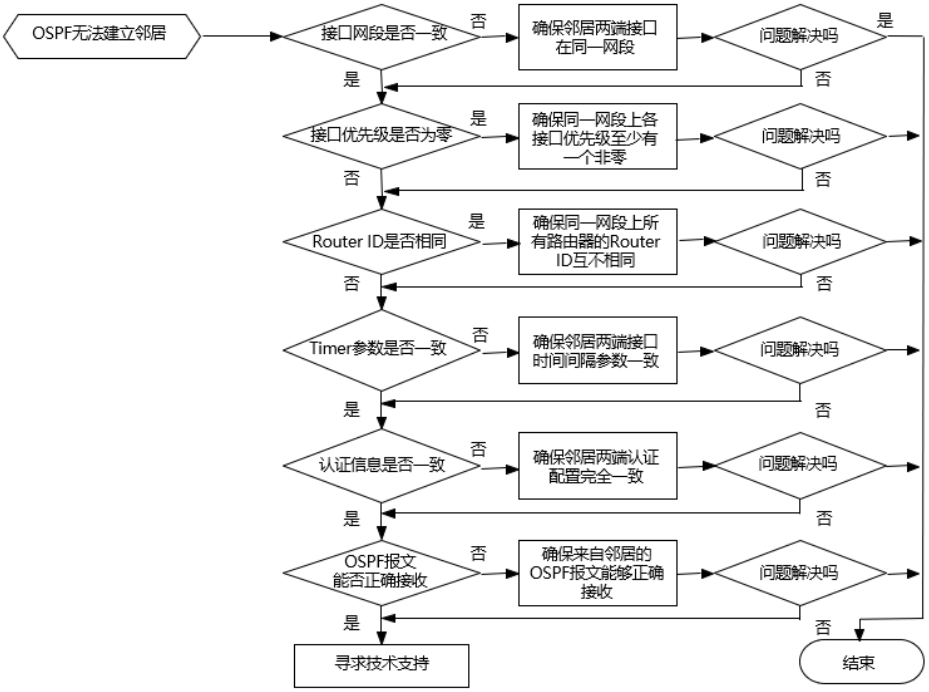
HCRSE118-故障案例分析

前言

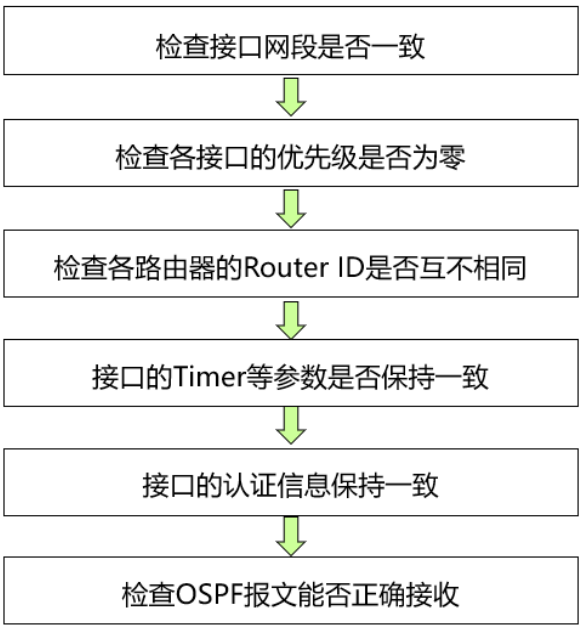
- 本课程介绍了路由器IP路由协议的故障处理流程和故障处理方法。
- 结合案例，分别针对OSPF、ISIS 、BGP协议和MPLS VPN常见的故障做深入探讨和分析，使大家能够进一步强化对路由协议故障排除的了解。

故障处理流程



- 对于 OSPF 来说，最为常见的 OSPF 故障即为 OSPF 的邻居建立失败的故障，
- 如果配置路由器后发现 OSPF 无法建立邻接关系。请使用如图所示的故障诊断流程。

故障处理步骤



- 步骤 1：检查接口网段是否一致
- 建立 OSPF 邻居时，Broadcast 和 NBMA 接口应该在同一网段，建立邻居的链路两端可以 **ping** 通，且接口所属区域的区域 ID、区域类型（包括 Nssa，Stub，Normal Area 等）应保持一致。
- 步骤 2：检查各接口的优先级是否有非零
- Broadcast 和 NBMA 类型的网段，各接口的优先级至少有一个是非零的，以确保能够正确的选举出 DR。否则各邻居只能达到 2-Way 的邻居状态。可以通过 **display ospf interface** 等命令查看接口的优先级。
- 步骤 3：检查各路由器的 Router ID 是否互不相同
- 同一 AS 内所有路由器的 Router ID 应该互不相同，否则会产生无法预料的路由振荡。可以通过 **display ospf brief** 等命令查看路由器的 Router ID。
- 步骤 4：接口的 Timer 等参数是否保持一致
- **ospf timer hello** 命令用来设置接口发送 Hello 报文的时间

间隔。缺省情况下，P2P（Point-to-Point）、Broadcast 类型接口发送 Hello 报文的时间间隔的值为 10 秒；P2MP（Point-to-Multipoint）、NBMA 类型接口发送 Hello 报文的时间间隔的值为 30 秒。

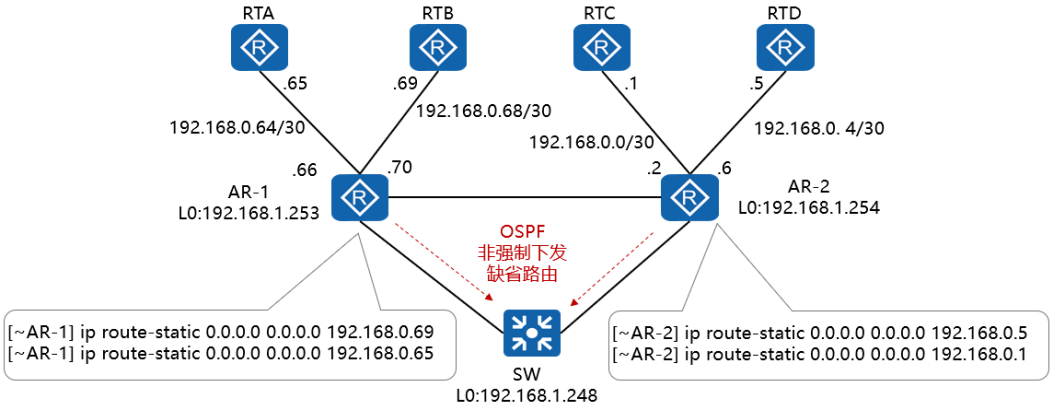
- **ospf timer dead** 命令用来设置 OSPF 的邻居失效时间，缺省情况下，P2P、Broadcast 类型接口的 OSPF 邻居失效时间为 40 秒，P2MP、NBMA 类型接口的 OSPF 邻居失效时间为 120 秒。
- 创建邻居时，对应接口时间间隔参数必须保持一致，否则无法建立邻居。可以通过 **display ospf interface verbose** 等命令查看。
- 步骤 5：接口的认证信息保持一致
- OSPF 在 Area 下和接口下分别有认证信息的配置。
- OSPF 认证的基本原则是：
- 如果接口下配置了认证，则使用接口下的认证；
- 如果接口下配置为 Null，则该接口不进行认证；
- 如果接口下没有配置认证（配置为 Null 不是没有配置认证），则采用 Area 下配置的认证；
- 如果 Area 也没有配置认证，则不认证。
- 创建邻居时，只有两端的认证配置完全一致时才能达到 Full 状态。
- 步骤 6：检查 OSPF 报文能否正确接收
- 有时 OSPF 的报文无法正确接收，原因有很多，先检查链路层是否畅通。可以打开 OSPF 的 Debug 开关查看报文的收发情况。Debug 命令有 **debugging ospf packet**、**debugging ospf event** 等，还可以通过 **display ospf error** 命令查看各种 OSPF 的错误计数。
- 如果 OSPF 的报文一切正常，请查看接口的 GTSM 配置是否正确。如果仅配置私网策略或仅配置公网策略，且将未匹配 GTSM 策略的报文的缺省动作设置为 pass，则其他实例的

OSPF 报文有可能被错误地丢弃。

- 打开 IP 报文的 Debug 信息来确认 IP 层是否转发成功。

通过 **debugging ip packet** 命令打开 IP 报文的 Debug 信息，可以增加 ACL Filter 对 Debug 信息过滤。

拓扑介绍



- 某企业网络出口 AR-1 和 AR-2 各有两个上行的 GE 接口，并分别定义两条缺省路由引导上行流量。
- 两台 AR 各自都在 OSPF 中非强制下发缺省路由给汇聚交换机 SW。

故障描述 (1)

- 正常时SW有两条OSPF默认外部路由分别指向两台AR。
- 如下情况时，SW上只有一条OSPF默认路由指到两台AR中的一台：
  - 在AR-1上删除下一跳为192.168.0.69 的缺省路由，保留下一跳为192.168.0.65的缺省路由。AR-2上还是两条缺省静态路由上行。

```
[~AR-1]undo ip route-static 0.0.0.0 0.0.0.0 192.168.0.69
```

故障描述 (2)

- 如在AR-1上删除下一跳为192.168.0.65 的缺省路由，保留下一跳为192.168.0.69的缺省路由。AR-2上还是两条缺省静态路由上行。则不存在上述问题。在SW上仍然有两条缺省路由分别指向两台AR。
- 类似地，如果在AR-2删除下一跳为192.168.0.5的缺省路由，保留下一跳为192.168.0.1的缺省路由。AR-1上还是两条缺省静态路由上行。SW上只有一条OSPF默认路由指到两台AR中的一台。

```
[~AR-2]undo ip route-static 0.0.0.0 0.0.0.0 192.168.0.5
```

- 如在AR-2上删除下一跳为192.168.0.1 的缺省路由，保留下一跳为192.168.0.5的缺省路由。AR-1上还是两条缺省静态路由上行。则不存在上述问题。在SW上仍然有两条缺省路由分别指向两台AR。

故障分析 (1)

- OSPF路由学习不正确可以从LSDB入手。在SW上查看OSPF缺省路由的LSA。

```
[~SW]display ospf lsdb ase
            OSPF Process 1 with Router ID 192.168.1.248
                  Link State Database
  Type      : External
  Ls id     : 0.0.0.0
  Adv rtr   : 192.168.1.254
  Ls age    : 482
  Len       : 36
  Options   :  E
  seq#      : 80000004
  chksum    : 0xeb78
  Net mask  : 0.0.0.0
  TOS 0  Metric: 1
  E type    : 2
  Forwarding Address : 0.0.0.0
  Tag       : 1
  Priority  : Low
```

故障分析 (2)

- 在SW上查看OSPF缺省路由的LSA。
- 可以看出FA地址被AR-1置错，此时SW上只有一条默认OSPF路由指向AR-2(192.168.1.254)。在最终比较相同的5类ASE LSA（E2、相同cost）时，会参考比较intra cost；cost值较小的路由将优选。

```
Type       : External
Ls id      : 0.0.0.0
Adv rtr    : 192.168.1.253
Ls age     : 149
Len        : 36
Options    :  E
seq#       : 8000000b
chksum     : 0xa50e
Net mask : 0.0.0.0
TOS 0  Metric: 1
E type     : 2
Forwarding Address : 192.168.0.65
Tag        : 1
Priority   : Low
```

- 按照RFC2328选路原则，在最终比较相同的5类ASE LSA（E2、相同cost）时，会参考比较intra cost，即到ASBR或者FA的cost值（FA为0，迭代ASBR；FA非0，迭代FA）；cost值较小的路由将优选（虽然最终不会将此cost加入到路由表cost值）。

故障分析 (3)

- 从如上故障现象说明过程中，我们发现SW上之所以出现OSPF路由学习不是预期得结果，根本的原因是AR将Forwarding Address设置错误。
- 下面来具体说明一下VRP平台填写5类LSA的FA地址及其路由计算规则。
  - FA填写为0.0.0.0时
    - 当一个5类LSA中的FA为0.0.0.0时，接收该LSA的路由器按照Adv Rtr（也就是ASBR）来计算下一跳。
  - FA填写为非0.0.0.0时，同时满足如下条件时，ASBR会在5类LSA的FA域内填写非0.0.0.0的转发地址，接收LSA的路由器按照该非0.0.0.0地址计算下一跳。
    - ASBR的下一跳接口路由可达。
    - ASBR与外部网络连接的下一跳接口没有被设置为被动接口。
    - ASBR与外部网络连接的下一跳接口不是OSPF P2P或P2MP类型的。
    - ASBR与外部网络连接的下一跳接口地址是落在OSPF协议中发布的网络范围之内。
  - 不满足如上四点条件的，FA都填写为0.0.0.0。

- 从如上故障现象说明过程中，我们发现SW上之所以出现OSPF路由学习不是预期得结果，根本的原因是AR将Forwarding Address设置错误。

- 下面来具体说明一下 VRP 平台填写 5 类 LSA 的 FA 地址及其路由计算规则。
- FA 填写为 0.0.0.0 时
- 当一个 5 类 LSA 中的 FA 为 0.0.0.0 时，接收该 LSA 的路由器按照 Adv Rtr（也就是 ASBR）来计算下一跳。
- FA 填写为非 0.0.0.0 时，同时满足如下条件时，ASBR 会在 5 类 LSA 的 FA 域内填写非 0.0.0.0 的转发地址，接收 LSA 的路由器按照该非 0.0.0.0 地址计算下一跳。
- OSPF 在 ASBR 与外部网络连接的下一跳接口启动。
- ASBR 与外部网络连接的下一跳接口没有被设置为被动接口。
- ASBR 与外部网络连接的下一跳接口不是 OSPF P2P 或 P2MP 类型的。
- ASBR 与外部网络连接的下一跳接口地址是落在 OSPF 协议中发布的网络范围之内。
- 不满足如上四点条件的，FA 都填写为 0.0.0.0。

故障处理

- 现在，我们找到了SW上OSPF路由计算之所以不是预期结果的原因，也知道了AR的OSPF 5类LSA填写FA地址的规则。因此比较简单的解决办是破坏FA填写非0地址的条件。三种解决方法如下：
  1. 检查两台AR的数据配置发现：AR-1（192.168.1.253）上OSPF里network了192.168.0.65/30网段，而192.168.0.68/30则没有；AR-2（192.168.1.254）上OSPF里network了192.168.0.0/30网段，而192.168.0.4/30则没有。所以只要在两台AR上OSPF内将对应静态路由下一跳网段的network配置给undo掉就可以了。
  2. 不用删除network配置，在OSPF视图下的配置不用更改，在两台AR被配置了network命令的接口下，配置ospf network-type p2p，对端接口也需要如此修改。
  3. AR上将对应接口设置为silence接口，或者让两AR的上行链路IP地址在下挂的SW上路由可达。
- AR如上修改后，SW上OSPF缺省路由计算就不会出现问题了。

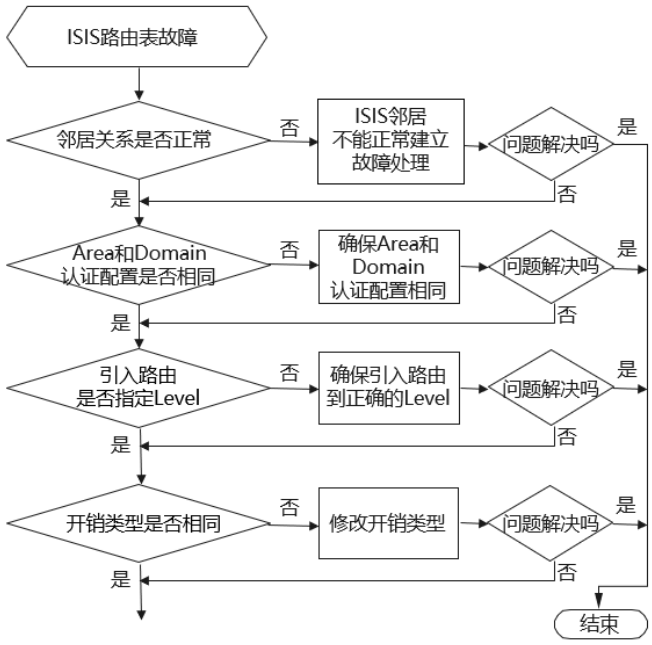
思考题

1. LSDB不显示引入的外部路由，如何处理?
2. ABR不能聚合区域网络地址，如何处理?
3. OSPF路由的相关LSA包含在LSDB中但在路由表中查找不到，如何处理?

- 问题：LSDB 不显示引入的外部路由，如何处理？
- 答：可能的原因如下。
- 使用 **display ospf interface** 命令查看运行 OSPF 协议的接口，确保接口不处于 Down 状态。
- 使用 **display ospf brief** 命令查看引入外部路由的路由器是否不属于 Stub 区域。
- 若外部路由是从邻居学到的，使用 **display ospf peer** 命令检查邻居状态是否为 Full。
- 查看是否配置了 **lsdb-overflow-limit** 命令，且外部路由的总数超过了 Over-Flow-Limit 的最大限制。
- 使用 **display ospf asbr-summary** 命令查看是否配置了 asbr-summary 命令对外部路由进行了聚合。
- 问题： ABR 不能聚合区域网络地址，如何处理？
- 答：可能的原因如下。
- 使用 **display current configuration** 命令确定区域中的网段地址连续。
- 如果不连续，则将它们分为几组连续的网段地址。
- 使用 **abr-summary** 命令在 ABR 上为每组连续的网段地址配置 ABR-Summary。
- 检查 Area view 下的 **filter { acl | ip-prefix** *prefix* **| route-policy** *route-policy-name* **} { import | export }**命令，确保 ABR 聚合的 LSA 没有被过滤掉。
- 问题： OSPF 路由的相关 LSA 包含在 LSDB 中但在路由表中查找不到，如何处理？

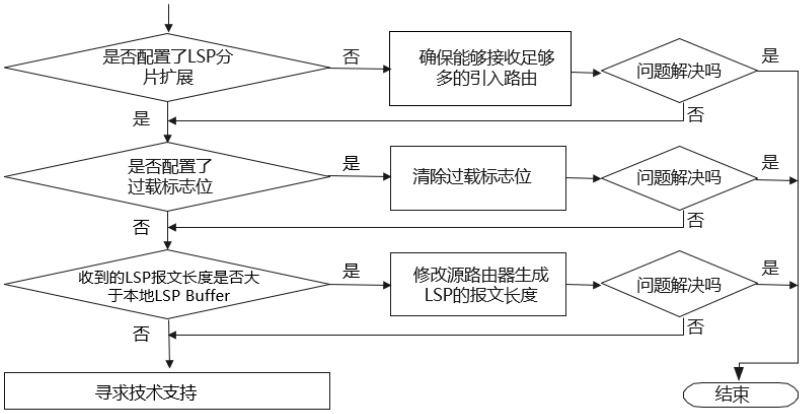
- 答：可能的原因如下。
- 查看 IP 地址是否配置正确。
- 查看转发地址是否已知。
- 查看是否正确路由聚合或路由引入。
- 查看是否配置了发布路由列表。
- 查看骨干区域是否中断。

故障诊断流程 (1)



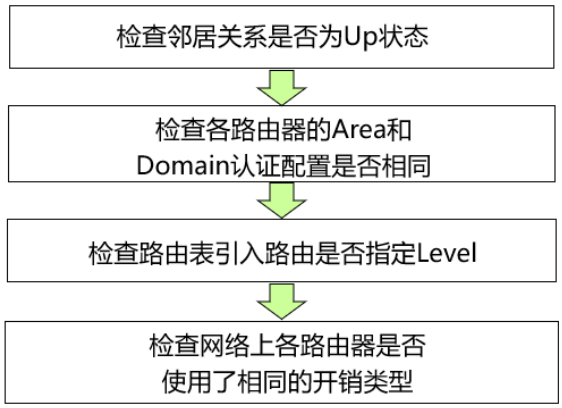
- ISIS 协议同 OSPF 均属于 IGP 协议，但 ISIS 协议在扩展性上有明显优势（例如对 IPv6 的支持），因此 ISIS 得到越来越广泛的应用。
- ISIS 的故障处理可参考如图所示的故障诊断流程

故障诊断流程 (2)



- ISIS 协议同 OSPF 均属于 IGP 协议，但 ISIS 协议在扩展性上有明显优势（例如对 IPv6 的支持），因此 ISIS 得到越来越广泛的应用。
- ISIS 的故障处理可参考如图所示的故障诊断流程

故障处理步骤 (1)

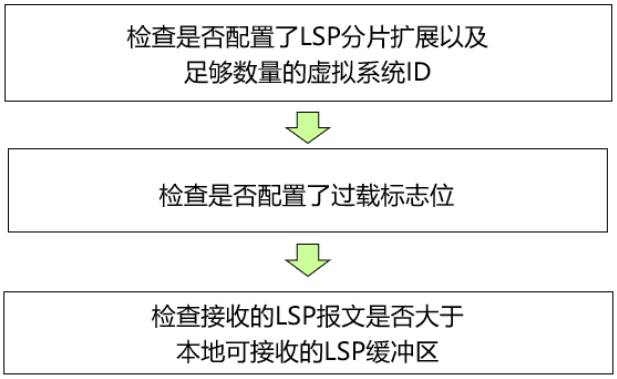


- 步骤 1：检查邻居关系是否为 Up 状态
- 使用 **display isis peer** 命令查看邻居关系是否为 Up 状态。
- 如果邻居状态为 Down，请参考 IS-IS 邻居不能正常建立

故障处理。

- 步骤 2：检查各路由器的 Area 和 Domain 认证配置是否相同
- 使用 **display isis lsdb** 命令查看邻居两端的 LSDB 内容是否一致。
- 如果 LSDB 没有同步，查看 Area 和 Domain 认证配置是否相同。
- 步骤 3：检查路由表引入路由是否指定 Level
- 如果引入路由到 Level-1 或 Level-1-2 路由表，在 IS-IS 视图下使用 **display this** 命令查看是否指定了级别。
- 步骤 4：检查网络上各路由器是否使用了相同的开销类型

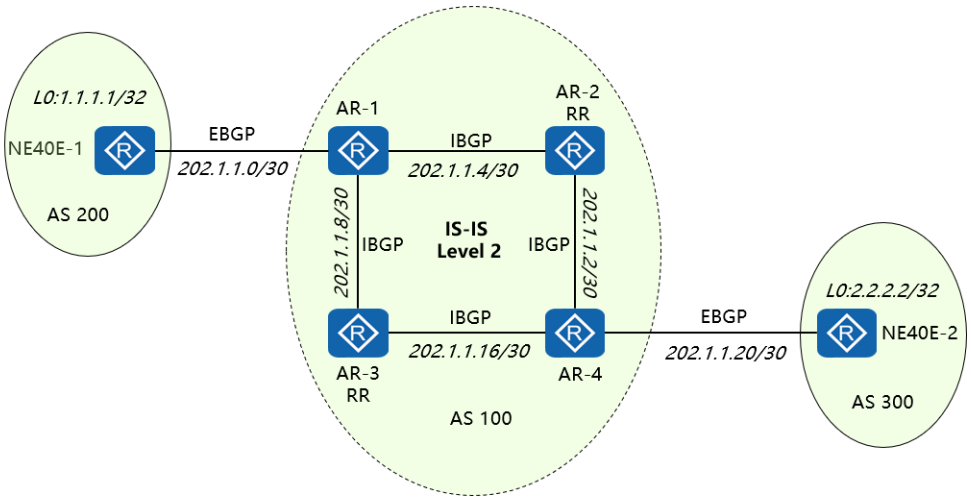
故障处理步骤 (2)



- 步骤 5：检查是否配置了 LSP 分片扩展以及足够数量的虚拟系统 ID
- 使用 **display isis statistics** 命令查看初始系统已使用的 LSP 分片数量，如果达到 256，则需要配置 LSP 分片扩展以及足够数量的虚拟系统 ID。
- 步骤 6：检查是否配置了过载标志位

- 如果配置了过载标志位，则该设备产生的 LSP 会通告其他设备本系统数据库过载，无法转发报文。其他设备就不会再将需要该设备转发的报文发送给它，除非报文的目的地址是该设备的直连地址。
- 使用 **undo set-overload** 命令清除过载标志位。
- 步骤 7：检查接收的 LSP 报文是否大于本地可接收的 LSP 缓冲区
- 如果对端发送 LSP 的长度大于本地可接收的 LSP 缓冲区，则本地 IS-IS 会丢弃这些 LSP 报文。
- 使用 lsp-length 命令修改产生 LSP 报文的长度或者接收 LSP 报文的长度。

拓扑介绍



- 某大型企业园区组网如图：
- NE40E-1 属于 AS 200，NE40E-2 属于 AS 300。
- AS 100 内的四台路由器建立 IBGP 邻居关系。AR-2 和 AR-3 是 BGP 路由反射器（RR），为 AR-1 和 AR-4 反射路由。
- AR-1 和 AR-4 之间没有直连链路，因此它们之间的 BGP 流量必须经过 RR 转发。
- AS 200 中 NE40E-1 通过主路径 AR-1 – AR-3 – AR-4 把

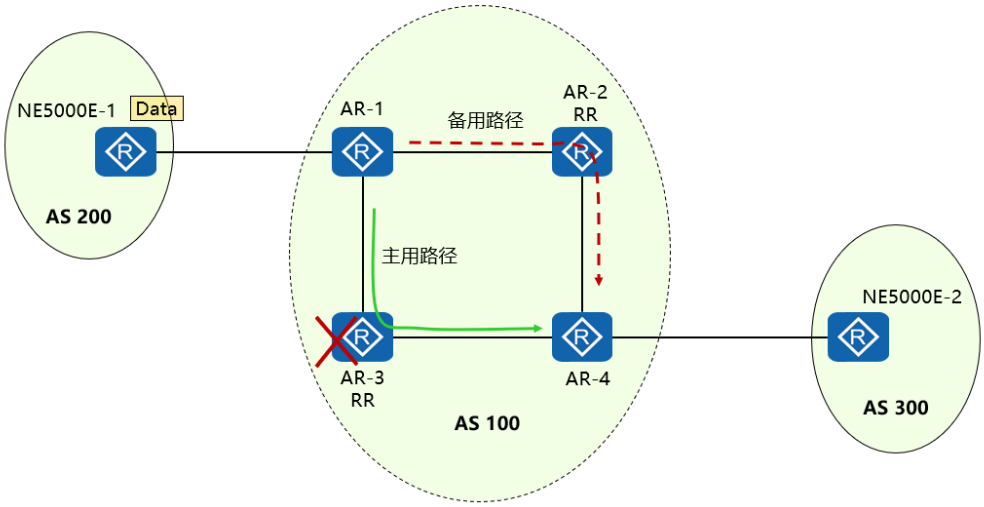
数据发往目的 NE40E-2。

- 路径 AR-1 – AR-2 – AR-4 为备用路径 。
- 通过调整 IGP 的 Cost 值使路由器优选 AR-1 – AR-3 – AR-4 路径转发 BGP 流量。

故障描述

- 当AR-3发生故障，设备重启。IS-IS快速收敛，AR-1引导BGP流量到AR-2转发。
- 当AR-3完成重启后，AR-1到AR-4的BGP流量暂时中断，几分钟后流量又恢复。

故障模拟



故障分析 (1)

- AR-3 中断，检查 AR-1 上路由表，此时 AR-1 选择了 ISIS 的备用路由 AR-2(202.1.1.6)。

```
[~AR-1]display ip routing-table 2.2.2.2 verbose
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1

    Destination:  2.2.2.2/32
       Protocol:  BGP              Process ID: 0
     Preference:  255                    Cost: 0
        NextHop:  10.1.1.4         Neighbour: 10.1.1.2
          State:  Active Adv Relied      Age: 00h00m01s
            Tag:  0                  Priority: low
          Label:  NULL               QoSInfo: 0x0
     IndirectID:  0x2
  RelayNextHop:  202.1.1.6          Interface: Ethernet0/0/0
       TunnelID:  0x0                    Flags: RD
```

故障分析 (2)

- AR-3重启完成，再次检查AR-1上路由表，此时AR-1选择了ISIS的主用路由AR-3(202.1.1.10)。

```
[~AR-1]display ip routing-table 2.2.2.2 verbose
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1

Destination: 2.2.2.2/32
       Protocol: BGP                Process ID: 0
     Preference: 255                      Cost: 0
        NextHop: 10.1.1.4           Neighbour: 10.1.1.2
          State: Active Adv Relied        Age: 00h04m34s
            Tag: 0                    Priority: low
          Label: NULL                 QoSInfo: 0x0
     IndirectID: 0x2
  RelayNextHop: 202.1.1.10          Interface: Ethernet0/0/4
       TunnelID: 0x0                      Flags: RD
```

故障分析 (3)

- 然而此时在AR-3上并未完成BGP的收敛，结果导致数据包在AR-3上被丢弃。

```
[~AR-3]display isis peer

                    Peer information for ISIS(1)

System Id          Interface   Circuit Id            State  HoldTime Type     PRI
------------------------------------------------------------------------------
0001.0001.0001  Eth0/0/2   0003.0003.0003.01 Up    2s         L2       64
0004.0004.0004  Eth0/0/3   0003.0003.0003.02 Up    1s         L2       64

Total Peer(s): 2

<AR-3>display bgp peer

 BGP local router ID : 10.1.1.3
 Local AS number : 100
 Total number of peers : 2              Peers in established state : 0

  Peer       V   AS  MsgRcvd  MsgSent  OutQ  Up/Down       State PrefRcv

  10.1.1.1  4   100     0          0          0   00:00:26      Idle      0
  10.1.1.4  4   100     0          0          0   00:00:26      Idle      0
```

故障分析 (4)

- 至此，故障产生的原因已明确：
  - AR-3网元重启恢复以后，在短短的几秒钟内，AR-1、AR-4与AR-3的IS-IS邻居状态建立且数据库同步完成， AR-1的FIB被刷新，送到NE40E-2的流量被AR-1送到AR-3上，但是由于BGP收敛较慢，短时间内，AR-3还来不及学习到有关NE40E-2的BGP路由信息，所以， AR-3将把AR-1发来的去往NE40E-2的数据包丢弃。于是临时的路由黑洞产生了。

- AR-3 恢复以后，在短短的几秒钟内，AR-1、AR-4 与 AR-3 的 IS-IS 邻居状态建立且数据库同步完成， AR-1 的 FIB 被刷新，送到 NE40E-2 的流量被 AR-1 送到 AR-3 上，但是由于 BGP 收敛较慢，短时间内，AR-3 还来不及学习到有关 NE40E-2 的 BGP 路由信息，所以， AR-3 将把 AR-1 发来的去往 NE40E-2 的数据包丢弃。于是临时的路由黑洞产生了。

故障处理

- 华为设备实现了ISIS的一个功能，通过设置overload位来避免临时的路由黑洞，命令如下：
  - set-overload [on-startup [wait-for-bgp [timeout1]] [allow { interlevel | external } *]
    - wait-for-bgp：系统启动时设置过载标志位，在BGP收敛后取消。如果BGP没有发信号通知IS-IS已收敛完成，IS-IS将在指定的超时时间或缺省的10分钟后（没有指定超时时间）取消过载标志位。
    - interlevel：当配置allow时，允许发布从不同层次IS-IS学来的IP地址前缀。
    - external：当配置allow时，允许发布从其它协议学来的IP地址前缀。
- 在AR-3上配置该命令，可以解决上述故障。

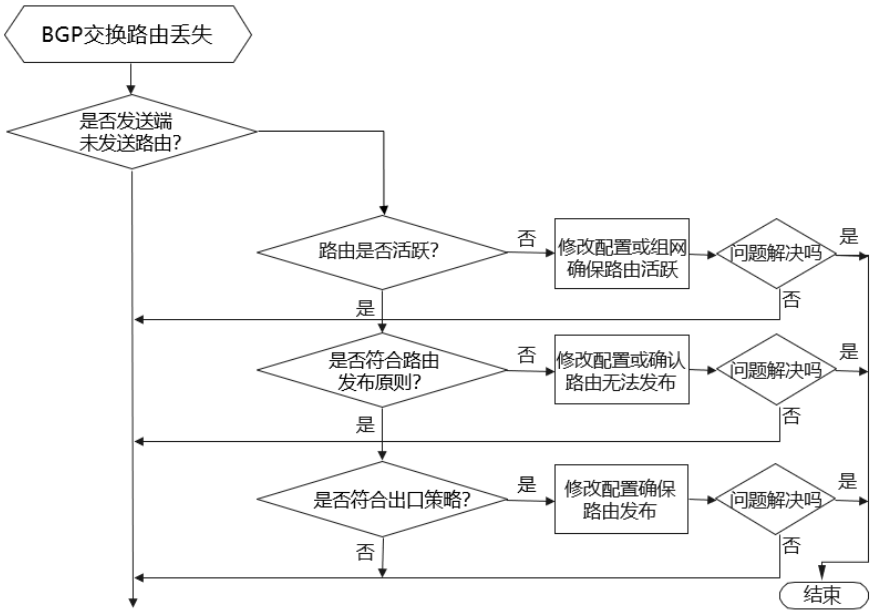
- 华为设备实现了 ISIS 的一个功能，通过设置 overload 位来避免临时的路由黑洞，命令如下：
set-overload [**on-startup** [**wait-for-bgp** [*timeout1*]] [**allow** { **interlevel** | **external** } *]
- wait-for-bgp：系统启动时设置过载标志位，在 BGP 收敛后取消。如果 BGP 没有发信号通知 IS-IS 已收敛完成，IS-IS 将在指定的超时时间或缺省的 10 分钟后（没有指定超时时间）取消过载标志位。
- interlevel：当配置 allow 时，允许发布从不同层次 IS-IS 学来的 IP 地址前缀。
- external：当配置 allow 时，允许发布从其它协议学来的 IP 地址前缀。
- 在 AR-3 上配置该命令，可以解决上述故障。

思考题

1. 路由器与其它路由器物理链路相通，但通过display isis peer命令，不显示对端邻居，如何处理故障?
2. Level-1路由器不能生成区域外的缺省路由，如何分析故障?
3. IS-IS不能够正确学到路由，可能原因有哪些?

- 问题：路由器与其它路由器物理链路相通，但通过 display isis peer 命令，不显示对端邻居，如何处理故障？
- 答：可能由多种原因造成。两端路由器层次不同、区域号不同、接口认证类型和密码不同，或者 System ID 配置重复，都有可能造成邻接关系建立不起来。
- 问题：Level-1 路由器不能生成区域外的缺省路由，如何分析故障？
- 答：Level-1 路由器必须和位于本区域的 Level-1-2 路由器建立 Level-1 的邻接关系，才能有到区域外的路由。位于区域边界的 Level-1-2 路由器如果有不同区域的 Level-2 邻居，会在生成的 LSP 中设置 ATT（Attachment）标志位，说明该路由器与其他区域相连，有通向区域外的路由。这时所有位于同一区域的 Level-1 路由器收到该 LSP 后，就会生成一条指向 0.0.0.0 0 的缺省路由。
- 问题：IS-IS 不能够正确学到路由，可能原因有哪些?
- 答：可能的原因有：
- 邻居不能够正常建立。
- 两端的 Cost 值类型不一致。
- IPv4 和 IPv6 拓扑结构不同导致没有下一跳。
- 路由被路由策略过滤掉了，不能加入到 URT 中。
- LSP 分片被填满，导致 Neighbor TLV 丢失。如果引入路由数量过多，已使用的 LSP 分片数量达到 255 时，必须配置 LSP 分片扩展。
- 路由器配置的 Area 或 Domain 认证不通过，导致 LSDB 不同步。

故障诊断流程 (1)

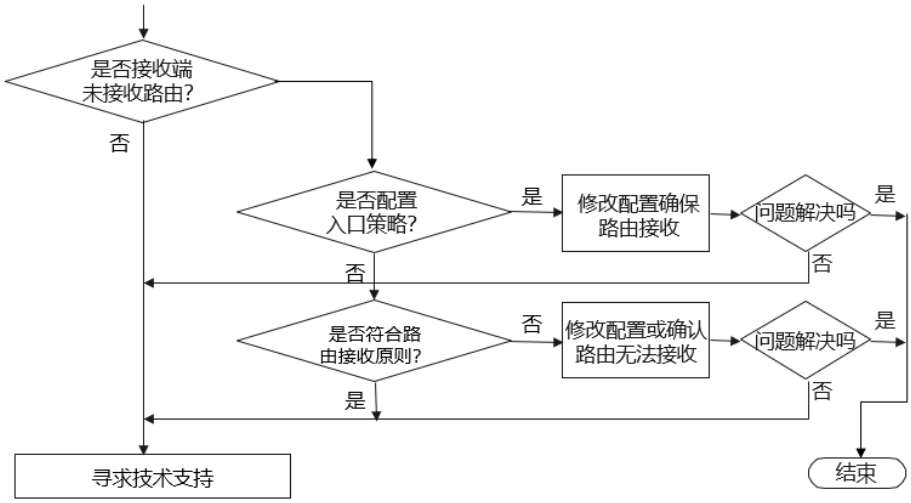


- BGP 是一种用于自治系统 AS 之间的动态路由协议。主要用于交换 AS 之间的可达路由信息。与 IGP 相比，BGP 具有如下特性：
- BGP 是一种外部网关协议，不同于 IGP 的计算和发现路由的能力，BGP 的主要功能在于在 AS 之间选择最佳路由和控制路由的传播。
- BGP 提供了丰富的路由策略，能够对路由实现灵活的过滤和选择。
- BGP 提供了防止路由振荡的机制，有效提高了 Internet 网络的稳定性。
- BGP 相较于 IGP 更易于扩展，能够适应网络新的发展。

- BGP 的故障基本上也可以分为 BGP 邻居故障和 BGP 路由学习故障，BGP 路由学习故障的处理流程基本如图所示：
- BGP 作为一种控制路由传递的协议，路由学习故障的检查，基本可以分成两个部分，即：路由发送的问题和路由接收的问题

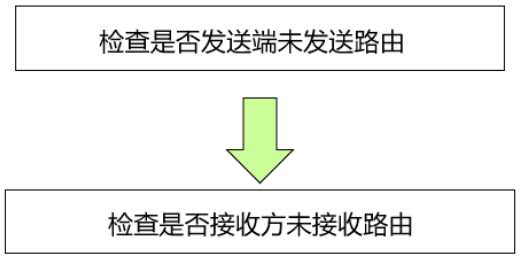
- 如果检查发送端没有问题，则转入接收端检查路由接收是否存在问题。

故障诊断流程 (2)



- 在 BGP 路由发送和接收都不存在问题的条件下，如果 BGP 路由仍然不能正常学习到，可以拨打华为 800 电话寻求技术支持。

故障处理步骤



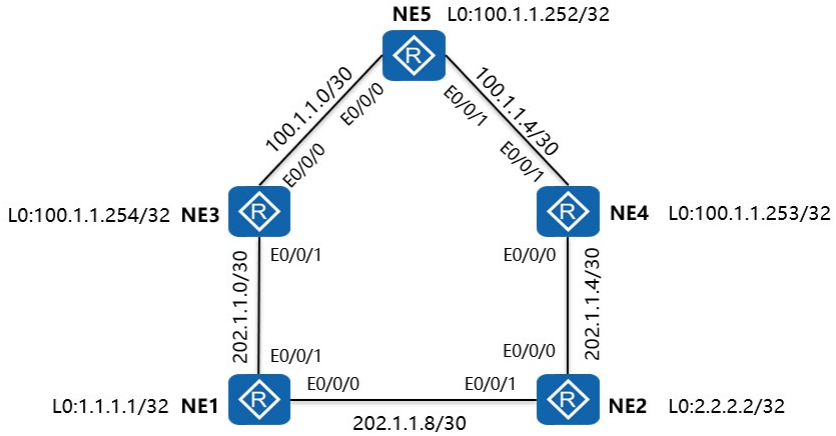
- 步骤 1 检查是否发送端未发送路由

- 在发送端使用 **display bgp routing-table peer***peer-addre ss***advertised-routes** 命令查看路由是否发送。
- 如果发送端未发送路由则进行如下处理：
- 检查本地路由是否活跃。使用 **display bgp routing-table** 命令查看路由是否活跃，即检查路由上是否*>标记。如果不活跃，可能的原因是 Next_Hop 不可达或本地存在其它优选的路由。
- 检查是否不符合发布路由的原则。被聚合抑制的路由不对外发布，使用 **display bgp routing-table** 命令查看路由有 s 标记；被 Dampening 抑止的路由不对外发布，使用 **display b gp routing-table** 命令查看路由有 d 标记；从 IBGP 对等体学到的路由不向 IBGP 对等体转发。
- 检查是否配置了出口策略过滤了路由发布。BGP 可以使用的过滤器包括以下几种：前缀列表过滤器：IP-Prefix List；路径列表过滤器：AS_Path Filter；团体属性列表过滤器：Co mmunity Filter；Route-Policy。这些过滤器既可以应用于从对等体接收的路由信息，也可以应用于向对等体发布的路由信息。
- 可以使用 **display current-configuration configuration bgp** 命令查看配置信息。
- 步骤 2 检查是否发送端未发送路由
- 在接收端使用 **display bgp routing-table peer** *peer-addre ss* **received-routes** 命令查看路由是否接收。
- 如果接收端未接收路由，则进行如下处理：
- 检查是否配置了入口策略过滤了路由接收。使用 **display current-configuration configuration bgp** 命令查看配置信息。
- 检查是否不符合接收路由的原则。满足如下条件的路由将被拒绝接收：未配置 **peer allow-as-loop** 命令，并且本地 A S 号出现在所接收的路由的 AS_Path 属性中；配置了 **peer all ow-as-loop [~** *number* **]**命令，在所接收的路由的 AS_Path 属性中，本地 AS 号重复出现的次数大于配置的 number 值（缺

省值为 1）；从 EBGP 对等体收到路由中 AS_Path 属性中的第一个 AS 号不是对方的 AS 号；Originator_ID 和本地 Router ID 相同，或者为不合法值 0.0.0.0；反射器收到的路由中 Cluster-List 中含有本地 Cluster-ID；Aggregator 是不合法值 0.0.0.0；Next_Hop 是本地的接口地址；从直连的 EBGP 对等体收到的路由的 Next_Hop 不可达；配置了 peer route-limit alert-only，超限后收到的路由都将拒绝。

- 如果检查结束，故障仍然无法排除，请联系华为的技术支持工程师。

拓扑介绍



- 上述拓扑为城域网出口拓扑，NE1 和 NE2 为城域网出口路由器（AS200），NE3、NE4、NE5 为省骨干路由器（AS100）。NE1 和 NE2 通过 network 方式向 EBGP 邻居 NE3 和 NE4 传递路由，NE3 和 NE4 分别与 NE5 建立 IBGP 邻居关系，NE5 作为 RR，与 NE3 和 NE4 建立客户对等体。NE3 和 NE4 上配置虚拟下一跳，在向 IBGP 邻居传递路由时，将 BGP 路由更改为虚拟下一跳地址 202.105.0.5。

故障描述

- 当NE1和NE3之间的连接中断时，网络在NE3和NE5之间出现环路问题。

```
[~NE3]tracert 202.1.1.11
 traceroute to  202.1.1.11(202.1.1.11), max hops: 30 ,packet length: 40
 1 100.1.1.2 40 ms  30 ms  50 ms
 2 100.1.1.5 60 ms 100.1.1.1 50 ms  50 ms
 3 100.1.1.2 60 ms  60 ms  80 ms
 4  *  *  *
 5  *  * 100.1.1.2 90 ms
 6 100.1.1.1 80 ms  90 ms  100 ms
 7 100.1.1.2 110 ms  110 ms  *
```

- 当 NE1 和 NE3 之间链路出现中断时，此时如果 NE3 访问 202.1.1.0/24 网段中的某一个 IP 地址会出现环路问题。假设访问 202.1.1.11 会出现如图所示现象。

故障分析 (1)

- 检查设备上BGP配置，了解路由传递过程。
- NE1配置检查，发现在NE1上采取的是黑洞路由 + network的方式向EBGP邻居NE3进行路由的通告。NE2的配置方式类似。

```
[~NE1]display  current-configuration  configuration bgp
bgp 200
 ......
 ipv4-family unicast
  undo synchronization
  network 202.1.1.0
  peer 2.2.2.2 enable
  peer 2.2.2.2 next-hop-local
  peer 202.1.1.1 enable

[~NE1]display current-configuration  | include  route-static
 ip route-static 202.1.1.0 255.255.255.0 NULL0
```

- 202.1.1.0/24 模拟为某网段用户地址池

故障分析 (2)

- 检查NE3的BGP配置。

```
[~NE3]display current-configuration  configuration bgp
#
bgp 100
 peer 100.1.1.252 as-number 100
 peer 100.1.1.252 connect-interface LoopBack0
 peer 202.1.1.2 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 100.1.1.252 enable
  peer 100.1.1.252 route-policy vir-add-bgp export
  peer 202.1.1.2 enable
#
Return
```

故障分析 (3)

- 在NE3上通过策略，强制更改了BGP的下一跳为虚拟地址202.105.0.5。

```
[~NE3]display  route-policy  vir-add-bgp
Route-policy : vir-add-bgp
  permit : 10
    Apply clauses :
      apply ip-address next-hop 202.105.0.5
[~NE3]display  this
#
 sysname NE3
isis 1
Import-route static
preference 100
.....
 ip route-static 202.105.0.5 255.255.255.255 100.1.1.2
```

- 虚拟地址的下一跳为 NE5 与 NE3 的互联地址 100.1.1.2

故障分析 (4)

- 检查NE3上的路由表：

```
[~NE3]display ip routing-table 202.1.1.0 verbose
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1

Destination: 202.1.1.0/24
        Protocol: BGP                 Process ID: 0
      Preference: 255                       Cost: 0
         NextHop: 202.105.0.5       Neighbour: 100.1.1.252
           State: Active Adv Relied         Age: 00h00m00s
             Tag: 0                    Priority: low
           Label: NULL                 QoSInfo: 0x0
      IndirectID: 0x5
 RelayNextHop: 100.1.1.2             Interface: Ethernet0/0/1
        TunnelID: 0x0                    Flags: RD
```

- 在 NE3 上，因为 NE1 和 NE3 之间的连接中断，所以 NE3 上的路由来自于 RR 反射的 NE4 通告的路由，路由出接口指向 NE5。

故障分析 (5)

- 检查NE5上的路由表，其中一条路由指回了NE3，路由环路。

```
[~NE5]display  ip routing-table 202.1.1.0
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1
Destination/Mask    Proto  Pre  Cost  Flags NextHop         Interface

     202.1.1.0/24   BGP    255  0       RD   202.105.0.5    Ethernet0/0/0
                    BGP    255  0       RD   202.105.0.5    Ethernet0/0/1
[~NE5]display  ip routing-table  202.105.0.5
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1
Destination/Mask  Proto  Pre  Cost   Flags NextHop     Interface

   202.105.0.5/32 ISIS     15   74     D    100.1.1.5     Ethernet0/0/0
                  ISIS     15   74     D    100.1.1.1     Ethernet0/0/1
```

故障处理

1. 通过查看设备上的路由信息，发现环路形成的根本原因在于NE3将去往虚拟地址202.105.0.5静态路由通过ISIS通告给了NE5。
2. 因为NE3上配置静态路由时，使用的是下一跳地址是100.1.1.2，因此，在链路Ethernet0/0/0 DOWN的情况下，仍然将静态路由引入到了ISIS。
3. 将NE3上指向虚拟地址的静态路由下一跳地址改为202.1.1.2。

```
[~NE3]ip route-static 202.105.0.5 255.255.255.255 202.1.1.2
```

思考题

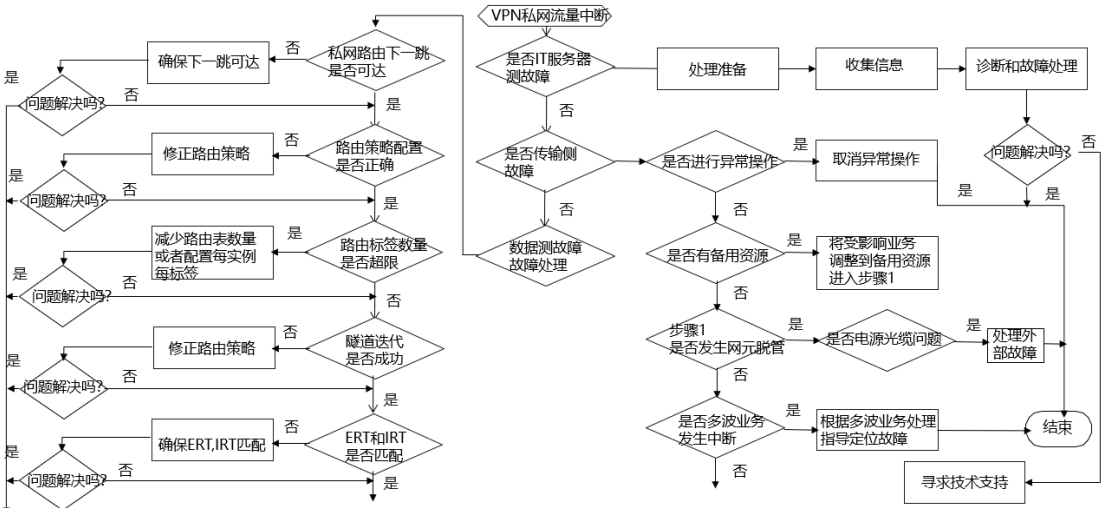
1. 为什么改变BGP邻居能力的配置，BGP连接会断开?
2. 为什么shutdown接口后BGP邻居不是立即断开?

- 问题：为什么改变 BGP 邻居能力的配置，BGP 连接会断开?
- 答：由于 BGP 协议目前不支持动态能力协商，BGP 邻居能力配置变化时，BGP 的连接会自动断开，然后重新进行邻居能力协商。如下面的情况：
- 使能或禁止发送标签路由能力（Label-Route-Capability）。
- 使能或禁止某地址族下的 BGP Peer。如在 VPNv4 地址族下执行 **peer enable/undo peer enable**，会影响其他地址族下该 peer 的 BGP 连接断开自动重新协商。
- 使能 GR 能力。
- 问题：为什么 shutdown 接口后 BGP 邻居不是立即断开?
- 答：规格上只有直连 EBGP 邻居，且在 BGP 下配置了命令 **ebgp-interface-sensitive**（缺省为配置该命令）的前提下，直连 EBGP 邻居才会在 **shutdown** 接口后立即断掉连接。其它情况其它类型的 BGP 邻居都会等 Hold Timer 超时。

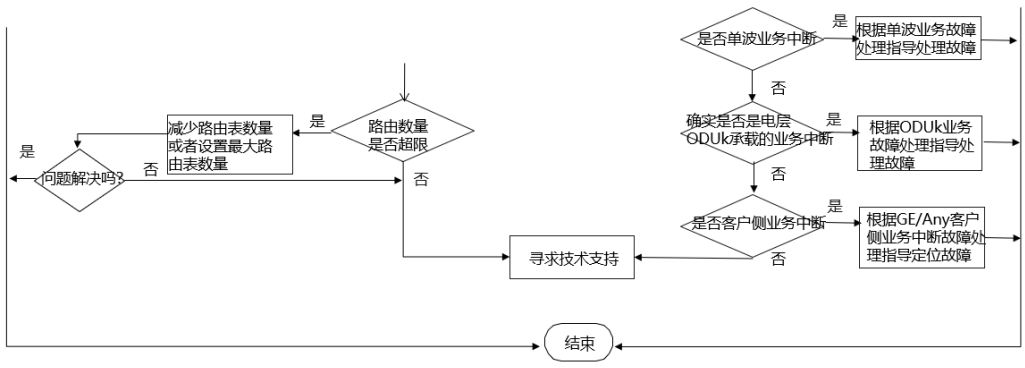
常见故障原因

- 路由下一跳不可达导致路由不活跃。
- 路由策略配置不当导致路由无法发布/接收。
- 标签超限导致私网路由无法发布。
- 私网路由迭代不到隧道导致路由不活跃。
- Export-RT/Import-RT不匹配导致路由无法交叉到私网路由表中。
- 路由超限导致收到的路由被丢弃。

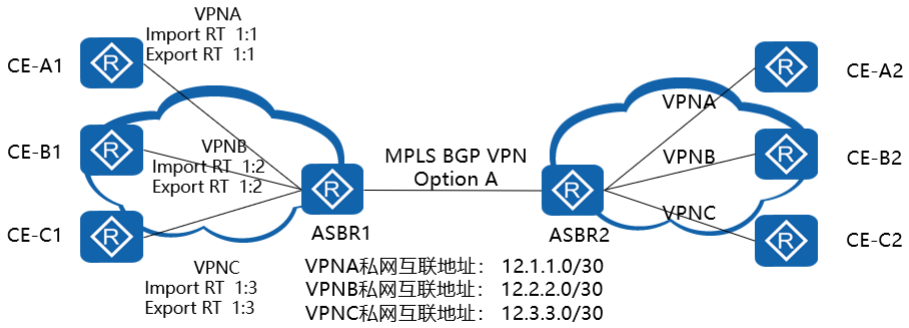
故障诊断流程



故障诊断流程

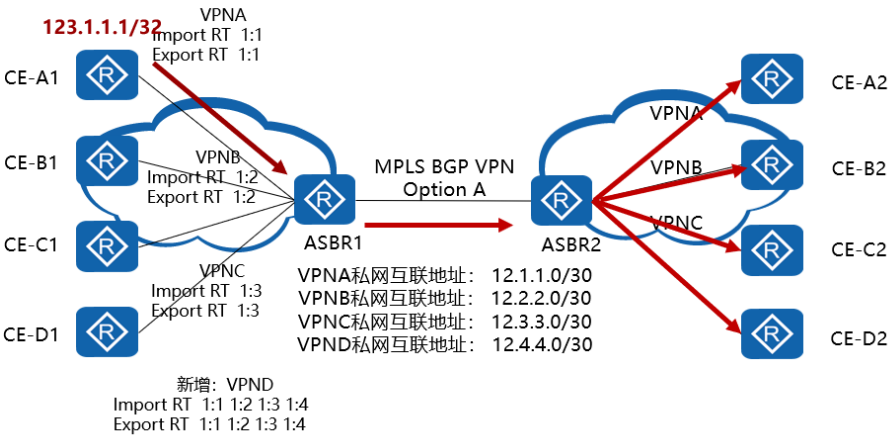


拓扑介绍



- 某公司网络原先有 vpna、vpnb、vpnc 三个 L3VPN 实例，其 route-distinguisher 分别为 1:1、1:2、1:3，vpn-target 分别为 1:1、1:2、1:3，目的是三个 VPN 相互隔离不允许互访。如图，ASBR1 的 vpna、vpnb、vpnc 下分别下挂 CE-A1、CE-B1、CE-C1，ASBR2 的 vpna、vpnb、vpnc 下分别下挂 CE-A2、CE-B2、CE-C2。ASBR1 和 ASBR2 通过 BGP 私网邻居做 OptionA 跨域。CE-A1 发布的路由，只有 CE-A2 能收到，CE-B2 和 CE-C2 不会收到，因此实现了三个 VPN 实例的隔离。

故障描述



- 客户因为业务扩展新增一个 vpnd，需要实现 vpnd 与 vpn

a、vpnb、vpnc 互访，而 vpna、vpnb、vpnc 之间仍保持隔离，因此设计 vpnd 的 route-distinguisher 为 1:4，vpn-target 为 1:1 1:2 1:3 1:4。在 ASBR1 和 ASBR2 之间同样做 OptionA 跨域。但是，当客户新增 vpnd 之后，发现 CE-B2 和 CE-C2 也学到了 CE-A1 的路由。不仅如此，所有 VPN 实例的路由在 Option A 跨域之后都能学到其他 VPN 实例的路由，原先设计的隔离失效了。

故障分析（1/6）

- 首先选取一条有问题的路由(123.1.1.1/32)，确认ASBR1上路由的本地交叉是否正常。

```
<~ASBR1>display bgp vpnv4 all routing-table 123.1.1.1 32
 BGP local router ID : 1.1.1.1
Local AS number : 100
Total routes of Route Distinguisher(1:1): 1
BGP routing table entry information of 123.1.1.1/32:
From: 10.1.1.1 (10.1.1.1)
Route Duration: 00h02m09s
Direct Out-interface: Ethernet0/0/0
Original nexthop: 10.1.1.1
Qos information : 0x0
Ext-Community:RT <1 : 1>
AS-path 1000, origin igp, MED 0, pref-val 0, valid, external, best, select, pre
255
Not advertised to any peer yet
VPN-Instance vpna, Router ID 1.1.1.1:
Total Number of Routes: 1
BGP routing table entry information of 123.1.1.1/32:
From: 10.1.1.1 (10.1.1.1)
.....
```

故障分析（2/6）

```
Route Duration: 00h15m07s
 Direct Out-interface: Ethernet0/0/0
 Original nexthop: 10.1.1.1
 Qos information : 0x0
 AS-path 1000, origin igp, MED 0, pref-val 0, valid, external, best, select, act
ive, pre 255
 Advertised to such 2 peers:
   10.1.1.1
   12.1.1.2        // 通过vpna的OptionA邻居12.1.1.2发布给ASBR2
 VPN-Instance vpnd, Router ID 1.1.1.1:
 Total Number of Routes: 1
 BGP routing table entry information of 123.1.1.1/32:
 From: 10.1.1.1 (10.1.1.1)
 Route Duration: 00h15m07s
 Direct Out-interface: Ethernet0/0/0
 Original nexthop: 10.1.1.1
 Qos information : 0x0
 Ext-Community:RT <1 : 1>
 AS-path 1000, origin igp, MED 0, pref-val 0, valid, external, best, select, act
ive, pre 255
 Advertised to such 2 peers:
   40.1.1.1
   12.4.4.2        // 通过vpnd的OptionA邻居12.4.4.2发布给ASBR2
```

- vpna 的 ERT（出方向 vpn-target）是 1:1，vpnd 的 IRT（入方向 vpn-target）包含 1:1，所以，此路由本地交叉到 vpnd 没问题。对 ASBR1 来说 OptionA 邻居 ASBR2 相当于 CE，本地交叉到 vpnd 的路由发布给 peer 12.4.4.2 也没问题。

故障分析（3/6）

- 在ASBR2上查看123.1.1.1/32路由传递情况：

```
<~ASBR2>display bgp vpnv4 all routing-table 123.1.1.1 32
 BGP local router ID : 2.2.2.2
Local AS number : 200
Total routes of Route Distinguisher(1:1): 1
BGP routing table entry information of 123.1.1.1/32:
From: 12.1.1.1 (1.1.1.1)
Route Duration: 00h19m33s
Direct Out-interface: GigabitEthernet0/0/1
Original nexthop: 12.1.1.1
Qos information : 0x0
Ext-Community:RT <1 : 1>
AS-path 100 1000, origin igp, pref-val 0, valid, external, best, select, pre 255
Not advertised to any peer yet
Total routes of Route Distinguisher(1:4): 1
BGP routing table entry information of 123.1.1.1/32:
From: 12.4.4.1 (1.1.1.1)
Route Duration: 00h12m23s
Direct Out-interface: GigabitEthernet0/0/3
Original nexthop: 12.4.4.1
```

故障分析（4/6）

```
Qos information : 0x0
 Ext-Community:RT <1 : 1>, RT <1 : 2>,
            RT <1 : 3>, RT <1 : 4>
 AS-path 100 1000, origin igp, pref-val 0, valid, external, best, select, pre 25
5
 Not advertised to any peer yet
 VPN-Instance vpna, Router ID 2.2.2.2:
 Total Number of Routes: 2
 BGP routing table entry information of 123.1.1.1/32:
 From: 12.1.1.1 (1.1.1.1)
 Route Duration: 00h19m33s
 Direct Out-interface: GigabitEthernet0/0/1
 Original nexthop: 12.1.1.1
 Qos information : 0x0
 AS-path 100 1000, origin igp, pref-val 0, valid, external, best, select, active
, pre 255
 Advertised to such 2 peers:
   10.1.2.1                // 路由发布给CE-A2
   12.1.1.1
```

- 通过 vpna 的 peer 12.1.1.1 学到 vpna 的私网路由 123.1.

1.1/32，发布给 CE-A2

故障分析（5/6）

```
BGP routing table entry information of
123.1.1.1/32:
 From: 12.4.4.1 (1.1.1.1)
 Route Duration: 00h12m24s
 Direct Out-interface: GigabitEthernet0/0/3
 Original nexthop: 12.4.4.1
 Qos information : 0x0
 Ext-Community:RT <1 : 1>, RT <1 : 2>,
               RT <1 : 3>, RT <1 : 4>
// vpnd的路由本地交叉到vpna
 AS-path 100 1000, origin igp, pref-val 0,
valid, external, pre 255, not preferr
ed for peer type
 Not advertised to any peer yet
```

```
 VPN-Instance vpnd, Router ID 2.2.2.2:

 Total Number of Routes: 2
 BGP routing table entry information of
123.1.1.1/32:
 From: 12.4.4.1 (1.1.1.1)
 Route Duration: 00h11m08s
 Direct Out-interface:
GigabitEthernet0/0/3
 Original nexthop: 12.4.4.1
 Qos information : 0x0
 AS-path 100 1000, origin igp, pref-val 0,
valid, external, best, select, active
, pre 255
 Advertised to such 1 peers:
   40.1.2.1 // 路由发布给CE-D2
```

- 通过 vpnd 的 peer 12.4.4.1 学到 vpnd 的私网路由 123.1.1.1/32，本地交叉给 vpna（不优选），同时本地交叉给了 CE-B2,CE-C2,CE-D2。

故障分析（6/6）

```
 VPN-Instance vpnb, Router ID 2.2.2.2:

 Total Number of Routes: 1
 BGP routing table entry information of
123.1.1.1/32:
 From: 12.4.4.1 (1.1.1.1)
 Route Duration: 00h11m08s
 Direct Out-interface: GigabitEthernet0/0/3
 Original nexthop: 12.4.4.1
 Qos information : 0x0
 Ext-Community:RT <1 : 1>, RT <1 : 2>,
               RT <1 : 3>, RT <1 : 4>  // vpnd的
路由本地交叉到vpnb
 AS-path 100 1000, origin igp, pref-val 0,
valid, external, best, select, active
, pre 255
 Advertised to such 2 peers:
   20.1.2.1          // 路由发布给CE-B2
   12.2.2.1
```

```
 VPN-Instance vpnc, Router ID 2.2.2.2:

 Total Number of Routes: 1
 BGP routing table entry information of
123.1.1.1/32:
 From: 12.4.4.1 (1.1.1.1)
 Route Duration: 00h11m08s
 Direct Out-interface: GigabitEthernet0/0/3
 Original nexthop: 12.4.4.1
 Qos information : 0x0
 Ext-Community:RT <1 : 1>, RT <1 : 2>,
               RT <1 : 3>, RT <1 : 4>
// vpnd的路由本地交叉到vpnc
 AS-path 100 1000, origin igp, pref-val 0,
valid, external, best, select, active
, pre 255
 Advertised to such 2 peers:
   30.1.2.1 // 路由发布给CE-C2
   12.3.3.1
```

- 通过 vpnd 的 peer 12.4.4.1 学到 vpnd 的私网路由 123.1.1.1/32，本地交叉给 vpna（不优选），同时本地交叉给了 CE-B2,CE-C2,CE-D2。

故障处理

- vpnd的Export RT是1:1 1:2 1:3 1:4，vpna的Import RT是1:1，vpnb的Import RT是1:2，vpnc的Import RT是1:3，所以，此路由本地交叉正常。本地交叉到vpnb、vpnc的路由发布给CE-B2、CE-C2也没问题。
- ASBR1和ASBR2的路由交叉与发布都没问题。问题在于vpnd可与其他3个VPN相互交叉，导致ASBR2从 vpnd学到的路由交叉到了其他3个VPN中。同理，其他私网路由也都如此交叉到了其他VPN中。所以，各VPN相互都学到其他VPN的路由，原先设计的隔离失效。

解决方案1

- 解决方案1：通过ASBR1对vpnd的OptionA邻居ASBR2配置出口策略解决：

```
ip extcommunity-filter 1 permit rt 1:4
ip extcommunity-filter 2 permit rt 1:1 rt 1:2 rt 1:3
#
route-policy rp-asbr2-vpnd permit node 10
 if-match extcommunity-filter 1
#
route-policy rp-asbr2-vpnd deny node 20
 if-match extcommunity-filter 2
#
route-policy rp-asbr2-vpnd permit node 100
#
bgp 100
  peer 12.4.4.2 route-policy rp-asbr2-vpnd export
#
```

- ASBR2向ASBR1做相同配置
- 注：若只是单纯本地交叉场景没有PE设备，可以不配置route-policy rp-asbr2-vpnd的node 10和extcommunity-filter 1。

- ASBR1 对 vpnd 的 OptionA 邻居配置出口策略，只发布源自 vpnd 的路由（包括 VPNv4 路由通过 IRT 1:4 交叉到 vpnd 的路由和从其他 vpnd 私网邻居收到的路由），不发布源自其他 vpn 的路由（包括 VPNv4 路由通过 IRT 1:1 或 1:2 或 1:3 交叉到 vpnd 的路由和从其他 VPN 实例本地交叉到 vpnd 的路由）。这样，ASBR2 通过 vpnd 的 OptionA 邻居不会收到来自 CE-A1 的路由，也就不会将该路由交叉到其他 VPN 实例。
- 因为 ASBR1 上源自 vpnd 的路由包括 VPNv4 路由通过 IRT 1:4 交叉到 vpnd 的路由（携带 extcommunity <1:4>）和从其他 vpnd 私网邻居收到的路由。

解决方案2

- 解决方案2：采用OptionB方式进行跨域，ASBR之间通过VPNv4邻居发布路由。这样，ASBR1从CE-A1收到的路由向ASBR2发布时，只发布extcommunity为1:1的VPNv4路由，不会通过vpnd的私网邻居发布路由。因此，ASBR2不会收到来自vpnd的私网路由，只会收到extcommunity为1:1的VPNv4路由，根据Import RT做匹配，只能交叉到vpna和vpnd，不会交叉到vpnb和vpnc。

思考题

1. L3VPN接入MPLS，怎么形成负载分担?
2. VPN的标签分配模式有几种，区别是什么?

- 问题：L3VPN 接入 MPLS,怎么形成负载分担？
- 答：L3VPN 接入 MPLS，默认情况下是不进行负载分担的。如果需要形成负载分担，则需要配置如下命令：
- **tunnel select-seq { cr-lsp | lsp }** * **load-balance-number** */ oad-balance-number*。
- 问题：VPN 的标签分配模式有几种，区别是什么？
- 答：分配模式有两种：
- Apply-label per-route（默认）。
- Apply-label per-instance。
- 区别：
- 每路由分配标签比较耗费设备资源，在路由量很大的时候，会超出产品规格，无法为每条路由分配标签，造成转发不通。
- 每实例分配标签节省设备资源，每个 VPN 只分配一个标签。
- 两种模式一般情况下取得的效果一致，但是建议采用每

实例分配模式。

•