IBGP路由。
- AS内部IGP的Metric相同。
- AS_Path属性完全相同。

配置BGP路由负载分担



以左侧拓扑为例，R1上两条BGP路由在不做任何路由策略、配置的情况下，前8条优选规则无法比较出优选路由。因此可以配置IBGP路由的负载分担。

配置BGP路由负载分担后

IP路由表中出现了到达10.0.45.0/24的等价路由

```
[R1]display ip routing-table 10.0.45.0 24
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 2
Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

10.0.45.0/24        IBGP    255  0         RD    10.0.2.2        GigabitEthernet0/0/0
                    IBGP    255  0         RD    10.0.3.3        GigabitEthernet0/0/1
```

BGP路由表中依旧只有一条最优的路由

```
[R1]display bgp routing-table

 BGP Local router ID is 10.0.1.1
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete
 Total Number of Routes: 2
      Network  NextHop        MED        LocPrf     PrefVal Path/Ogn

 *>i  10.0.45.0/24    10.0.2.2        0          100        0       45?
 * i                  10.0.3.3        0          100        0       45?
```

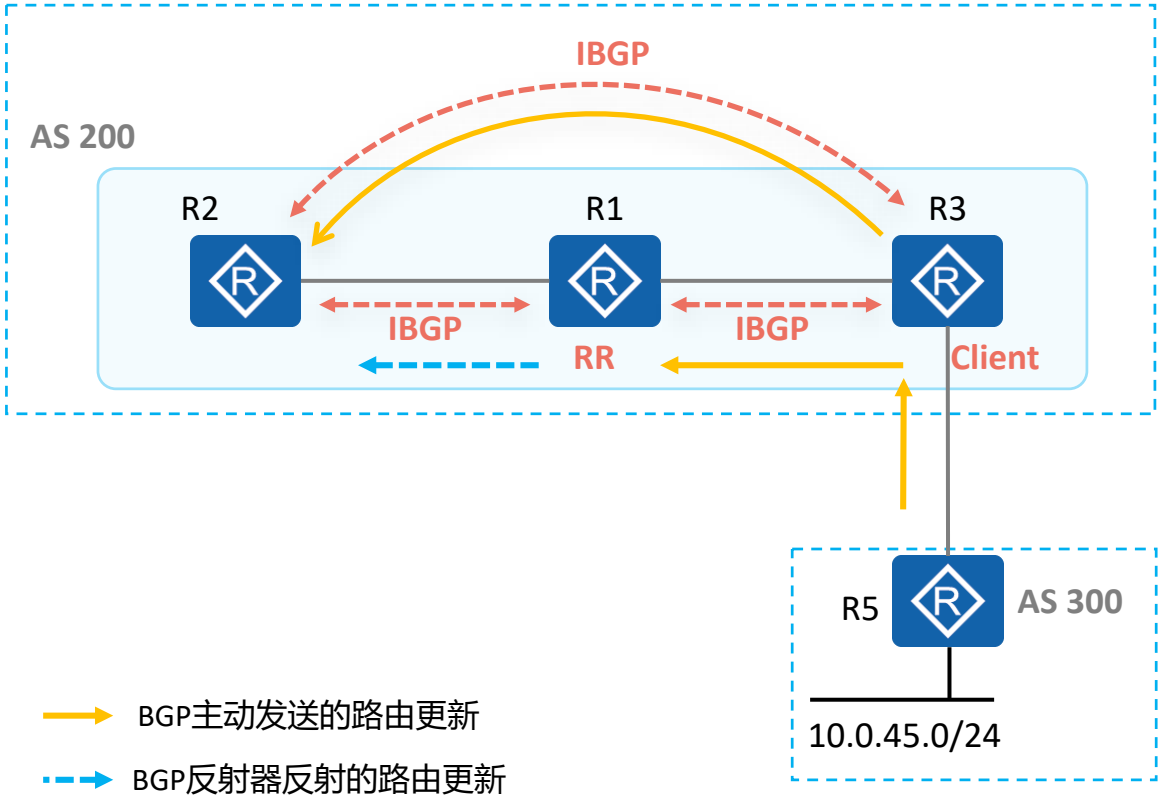
BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. **优选Cluster_List最短的路由。**
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. 优选具有最小IP地址的对等体通告的路由。

优选Cluster_List最短案例 (1)



对拓扑做如下修改：

- 只在R5上将10.0.45.0/24发布到BGP
- 配置R1为RR，R3为R1的客户端。
- R2、R3之间基于环回口建立IBGP对等体关系

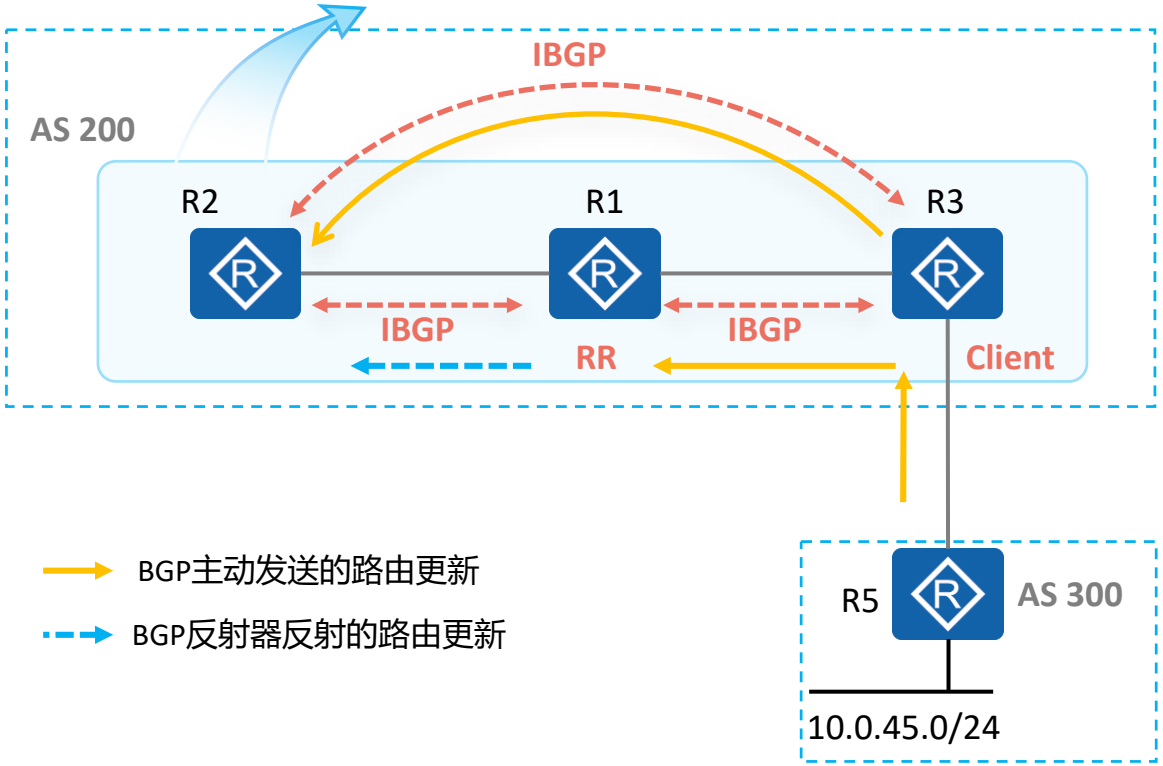
R2上将收到R3通告的BGP路由10.0.45.0/24、R1反射的BGP路由10.0.45.0/24。

默认配置下，前面介绍的规则无法比较出优选路由，此时将根据Cluster_List进行优选。

优选Cluster_List最短案例 (2)

```
Total Number of Routes: 2
     Network         NextHop   MED    LocPrf    PrefVal Path/Ogn

*>i  10.0.45.0/24    10.0.3.3       0        100       0     300?
* i                  10.0.3.3       0        100       0     300?
```



从BGP路由表中无法看出优选的是R1反射的BGP路由还是R3通告的BGP路由，此时可以通过命令**display bgp routing 10.0.45.0 24**查看BGP路由详细信息。

优选Cluster_List最短案例 (3)

BGP routing table entry information of 10.0.45.0/24:

From: 10.0.1.1 (10.0.1.1)

Route Duration: 00h03m10s

Relay IP Nexthop: 10.0.12.1

Relay IP Out-Interface: GigabitEthernet0/0/0

Original nexthop: 10.0.3.3

Qos information : 0x0

AS-path 300, origin incomplete, MED 0, localpref 100, pref-val 0, valid, internal, pre 255, IGP cost 2, **not preferred for Cluster List**

Originator: 10.0.3.3

Cluster list: 10.0.1.1

Not advertised to any peer yet

- 经由R1反射的路由不是最优路由，原因也被标出：not preferred for Cluster List
- R3直接通告给R2的BGP路由因为没有经过路由反射器，不存在Cluster_List属性，即被认为Cluster_List长度为0，小于由R1反射的BGP路由其Cluster_List长度（1），所以R3通告的BGP路由为优选路由。

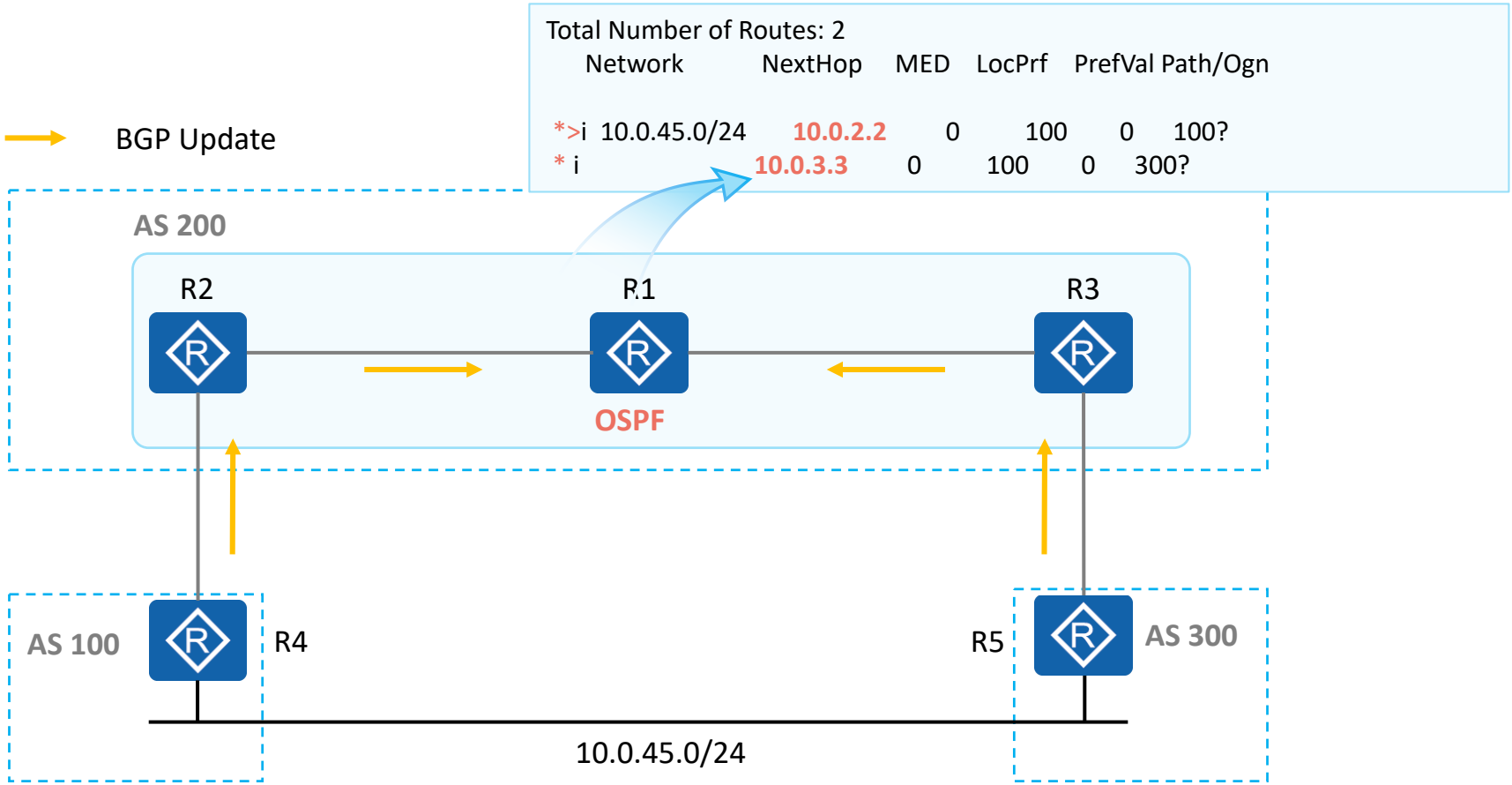
BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. **优选Router ID（Orginator_ID）最小的设备通告的路由。**
11. 优选具有最小IP地址的对等体通告的路由。

优选Router ID最小 (1)



在我们的讲解拓扑中，默认配置下R1从R2、R3都会收到BGP路由10.0.45.0/24，并且前面的优选规则无法比较出优选路由，最终将会根据本条规则，优选Router ID最小的对等体通告的BGP路由，在本案例中也就是R2通告的BGP路由。

优选Router ID最小 (2)

BGP routing table entry information of 10.0.45.0/24:

From: 10.0.3.3 (10.0.3.3)

Route Duration: 00h40m15s

Relay IP Nexthop: 10.0.13.3

Relay IP Out-Interface: GigabitEthernet0/0/1

Original nexthop: 10.0.3.3

Qos information : 0x0

AS-path 300, origin incomplete, MED 0, localpref 100, pref-val 0, valid, internal, pre 255, IGP cost 1, **not preferred for router ID**

Not advertised to any peer yet

查看R1的BGP路由表详细信息，来自10.0.3.3的BGP路由因Router ID原因没有被优选：

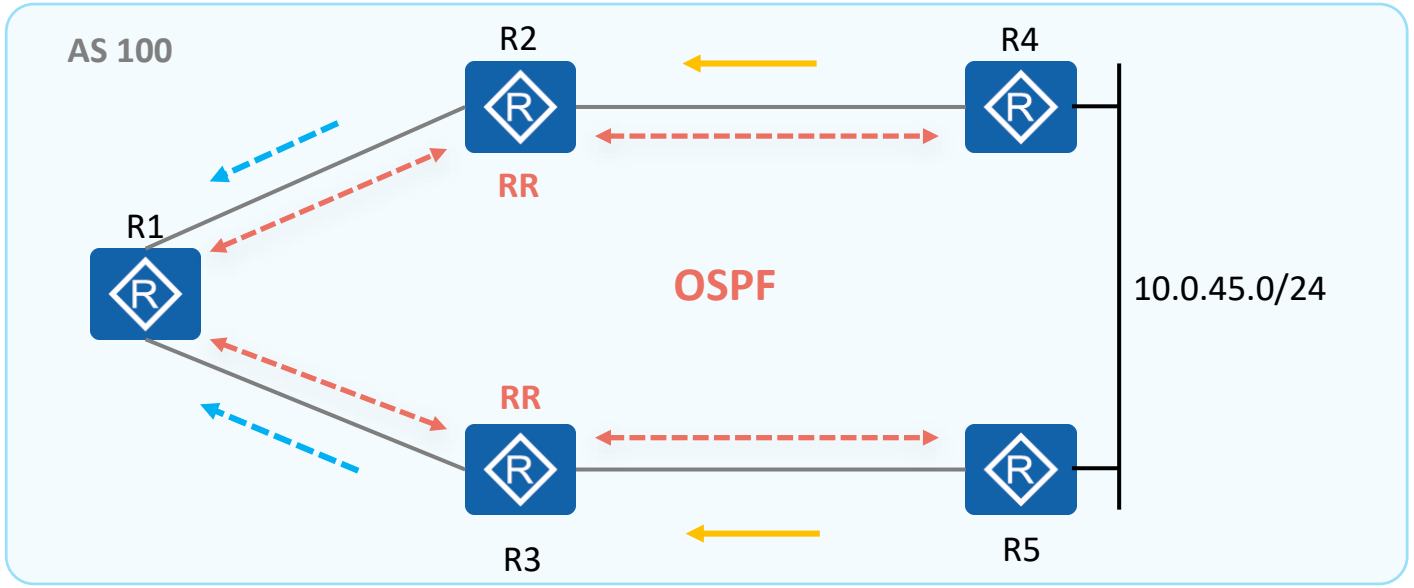
not preferred for router ID

优选Orginator_ID最小 (1)

BGP主动发送的路由更新

BGP反射器反射的路由更新

IBGP

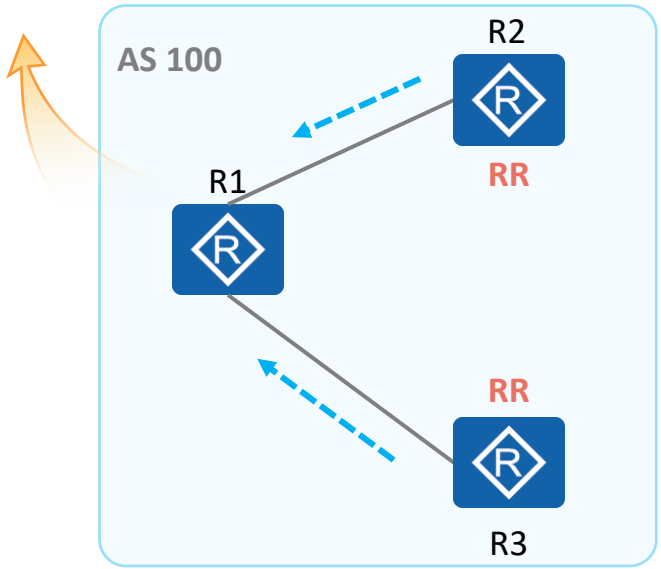


如果BGP路由携带Originator_ID属性，则在本条规则的优选过程中，将比较Originator_ID的大小，并优选Originator_ID最小的BGP路由。

优选Orginator_ID最小 (2)

```
Total Number of Routes: 2
     Network         NextHop     MED    LocPrf    PrefVal Path/Ogn

 *>i  10.0.45.0/24    10.0.4.4       0        100       0        ?
 * i                  10.0.5.5       0        100       0        ?
```



BGP主动发送的路由更新

BGP反射器反射的路由更新

```
BGP routing table entry information of 10.0.45.0/24:
From: 10.0.3.3 (10.0.3.3)
Route Duration: 00h33m15s
Relay IP Nexthop: 10.0.13.3
Relay IP Out-Interface: GigabitEthernet0/0/1
Original nexthop: 10.0.5.5
Qos information : 0x0
AS-path Nil, origin incomplete, MED 0, localpref 100, pref-val 0, valid, internal, pre
255, IGP cost 2, not preferred for router ID
Originator:  10.0.5.5
Cluster list: 10.0.3.3
Not advertised to any peer yet
```

R3反射过来的BGP路由未被优选，原因标注的还是Router ID，这里的Router ID其实是指Originator ID（其中携带的内容为原始路由发布者的Router ID）。

BGP路由优选规则

当到达同一个目的网段存在多条路由时，BGP通过如下的次序进行路由优选：

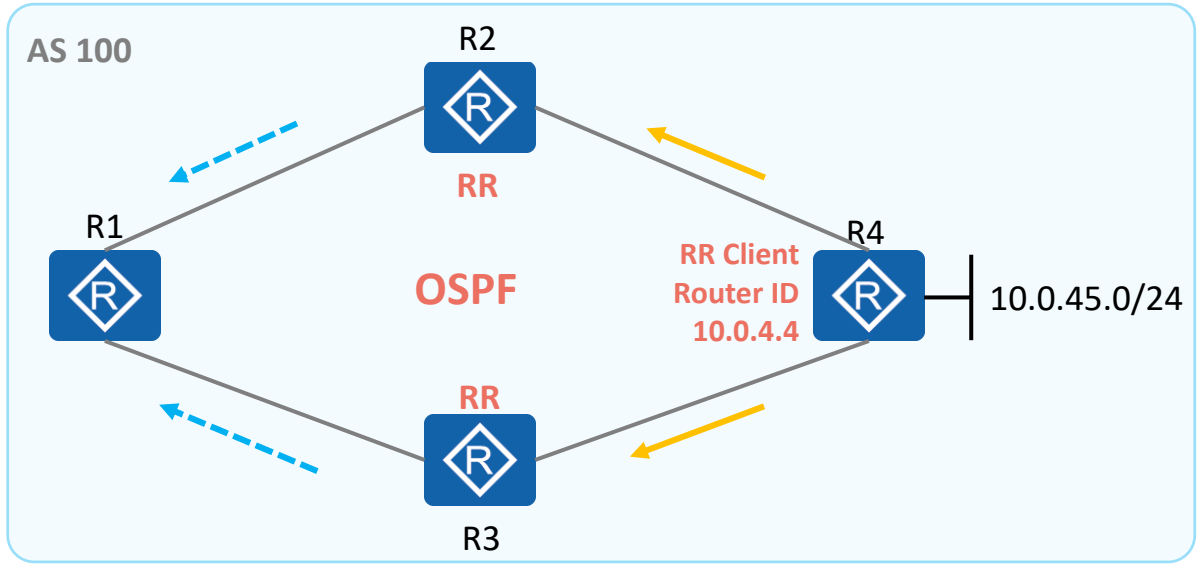
丢弃下一跳不可达的路由。

1. 优选Preferred-Value属性值最大的路由。
2. 优选Local_Preference属性值最大的路由。
3. 本地始发的BGP路由优于从其他对等体学习到的路由，本地始发的路由优先级：优选手动聚合>自动聚合>network>import>从对等体学到的。
4. 优选AS_Path属性值最短的路由。
5. 优选Origin属性最优的路由。Origin属性值按优先级从高到低的排列是：IGP、EGP及Incomplete。
6. 优选MED属性值最小的路由。
7. 优选从EBGP对等体学来的路由（EBGP路由优先级高于IBGP路由）。
8. 优选到Next_Hop的IGP度量值最小的路由。
9. 优选Cluster_List最短的路由。
10. 优选Router ID（Orginator_ID）最小的设备通告的路由。
11. **优选具有最小IP地址的对等体通告的路由。**

优选具有最小IP地址的对等体 (1)

BGP主动发送的路由更新

BGP反射器反射的路由更新

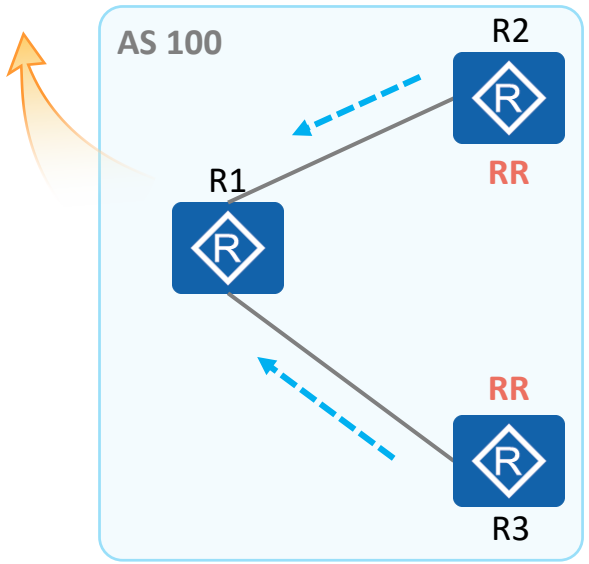


- 当前面所有规则都无法比较出优选路由时，此时会根据对等体地址大小来进行优选，对等体地址较小者发送的路由较优。
- 修改前一条规则的验证拓扑，R2、R3都与R4相连，R4作为RR客户端，只在R4上将路由发布到BGP，此时R2、R3反射的BGP路由将拥有相同的Originator ID：10.0.4.4。

优选具有最小IP地址的对等体 (2)

```
Total Number of Routes: 2
     Network         NextHop     MED    LocPrf    PrefVal Path/Ogn

 *>i 10.0.45.0/24    10.0.4.4      0       100       0       ?
 * i                 10.0.4.4      0       100       0       ?
```



BGP主动发送的路由更新

BGP反射器反射的路由更新

```
BGP routing table entry information of 10.0.45.0/24:
From: 10.0.3.3 (10.0.3.3)
Route Duration: 00h01m07s
Relay IP Nexthop: 10.0.12.2
Relay IP Out-Interface: GigabitEthernet0/0/0
Original nexthop: 10.0.4.4
Qos information : 0x0
AS-path Nil, origin incomplete, MED 0, localpref 100, pref-val 0, valid, intern
al, pre 255, IGP cost 2, not preferred for peer address
Originator:  10.0.4.4
Cluster list: 10.0.3.3
Not advertised to any peer yet
```

R3反射过来的BGP路由未被优选，原因为对等体地址较大：来自R2反射的路由对等体地址为10.0.2.2，而R3反射的路由对等体地址为10.0.3.3，因此未被优选。

思考题

1. （简答题）从EBGP对等体收到的BGP路由通告给IBGP对等体时如何修改next_hop属性值为自身更新源地址?
2. （判断题）当前三条优选规则相同的情况下，BGP会比较AS_Path长度，当AS_Path长度相同时会比较AS号的大小。

本章总结

- BGP采用路径属性进行路由优选，这让BGP拥有丰富的可选比较项，可以在不同的场景下根据不同的路径属性选择出最适合的路由。
- BGP定义了一套详细的最优路径选择算法，这使得路由器能够在任何复杂的、高冗余性的网络环境下选择出最优的路径，这套算法也被称作BGP路由优选规则，或者BGP选路原则。
- BGP选路原则在实际中频繁应用，需要熟练掌握。


谢 谢
www.huawei.com