RFC5549 v4 NLRI with v6 NH (ユースケースと実装状況)

Shishio Tsuchiya
shtsuchi@arista.com

Agenda

- RFC5549とは？
- ユースケース
- 実装状況

RFC5549 v4 NLRI with v6 NH とは

- Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop
- IPv4のNLRIをIPv6ネクストホップで運ぶ技術
- 2009年5月にRFC化

RFC5549 Capability Negotiation

```
+-------------------------------+
| Capability Code (1 octet:5 ) |
+-------------------------------+
| Capability Length (1 octet)  |
+-------------------------------+
| Capability Value (variable)  |
+-------------------------------+
```

```
+-----------------------------------------------------+
| NLRI AFI (2 octets)   1:IPv4                        |
+-----------------------------------------------------+
| NLRI SAFI(2 octets)   1:Unicast 2:Multicast         |
|                       4:Label 128:MPLS-VPN          |
+-----------------------------------------------------+
| Nexthop AFI(2 octets)  2:IPv6                       |
+-----------------------------------------------------+
```

- RFC5492 のCapability Code 5 Extended Next Hop Encodingでネゴシエーション
- Valueにてサポート可能なAFI/SFIおよび Nexthop AFIを通知(IPv6:2)

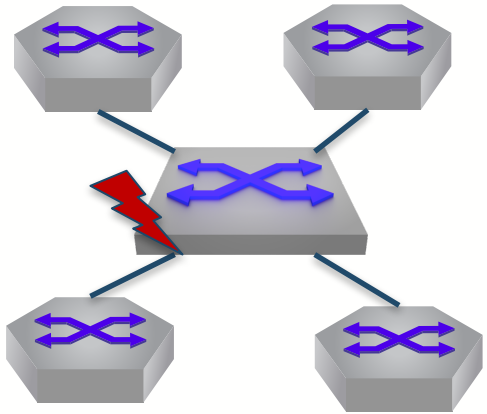
RFC5549 UPDATE/WITHDRAWN

```
+---------------------------------------------------+
| Address Family Identifier (2 octets)              |
+---------------------------------------------------+
| Subsequent Address Family Identifier (1 octet)    |
+---------------------------------------------------+
| Length of Next Hop Network Address (1 octet)      |
+---------------------------------------------------+
| Network Address of Next Hop (variable)            |
+---------------------------------------------------+
| Reserved (1 octet)                                |
+---------------------------------------------------+
| Network Layer Reachability Information (variable) |
+---------------------------------------------------+
```

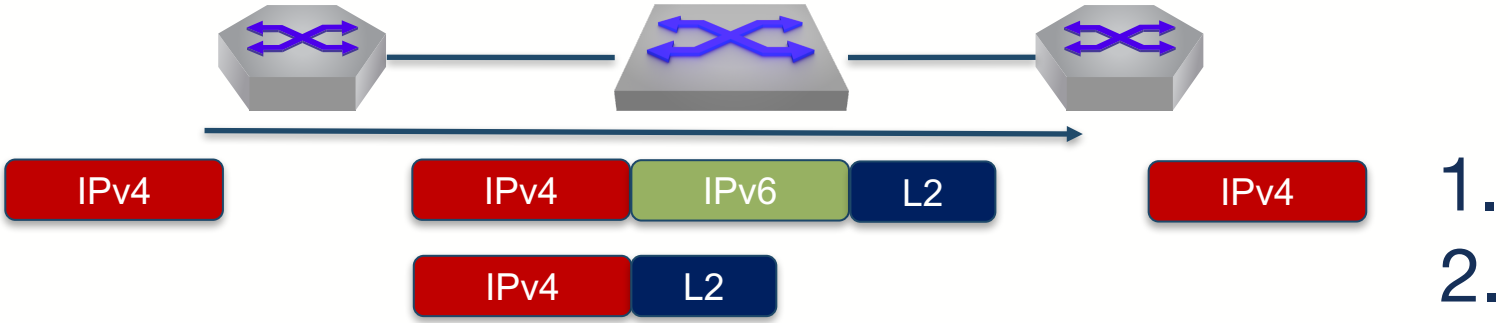
- RFC4760 のMPBGPのMP_REACH_NLRIとMP_UNREACH_NLRIをそのまま使用可能

ユースケース IXでの使用

- NANOG56でAMS-IXがIPアドレスの枯渇を説明(2012年)　2016-2017年
  1. 新しいアドレスのアサイン
  2. アドレス分割/dot1qの分割/ルートサーバー活用
  3. RFC1918（プライベートアドレス）の使用
  4. RFC3927（リンクローカル）の使用
  5. RFC5549の使用



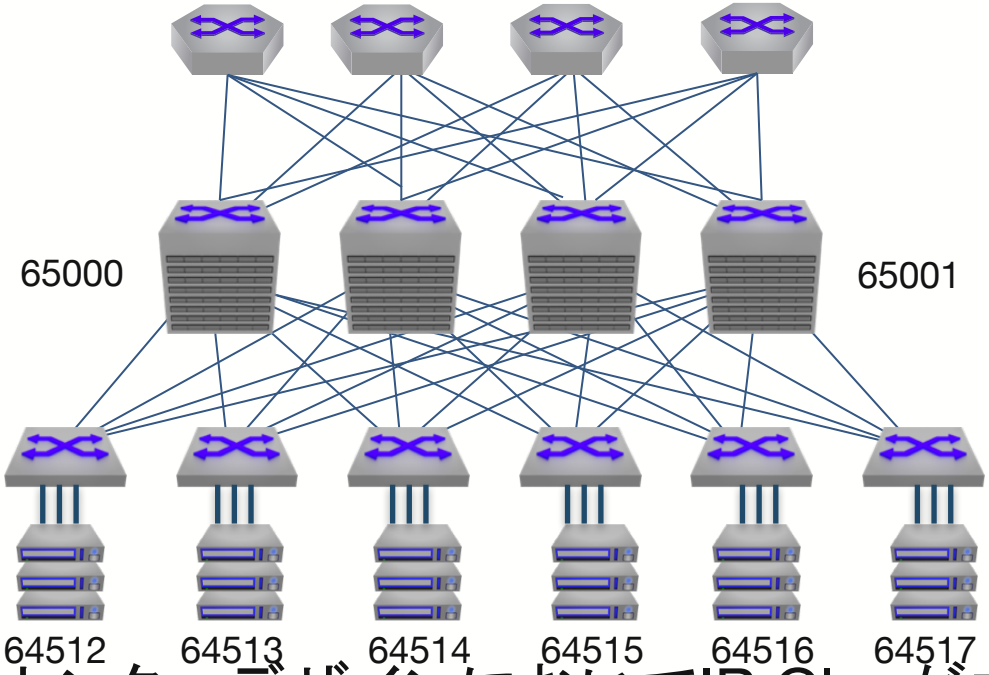
フォールディングの考慮



1. Softwire Mesh(4in6) RFC5565 IPv4でカプセル
   - 常に40バイトのオーバーヘッドがついてしまう
2. 直接L2（MACアドレス）をマップ

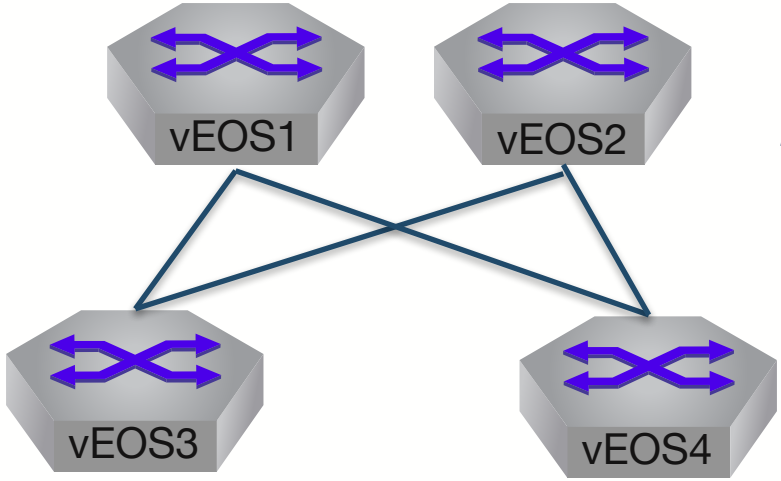
- Quaggaにて実装、公開
  - https://ams-ix.net/downloads/RFC5549/

ユースケース IPv6オンリーデータセンター



- データセンターデザインにおいてIP Closが主流に
- 同時期にIPv6オンリーデータセンターなども登場

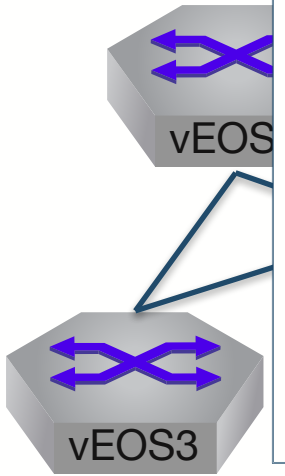
デモ



AS:6450x
IPv4アドレス:10.255.255.x/32
IPv6アドレス:2001:db8:ffff:ffff::x/128

- 各リンクにはIPv6アドレスのみを利用
- ループバックのIPv4アドレスをアドバタイズする

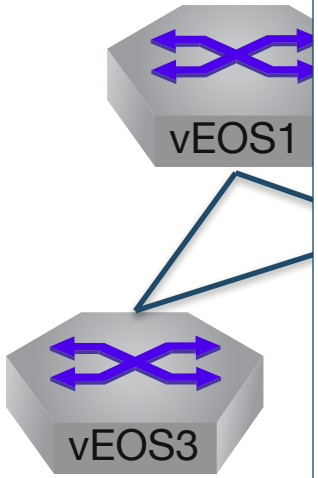
config




```
router bgp 64503
   no bgp default ipv4-unicast
   bgp default ipv4-unicast transport ipv6
   maximum-paths 128
   neighbor 2001:db8:10:13::1 remote-as 64501
   neighbor 2001:db8:10:13::1 maximum-routes 12000
   neighbor 2001:db8:10:23::2 remote-as 64502
   neighbor 2001:db8:10:23::2 maximum-routes 12000
   !
   address-family ipv4
      neighbor 2001:db8:10:13::1 activate
      neighbor 2001:db8:10:13::1 next-hop address-family ipv6 originate
      neighbor 2001:db8:10:23::2 activate
      neighbor 2001:db8:10:23::2 next-hop address-family ipv6 originate
      network 10.255.255.3/32
   !
```

- **bgp default ipv4-unicast transport ipv6**
  - 全てのIPv6ネイバーでIPv4ユニキャストの受信を可能に
- **neighbor [neighborアドレス] next-hop address-family ipv6 originate**
  - 指定ネイバーでNexthop IPv6送信を行う

capability

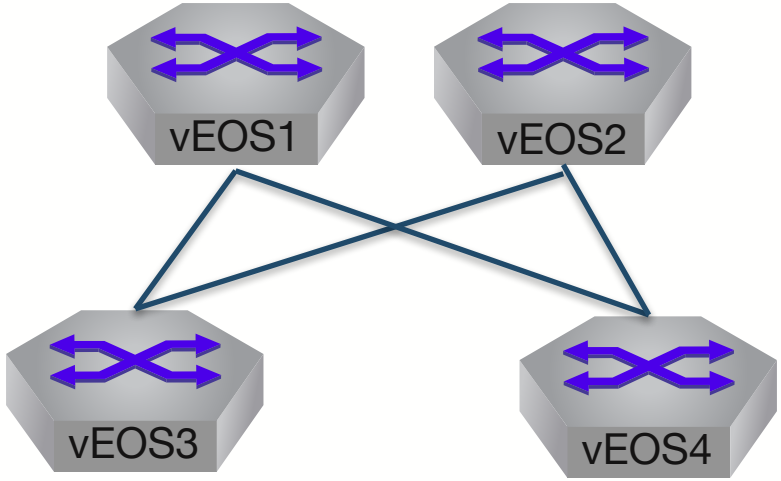



```
vEOS3#show ipv6  bgp neighbors 2001:db8:10:13::1
BGP neighbor is 2001:db8:10:13::1, remote AS 64501, external link
  BGP version 4, remote router ID 10.255.255.1, VRF default
  Negotiated BGP version 4
  Last read 00:00:03, last write 00:00:10
  Hold time is 180, keepalive interval is 60 seconds
  Configured hold time is 180, keepalive interval is 60 seconds
  Connect timer is inactive
  Idle-restart timer is inactive
  BGP state is Established, up for 02:04:10
  Number of transitions to established: 1
  Last state was OpenConfirm
  Last event was RecvKeepAlive
  Neighbor Capabilities:
    Multiprotocol IPv4 Unicast: advertised and received and negotiated
    Multiprotocol IPv6 Unicast: advertised and received and negotiated
    Extended Next-Hop Capability:
      IPv4 Unicast: advertised and received and negotiated
    Four Octet ASN: advertised and received
    Route Refresh: advertised and received and negotiated
    Send End-of-RIB messages: advertised and received and negotiated
```

show ip bgp

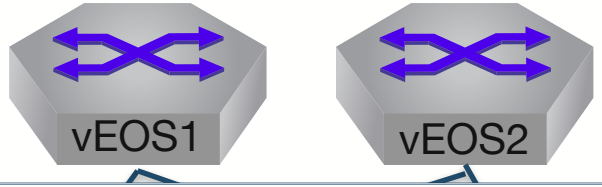
```
vEOS3#show ip bgp summary
BGP summary information for VRF default
Router identifier 10.255.255.3, local AS number 64503
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State  PfxRcd PfxAcc
  2001:db8:10:13::1 4  64501            161       135    0    0 02:06:55 Estab  2      2
  2001:db8:10:23::2 4  64502            161       137    0    0 02:06:53 Estab  2      2
vEOS3#show ip bgp
BGP routing table information for VRF default
Router identifier 10.255.255.3, local AS number 64503
Route status codes: s - suppressed, * - valid, > - active, # - not installed, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup, L - labeled-unicast
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
 * >      10.255.255.1/32        2001:db8:10:13::1     0       100     0       64501 i
 * >      10.255.255.2/32        2001:db8:10:23::2     0       100     0       64502 i
 * >      10.255.255.3/32        -                     0       0       -       i
 * >Ec    10.255.255.4/32        2001:db8:10:13::1     0       100     0       64501 64504 i
 *  ec    10.255.255.4/32        2001:db8:10:23::2     0       100     0       64502 64504 i
```

show ip bgp summary



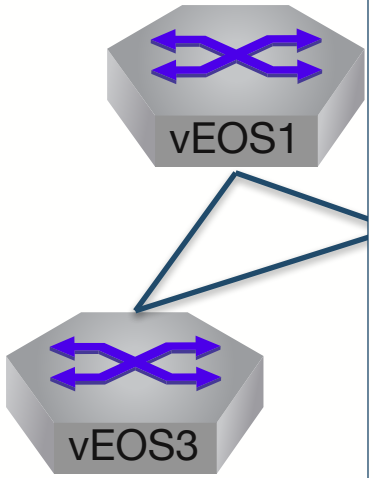

```
vEOS3#show ip bgp summary
BGP summary information for VRF default
Router identifier 10.255.255.3, local AS number 64503
Neighbor Status Codes: m - Under maintenance
  Neighbor         V  AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State  PfxRcd PfxAcc
  2001:db8:10:13::1 4  64501            161       135    0    0 02:06:55 Estab  2      2
  2001:db8:10:23::2 4  64502            161       137    0    0 02:06:53 Estab  2      2
```

show ip bgp




```
vEOS3#show ip bgp
BGP routing table information for VRF default
Router identifier 10.255.255.3, local AS number 64503
Route status codes: s - suppressed, * - valid, > - active, # - not installed, E - ECMP head, e - ECMP
                    S - Stale, c - Contributing to ECMP, b - backup, L - labeled-unicast
Origin codes: i - IGP, e - EGP, ? - incomplete
AS Path Attributes: Or-ID - Originator ID, C-LST - Cluster List, LL Nexthop - Link Local Nexthop

          Network                Next Hop              Metric  LocPref Weight  Path
 * >      10.255.255.1/32        2001:db8:10:13::1     0       100     0       64501 i
 * >      10.255.255.2/32        2001:db8:10:23::2     0       100     0       64502 i
 * >      10.255.255.3/32        -                     0       0       -       i
 * >Ec    10.255.255.4/32        2001:db8:10:13::1     0       100     0       64501 64504 i
 *  ec    10.255.255.4/32        2001:db8:10:23::2     0       100     0       64502 64504 i
```

show ip route




```
vEOS3#show ip route

VRF name: default
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B I - iBGP, B E - eBGP,
       R - RIP, I L1 - ISIS level 1, I L2 - ISIS level 2,
       O3 - OSPFv3, A B - BGP Aggregate, A O - OSPF Summary,
       NG - Nexthop Group Static Route, V - VXLAN Control Service

Gateway of last resort is not set

 C     10.10.1.0/24 [0/1]
        via Management1, directly connected
 B E   10.255.255.1/32 [200/0]
        via 2001:db8:10:13::1, Ethernet1
 B E   10.255.255.2/32 [200/0]
        via 2001:db8:10:23::2, Ethernet2
 C     10.255.255.3/32 [0/0]
        via Loopback0, directly connected
 B E   10.255.255.4/32 [200/0]
        via 2001:db8:10:13::1, Ethernet1
        via 2001:db8:10:23::2, Ethernet2
```

実装状況

- Arista EOS
  - **RFC 5549: IPv4 unicast NLRI with IPv6 next-hop support**
    https://eos.arista.com/eos-4-17-0f/rfc-5549-ipv4-unicast-with-ipv6-next-hop-support/
  - **IPv4 Addressless Forwarding on IPv6 addressed Interfaces for BGP RFC5549 support**

    https://eos.arista.com/eos-4-17-0f/ip-addressless-forwarding-changes-for-bgp-v6-nexthop-for-v4-routes/

Cisco NX-OS/Cumulus Linuxなど

まとめ

- データセンターへの需要を元に少しRFC5549の実装も増え始めてきた
  - 同じ目的で独自の実装もありそうなので、要注意
- AMS-IXの様にアドレスは枯渇してきた？
  - 昔聞いた時は大丈夫でしょ。みたいな感じだった
- トラブルシューティング手法など色々検討が必要?
  - 逆にアドレスに意味を持たせられるので楽かも？

ARISTA