BGP案例分析：一些容易被忽略的因素

目录

硬件平台

路由器以及多层交换机设备

软件版本

运行IOS/IOS-XR的设备

案例介绍

BGP众多的属性特征与策略为进行灵活的路由控制提供了必要的基础，那么我们先来简单回顾一下常用的主要属性：

属性名称	类别
Origin	Well-known mandatory
As_Path	Well-known mandatory
Next_Hop	Well-known mandatory
Local_Pref	Well-known discretionary
Atomic_Aggregate	Well-known discretionary
Aggregator	Optional transitive
Community	Optional transitive
Multi_Exit_Disc(MED)	Optional non-transitive
Originatior_ID	Optional non-transitive
Cluster_List	Optional non-transitive

通常在AS内部一般用Local_Pref进行策略部署，而对于inter-AS而言As-Path是比较常用的策略，可以实现多种路由选路调整，结合团体属性以及符合规则的正则表达式我们可以为整个BGP网络设置灵活的策略体系。但是在一些情况下BGP策略部署可能与设计的预期有所不同，让我们来看看下面两个案例吧（由于IP地址及BGP信息涉及客户隐私，因此以下案例信息并不来自于客户真实设备，所有信息均来自于实验室设备，供学习参考）：

案例 1

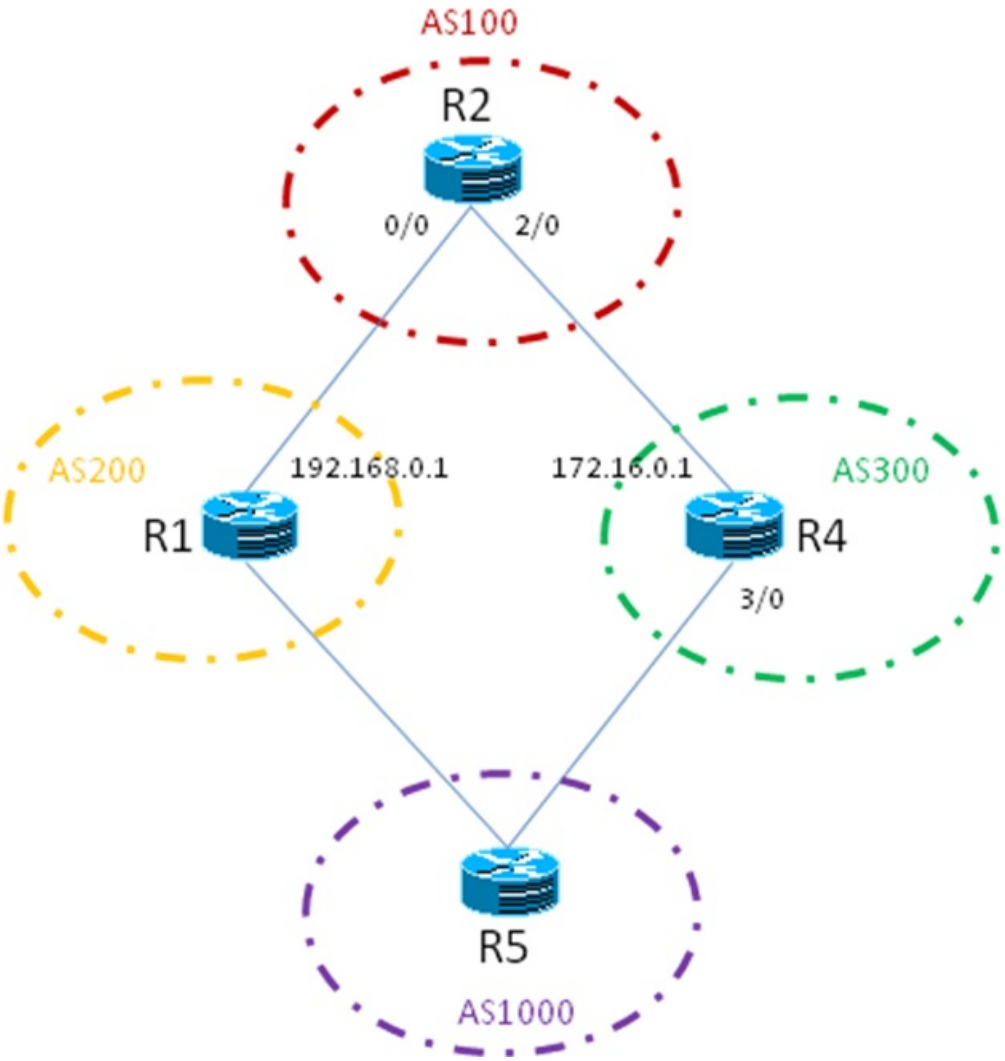
网络结构如下，其中R2为可控制路由器，R1和R4均为上联ISP路由器。网络基本设置如下：

1# R2-R1, R2-R4,R1-R5以及R4-R5之间均为EBGP邻居

2# AS200与AS300都能学习到AS1000的路由并且传递给AS100，在这个过程中都没有策略调整BGP路由属性。

3# R2与R1以及R2与R4之间均为10GE接口，R2在接收来自R4与R1的EBGP路由时也没有任何的route-map进行属性调整。

4# R4的路由器ID为172.16.0.1。R1的路由器ID为192.168.0.1



那么根据BGP选路原则，R2有可能优选R4最为到达R5 BGP路由的下一跳（因为R4拥有更低的路由器ID）。因此去往R5的路由选择如下线路R2—R4—R5。

但几个月后客户发现R2与R1之间的流量明显上升，而R2与R4之间的流量有所下降，检查路由表发现有大量之前选择R4的路由现在选择了R1，经过排查分析，案例中所涉及的R5,R1,R4以及R2对于这部分来自AS1000的路由都没有做任何调整。

问题分析思路

最初R2上的路由如下，他的确显示来自AS1000的路由会优选R4(AS300)。

```
R2#sho ip bgp
BGP table version is 9, local router ID is 2.2.2.2
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*  20.20.20.0/32    192.168.0.1                            0 200 1000 i
*>                  172.16.0.1                             0 300 1000 i
*  30.30.30.0/32    192.168.0.1                            0 200 1000 i
*>                  172.16.0.1                             0 300 1000 i
R2#
```

然而现在则优选R1(AS200)

```
R2#sho ip bgp
BGP table version is 11, local router ID is 2.2.2.2
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*  20.20.20.0/32    172.16.0.1                             0 300 1000 i
*>                  192.168.0.1                            0 200 1000 i
*  30.30.30.0/32    172.16.0.1                             0 300 1000 i
*>                  192.168.0.1                            0 200 1000 i
```

现在我们再来看一下关于具体BGP路由20.20.20.0/32的一些信息。

```
R2#sho ip bgp 20.20.20.0
BGP routing table entry for 20.20.20.0/32, version 11
Paths: (2 available, best #2)
Flag: 0x820
  Advertised to update-groups:
     2
  300 1000, (received & used)
    172.16.0.1 from 172.16.0.1 (4.4.4.4)
      Origin IGP, localpref 100, valid, external
  200 1000, (received & used)
    192.168.0.1 from 192.168.0.1 (1.1.1.1)
      Origin IGP, localpref 100, valid, external, best
```

对于来自R1(192.168.0.1)以及R4(172.16.0.1)的路由20.20.20.0。分析如下：

- 相同的weight权重（默认设置）
- 相同的as-path长度
- 相同的local-pref
- 相同的Original属性（同为external）
- 由于as-path的第一个AS号不同，因此默认情况下MED并不比较
- 到达IGP下一跳相同的cost（下图）

```
R2#sho ip rout 172.16.0.1
Routing entry for 172.16.0.0/30
  Known via "connected", distance 0, metric 0 (connected, via interface)
  Routing Descriptor Blocks:
  * directly connected, via Ethernet2/0
      Route metric is 0, traffic share count is 1

R2#sho ip rout 192.168.0.1
Routing entry for 192.168.0.0/30
  Known via "connected", distance 0, metric 0 (connected, via interface)
  Routing Descriptor Blocks:
  * directly connected, via Ethernet0/0
      Route metric is 0, traffic share count is 1
```

那么按照BGP选路原则路由器是否应该就选择拥有较小路由器ID的邻居作为下一跳呢？其实在进入这一项对比之前BGP会优选最早学习到的路由（也就是存在时间最长的路由），而这点正是我们这个案例的问题所在。

由于互联网络BGP路由有较多不定因素，因此对于某个具体条目的稳定性较难确定，所以对于本例来说R1-R5, R4-R5之间的线路稳定性将会决定R2对于来自R5(AS1000)的BGP路由选择。而这个因素并不是AS100自身所能控制的。因此R4-R5之间的链路震荡将导致路由优选到较稳定的R1-R5链接上。

问题总结

本案例的关键点在于R2(AS100)没有对来自R1(AS200)以及R4(AS300)学习的路由进行有效控制，最终使得路由选择受到非可控因素的影响。可考虑在R2上部署一些BGP策略，如调整相应的weight权重或者local_pref属性就可以实现灵活的策略控制和流量调整。

案例 2

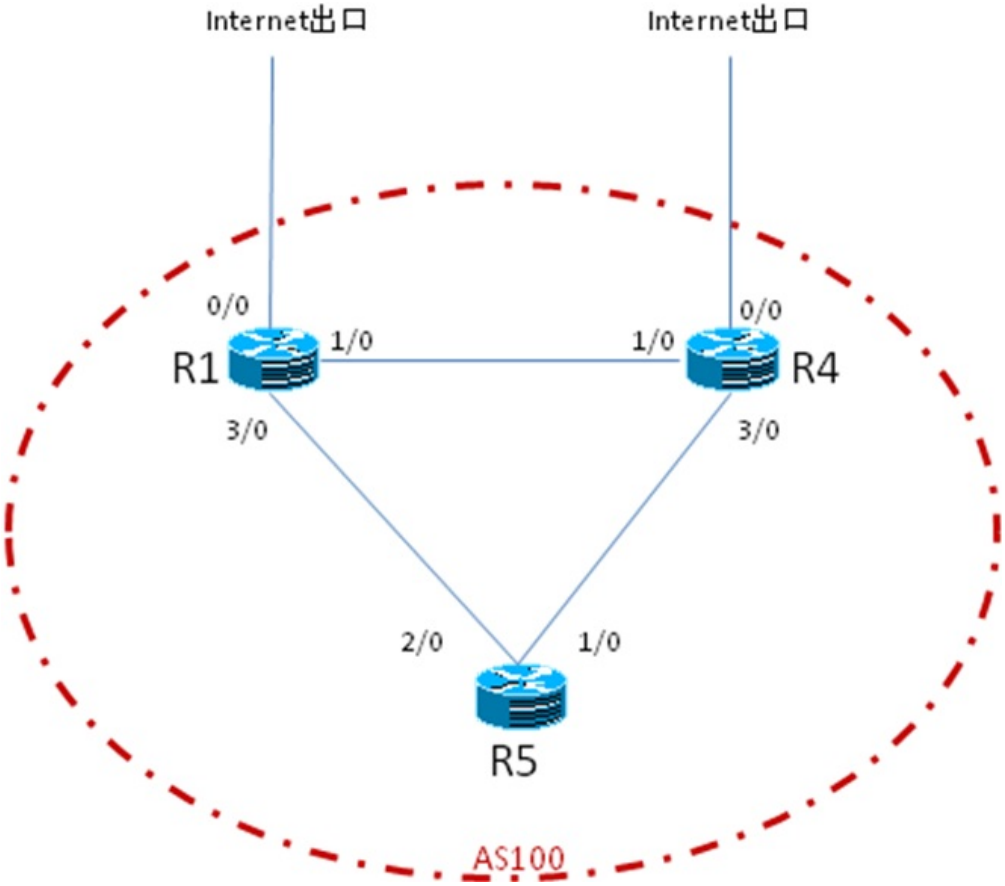
网络结构如下：

1# AS100拥有R1和R4两个出口路由器。AS100内部均为IBGP邻居

2# R1与R4上均设置默认路由作为出口，并通过IBGP进行分发

3# R1与R4作为AS100的核心汇聚路由器负责聚合AS100内的地址段，然后发布给EBGP邻居

4# R1与R4互相备份，当上行链路中断时流量自动切换到另一台设备



出于浮动路由的考虑R1与R4上设置的静态汇聚路由和默认路由均设置AD值为250。

```
ip route 0.0.0.0 0.0.0.0 192.168.0.6 250
ip route 30.0.0.0 255.255.0.0 Null0 250
ip route 40.0.0.0 255.255.0.0 Null0 250
```

然后通过BGP下的network进行分发：

```
router bgp 100
 no synchronization
 bgp log-neighbor-changes
 network 0.0.0.0
 network 30.0.0.0 mask 255.255.0.0
 network 40.0.0.0 mask 255.255.0.0
```

R1与R4相互学习，根据BGP选路原则，路由器会优选本地产生的路由，比如通过network或者aggregate产生的（本地产生的weight值为默认为32768，远端学习到的weight默认为0）。因此R1和R4仍然优选自身设置的路由。下图为R1上的路由信息。

```
R1#sho ip bgp
BGP table version is 4, local router ID is 1.1.1.1
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
* i0.0.0.0          192.168.0.6              0    100      0 i
*>                  192.168.0.2              0         32768 i
* i30.0.0.0/16      4.4.4.4                  0    100      0 i
*>                  0.0.0.0                  0         32768 i
* i40.0.0.0/16      4.4.4.4                  0    100      0 i
*>                  0.0.0.0                  0         32768 i
```

```
R1#sho ip bgp 0.0.0.0
BGP routing table entry for 0.0.0.0/0, version 2
Paths: (2 available, best #2)
  Advertised to update-groups:
     1
  Local, (received & used)
    192.168.0.6 (metric 20) from 4.4.4.4 (4.4.4.4)
      Origin IGP, metric 0, localpref 100, valid, internal
  Local
    192.168.0.2 from 0.0.0.0 (1.1.1.1)
      Origin IGP, metric 0, localpref 100, weight 32768, valid, sourced, local, best
R1#sho ip bgp 30.0.0.0
BGP routing table entry for 30.0.0.0/16, version 3
Paths: (2 available, best #2)
  Advertised to update-groups:
     1
  Local, (received & used)
    4.4.4.4 (metric 11) from 4.4.4.4 (4.4.4.4)
      Origin IGP, metric 0, localpref 100, valid, internal
  Local
    0.0.0.0 from 0.0.0.0 (1.1.1.1)
      Origin IGP, metric 0, localpref 100, weight 32768, valid, sourced, local, best
R1#
```

当R1上行链路中断时，默认路由切换到R4。

```
R1#sho ip bgp 0.0.0.0
BGP routing table entry for 0.0.0.0/0, version 5
Paths: (1 available, best #1)
  Not advertised to any peer
  Local, (received & used)
    192.168.0.6 (metric 20) from 4.4.4.4 (4.4.4.4)
      Origin IGP, metric 0, localpref 100, valid, internal, best
```

但是当R1链路恢复时路由仍然走R4，不能切换回R1。

```
R1#sho ip rout 0.0.0.0
Routing entry for 0.0.0.0/0, supernet
  Known via "bgp 100", distance 200, metric 0, candidate default path, type internal
  Advertised by bgp 100 (self originated)
  Last update from 192.168.0.6 00:02:57 ago
  Routing Descriptor Blocks:
  * 192.168.0.6, from 4.4.4.4, 00:02:57 ago
      Route metric is 0, traffic share count is 1
      AS Hops 0, BGP network version 0
```

问题分析思路

本案例的核心问题在于浮动路由的AD为250，而IBGP学习路由的默认AD值为200。如果在路由器上AD为250的这种浮动路由已经存在，并且network到IBGP中，那么通过IBGP学习到的路由就无法被优选（因为BGP将优选来自本地的路由），自然不会影响到我们的静态路由。

但是当静态路由消失时，比如链路中断，此时BGP表中的local也不存在了，那么从对端学习过来的IBGP路由就会生效，并最终放到路由表中。当静态路由再恢复的时候由于AD值小于IBGP，因此无法在对比时被优选。所以在R1上行链路恢复时默认路由无法切换回来。

既然找到了原因那么解决方法就出来了。把静态路由的AD设置为180或者其他低于200的值。

```
R1#sho ip rout 0.0.0.0
Routing entry for 0.0.0.0/0, supernet
  Known via "static", distance 180, metric 0, candidate default path
  Advertised by bgp 100
  Routing Descriptor Blocks:
  * 192.168.0.2
      Route metric is 0, traffic share count is 1
```

问题总结

AD值在路由协议之间进行对比，而各路由协议内有自身的对比规则，比如本例中BGP优选来自于Local的路由。而IBGP学习到的路由没有能够在BGP内部被优选，因此无法进入到AD对比环节，这就是最初问题没有暴露出来的原因。

关于浮动路由的AD值设置需要考虑整体的路由策略，因此如何去选择一个平衡的数值就比较重要。对于本例而言，在R1与R4上的静态路由是指导转发的，而相互学习的IBGP路由是备份，因此这样的默认路由AD值理应低于IBGP AD。 那么回过头来我们再看看那些汇聚路由的AD是否合理呢：

```
ip route 30.0.0.0 255.255.0.0 Null0 250
ip route 40.0.0.0 255.255.0.0 Null0 250
```

关于这样的BGP汇聚路由也建议设置AD的时候小于IBGP AD，主要原因如下：

1# 汇聚路由通常是发送给EBGP邻居使用，在本AS内不具备实际转发效应（由明细路由完成），但是一旦产生路由黑洞（汇聚就容易产生黑洞，即那些没有明细路由但被包含在汇聚路由网段内的地址），建议让去往黑洞的数据直接在本机丢弃（通过null 0实现），不要再转发至别的路由器。因此汇聚路由在本机生效有利于减少网络中的垃圾流量。

2# 按照之前的原理分析当R1与R4已经形成IBGP时，设置新的AD为250的汇聚路由必须在R1和R4两台设备上同时配置，并确保同时生效。假设首先在R1 上配置，那么R1会生效，并且发送给R4, 那么在R4通过IBGP学习到之后，再行配置的AD为250的汇聚路由就不会生效，因此网络中其实只有R1在通告该汇聚路由。一旦R1路由器故障，那么R4需要重新进行路由通告，从而增加网络的收敛时间。

经验总结

BGP有众多的属性特征从而可以部署灵活的策略体系，对于出现预期外的问题或者现象根据策略规则，进行细节分析都能找到问题的线索。一些不常用的属性和因素在某些情况下会成为问题的关键，本例可供大家学习参考，希望对您的BGP网络部署和排错有一些帮助。

相关命令

show ip bgp sum
show ip bgp neighbors x.x.x.x
show ip bgp neighbors x.x.x.x advertised-routes
show ip bgp neighbors x.x.x.x routes
show ip bgp neighbors x.x.x.x received-routes
show ip bgp x.x.x.x
show ip route