BGP概述及基础配置

朱仕耿

www.huawei.com

Author / Email : Zhushigeng 261992 / zhushigeng@huawei.com

Version 1.1

HUAWEI

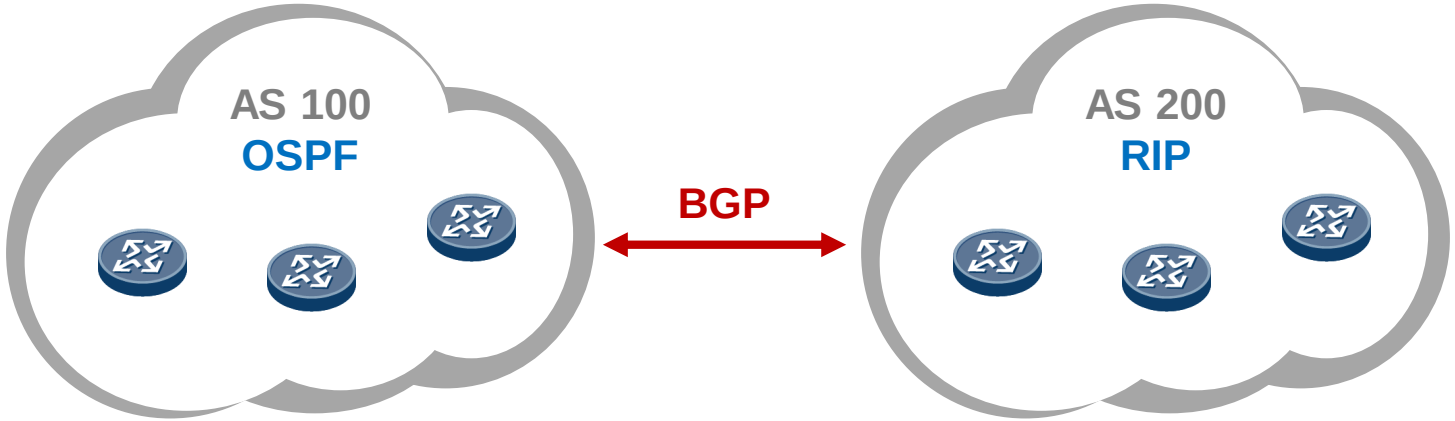
课程目标

- 理解IGP与EGP的区别及工作场景。
- 理解BGP基本概念（AS的概念、协议特征、报文类型、状态机、对等体类型、同步规则、路由黑洞问题、IBGP水平分割规则、BGP的各种表项等）。
- 理解BGP的基础工作机制。
- 掌握BGP的基础配置。

目录

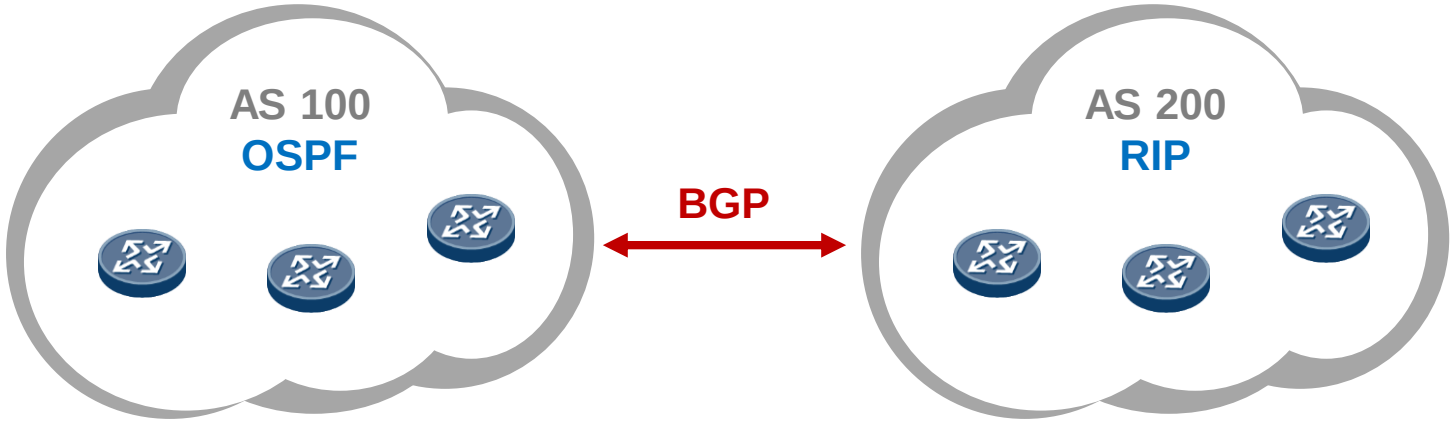
HUAWEI

IGP与BGP的工作场景



AS的概念

- 自治系统（Autonomous System，AS），指的是在同一个组织管理下、使用相同策略的设备的集合。
- 不同AS通过AS号区分，AS号取值范围1 - 65535，其中64512 - 65535是私有AS号。IANA负责AS号的分发。
- 中国电信163 AS号：4134。
- 中国电信CN2 AS号：4809。
- 中国网通AS 号：9929。

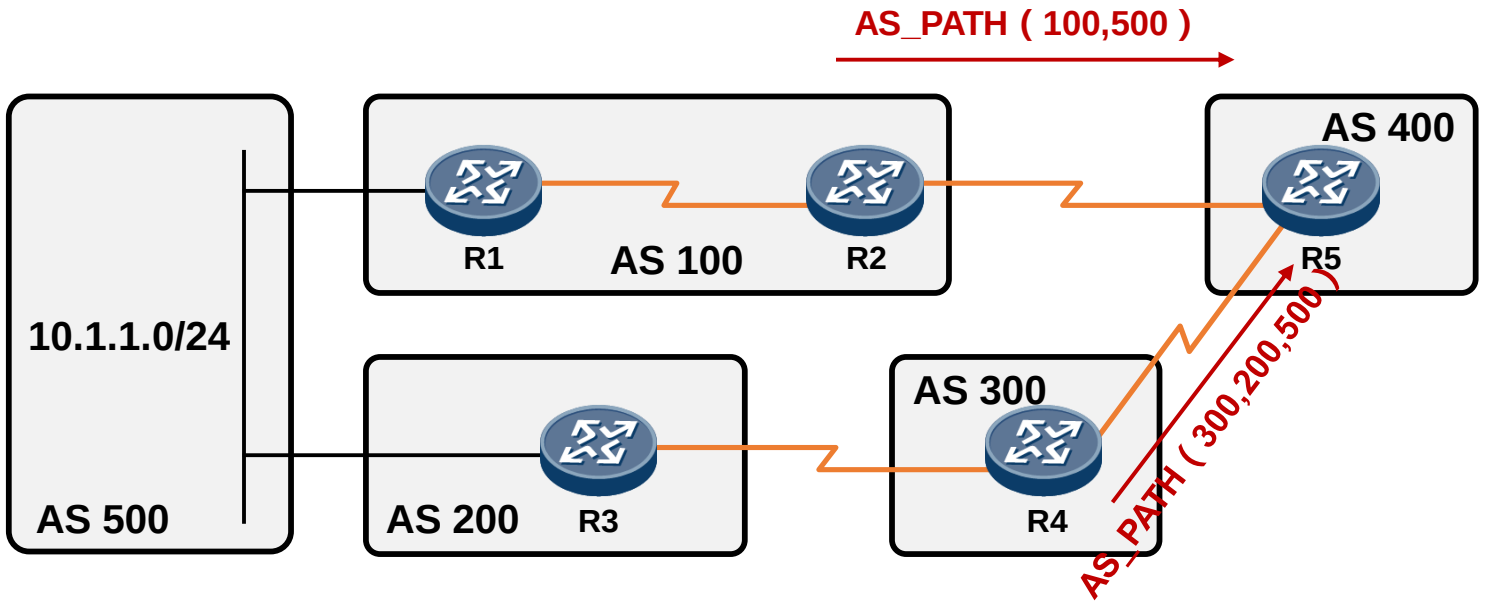


BGP概述

- 边界网关协议（Border Gateway Protocol，BGP）是一种实现自治系统AS之间的路由可达，并选择最佳路由的矢量性协议。早期发布的三个版本分别是BGP-1（RFC1105）、BGP-2（RFC1163）和BGP-3（RFC1267），1994年开始使用BGP-4(RFC1771)，2006年之后单播IPv4网络使用的版本是BGP-4（RFC4271），其他网络使用的版本是MP-BGP（RFC4760）。
- BGP的特点：
  - BGP能够承载大批量的路由信息，能够支撑大规模网络。
  - BGP提供了丰富的路由策略，能够灵活的进行路由选路，并能指导邻居按策略发布路由。
  - BGP能够支撑MPLS/VPN的应用，传递客户VPN路由。
  - BGP提供了路由聚合和路由衰减功能用于防止路由振荡，有效提高了网络的稳定性。
  - BGP使用TCP作为其传输层协议（端口号为179），并支持BGP与BFD联动、BGP Tracking、BGP Auto FRR和BGP GR和NSR，提高了网络的可靠性。

BGP的路径矢量特征

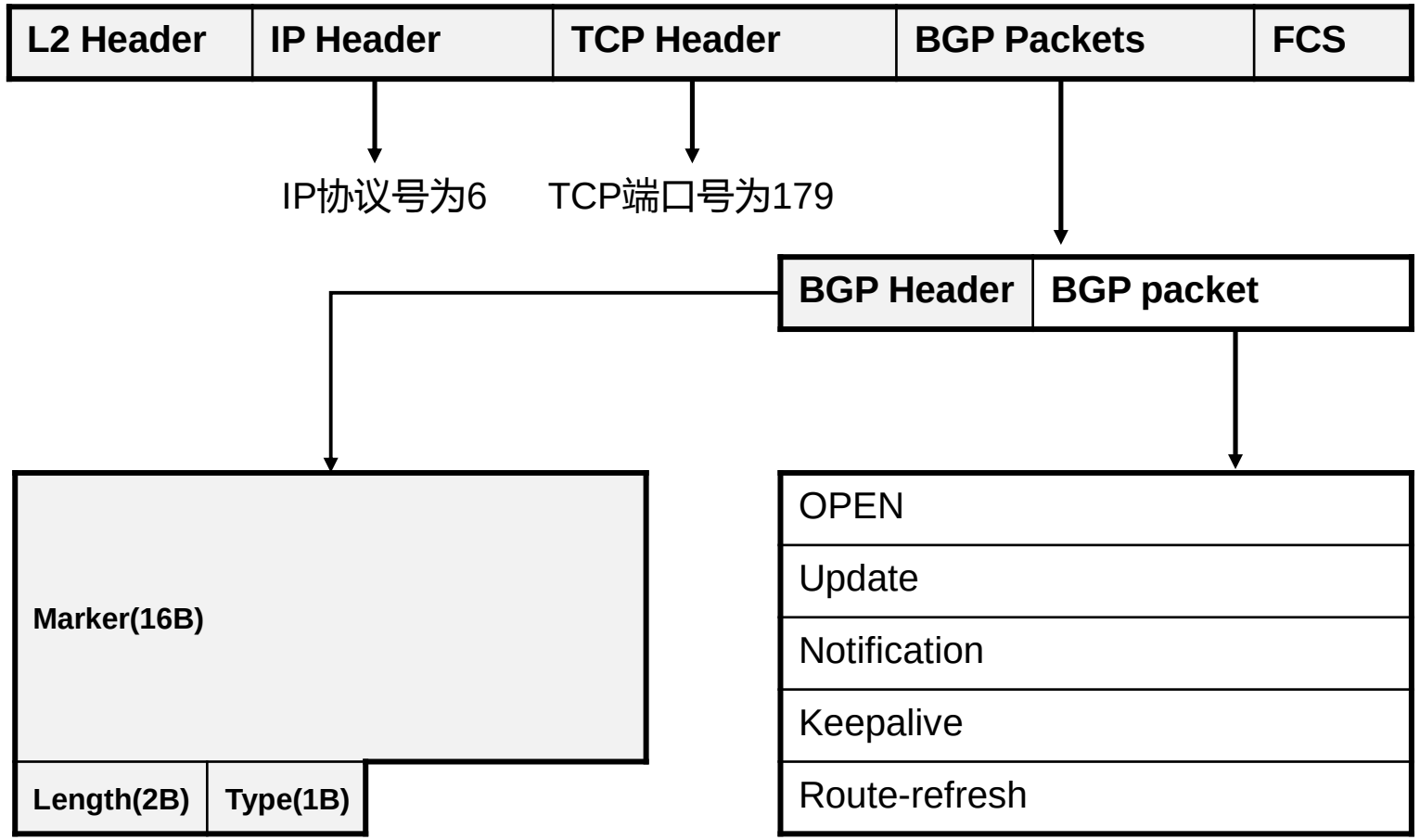
- BGP通常被称为路径矢量路由协议（Path-Vector Routing Protocol）。
- 每条BGP路由都携带着多种路径属性（Path attribute），在各种路径属性中，AS_Path属性是非常关键的一个。AS_Path属性记录了BGP路由传递过程中所经过的AS号，实际上它是一个AS号的列表。
- BGP路由器不接受AS_Path中包含其自身AS号的路由更新。AS_Path属性值的长短（AS号的个数）会作为一个比较的依据，影响BGP路由选择的决策。



BGP协议特征

- BGP使用TCP为传输层协议，TCP端口号179。路由器之间的BGP会话基于TCP连接而建立。
- 运行BGP的路由器被称为BGP发言者（BGP Speaker），或BGP路由器。
- 两个建立BGP会话的路由器互为对等体（Peer）。BGP对等体之间交换BGP路由表。
- BGP路由器只发送增量的BGP路由更新，或进行触发更新（不会周期性更新）。
- BGP具有丰富的路径属性和强大的路由策略工具。
- BGP能够承载大批量的路由前缀，用于大规模的网络中。

BGP报文类型



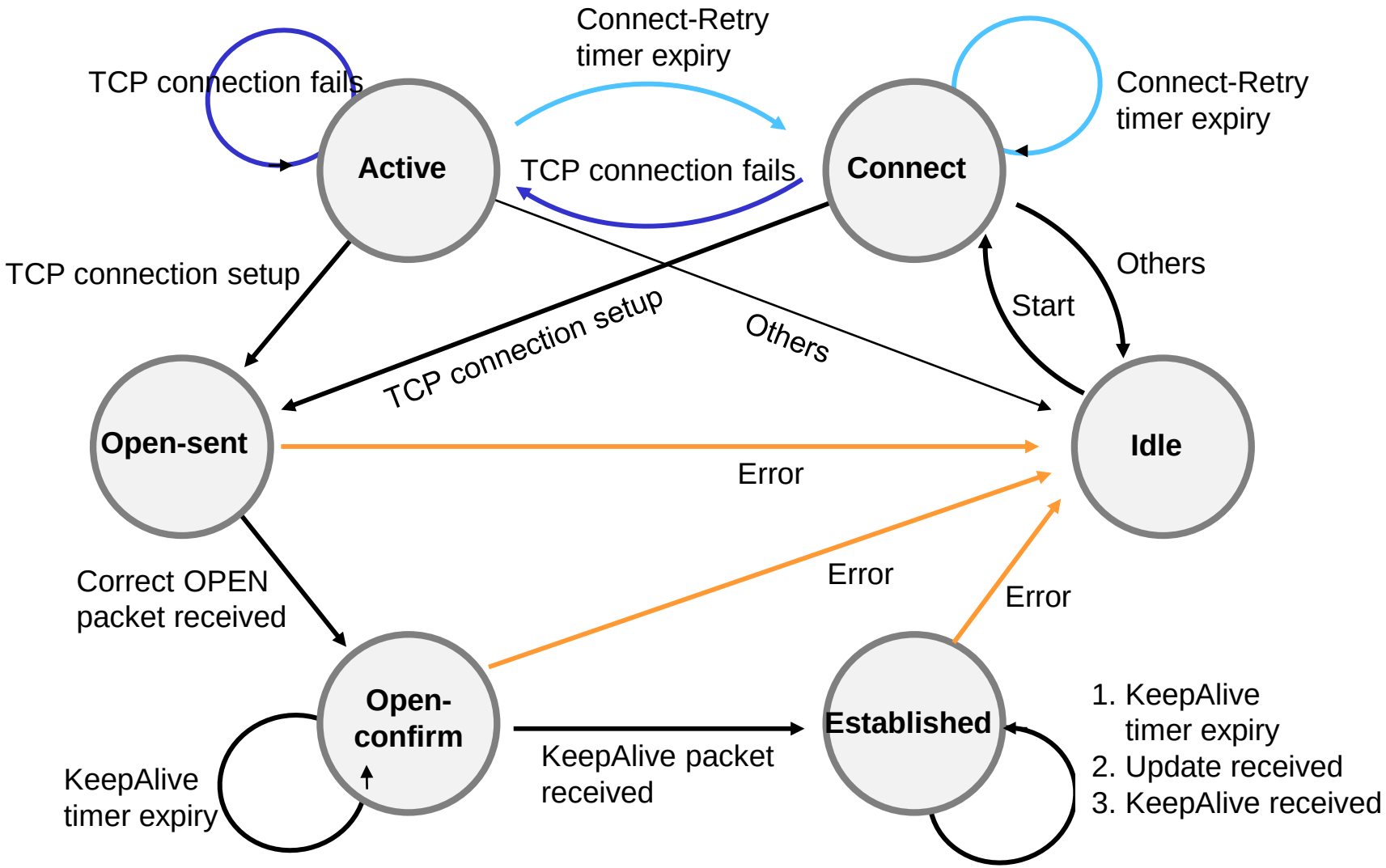
BGP报文类型

报文名称	作用是什么	什么时候发包
Open	协商BGP邻居的各项参数，建立邻居关系。	BGP对等体之间需先建立TCP连接，如果TCP连接成功，那么BGP向对等体发送Open报文。
Update	用于发送BGP路由信息。	连接建立后，有路由需要发送或路由变化时，发送UPDATE通告对端路由信息。
Notification	报告错误，中止对等体关系。	当BGP在运行中发现错误时，要发送NOTIFICATION报文通告BGP对端。
Keepalive	维持BGP对等体关系。	定时发送Keepalive报文以保持BGP对等体关系的有效性。
Route-refresh	用于在改变路由策略后请求对等体重新发送路由信息。只有支持路由刷新能力的BGP设备会发送和响应此报文。	当路由策略发生变化时，触发请求对等体重新通告路由。

BGP的状态机

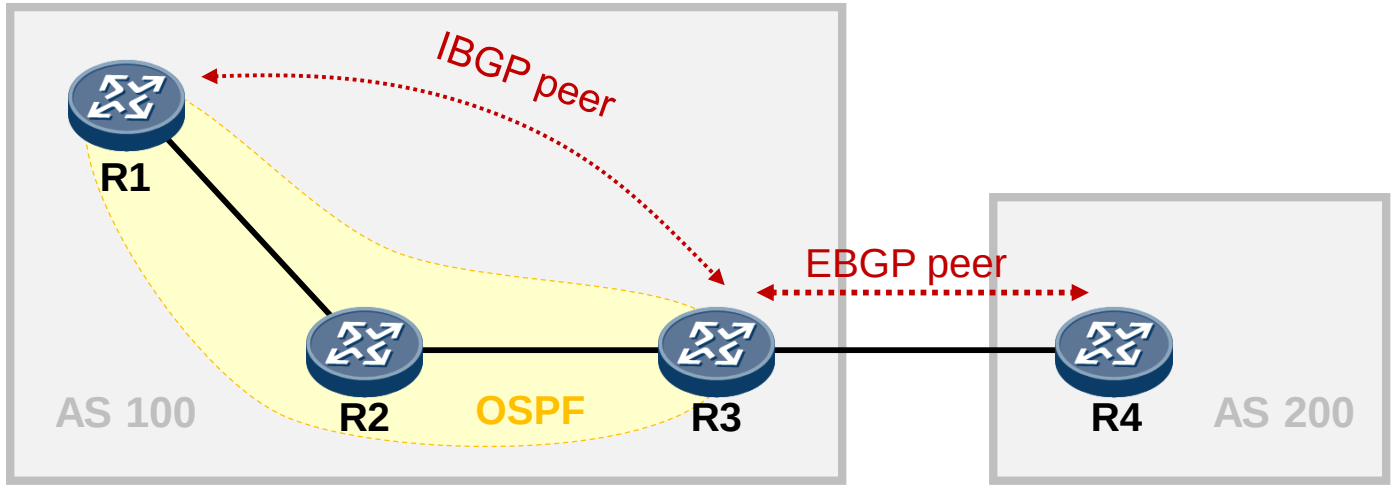
Peer状态名称	发什么包	在做什么
Idle	尝试建立TCP连接	开始准备TCP的连接并监视远程peer启动TCP连接，启用BGP时，要准备足够的资源。
Connect	发TCP包	正在进行TCP连接，等待完成中，认证都是在TCP建立期间完成的。如果TCP连接建立失败则进入Active状态，反复尝试连接。
Active	发TCP包	TCP连接没建立成功，反复尝试TCP连接。
OpenSent	发Open包	TCP连接建立已经成功，开始发送Open包，Open包携带参数协商对等体的建立。
OpenConfirm	发Keepalive包	参数、能力特性协商成功，自己发送Keepalive包，等待对方的Keepalive包。
Established	发Update包	已经收到对方的Keepalive包，双方能力特性经协商发现一致，开始使用Update通告路由信息。

BGP的状态机



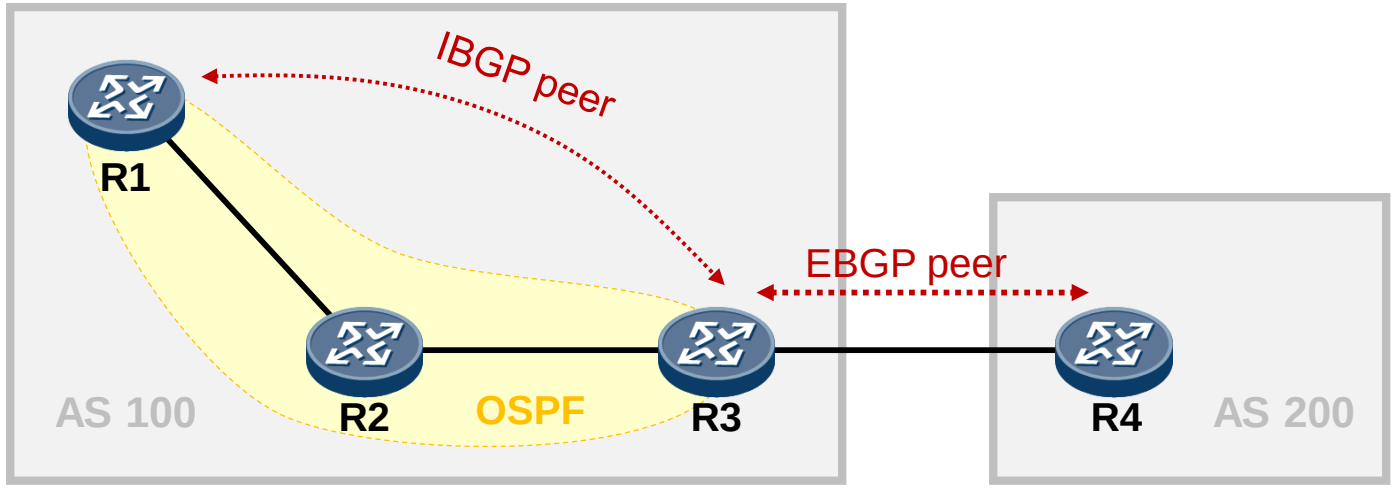
BGP Peer

- 运行BGP的路由器被称为BGP发言者，或者BGP路由器。
- BGP对等体也叫BGP邻居，与OSPF、RIP等协议不同，BGP的会话是基于TCP建立的。建立BGP对等体关系的两台路由器并不要求必须直连。
- BGP存在两种对等体关系类型：EBGP及IBGP。针对这两种对等体类型，BGP处理路由的操作存在较大差异。



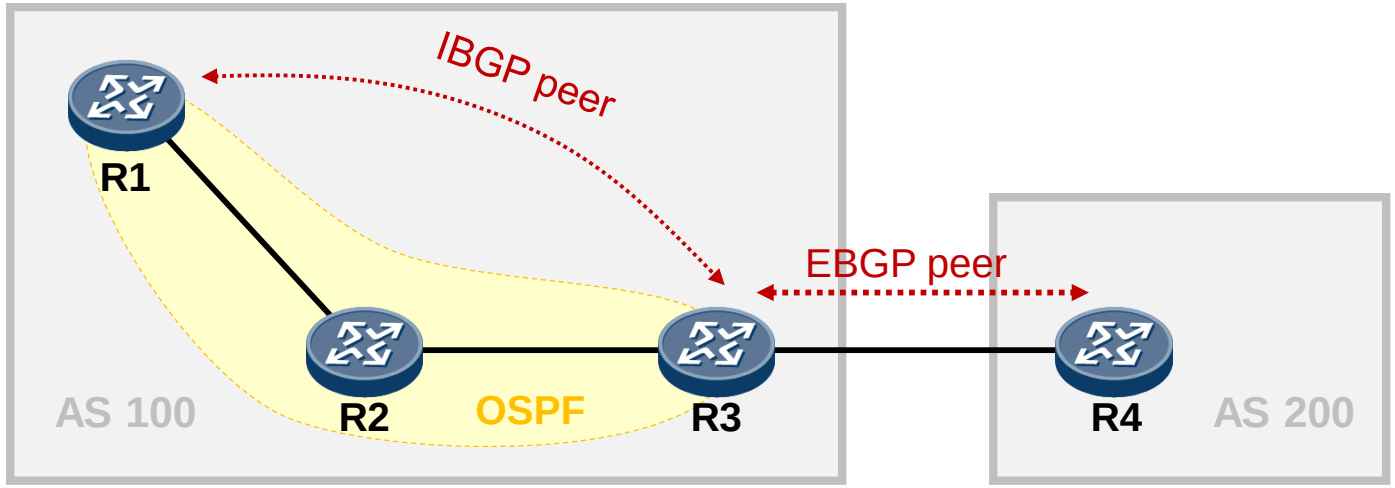
BGP Peer

- EBGP（External BGP）：位于不同自治系统的BGP路由器之间的BGP邻接关系。
- 两台路由器之间要建立EBGP对等体关系，必须满足两个条件：
  - 两个路由器所属AS不同（也即AS号不同）。
  - 在配置BGP时，Peer命令所指定的对等体IP地址要求路由可达，并且TCP连接能够正确建立。

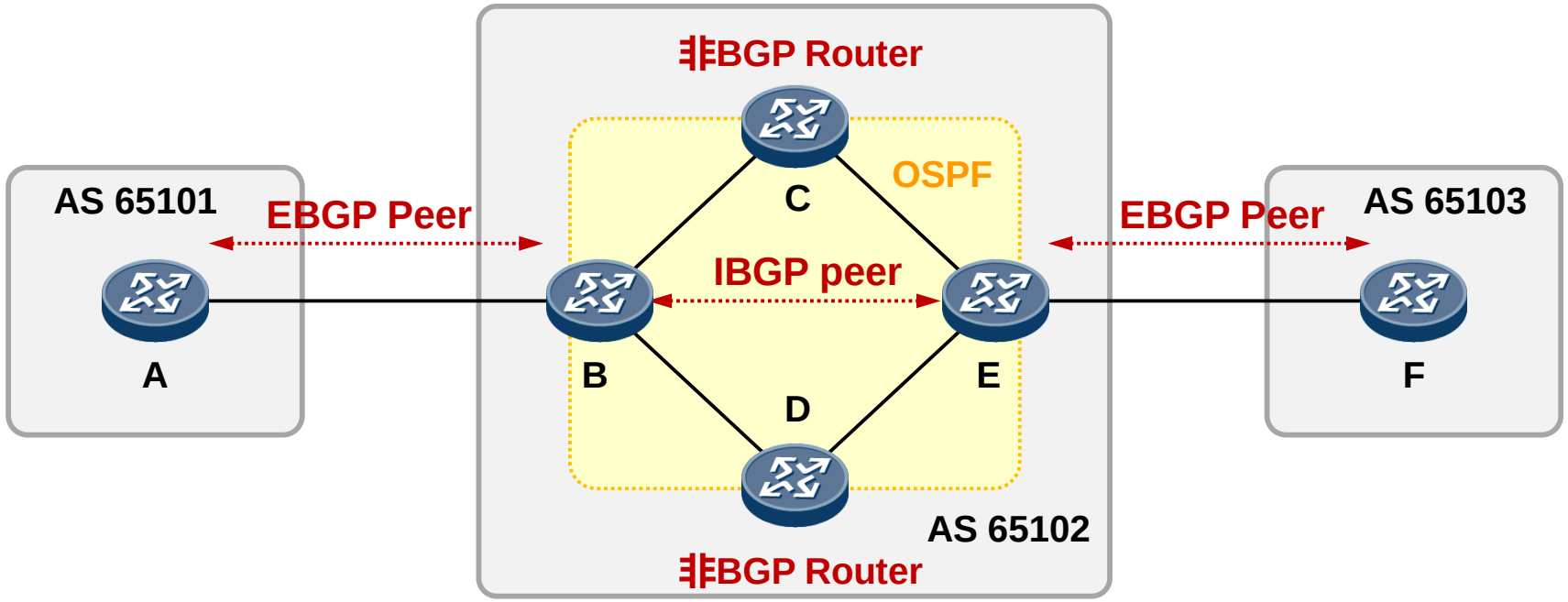


BGP Peer

- IBGP（Internal BGP）：位于相同自治系统的BGP路由器之间的BGP邻接关系。
- 两台路由器之间要建立IBGP对等体关系，必须满足两个条件：
  - 两个路由器所属AS需相同（也即AS号相同）。
  - 在配置BGP时，Peer命令所指定的对等体IP地址要求路由可达，并且TCP连接能够正确建立。



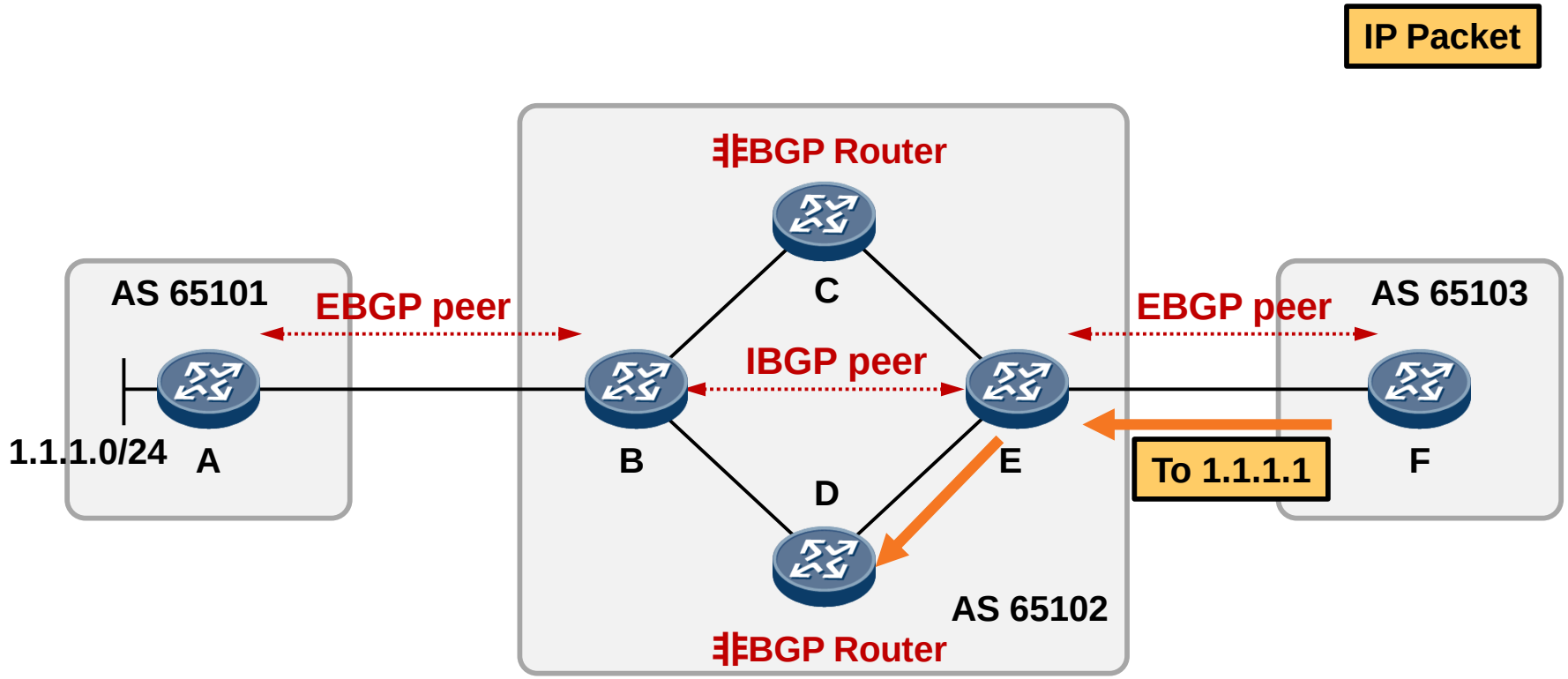
传输AS中的路由黑洞问题



在传输AS（Transit AS）65102中，BCDE四台路由器运行了OSPF，确保AS内部路由实现互通。

B与E运行BGP，并且两者建立IBGP对等体关系（两者并非直连，但是对于BGP，这是允许的，仅需确保两者之间能够正确建立TCP连接即可）。C与D并未运行BGP。

传输AS中的路由黑洞问题



A将本地路由1.1.1.0/24通告到BGP，最终F能够通过BGP学习到该条路由。C、D由于并未运行BGP，因此无法通过BGP学习到1.1.1.0/24路由。

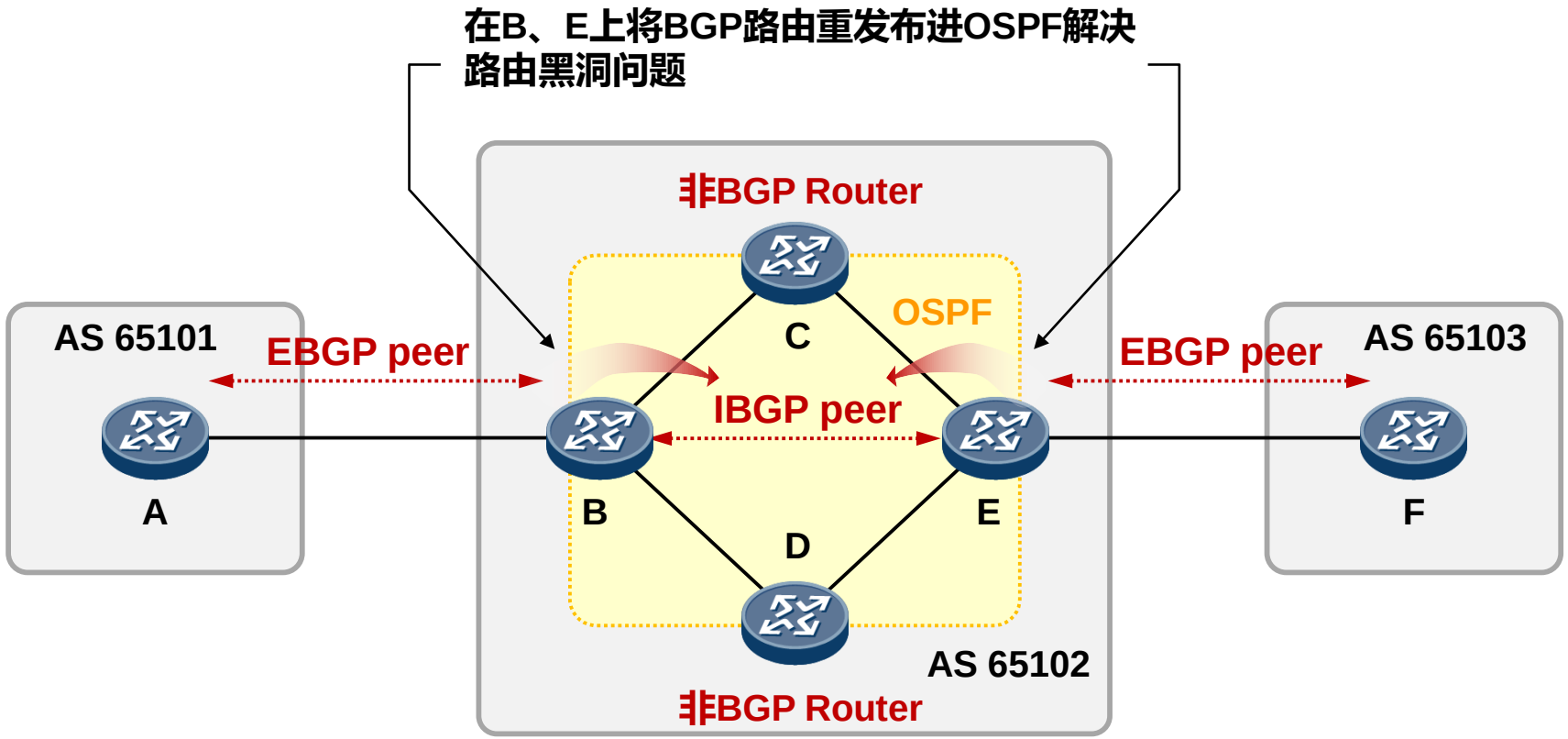
如此一来，F发往1.1.1.0/24网络的数据包在到达C/D后将被丢弃，在C及D路由器这里，就出现了黑洞。

BGP同步（synchronization）

- 若路由器从IBGP对等体学习到一条BGP路由，它不能使用该条路由，更不能将路由传递给自己的EBGP对等体，除非它又从IGP学习到该条路由，这就是BGP的同步规则。
- 同步规则的存在，可以防止数据在传输AS内由于转发设备没有目标网络的路由而被丢弃的问题，也就是所谓的黑洞问题。
- 为了使得BGP路由能够正常交互，我们就不得不在该传输AS内所有路由器上都运行BGP，且构建全互联的IBGP对等体关系；或者在AS边界上将BGP路由引入IGP。显然这两种方法各有利弊，尤其是后者，盲目地将BGP路由引入IGP是非常危险的。
- 同步规则的存在意义是避免出现黑洞问题，而如果AS内路由问题已经解决，那么同步规则也就没有必要再开启了。华为路由交换产品缺省关闭同步规则。

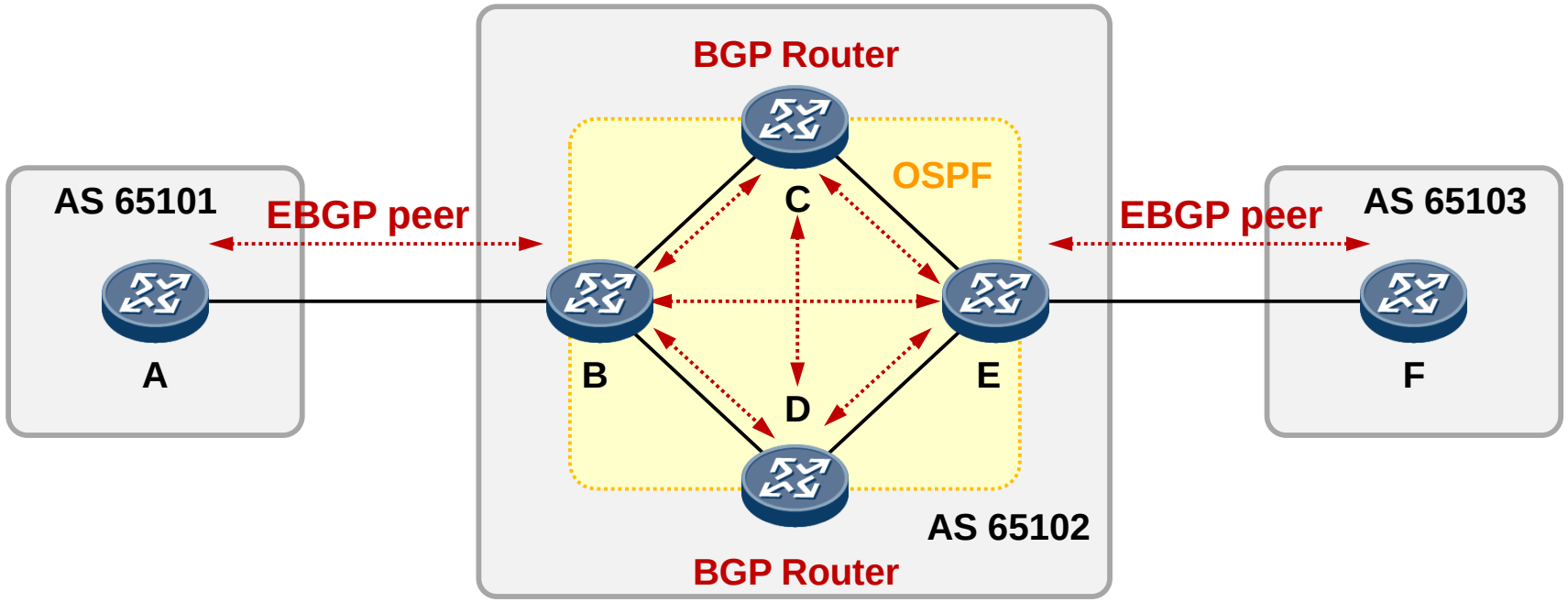
传输AS中的黑洞问题 解决办法1

- 将BGP路由引入到IGP，并关闭同步规则。



传输AS中的黑洞问题 解决办法2

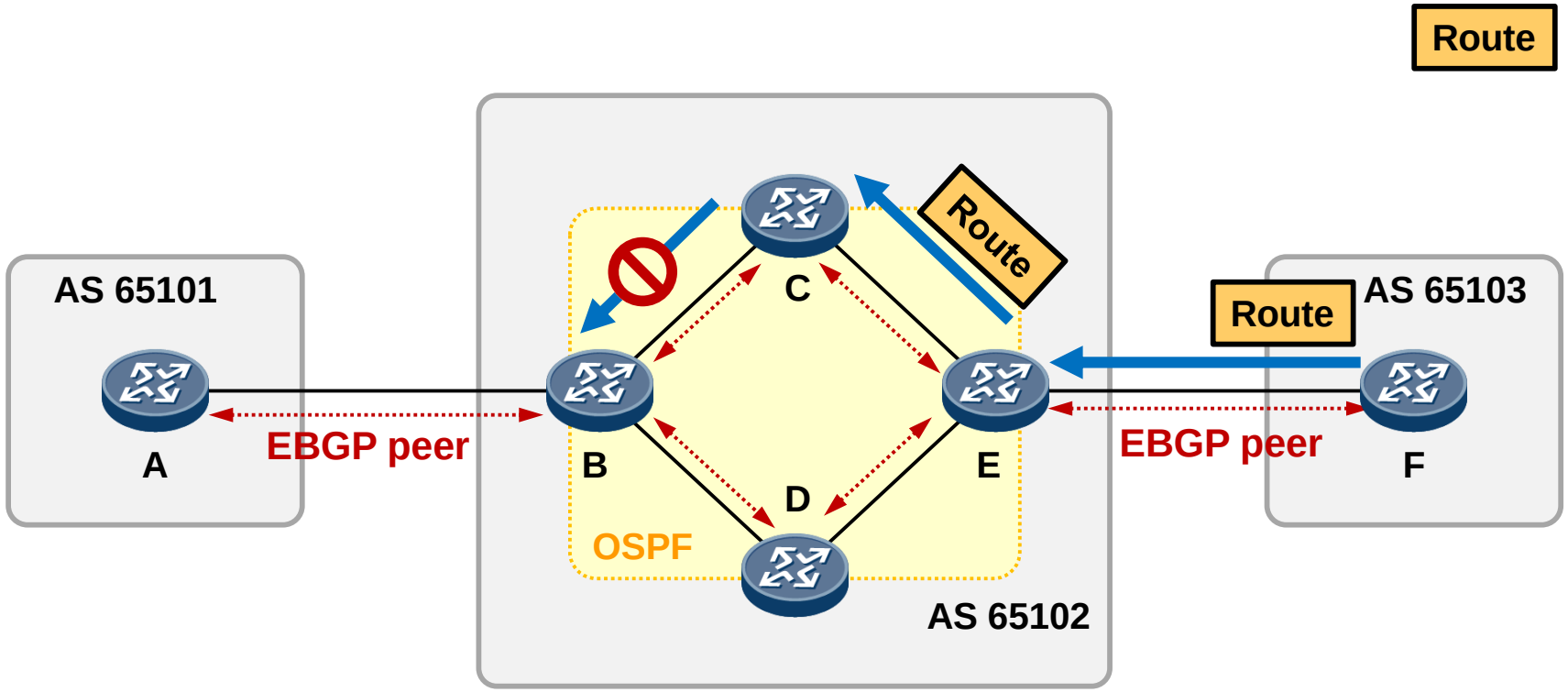
- 传输AS内所有路由器均运行BGP，实现BGP对等体关系的全互联，并关闭同步规则。



IBGP水平分割规则

- BGP路由在AS之间的防环依赖于AS_Path路径属性，当路由器收到BGP路由后，发现该路由所携带的AS_Path属性中出现了其自己所处的AS号，则路由器认为出现了路由环路，它将忽略该条路由。
- AS_Path属性仅在路由离开AS时才会被更改，而BGP路由在AS内部传递时，路由的AS_Path属性值不会发生改变，如此一来，IBGP路由的防环就无法依赖AS_Path了。
- 为了防止BGP路由在AS内部传递时发生环路，BGP要求：“路由器不能将自己从IBGP对等体学习到的路由再传递给其他IBGP对等体”，这就是IBGP水平分割规则。
- 由于IBGP水平分割原则的存在，BGP要求AS内须保证IBGP对等体关系的全互联，因为只有这样，才能够确保每一个路由器都能学习到路由。

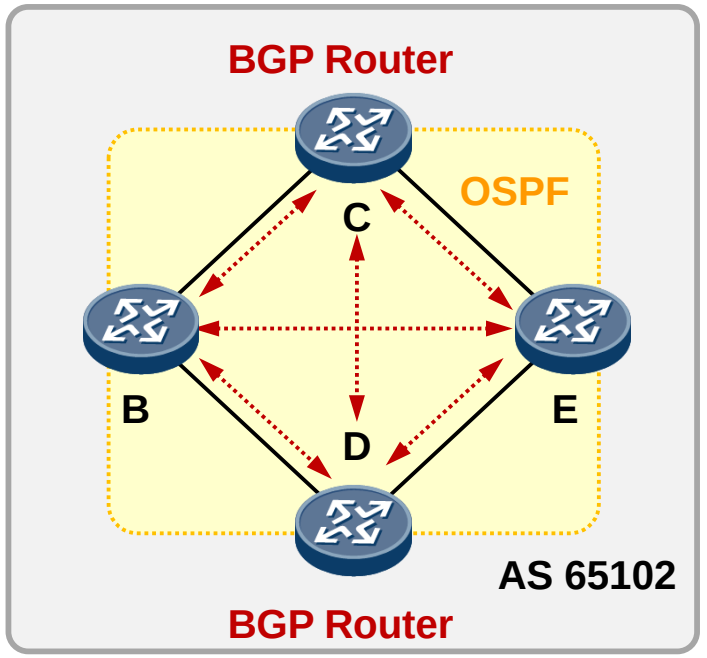
IBGP水平分割规则



C从E学习到的IBGP路由，由于水平分割规则的限制，不能够传递给B路由器，这将导致B无法学习到F通告的BGP路由。

IBGP水平分割规则

- 考虑到IBGP水平分割规则的限制，为了使得AS内的路由器都能够正常学习到BGP路由，我们不得不建立一个全互联的IBGP对等体关系（如图所示）。
- 然而在AS内的所有BGP路由器之间维护全互联的IBGP对等体关系是需要耗费大量资源的，网络的可扩展性、可维护性也非常差。解决方案：
  - **路由反射器**
  - **联邦**



BGP路由通告规则

- 当存在多条路径时，路由器只选取最优（Best）的BGP路由来使用（没有激活负载均衡的情况下）。
- BGP只把自己使用的路由，也就是自己认为最优的路由传递给对等体。
- 路由器从EBGP对等体获得的路由会传递给它所有的BGP对等体（包括EBGP和IBGP对等体）。
- 路由器从IBGP对等体获得的路由不会传递给它的IBGP对等体（存在反射器RR的情况除外）。
- 路由器从IBGP对等体获得的路由是否通告给它的EBGP对等体要视IGP和BGP同步的情况来决定。

BGP相关的几张表

名称	查看命令	说明
BGP邻居表	display bgp peer	列出本设备的BGP对等体，以及对等体的状态等信息。
BGP路由表	display bgp routing-table	列出本设备发现的所有BGP路由，如果到达同一个目的地存在多条路由，则将路由都进行罗列，但每个目的地只会优选一条路由。
路由表	display ip routing-table	设备的路由表，被优选的BGP路由会被加载到路由表中使用。

BGP邻居表

```
<Quidway> display bgp peer
 BGP Local router ID : 1.2.3.4
 local AS number : 10
 Total number of peers : 2
 Peers in established state : 1

 Peer       V    AS    MsgRcvd   MsgSent   OutQ     Up/Down       State          PrefRcv

 1.1.1.1    4    100   0         0         0        00:00:07      Idle           0
 1.2.5.6    4    200   32        35        0        00:17:49      Established    181
```

BGP当前的状态，当网络稳定时，一般状态需为Established

本端从对等体上收到路由前缀的数目

BGP路由表

```
<Quidway> display bgp routing-table
 BGP Local router ID is 1.1.1.2
 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
 *    1.1.1.0/24         1.1.1.1        0          0                     100 ?
 *    1.1.1.2/32         1.1.1.1        0          0                     100 ?
 *>   5.1.1.0/24         1.1.1.1        0          0                     100 ?
 *>   100.1.1.0/24       1.1.1.1        0          0                     100 ?
```

路由的网络号及掩码

BGP路由的下一跳IP地址

BGP路由度量值

BGP路由本地优先级

BGP路由协议首选值

BGP路由AS路径号及Origin属性

目录

HUAWEI

BGP基础配置

- **启动BGP进程，并指定BGP Router-ID：**

  [Router] **bgp** *as-num*

  [Router-bgp] **router-id** *x.x.x.x*

  - As-num参数为设备所处的AS号；
  - 为了增加网络的可靠性，建议将BGP Router-ID手工配置为设备Loopback接口的地址。如果没有配置，则BGP会自动选取系统视图下的Router-ID作为BGP协议的Router-ID。系统视图下的Router-ID选择规则，请参见命令router-id中的描述。

BGP基础配置

- **配置BGP对等体：**

  [Router-bgp] **peer** *x.x.x.x* **as-number** *as-num*

  - Peer关键字后面的x.x.x.x为对等体的IP地址，本设备与该IP地址之间必须路由可达。
  - 在BGP中，对等体需要通过peer命令手工指定，无法像IGP那样通过协议自动发现。
  - AS号码决定了与对等体建立的是EBGP会话还是IBGP会话。
- **（可选）指定用于建立BGP会话（也就是TCP连接）的源接口或源地址：**

  [Router-bgp] **peer** *x.x.x.x* **connect-interface** *intf* [*ipv4-src-address*]

  - 缺省情况下，BGP使用报文出接口的IP地址作为与对等体建立会话的源地址。

BGP引入IGP路由

- 需注意的是：BGP本身不发现路由，因此需要将其他路由引入到BGP路由表。
- BGP引入路由时支持Import和Network两种方式：
  - Import方式是按协议类型，将RIP、OSPF、ISIS等协议的路由引入到BGP路由表中。Import方式还可以引入静态路由和直连路由。
  - Network方式是逐条将IP路由表中已经存在的路由引入到BGP路由表中。
- BGP在引入IGP的路由时，可以使用路由策略进行路由过滤和路由属性设置。

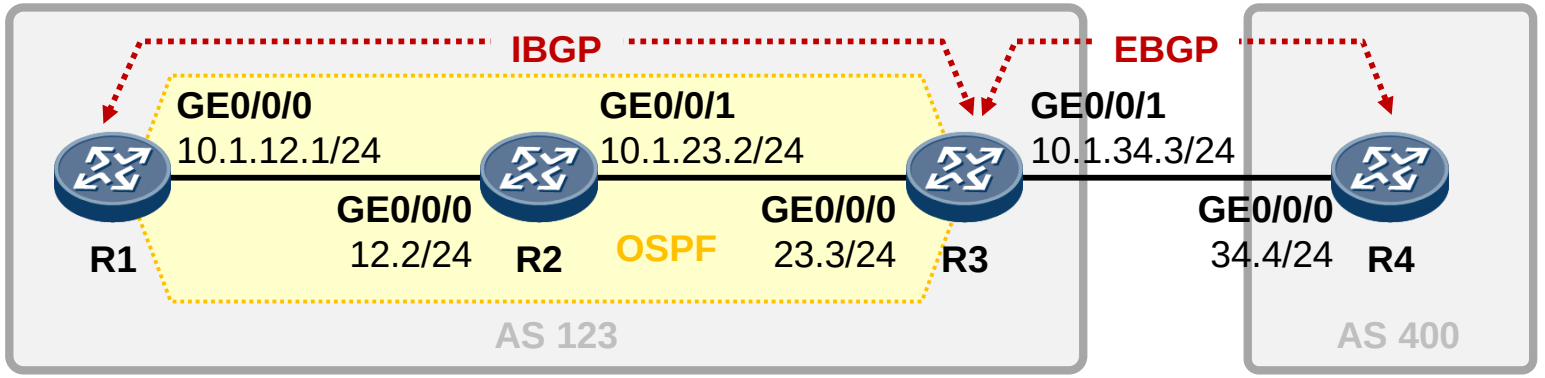
BGP基础配置

- **使用network命令将路由通告到BGP：**

  [router-bgp] **network** *ipv4-address* [*mask* | *mask-length*]

  - 使用network命令将路由表中的路由通告到BGP。路由必须存在于路由表中才能够被成功通告到BGP。
  - 如果上述命令没有指定mask或mask-length参数，则按有类地址处理。指定了mask，则仅当路由选择表中有与该网络完全匹配的条目时才被通告出去。
  - BGP中的network命令与IGP中的network是大不相同的，BGP中的network命令用于将路由通告到BGP，而不是在接口上激活BGP。
  - 在执行上述命令时，可以关联route-policy从而更为灵活的控制所引入的路由。

BGP基本配置 示例



1. R1、R2及R3属于AS123，R4属于AS 400；
2. AS123内的R1、R2及R3运行OSPF，通告各自直连接口（包括三台设备的Loopback0接口），注意OSPF域的工作范围；所有设备的Loopback0接口地址为x.x.x.x/32，其中x为设备编号（如R1的接口地址为1.1.1.1/32）。
3. R3与R4之间建立EBGP对等体关系，R2暂时不运行BGP，R1与R3之间建立IBGP对等体关系，所有的BGP对等体关系基于直连接口建立；R4将直连路由4.4.4.4/32通告到BGP，要求R1能学习到BGP路由4.4.4.4/32；
4. 修改BGP配置，使得R1与R3基于Loopback0接口建立IBGP对等体关系。

BGP基本配置 示例

- *Note：此处省略OSPF配置。*
- **R1的配置如下：**

  [R1] bgp 123

  [R1-bgp] router-id 1.1.1.1

  [R1-bgp] peer 10.1.23.3 as-number 123
- **R3的配置如下：**

  [R3] bgp 123

  [R3-bgp] router-id 3.3.3.3

  [R3-bgp] peer 10.1.12.1 as-number 123

  [R3-bgp] peer 10.1.34.4 as-number 400
- **R4的配置如下：**

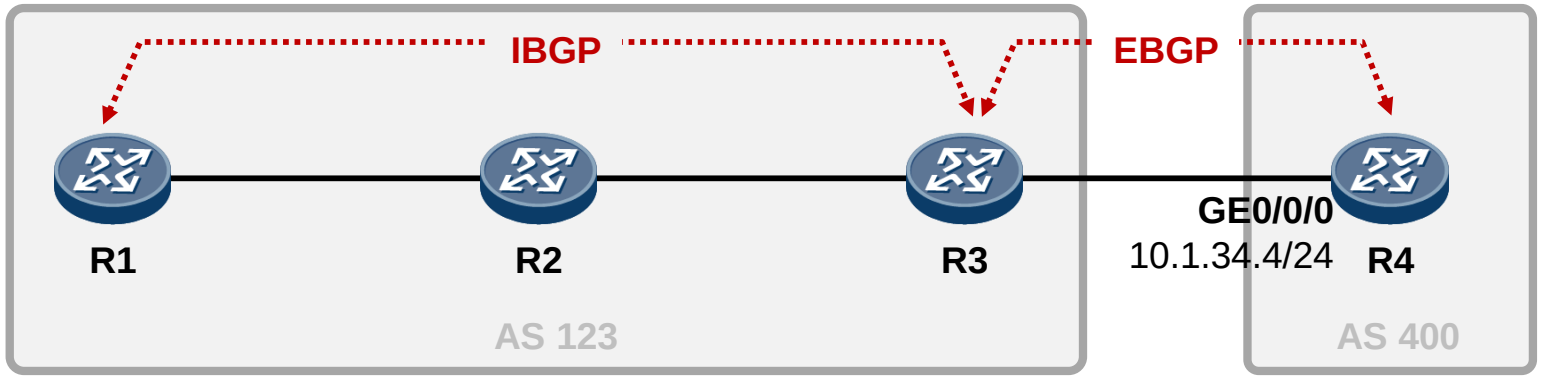
  [R4] bgp 400

  [R4-bgp] router-id 4.4.4.4

  [R4-bgp] peer 10.1.34.3 as-number 123

  [R4-bgp] network 4.4.4.4 32

查看R3的BGP表



[R3] display bgp routing-table

BGP Local router ID is 3.3.3.3

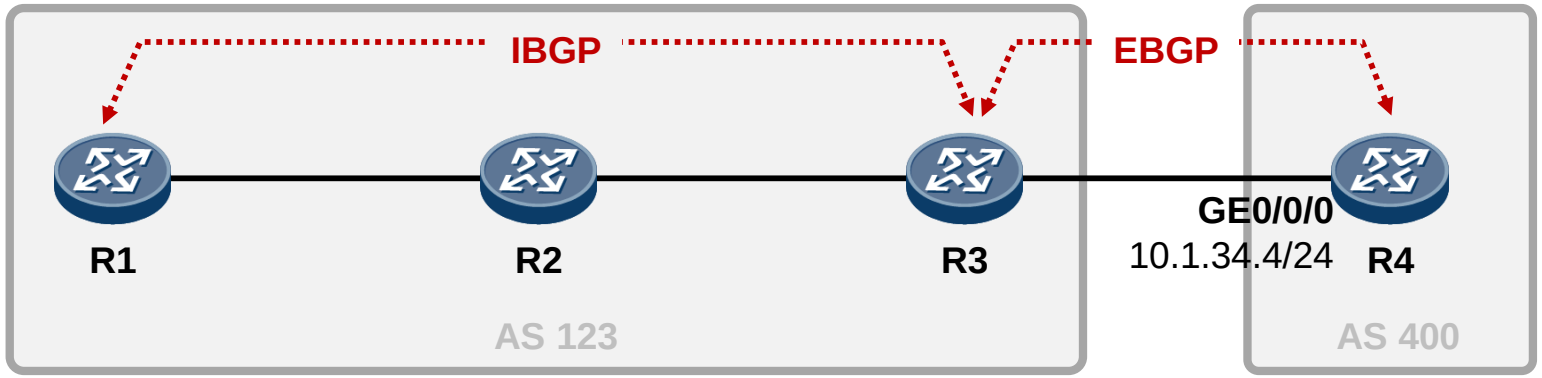
Status codes: * - valid, > - best, d - damped,
h - history, i - internal, s - suppressed, S - Stale
Origin : i - IGP, e - EGP, ? - incomplete

Total Number of Routes: 1

	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
*>	4.4.4.4/32	10.1.34.4	0		0	400 i

R3已经学习到R4传递过来的BGP路由4.4.4.4/32，并且该路由被优选（Best），路由前面有“>”标记。

查看R3的路由表

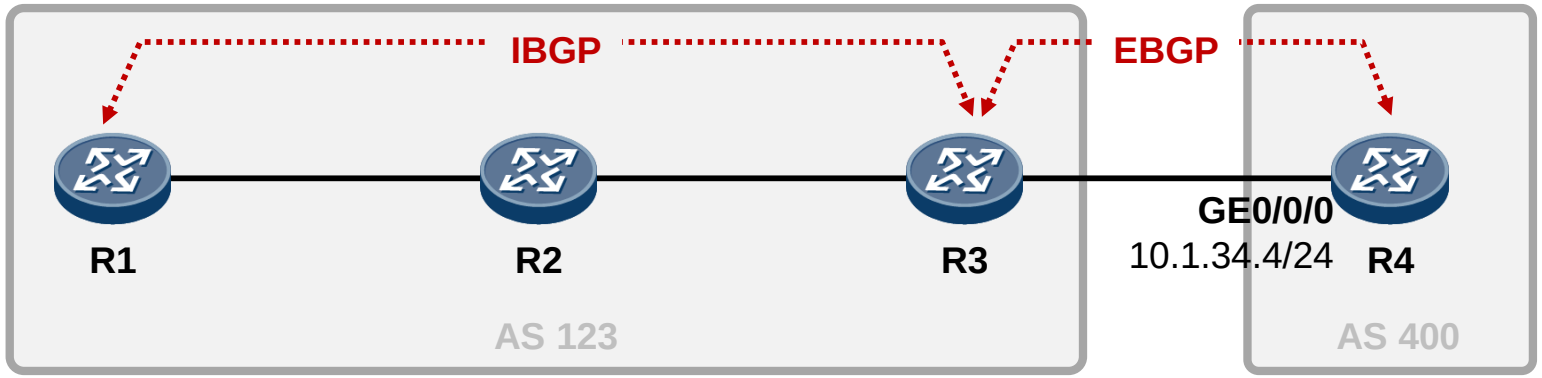


[R3] display ip routing-table protocol bgp

Destination/Mask	Proto	Pre	Cost	Flags	NextHop	Interface
4.4.4.4/32	EBGP	255	0	D	10.1.34.4	GigabitEthernet0/0/1

R3将优选的BGP路由加载到全局路由表中使用。

查看R1的BGP表



[R1] display bgp routing-table

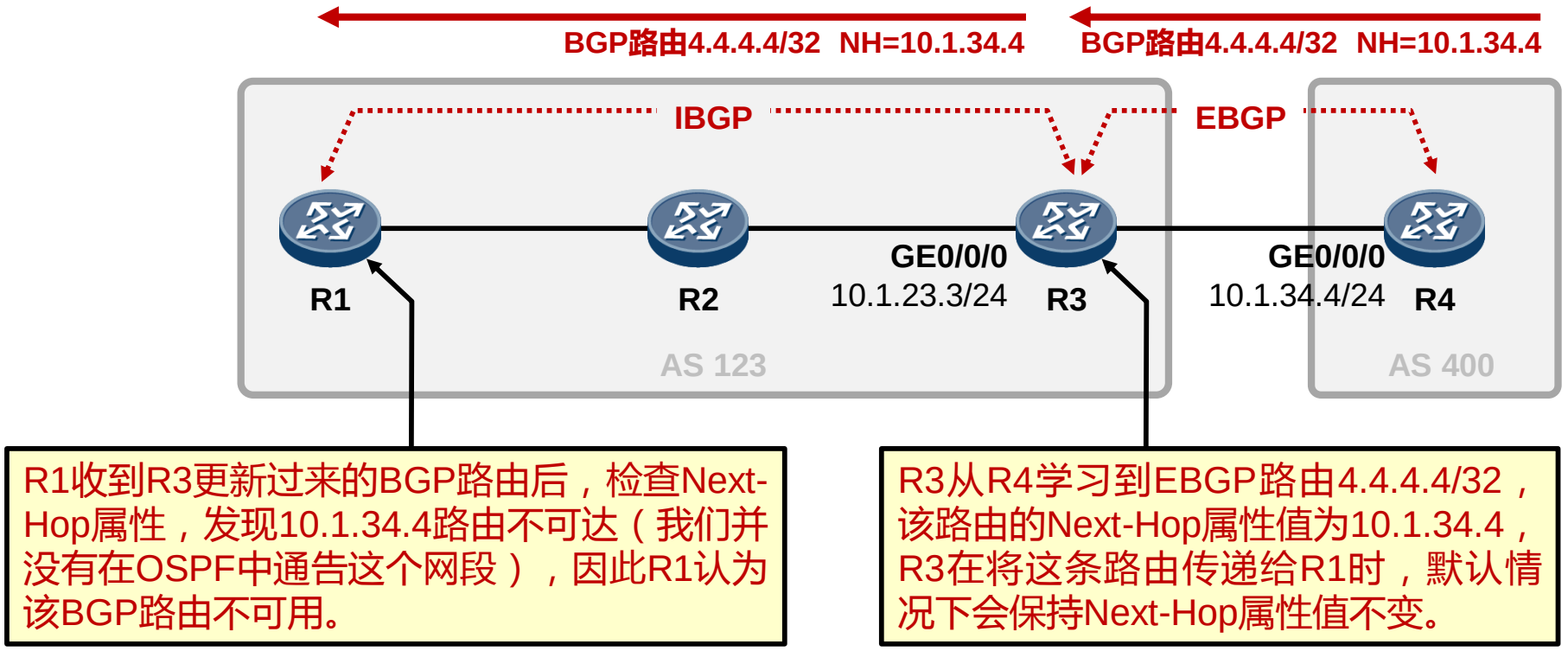
	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
i	4.4.4.4/32	10.1.34.4	0	100	0	400i

R1学习到R3传递过来的BGP路由4.4.4.4/32后，发现路由的Next-Hop（下一跳）是10.1.34.4，但是R1在本地路由表中没有到达10.1.34.0/24网段的路由，因此BGP路由Next-Hop不可达，该BGP路由也就不能被优选，更不能被加载到R1的路由表。 留意到路由前面没有*号，这表示路由不可用。

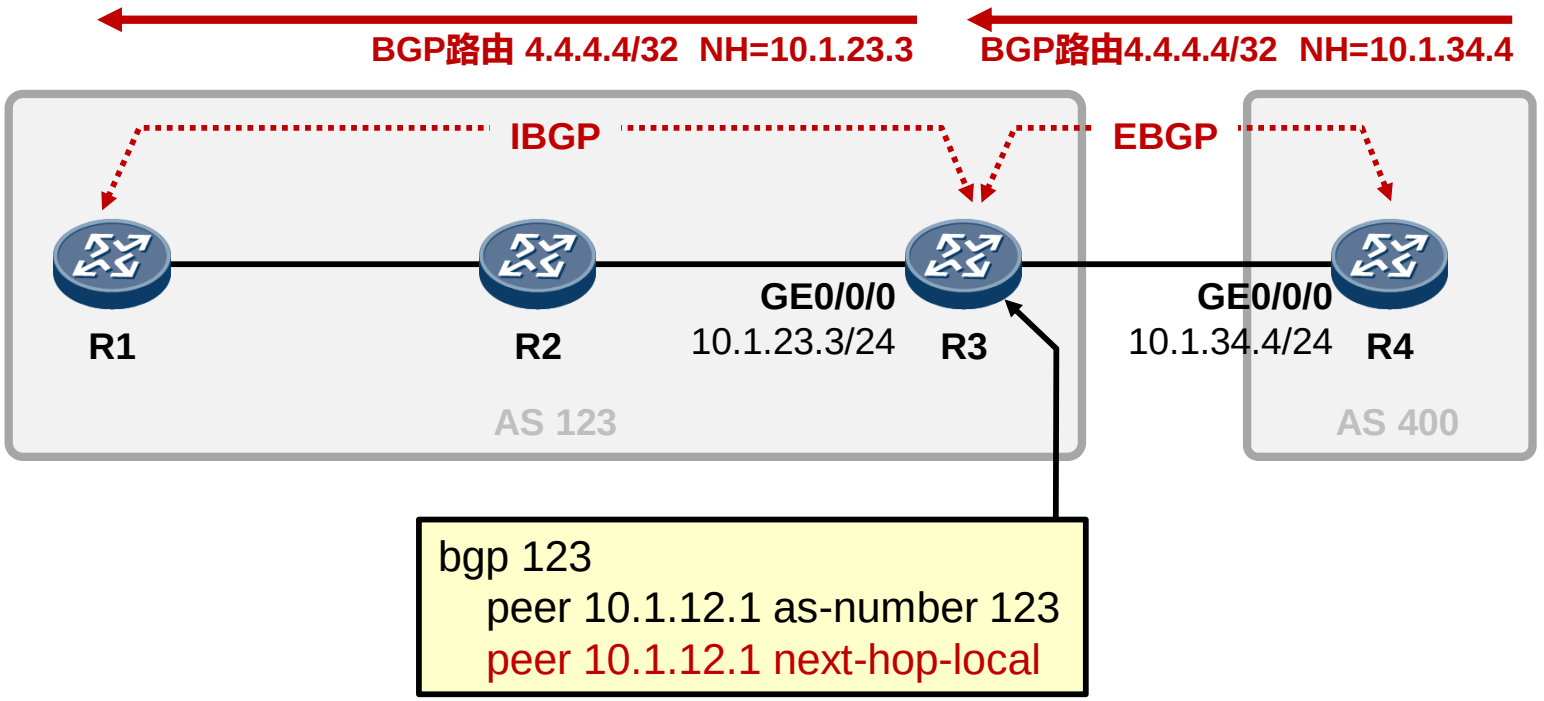
BGP基本配置-修改BGP路由下一跳为自身

- BGP是AS-by-AS的路由协议，而不是Router-by-Router的路由协议。在BGP中，Next-Hop属性值并不意味着是下一台路由器，而是到达下一个AS的IP地址。
- 当路由器将一条BGP路由传递给自己的EBGP对等体时，缺省情况下该路由的Next-Hop属性值为其更新源地址，一般为自己的接口IP地址。
  - 本例中R4将4.4.4.4/32路由通告给R3，该路由的Next-Hop属性值为10.1.34.4，也就是R4的地址。
- 当路由器将一条学习自EBGP对等体的路由传递给自己的IBGP对等体时，它会保持路由原有的Next-Hop属性不变。使用peer x.x.x.x next-hop-local命令可以将路由的Next-Hop修改为自己。
  - 本例中R3将学习自R4的4.4.4.4/32路由传递给IBGP对等体R1，保持路由的Next-Hop属性不变，依然为10.1.34.4，而此时10.1.34.4并不为R1所知，因此导致R1虽然学习到了这条BGP路由，却不可使用它。解决办法之一是，在R3上，针对R1配置peer 10.1.12.1 next-hop-local。

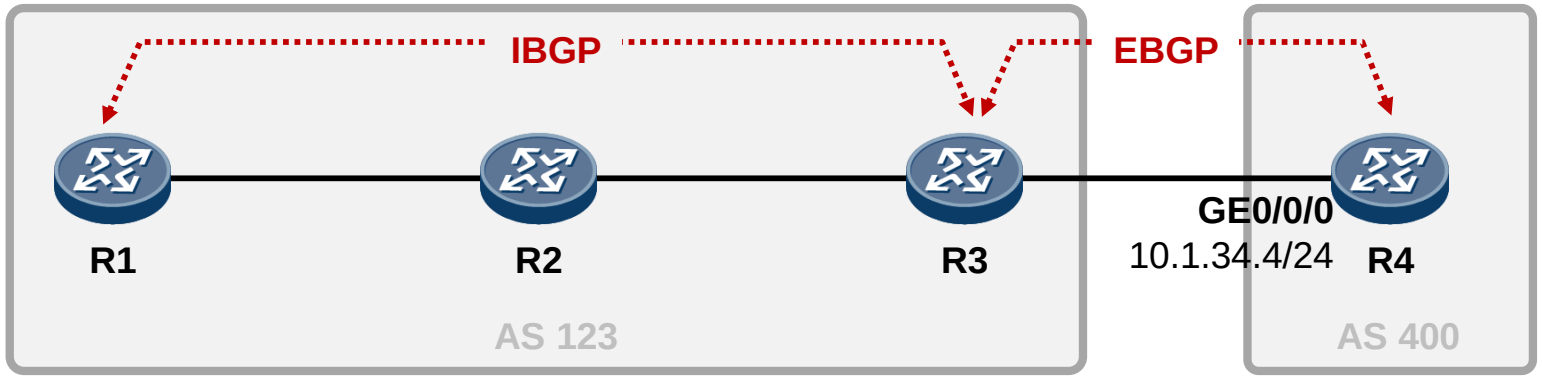
R3没有配置peer 10.1.12.1 next-hop-local命令前



BGP基本配置-修改BGP路由下一跳为自身



查看R1的BGP表

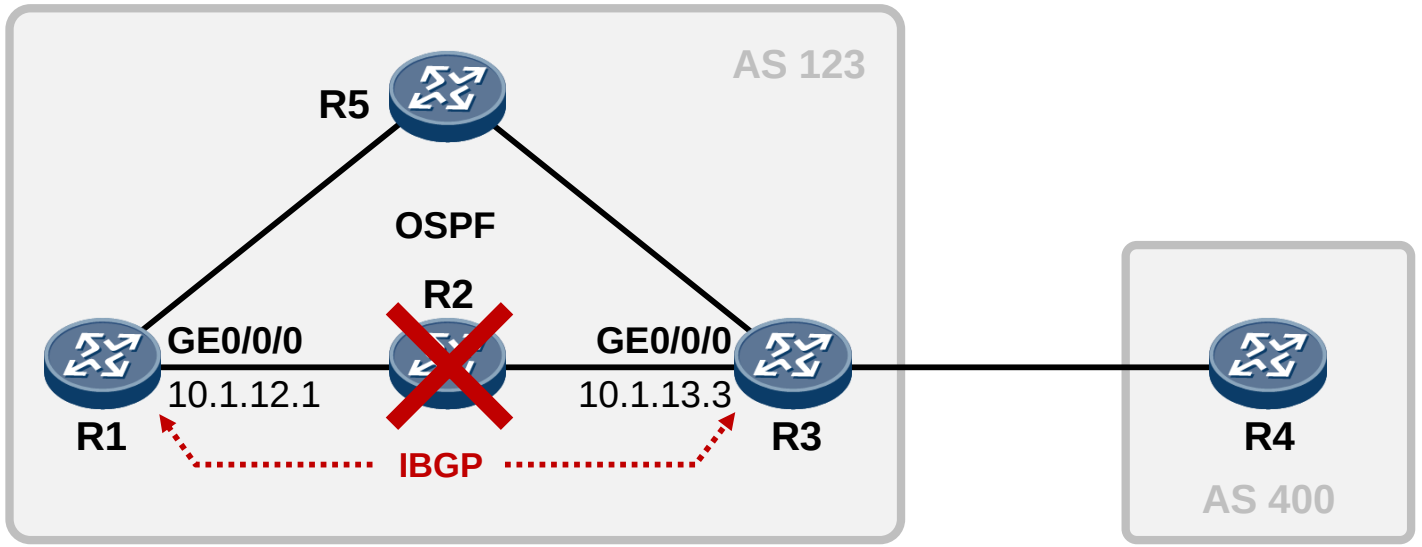


[R1] display bgp routing-table

	Network	NextHop	MED	LocPrf	PrefVal	Path/Ogn
i	4.4.4.4/32	10.1.23.3	0	100	0	400i

R3将EBGP路由4.4.4.4/32传递给R1时，将路由的下一跳属性值改成了自身的IP，而这个IP对于R1来说是路由可达的，因此路由优化了，并被装载进了路由表。

指定更新源IP



一般而言在AS内部，网络具备一定的冗余性。在R1与R3之间，如果采用直连接口建IBGP邻居关系，那么一旦接口或者直连链路发生故障，BGP会话也就断了，但是事实上，由于冗余链路的存在，R1与R3之间的IP连通性其实并没有DOWN（仍然可以通过R5到达彼此）。

BGP基本配置-指定更新源IP

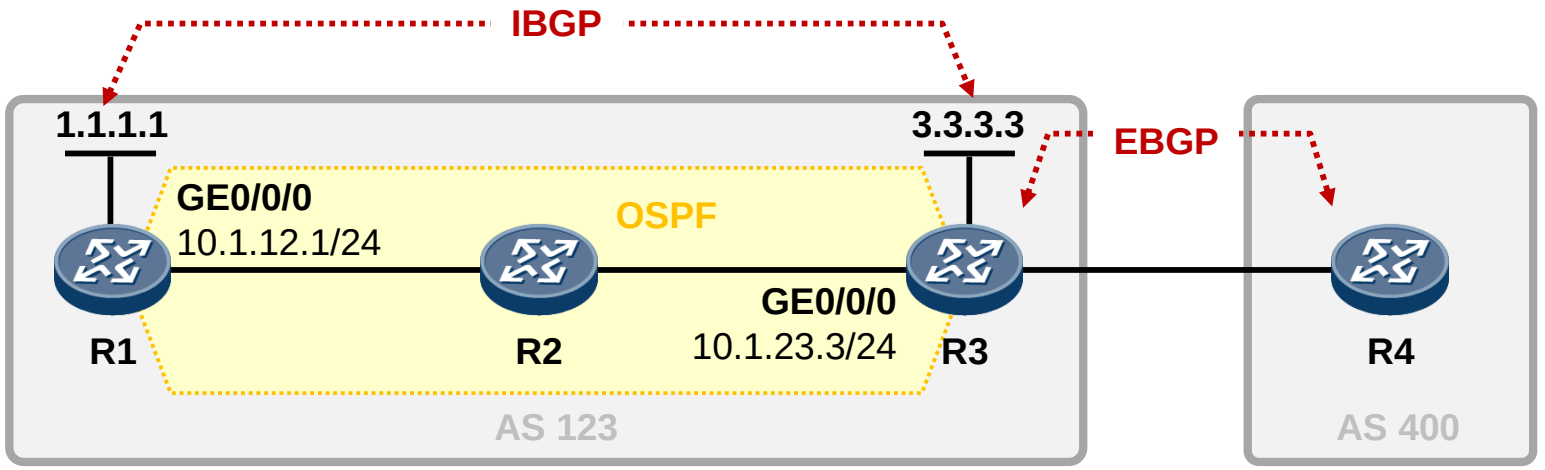
- 指定BGP连接所使用的建立TCP连接会话的源接口和源地址

  [Router-bgp] **peer** *x.x.x.x* **connect-interface** *intf* [*ipv4-src-address*]

  缺省情况下，BGP使用报文出接口的IP地址作为与对等体建立会话的源地址。

- 在部署IBGP对等体关系时，建议使用Loopback地址作为更新源地址。Loopback接口非常稳定，而且可以借助AS内的IGP和冗余拓扑来保证可靠性。
- 在部署EBGP对等体关系时，通常使用直连接口的IP地址作为源地址，如若使用Loopback接口建立EBGP对等体关系，则应注意EBGP多跳问题。

BGP基本配置-指定更新源IP



R1的关键性配置（修改）如下：

```
Interface loopback0
   ip address 1.1.1.1 32
bgp 123
   peer 3.3.3.3 as-number 123
   peer 3.3.3.3 connect-interface LoopBack 0
```

R1的关键性配置（修改）如下：

```
Interface loopback0
   ip address 3.3.3.3 32
bgp 123
   peer 1.1.1.1 as-number 123
   peer 1.1.1.1 connect-interface LoopBack 0
   peer 1.1.1.1 next-hop-local
```

注意，务必要在OSPF中通告R1及R3的Loopback0接口路由

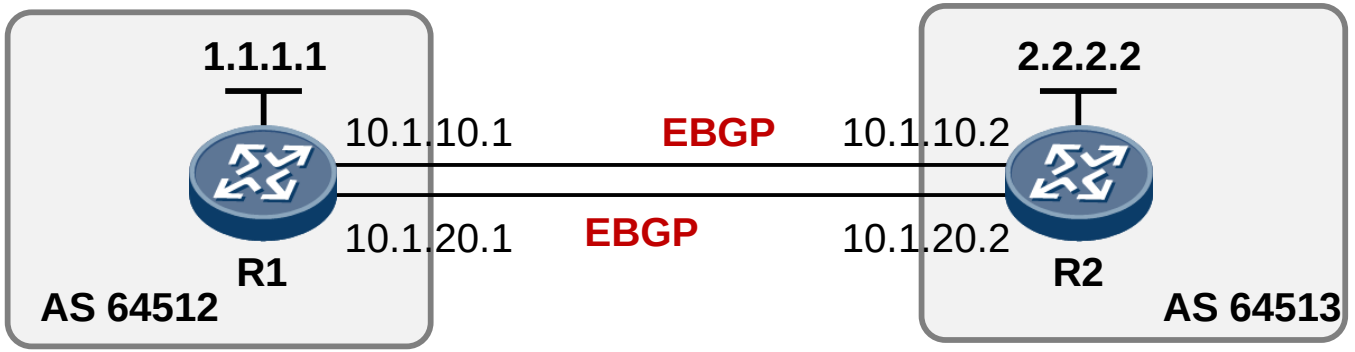
BGP基本配置-EBGP多跳

- 通常EBGP的对等体关系基于直连接口建立。如果EBGP的对等体关系并非基于直连接口建立，而是基于Loopback接口，又或者EBGP对等体不是直连的（中间隔着其他设备），那么要注意EBGP多跳的问题：在EBGP之间，所发送的BGP报文默认的TTL为1，因此如果EBGP对对等体之间存在多跳，则需修改最大跳数限制。
- 命令如下：

  [Router-bgp] **peer** *ipv4-address* **ebgp-max-hop** [*hop-count*] ,

  如果配置上述命令时没有指定参数hop-count，则为255。

BGP基本配置-EBGP多跳



R1的关键性配置如下：

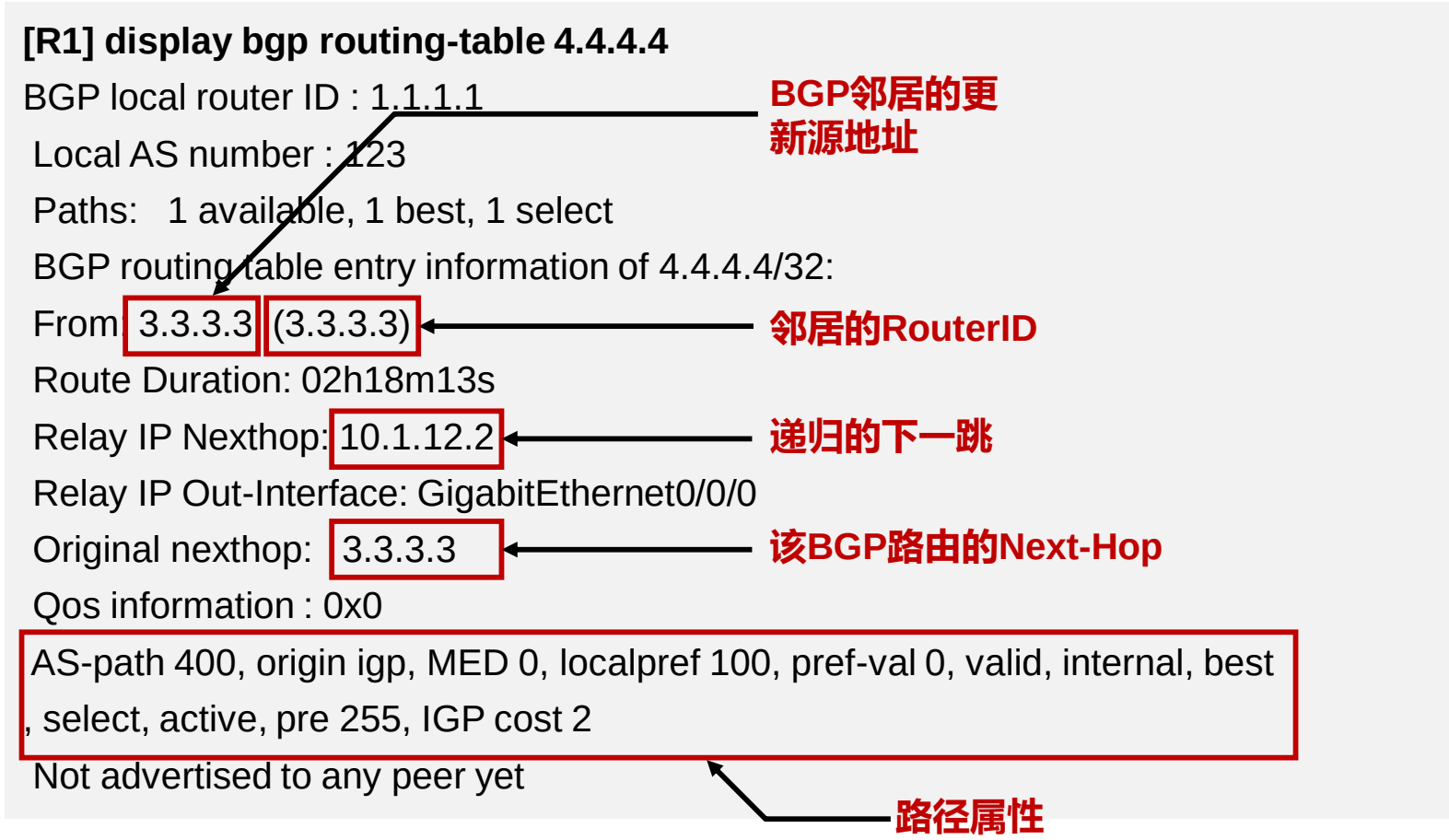
```
bgp 64512
  peer 2.2.2.2 as-number 64513
  peer 2.2.2.2 ebgp-max-hop 2
  peer 2.2.2.2 connect-interface loopback0
!
ip route-static 2.2.2.2 32 10.1.10.2
ip route-static 2.2.2.2 32 10.1.20.2 80
```

R2的关键性配置如下：

```
bgp 64513
  peer 1.1.1.1 as-number 64512
  peer 1.1.1.1 ebgp-max-hop 2
  peer 1.1.1.1 connect-interface loopback0
!
ip route-static 1.1.1.1 32 10.1.10.1
ip route-static 1.1.1.1 32 10.1.20.1 80
```

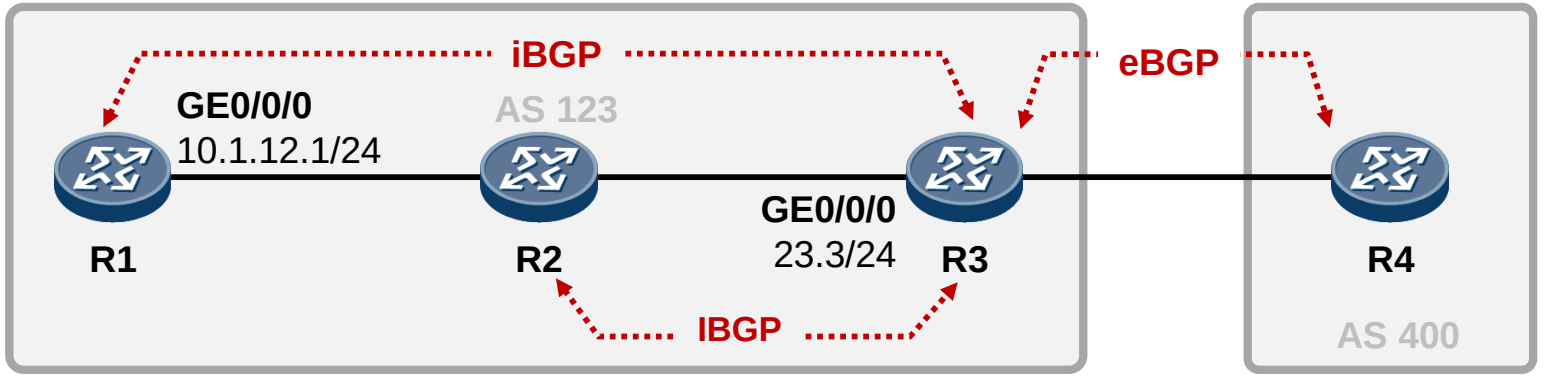
R1及R2之间存在多条物理链路，这些链路为两者间的通信提供了冗余。现在需要在R1及R2之间基于Loopback接口建立EBGP对等体关系，则必须修改最大跳数限制。

查看BGP条目详细信息



完成这个实验

- BGP对等体关系如图所示，完成该实验，使R1能够ping通R4的Loopback0接口地址。



Thank you

www.huawei.com