iBGP multipath在MPLS 环境下等价路由负载不均衡问题的排障

目录

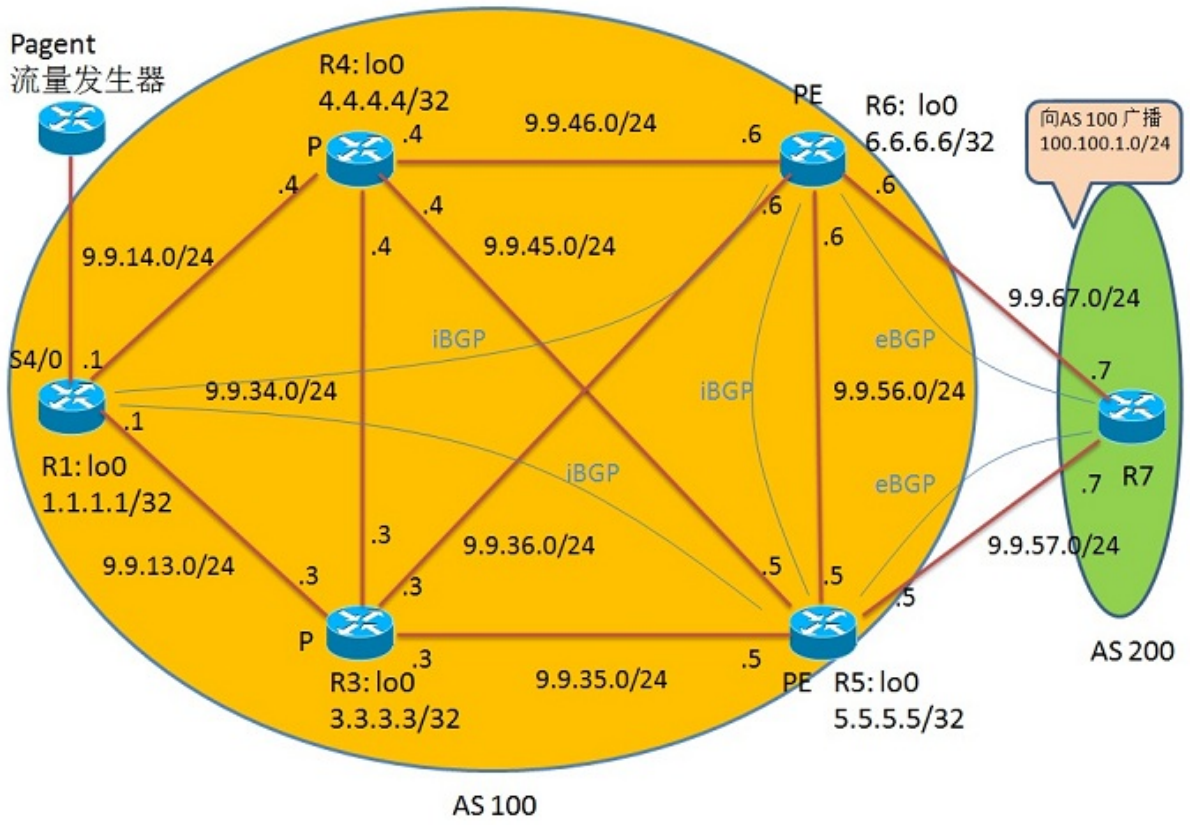
硬件平台

Cisco 12000 系列路由器

软件版本

12.0S，12.0SY

案例简介



上图是运营商中常见的一种网络拓扑，可见于城域网或省网。以上图为例，AS100 为城域网

，AS200为骨干网。AS100与AS200通过两条路径互连，设计上希望从AS100去往AS200的出口流量在两个出口上实现负载均担。常见的一种做法是在AS100内部的iBGP上配置iBGP multipath, 这样对每个AS200广播过来的路由，AS100内部的iBGP路由器对它有两个路径，分别指向两个出口路由器。以上图为例，AS200向AS100广播一个100.100.1.0/24的路由，R1上的BGP 表显示这个路由的两条路径都被选中，它们的下一跳分别指向R5和R6的loopback地址。

```
R1#show ip bgp 100.100.1.0
BGP routing table entry for 100.100.1.0/24, version 43
Paths: (2 available, best #1)
Multipath: iBGP
Flag: 0x820
  Not advertised to any peer
  200
    5.5.5.5 (metric 20) from 5.5.5.5 (5.5.5.5)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath, best
  200
    6.6.6.6 (metric 20) from 6.6.6.6 (6.6.6.6)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath
```

关于路由协议配置方面，AS100 内部跑IS-IS, R1，R5和R6之间跑iBGP。R1和R5上主要的BGP配置情况如下，R6的配置与R5类似：

R1：

```
router bgp 100
 bgp router-id 1.1.1.1
 neighbor 5.5.5.5 remote-as 100
 neighbor 5.5.5.5 update-source Loopback0
 neighbor 6.6.6.6 remote-as 100
 neighbor 6.6.6.6 update-source Loopback0
 maximum-paths ibgp 4
```

R5：

```
router bgp 100
 bgp router-id 5.5.5.5
 neighbor 1.1.1.1 remote-as 100
 neighbor 1.1.1.1 update-source Loopback0
 neighbor 1.1.1.1 next-hop-self
 neighbor 6.6.6.6 remote-as 100
 neighbor 6.6.6.6 update-source Loopback0
 neighbor 6.6.6.6 next-hop-self
 neighbor 9.9.57.7 remote-as 200
 maximum-paths ibgp 4
```

在大部分情况下，城域网内部都开启了MPLS，以提供MPLS L2/L3 VPN，MPLS TE等服务。对于MPLS，通常只有IGP路由及各个路由器的loopback地址参与MPLS标签分发。这种情况下，有时候会发现R5，R6上出口流量并不特别均衡的现象。

R5，R6上的出口流量均分是基于流的、per destination的统计上的平均，在城域或省域流量的情况下，两边的出流量不应相差很大，但问题是在上述部署中有时会出现两边出流量相差很大的情况，最坏时甚至只有一个出口有流量，而另一个出口几乎没有流量。 如上图的例子，只有 R5上有去往AS200的流量，而R6上去往AS200的流量几乎为零。

故障诊断步骤

通过对端口流量速率和MPLS forwarding 转发表的观察，通过R1去往AS200的流量只经由以下两种路径：

R1--->R3--->R5--->R7 和 R1--->R4--->R5--->R7

路由器转发所用的hash算法，以及流量的内容和特性等本身的原因基本可以排除。首先检查CEF 表：

```
R1#show ip cef 100.100.1.0 255.255.255.0
100.100.1.0/24, version 86, epoch 0, per-destination sharing
0 packets, 0 bytes
  tag information from 5.5.5.5/32, shared, all rewrites owned
    local tag: 43
  via 6.6.6.6, 0 dependencies, recursive
    traffic share 1
    next hop 9.9.14.4, Ethernet2/0 via 6.6.6.6/32 (Default)
    valid adjacency (0x0172F288)
    tag rewrite with Et2/0, 9.9.14.4, tags imposed {18}
  via 5.5.5.5, 0 dependencies, recursive
    traffic share 1
    next hop 9.9.14.4, Ethernet2/0 via 5.5.5.5/32 (Default)
    valid adjacency (0x0172F288)
    tag rewrite with Et1/0, 9.9.13.3, tags imposed {19}

  Multilevel loadinfo missing some IGP paths
  0 packets, 0 bytes switched through the prefix
  tmstats: external 0 packets, 0 bytes
           internal 0 packets, 0 bytes
```

虽然CEF表显示100.100.1.0/24有分别经过R5 loopback0 5.5.5.5 和R6 loopback0 6.6.6.6两条递归路由，但是在12.0S 和12.0SY系列的IOS 中，对于这种指向多个bgp nexthop的BGP路由，如果只是CEF转发，则CEF会选择所有的BGP nexthop路径，然后为每个BGP nexthop选取一个IGP路径（即使每个BGP nexthop也有多个IGP等价路径）。

MPLS与CEF不同，MPLS forwarding只会选择一个BGP nexthop，然后在这个nexthop 的所有LSP上做负载分担，这也是上面"tag information from 5.5.5.5/32，shared"所表示的，它只选了R5 5.5.5.5。

由于AS100内部的所有路由器都开启了MPLS，通过以下的MPLS forwarding表可以看到，去往100.100.1.0/24的流量会分别从E2/0发送并压上标签18和从E1/0发送并压上标签19，但这两种情况都是属于去往R5 loopback0 5.5.5.5的LSP的。

```
R1#show mpls forwarding 100.100.1.0 24 detail
Local  Outgoing    Prefix            Bytes tag  Outgoing    Next Hop
tag    tag or VC   or Tunnel Id      switched   interface
43     18          100.100.1.0/24    0          Et2/0       9.9.14.4
        MAC/Encaps=14/18, MRU=1500, Tag Stack{18}
        AABBCC00CC00AABBCC00C9028847 00012000
        No output feature configured
    Per-destination load-sharing, slots:  0 2 4 6 8 10 12 14
       19          100.100.1.0/24    0          Et1/0       9.9.13.3
        MAC/Encaps=14/18, MRU=1500, Tag Stack{19}
```

```
        AABBCC00CB00AABBCC00C9018847 00013000
        No output feature configured


R1#show mpls forwarding-table
Local  Outgoing    Prefix           Bytes tag  Outgoing    Next Hop
tag    tag or VC   or Tunnel Id     switched   interface
41     Pop tag     3.3.3.3/32       0          Et1/0       9.9.13.3
42     Pop tag     4.4.4.4/32       0          Et2/0       9.9.14.4
43     18          5.5.5.5/32       0          Et2/0       9.9.14.4
       19          5.5.5.5/32       0          Et1/0       9.9.13.3
44     19          6.6.6.6/32       0          Et2/0       9.9.14.4
       20          6.6.6.6/32       0          Et1/0       9.9.13.3
```

解决方案

由于Cisco 12000 上多层递归路由的等价路径负载分担功能目前只在IOS XR上支持，而IOS上并不支持，为了满足去往AS200的流量分担于R5和R6出口的需求，可以考虑在AS100内部的BGP路由器上启动BGP send label功能，让BGP为BGP路由分配标签，这样MPLS forwarding 会用上所有的BGP nexthop的 LSP。

相关配置和验证过程如下：

R1，R5和R6上都启用bgp neighbor send-label。以下仅列出R1和R5上的相关配置，R6与R5类似。

R1:

```
router bgp 100
 bgp router-id 1.1.1.1
 neighbor 5.5.5.5 remote-as 100
 neighbor 5.5.5.5 update-source Loopback0
 neighbor 5.5.5.5 send-label
 neighbor 6.6.6.6 remote-as 100
 neighbor 6.6.6.6 update-source Loopback0
 neighbor 6.6.6.6 send-label
 maximum-paths ibgp 4
```

R5：

```
router bgp 100
 bgp router-id 5.5.5.5
 neighbor 1.1.1.1 remote-as 100
 neighbor 1.1.1.1 update-source Loopback0
 neighbor 1.1.1.1 next-hop-self
 neighbor 1.1.1.1 send-label
 neighbor 6.6.6.6 remote-as 100
 neighbor 6.6.6.6 update-source Loopback0
 neighbor 6.6.6.6 next-hop-self
 neighbor 9.9.57.7 remote-as 200
 maximum-paths ibgp 4
```

然后可以看到R5和R6都通过BGP为100.100.1.0/24分配了标签：

```
R1#show ip bgp 100.100.1.0
BGP routing table entry for 100.100.1.0/24, version 31
Paths: (2 available, best #2)
Multipath: iBGP
  Not advertised to any peer
  200
    6.6.6.6 (metric 20) from 6.6.6.6 (6.6.6.6)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath
      mpls labels in/out BGP-route(from LDP)/31
  200
    5.5.5.5 (metric 20) from 5.5.5.5 (5.5.5.5)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath, best
      mpls labels in/out BGP-route(from LDP)/18
```

检查CEF表可以发现类似"tag information from 5.5.5.5/32, shared"的这样一行没有了，CEF选择两个BGP nexthop的LSP：

```
R1#show ip bgp 100.100.1.0
BGP routing table entry for 100.100.1.0/24, version 31
Paths: (2 available, best #2)
Multipath: iBGP
  Not advertised to any peer
  200
    6.6.6.6 (metric 20) from 6.6.6.6 (6.6.6.6)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath
      mpls labels in/out BGP-route(from LDP)/31
  200
    5.5.5.5 (metric 20) from 5.5.5.5 (5.5.5.5)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath, best
      mpls labels in/out BGP-route(from LDP)/18
```

再从MPLS forwarding表也可以看到100.100.1.0/24有经E2/0和E1/0去往5.5.5.5以及经E2/0和E1/0去往6.6.6.6的共四个LSP路径，在R1上接收多条BGP路由的情况下，通过Pagent 流量发生器向R1注入模拟的流量，可以观察到流量均分在下述的路径上：

R1--->R3--->R5
R1--->R3--->R6
R1--->R4--->R5
R1--->R4--->R6

```
R1#show mpls for 100.100.1.0 24 detail
Local  Outgoing    Prefix            Bytes tag  Outgoing   Next Hop
tag    tag or VC   or Tunnel Id      switched   interface
None   Recursive   100.100.1.0/24    0
        Recursive rewrite via 5.5.5.5/32, Tag Stack{18}
         00012000
        No output feature configured
    Per-destination load-sharing, slots:  0 2 4 6 8 10 12 14
       Recursive   100.100.1.0/24    0
        Recursive rewrite via 6.6.6.6/32, Tag Stack{31}
         0001F000
        No output feature configured
    Per-destination load-sharing, slots:  1 3 5 7 9 11 13 15
```

```
R1#show mpls forwarding-table
Local  Outgoing    Prefix           Bytes tag  Outgoing    Next Hop
tag    tag or VC   or Tunnel Id     switched   interface
41     Pop tag     3.3.3.3/32       0          Et1/0       9.9.13.3
42     Pop tag     4.4.4.4/32       0          Et2/0       9.9.14.4
43     18          5.5.5.5/32       0          Et2/0       9.9.14.4
       19          5.5.5.5/32       0          Et1/0       9.9.13.3
44     19          6.6.6.6/32       0          Et2/0       9.9.14.4
       20          6.6.6.6/32       0          Et1/0       9.9.13.3
```

经验总结

在IOS 12.0 系列上，对于BGP multipath，在没有启用MPLS时，路由器会在所有BGP nexthop上做负载分担，但对每个BGP nexthop只会选取一个IGP路径。

当在全网启用了MPLS之后，路由器只会选取一个nexthop做转发。以上图为例，如果R1分别只直连到R5和R6，对于有多个nexthop的BGP路由，出流量可能只会走一边，比如只走R1---R5而不走R1---R6。

Cisco早已注意到了这种情况，从12.0(33)S02开始，如果路由器上同时配置了iBGP multipath和MPLS，则IOS会给出如下提示，建议启用BGP send-label：

```
R1#show mpls for 100.100.1.0 24 detail
Local  Outgoing    Prefix           Bytes tag  Outgoing    Next Hop
tag    tag or VC   or Tunnel Id     switched   interface
None   Recursive   100.100.1.0/24   0
        Recursive rewrite via 5.5.5.5/32, Tag Stack{18}
         00012000
        No output feature configured
    Per-destination load-sharing, slots:  0 2 4 6 8 10 12 14
       Recursive   100.100.1.0/24   0
        Recursive rewrite via 6.6.6.6/32, Tag Stack{31}
         0001F000
        No output feature configured
    Per-destination load-sharing, slots:  1 3 5 7 9 11 13 15

R1#show mpls forwarding-table
Local  Outgoing    Prefix           Bytes tag  Outgoing    Next Hop
tag    tag or VC   or Tunnel Id     switched   interface
41     Pop tag     3.3.3.3/32       0          Et1/0       9.9.13.3
42     Pop tag     4.4.4.4/32       0          Et2/0       9.9.14.4
43     18          5.5.5.5/32       0          Et2/0       9.9.14.4
       19          5.5.5.5/32       0          Et1/0       9.9.13.3
44     19          6.6.6.6/32       0          Et2/0       9.9.14.4
       20          6.6.6.6/32       0          Et1/0       9.9.13.3
```

如果遵照IOS给出的建议，在iBGP multipath加MPLS的环境中启用BGP send-label，则可以实现BGP路由在所有iBGP nexthop的所有LSP上的负载分担。

其他相关文档

iBGP Multipath Load Sharing

Load Sharing MPLS VPN Traffic

RFC3107