IXP 維運
BGP Session Culling

是方電訊股份有限公司
網路服務部　林盛琪 資深經理

June. 8, 2018
http://www.chief.com.tw

CHIEF | 是方電訊

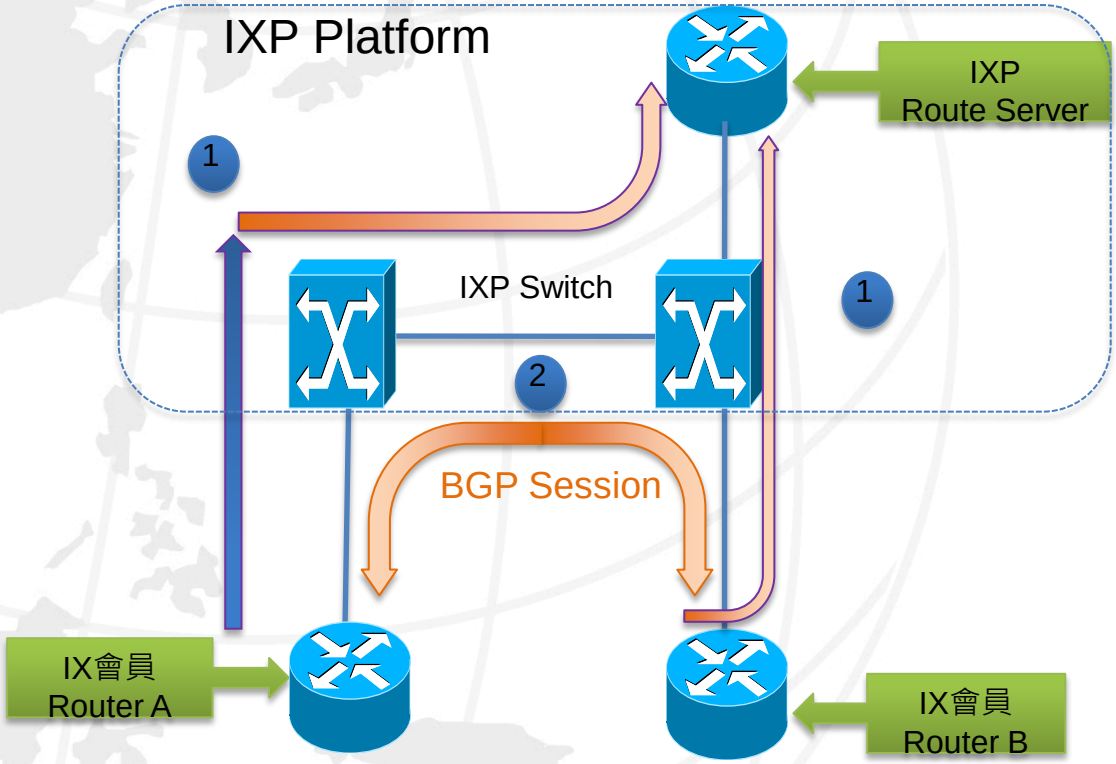
前言（一）

IXP 最常見的交換模式:

第1種狀況是:所有IX會員與Route Server建EBGP連線,交換路由,然後再產生流量.

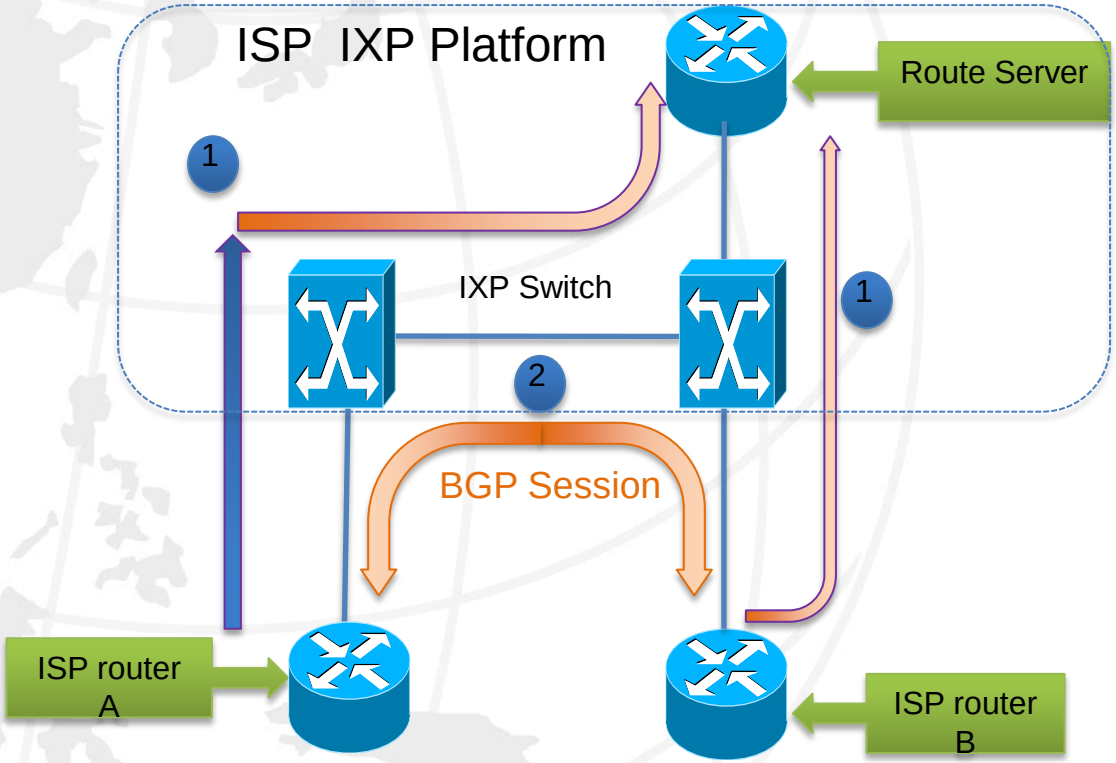
第2種狀況是: IXP 提供Layer 2連線,IX會員彼此之間自行建EBGP,並交換流量.

- 狀況(1)下, IXP要維護的時候,只要公告後, 切斷Route Server與其它IX會員 EBGP連線, 就可進行維護
- 不會影響客戶端 IX會員之間傳輸問題而導致大量丟包情形 .



前言(二)

- 狀況(2)下, 在IXP Platform 下面的IX會員自行建立EBGP Session交換路由及流量
- 當提供IXP 要進行設備維護時(可能是重啟 IX Switch ,更換Port或是更換設備)由於各IX會員時間配合的問題需等到BGP tear down 才能進行維護
- 否則會影響客戶端 IX會員之間傳輸問題並導致線路切換期間大量丟包情形



網際網路交換中心功能

- IX會員透過交換背板（switched fabric)連接彼此網路

  TPIX提供 1->100 互聯的乙太網路交換
- TPIX會員透過交換中心彼此連接其路由器並交換資料流
- 使用BGP 協定產生FIB(Forwarding Information Base)針對 TPIX會員提供路由可達資訊(reachability)

交換中心如何安排維修時間

From: 是方電訊客戶服務部(Chief Telecom Customer Service Dept.) [mailto:service@chief.com.tw]
Sent: Wednesday, October 04, 2017 11:01 PM
To: 'ykao@mail.nctu.edu.tw'; 'fangyulin@nctu.edu.tw'
Cc: 是方電訊客戶服務部 (service@chief.com.tw); Kevin_Lin
Subject: T201710040033--國立交通大學--[緊急通知]是方電訊 10/6 機房設備緊急優化維護作業

停機維護通知

一、維護原因：機房設備緊急優化維護作業
二、預訂作業日期：2017 年 10 月 6 日(星期五)
三、預訂作業時間：05:00~05:30
四、影響電路時間：中斷 10 分鐘
五、受影響範圍：TPIX Internet Service LYI-7534 (203.163.222.33)
六、通知日期：2017.10.4

客戶服務部

是方電訊股份有限公司
台北市 114 內湖區瑞光路 68 號 2 樓
電話：0800-365-070
070-1017-1800
(02) 8797-4657
傳真：(02) 8791-9290
Email：service@chief.com.tw

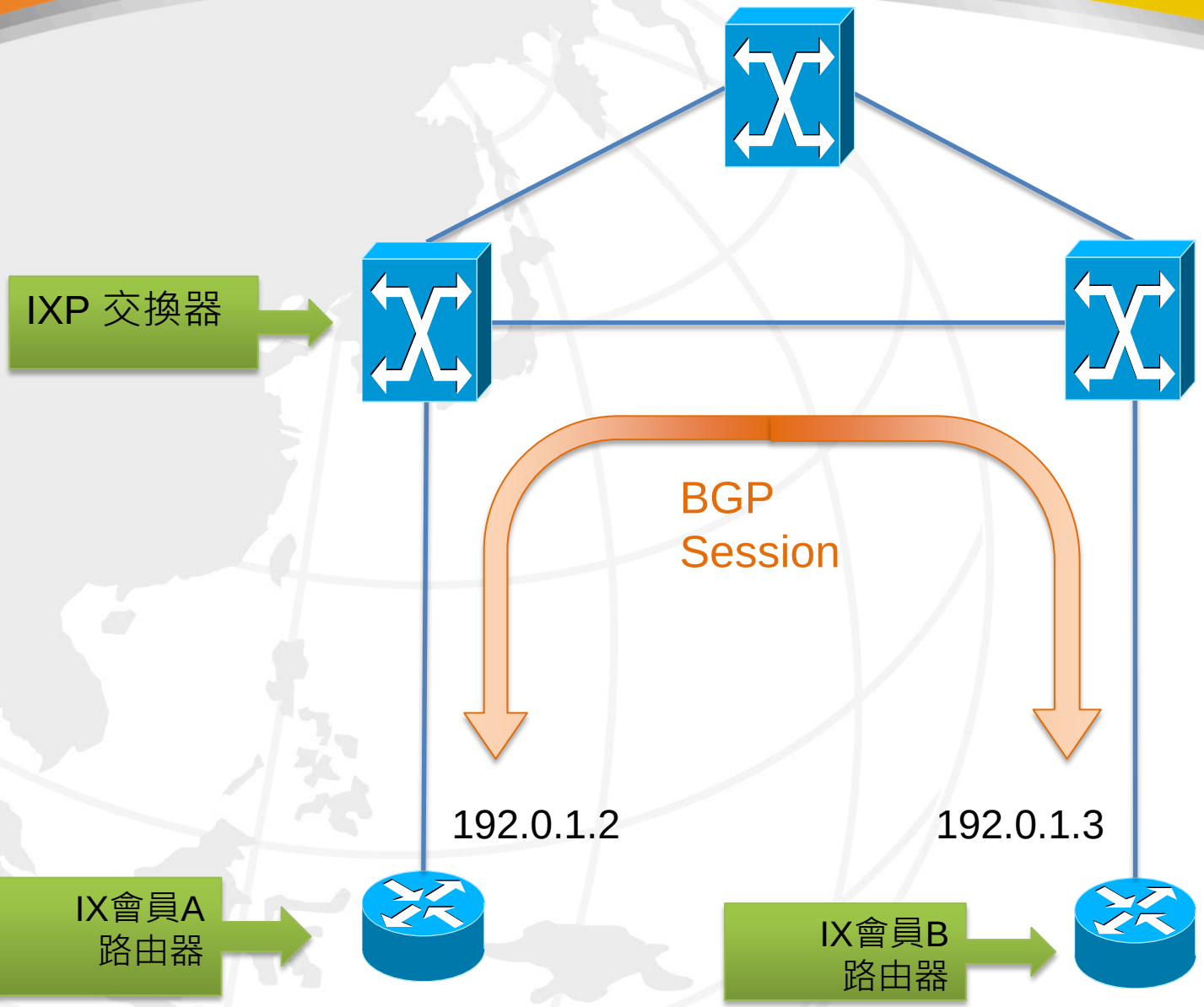
- 理想狀態為各個IX會員自行關閉BGP sessions ☺
- 但是交換中心維修工作時,IX會員可能會認為跟其無關,未必會配合,此時誰會為此事煩惱? ☹

交換中心維修工作

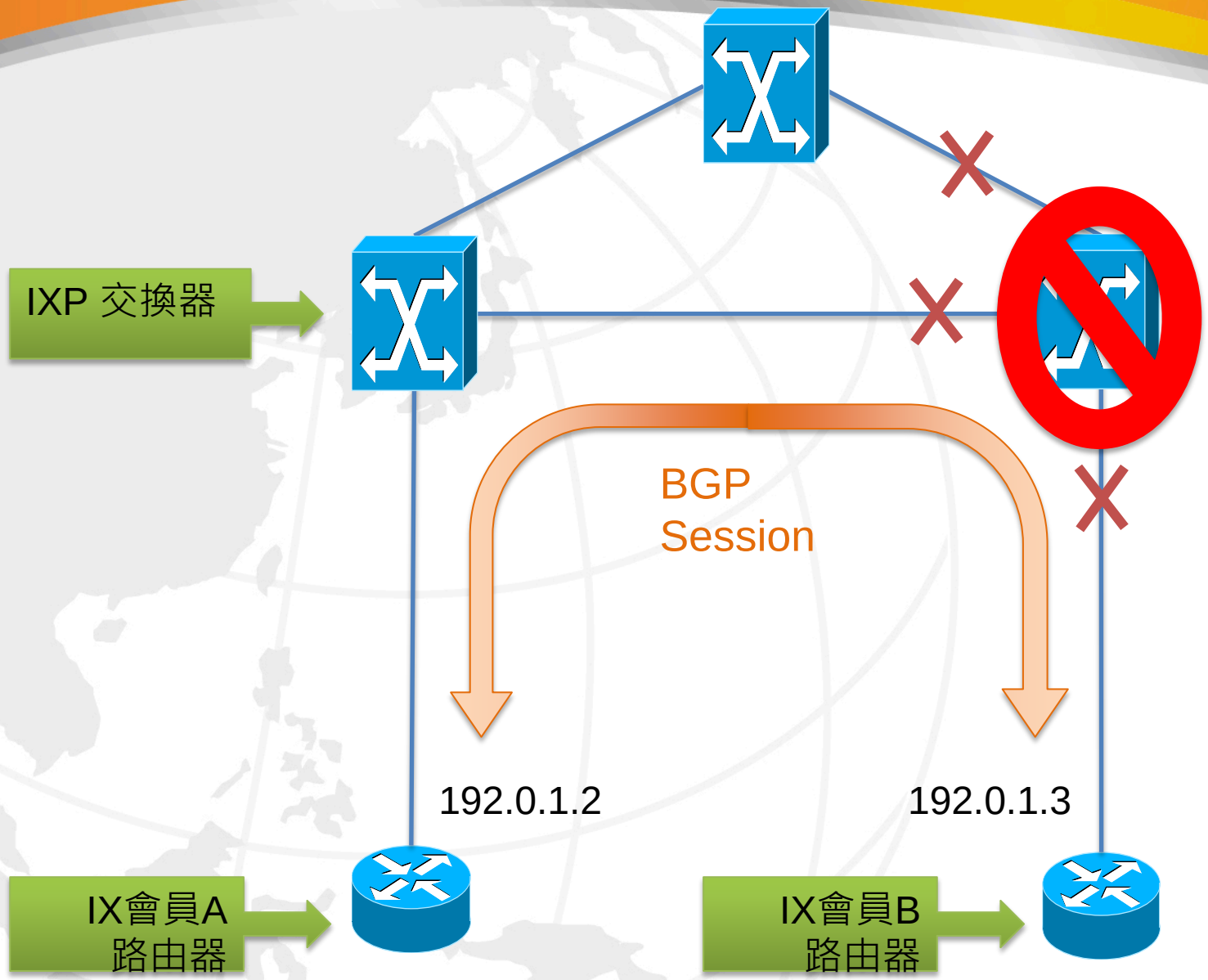
CHIEF | 是方電訊

- 交換中心例行的維修工作:
  - 交換器重新開機
  - 更換端口ODF
  - 設備更換
- 當交換中心進行維護工作時交換中心內的會員流量會有何種影響

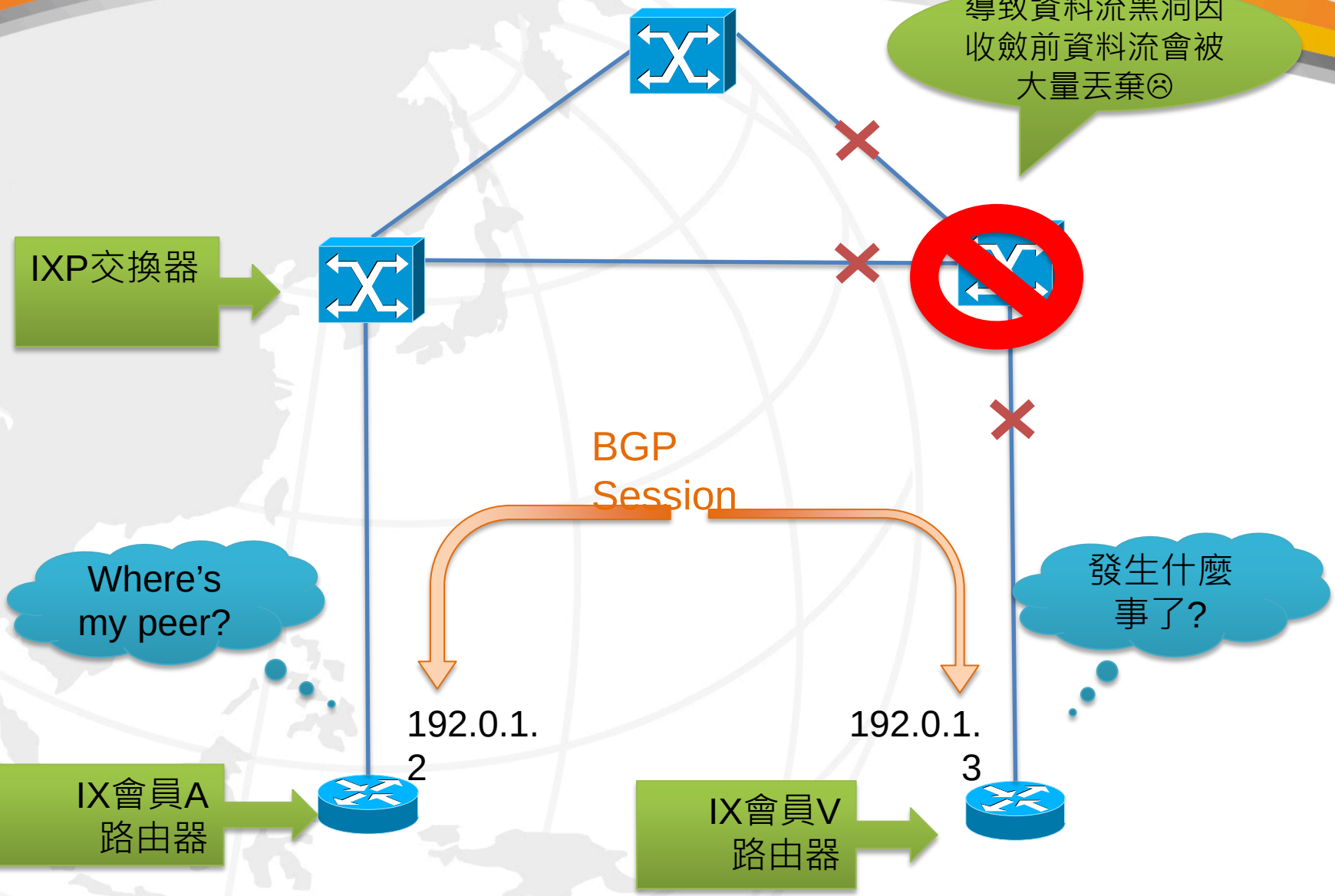
傳統的IX網路交換中心



IX進行維護工作



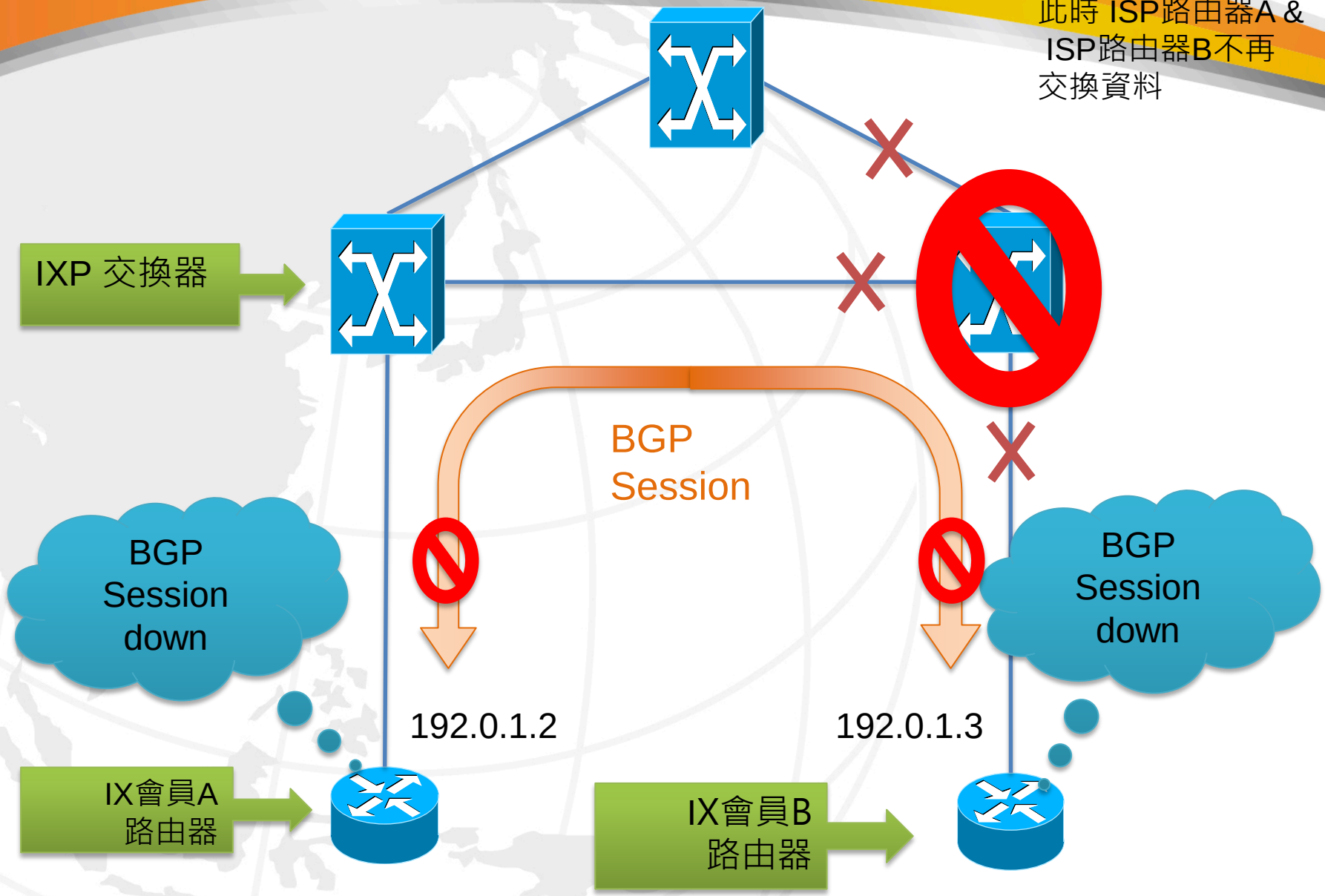
等待 BGP收斂(hold timers)



BGP收斂約30-180秒

CHIEF | 是方電訊

此時 ISP路由器A & ISP路由器B不再交換資料



經驗值的改善

CHIEF | 是方電訊

依據過往經驗, 交換中心(IXP 交換器)維修工作導致約 90+ 秒交換中心資料流被丟棄**– 會造成會員不好的使用經驗** ☹

解決方案:

1. 於維修工作開始時**剔除(culling) BGP sessions**
2. 等待BGP收斂後,流量黑洞現象消失 (約3-5 分鐘)
3. 直接進行維修工作

但無法控制IX客戶的路由器時,要如何進行剔除(BGP session culling)

解答: L4 ACLs on IXP port!

L4 BGP ACLs 設定

CHIEF | 是方電訊

```
ipv6 access-list acl-ipv6-permit-all-except-bgp
    10 deny tcp 2001:db8:2::/64 eq bgp 2001:db8:2::/64
    20 deny tcp 2001:db8:2::/64 2001:db8:2::/64 eq bgp
    30 permit ipv6 any any
!
ip access-list acl-ipv4-permit-all-except-bgp
    10 deny tcp 192.0.2.0/24 eq bgp 192.0.2.0/24
    20 deny tcp 192.0.2.0/24 192.0.2.0/24 eq bgp
    30 permit ip any any
!
interface Ethernet33
    description IXP Participant Affected by Maintenance
    ip access-group acl-ipv4-permit-all-except-bgp in
    ipv6 access-group acl-ipv6-permit-all-except-bgp in
!
```

IXP 網段

雙向隔離BGP Session, 否則 BGP sessions 會重新建立

重複 IXP 內 的 IPv4 網段

改善後的結果

CHIEF | 是方電訊

- 移除資料流的時間小於 3分鐘
- 當IXP 維護準備好時,可以還原資料流
  - 這包括多次重啟,更新韌體, 或作業錯誤後

這就是RFC 8327

"Mitigating Negative Impact of Maintenance through BGP Session Culling"

1. 最理想狀態為 IX會員可以主動關閉他們的 BGP sessions
2. 但當 IX會員無法配合IXP 維運時,IX交換中心可以使用此技術 BGP Session Culling 來主動停止資料流以降低負面的影響

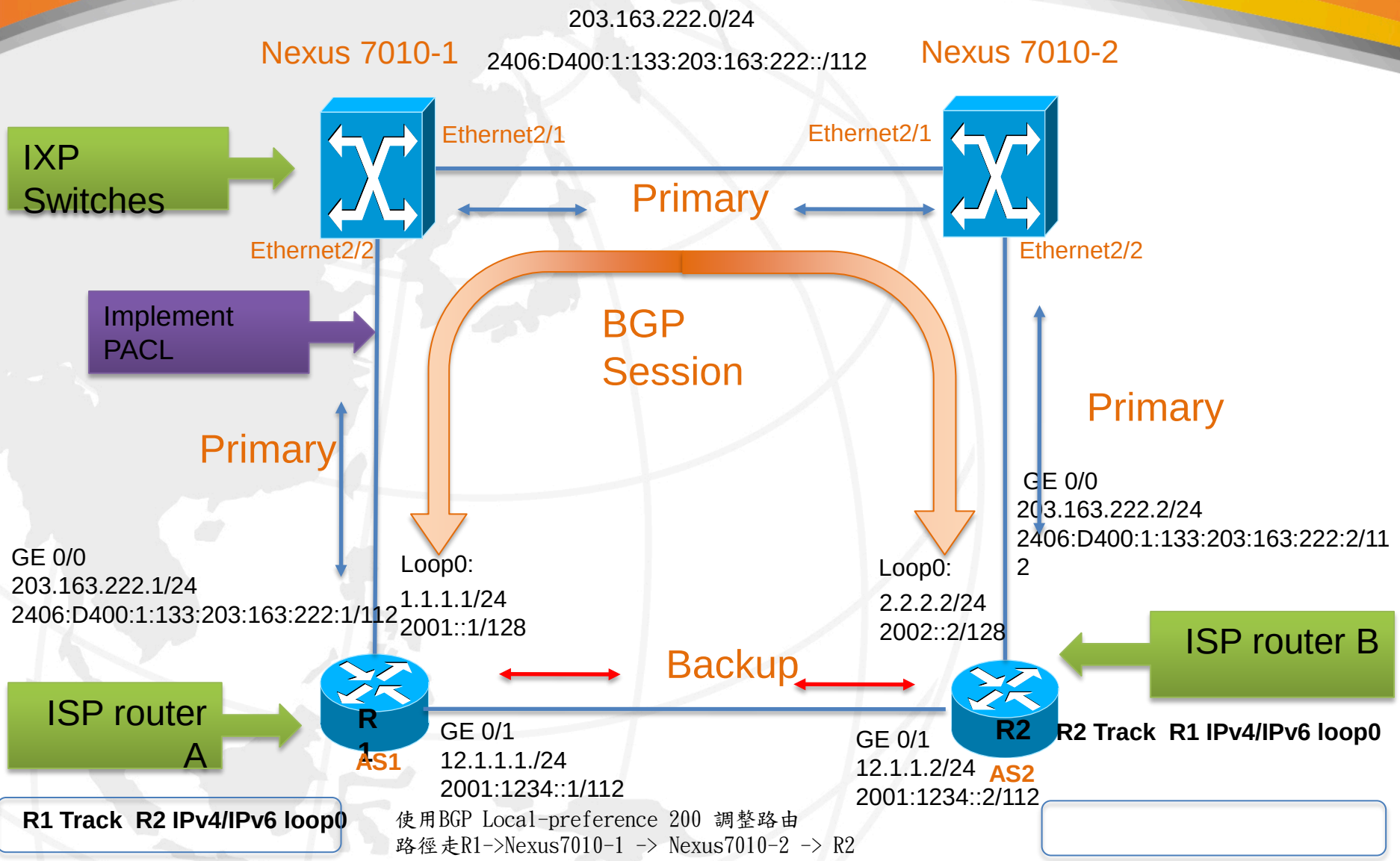
結語

- 使用此技術可以提升更好的IX網路品質
- 可以要求IXP 業者都採用 BGP Session Culling 在進行維護的時候!

- 資料來源提供

- https://www.youtube.com/watch?v=B8cpa4HEaKQ
- https://tools.ietf.org/html/rfc8327
- Reducing the impact of IXP maintenance (Will Hargrave // LONAP)London access point

BGP Session Culling LAB TEST架構圖



BGP-Session-Culling 測試說明

CHIEF | 是方電訊

IX Switch Eth 2/2使用L4 ACL來阻止IXP的BGP流量.

IX Peer連接IX Switch Interface : Vlan 10

IX Switch之間Interface : trunk

Peer 之間LAN Prefix/Mask:

Pv4 NETWORK: 203.163.222.0/24

IPv6 NETWORK: 2406:D400:1:133:203:163:222::/112

測試時間：2018-05-22

測試設備型號：Cisco Nexus7010 (若設備是N9K 須調整Tcam Resource)

Software Version：NXOS：版本6.2(10)

目前AS1與AS2 之間BGP流量Path :

(1) ISP router A(R1) AS1 ->IXP Switches(Nexus 7010-1)→ IXP Switches(Nexus 7010-2)
→ ISP router B(R2) AS2 →Primary
反之亦然!

(2) 使用L4 ACL來阻止IXP的BGP Session keepalive 後, BGP Session down,流量就會停止,改走Backup路徑
Backup路徑 : ISP router A (G0/1) → ISP router B (G0/1) ,
反之亦然!

(3) 驗證切換是否影響ISP之間流量傳輸 :
在ISP router A/ ISP router B 彼此之間使用IP SLA 監控對方 loop 0 interface
觀察執行L4 ACL後 ,BGP切換到Backup Path,監控到對方 loop 0 interface 連線,是否因切換而中斷

BGP-Session-Culling 測試前路徑說明 (1)

CHIEF | 是方電訊

IPv4部分:測試前,先觀察ISP1(AS1)/ISP2(AS2) 之間流量路徑:

```
R1#show ip bgp summary
BGP router identifier 1.1.1.1, local AS number 1
BGP table version is 11, main routing table version 11
2 network entries using 296 bytes of memory
3 path entries using 192 bytes of memory
3/2 BGP path/bestpath attribute entries using 408 bytes of memory
1 BGP AS-PATH entries using 24 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 920 total bytes of memory
BGP activity 5/1 prefixes, 14/8 paths, scan interval 60 secs

Neighbor        V           AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
12.1.1.2        4            2    1344    1341       11    0    0 20:06:22        1
203.163.222.2   4            2      10      10       11    0    0 00:01:28        1
```

```
R1#show track 1
Track 1
  IP SLA 1 state
  State is Up
    7 changes, last change 00:00:47
  Latest operation return code: OK
  Latest RTT (millisecs) 1
```

[說明] :
R1 track R2 IPv4 loop0 目前狀態是OK
有 7 changes

[說明] : 目前 R1到R2之間,有2個EBGP Neighbor .

```
R1#show ip bgp 2.2.2.2
BGP routing table entry for 2.2.2.2/32, version 11
Paths: (2 available, best #1, table default)
  Advertised to update-groups:
     2
  Refresh Epoch 2
  2
    203.163.222.2 from 203.163.222.2 (2.2.2.2)
      Origin IGP, metric 0, localpref 200, valid, external, best
  Refresh Epoch 3
  2
    12.1.1.2 from 12.1.1.2 (2.2.2.2)
      Origin IGP, metric 0, localpref 100, valid, external
```

[說明] : R1到R2 loop0 2.2.2.2 ,Best route是走203.163.222.2 (IPv4)

BGP-Session-Culling 測試前路徑說明 (2)

IPv6部分:測試前,先觀察ISP1(AS1)/ISP2(AS2) 之間流量路徑:

```
R1#show bgp ipv6 unicast summary
BGP router identifier 1.1.1.1, local AS number 1
BGP table version is 11, main routing table version 11
2 network entries using 344 bytes of memory
3 path entries using 264 bytes of memory
3/2 BGP path/bestpath attribute entries using 408 bytes of memory
1 BGP AS-PATH entries using 24 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 1040 total bytes of memory
BGP activity 5/1 prefixes, 14/8 paths, scan interval 60 secs

Neighbor        V           AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
2001:1234::2    4            2    1322    1328       11    0    0 19:38:47        1
2406:D400:1:133:203:163:222:2
                4            2      65      64       11    0    0 00:51:05        1
```

[說明] : 目前 R1到R2之間,有2個IPv6 EBGP Neighbor.

```
R1#show bgp ipv6 unicast 2002::2/128
BGP routing table entry for 2002::2/128, version 11
Paths: (2 available, best #1, table default)
  Advertised to update-groups:
     1
  Refresh Epoch 2
  2
    2406:D400:1:133:203:163:222:2 (FE80::F6CF:E2FF:FE42:2960) from 2406:D400:1:133:203:163:222:2 (2.2.2.2)
      Origin IGP, metric 0, localpref 200, valid, external, best
  Refresh Epoch 6
  2
    2001:1234::2 (FE80::F6CF:E2FF:FE42:2961) from 2001:1234::2 (2.2.2.2)
      Origin IGP, metric 0, localpref 100, valid, external
```

[說明] : R1到R2 loop0 2002::2/128 ,Best route是走 2406:D400:1:133:203:163:222:2 (IPv6)

```
R1#show track 2
Track 2
  IP SLA 2 state
  State is Up
    7 changes, last change 00:04:38
  Latest operation return code: OK
  Latest RTT (millisecs) 1
```

[說明] :
R1 track R2 IPv6 loop0 目前狀態是OK
有 7 changes

流量走的路徑也是 ISP router A(R1) AS1 ->IXP Switches(Nexus 7010-1)→ IXP Switches(Nexus 7010-2) → ISP router B(R2) AS2

BGP-Session-Culling 測試前路徑說明(3)

綜合IPv4/IPv6 2個部分:
測試前,ISP1(AS1)/ISP2(AS2) IPv4/IPv6之間流量路徑:

ISP router A(R1) AS1 ->IXP Switches(Nexus 7010-1)→ IXP Switches(Nexus 7010-2) → ISP router B(R2) AS2

```
R1#show track 1
Track 1
  IP SLA 1 state
  State is Up
    7 changes, last change 00:00:47
  Latest operation return code: OK
  Latest RTT (millisecs) 1
```

[說明] :
R1 track R2 IPv4 loop0 目前狀態是OK
有 7 changes

```
R1#show track 2
Track 2
  IP SLA 2 state
  State is Up
    7 changes, last change 00:04:38
  Latest operation return code: OK
  Latest RTT (millisecs) 1
```

[說明] :
R1 track R2 IPv6 loop0 目前狀態是OK
有 7 changes

L4 BGP ACLs on IXP Switch

Implement L4 BGP ACLs on IXP to Nexus 7000-1 Switch E2/2 設定如下:

```
ipv6 access-list acl-ipv6-permit-all-except-bgp
 10 deny tcp 2406:D400:1:133:203:163:222::/112 eq bgp 2406:D400:1:133:203:163:222::/112
 20 deny tcp 2406:D400:1:133:203:163:222::/112 2406:D400:1:133:203:163:222::/112  eq bgp
 30 permit ipv6 any any
!
ip access-list acl-ipv4-permit-all-except-bgp
 10 deny tcp 203.163.222.0/24  eq bgp 203.163.222.0/24
 20 deny tcp 203.163.222.0/24  203.163.222.0/24 eq bgp
 30 permit ip any any
!
interface Ethernet2/2
 description IXP Participant Affected by Maintenance
 switchport
 switchport access vlan 10
 ip port access-group acl-ipv4-permit-all-except-bgp in
 ipv6 port traffic-filter acl-ipv6-permit-all-except-bgp in
```

BGP-Session-Culling 測試結果說明 (1)

因為L4 BGP ACLs在N7010 E2/2 implement 的緣故,ISP1與ISP2經由IX Switch 之間BGP連線
BGP Peer 之間 Keepalive 被IX Switch ACL 過濾掉 , BGP Peer 之間無法收到Keepalive
BGP holdown timer timeout (3分鐘)後,, 導致BGP session down ,
R1收到Message說明如下:

```
*May 23 03:41:29.710: %BGP-3-NOTIFICATION: received from neighbor 203.163.222.2 4/0 (hold time expired) 0 bytes
*May 23 03:41:29.710: %BGP-5-ADJCHANGE: neighbor 203.163.222.2 Down BGP Notification received
*May 23 03:41:29.710: %BGP_SESSION-5-ADJCHANGE: neighbor 203.163.222.2 IPv4 Unicast topology base removed from session  BGP Notification received
*May 23 03:41:44.046: %BGP-3-NOTIFICATION: received from neighbor 2406:D400:1:133:203:163:222:2 4/0 (hold time expired) 0 bytes
*May 23 03:41:44.046: %BGP-5-ADJCHANGE: neighbor 2406:D400:1:133:203:163:222:2 Down BGP Notification received
*May 23 03:41:44.046: %BGP_SESSION-5-ADJCHANGE: neighbor 2406:D400:1:133:203:163:222:2 IPv6 Unicast topology base removed from session  BGP Notification rece
ived
```

[說明]:無論是IPv4/IPv6 , R1到R2 BGP Session經由R1 → IXP Switches (Nexus7000-1 –Nexus7000-2) ->R2
因L4 ACL關係 , BGP session中斷了

首先觀察 IPv4部分:

```
R1#show ip bgp summary
BGP router identifier 1.1.1.1, local AS number 1
BGP table version is 12, main routing table version 12
2 network entries using 296 bytes of memory
2 path entries using 128 bytes of memory
2/2 BGP path/bestpath attribute entries using 272 bytes of memory
1 BGP AS-PATH entries using 24 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 720 total bytes of memory
BGP activity 5/1 prefixes, 15/11 paths, scan interval 60 secs

Neighbor        V           AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
12.1.1.2        4            2    1444    1440       12    0    0 21:36:16        1
203.163.222.2   4            2       0       0        1    0    0 00:06:59 Active
```

[說明]: R1 ipv4 BGP NeighborR2 (203.163.222.2) 進入Active Status

BGP-Session-Culling 測試結果說明(2)

```
R1#show ip bgp 2.2.2.2
BGP routing table entry for 2.2.2.2/32, version 12
Paths: (1 available, best #1, table default)
  Not advertised to any peer
  Refresh Epoch 3
  2
    12.1.1.2 from 12.1.1.2 (2.2.2.2)
      Origin IGP, metric 0, localpref 100, valid, external, best
```

[說明] 目前R1到R2 loop0 (2.2.2.2) 流量, 已經改走Backup路徑

```
R1#show track 1
Track 1
  IP SLA 1 state
  State is Up
    7 changes, last change 00:37:16
  Latest operation return code: OK
  Latest RTT (millisecs) 1
```

[說明] 目前 R1 Track到R2 loop0 2.2.2.2 IPv4 部分,仍然是OK
且還是7 Change ,
說明R1 執行的IP SLA 機制並無偵測到有斷線的跡象
即使線路已經從Primary 線路切換到Backup線路,

BGP-Session-Culling 測試結果說明 (3)

其次,觀察 IPv6部分:

```
R1#show bgp ipv6 unicast summary
BGP router identifier 1.1.1.1, local AS number 1
BGP table version is 14, main routing table version 14
2 network entries using 344 bytes of memory
2 path entries using 176 bytes of memory
2/2 BGP path/bestpath attribute entries using 272 bytes of memory
1 BGP AS-PATH entries using 24 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 816 total bytes of memory
BGP activity 5/1 prefixes, 15/11 paths, scan interval 60 secs

Neighbor        V           AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
2001:1234::2    4            2    1376    1381       14    0    0 20:26:44        1
2406:D400:1:133:203:163:222:2
                4            2       0       0        1    0    0 00:14:34 Active
```

[說明]: R1 ipv6 BGP NeighborR2 (2406:D400:1:133:203:163:222:2) 目前進入Active Status

```
R1#show bgp ipv6 unicast 2002::2/128
BGP routing table entry for 2002::2/128, version 14
Paths: (1 available, best #1, table default)
  Not advertised to any peer
  Refresh Epoch 6
  2
    2001:1234::2 (FE80::F6CF:E2FF:FE42:2961) from 2001:1234::2 (2.2.2.2)
      Origin IGP, metric 0, localpref 100, valid, external, best
```

[說明] 目前R1到R2 loop0 2002::2/128 流量,已經改走Backup路徑

BGP-Session-Culling 測試結果說明 (4)


```
R1#show track 2
Track 2
  IP SLA 2 state
  State is Up
    7 changes, last change 05:59:58
  Latest operation return code: OK
  Latest RTT (millisecs) 1
```

[說明] 目前 R1 Track到R2 loop0 2002::2/128 IPv6 部分,仍然是OK
且仍然還是7 Change ,
說明R1 執行的IP SLA 機制並無偵測到有斷線的跡象
(即使線路已經從Primary 線路切換到Backup線路,)

BGP-Session-Culling 測試結果說明(5)


```
N7K-1# show ip access-list acl-ipv4-permit-all-except-bgp

IP access list acl-ipv4-permit-all-except-bgp
        statistics per-entry
        10 deny tcp 203.163.222.0/24 eq bgp 203.163.222.0/24 [match=21]
        20 deny tcp 203.163.222.0/24 203.163.222.0/24 eq bgp [match=12]
        30 permit ip any any [match=180]

IP access list acl-ipv4-permit-all-except-bgp
        statistics per-entry
        10 deny tcp 203.163.222.0/24 eq bgp 203.163.222.0/24 [match=24]
        20 deny tcp 203.163.222.0/24 203.163.222.0/24 eq bgp [match=12]
        30 permit ip any any [match=238]
N7K-1# show ip access-list acl-ipv4-permit-all-except-bgp

IP access list acl-ipv4-permit-all-except-bgp
        statistics per-entry
        10 deny tcp 203.163.222.0/24 eq bgp 203.163.222.0/24 [match=26]
        20 deny tcp 203.163.222.0/24 203.163.222.0/24 eq bgp [match=12]
        30 permit ip any any [match=244]
```

[說明]:在IX Switch Nexus 7010 下, show ip access-list acl-ipv4-permit-all-except-bgp 可看到 R1/R2 仍試圖送Hello Packet建立IPv4 BGP session

BGP-Session-Culling 測試結果說明 (6)


```
N7K-1# show access-lists acl-ipv6-permit-all-except-bgp

IPv6 access list acl-ipv6-permit-all-except-bgp
        statistics per-entry
        10 deny tcp 2406:d400:1:133:203:163:222:0/112 eq bgp 2406:d400:1:133:203:163:222:0/112
[match=3960]
        20 deny tcp 2406:d400:1:133:203:163:222:0/112 2406:d400:1:133:203:163:222:0/112 eq bgp
[match=2278]
        30 permit ipv6 any any [match=2505]
N7K-1#
N7K-1# show access-lists acl-ipv6-permit-all-except-bgp

IPv6 access list acl-ipv6-permit-all-except-bgp
        statistics per-entry
        10 deny tcp 2406:d400:1:133:203:163:222:0/112 eq bgp 2406:d400:1:133:203:163:222:0/112
[match=3960]
        20 deny tcp 2406:d400:1:133:203:163:222:0/112 2406:d400:1:133:203:163:222:0/112 eq bgp
[match=2278]
        30 permit ipv6 any any [match=2506]
```

[說明]:在IX Switch Nexus 7010 下, show ip access-list acl-ipv4-permit-all-except-bgp 可看到 R1/R2 仍試圖送Hello Packet建立IPv4 BGP session

林盛琪 資深經理
Mark_Lin@chief.com.tw
是方電訊股份有限公司

http://www.chief.com.tw