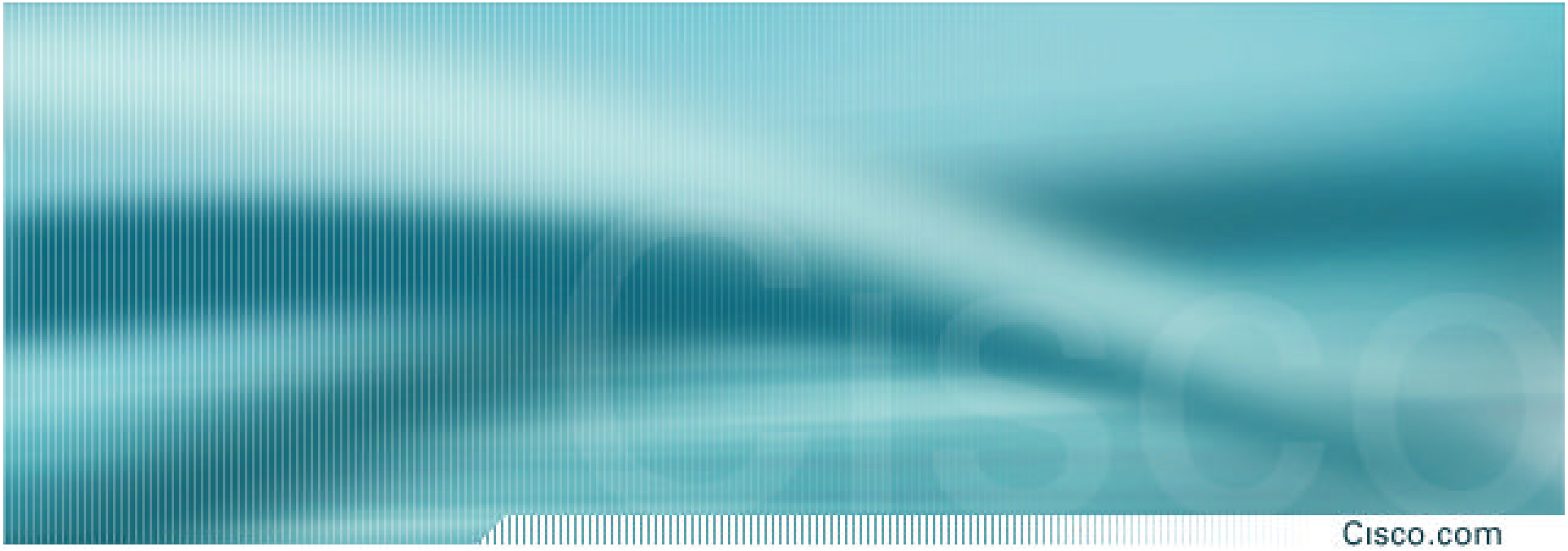


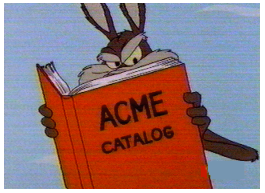
BGP Feature Update – 12.0S
July 2003

Mike Pennington

mpenning@cisco.com

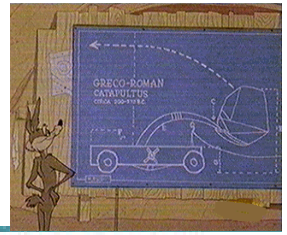
Cisco Systems - Denver, CO

Overview



- Overview
- Definition of Terms
- BGP Convergence optimization
- Issues w/ Static peer-groups
  - ✓ Peer Template Feature
  - ✓ Dynamic update-groups
- BGP multipath
  - ✓ eBGP Multipath
  - ✓ iBGP Multipath Feature
  - ✓ eiBGP Multipath Feature

Info & Standard Disclaimers

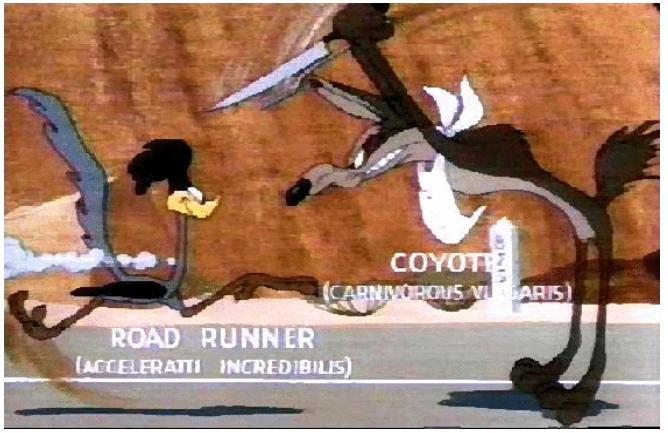


Cisco.com

- **What I do every day**
- **I am not the expert on these features**
- **Please test this stuff in your lab :-)**

Definition of Terms

- **FIB – Forwarding Information Base. This is essentially the CEF table on the main RP.**
- **LC FIB – A version of the FIB that is pushed from the RP to the LC DRAM.**
- **HW FIB – A version of the FIB that is formatted specifically for the HW Forwarding Engine**
- **IPC – Communication messaging between RP & LC**
- **BGP Update – A BGP message containing one or more prefixes that share a *common combination of attributes* (i.e. MED, Local-Pref, etc…)**



BGP Convergence Optimization

BGP Convergence Optimizations

- **12.0S - long history of BGP optimizations**
- **Some “earlier” (i.e. pre 21S) optimizations are listed below:**
  - ✓ **CSCdm56595 - BGP: reduce the initial routing churn at bootup**
  - ✓ **CSCdr50217 - BGP: Sending updates slow**
  - ✓ **CSCdt34187 - BGP should optimize update advertisement**

BGP Convergence Optimizations

- **CSCdm56595 - BGP: reduce the initial routing churn at bootup**
  - ✓ **Challenge: Installing routes into the routing table is very slow; BGP was computing bestpath as each prefix was received.**
  - ✓ **Solution: CSCdm56595 introduces BGP READ_ONLY / READ_WRITE modes. BGP should stay in READ_ONLY mode until all initial updates are received. Bestpath is only computed after transition to READ_WRITE mode.**
  - ✓ **Integrated: 12.0(5)S**

BGP Convergence Optimizations

- **CSCdr50217 - BGP: Sending updates slow**
  - ✓ **Challenge: IOS bug affecting how quickly BGP could dequeue updates from the BGP output queue**
  - ✓ **Solution: Drain BGP OutQ aggressively**
  - ✓ **Integrated: 12.0(14)S**

BGP Convergence Optimizations

Cisco.com

- **CSCdt34187 - BGP should optimize update advertisement**
  - ✓ **Challenge: BGP packs update messages inefficiently.**
  - ✓ **Solution: When packing update messages, use an “attribute cache” and sort NLRI by common attributes; then format and dequeue the updates.**
  - ✓ **Integrated: 12.0(18)S1**

BGP Convergence Optimizations

- **Even with the older performance improvements BGP flaps take ~ 8 minutes to recover in 12.0(21)S6 (145k prefixes, See next slide)**
- **CSCdw45143 - BGP: Fast Convergence**
  - ✓ **More convergence improvement over 21S**

BGP Convergence Optimizations

Cisco.com

- **CSCdw45143 - BGP: Fast Convergence**
  - ✓ **Challenge: BGP OUTQ was limited to 500 total messages between all peers**
  - ✓ **Solution: BGP OUTQ now significantly increased.**
  - ✓ **Integrated: 12.0(22)S**
  - ✓ **Beware of: CSCea03118 and other BGP issues in early 12.0(22)S / 23S. Use at least 12.0(22)S4 or 12.0(23)S2.**

BGP Convergence Optimizations

- **Convergence performance test parameters**
  - ✓ **GRP-B w/ 512Mb**
  - ✓ **Sending 145,000 NLRI to one peer (using an iBGP table from a live provider)**
  - ✓ **Measure time for OutQ = 0 after “clear ip bgp <nei>”**
- **Data**
  - ✓ **7.9 min (476 s) – 12.0(21)S6**
  - ✓ **3.7 min (222 s) – 12.0(24)S**

BGP Convergence Optimizations

- **Other platform considerations**
  - ✓ **GSR**
  - ✓ **7500**

BGP Convergence Optimizations

- **GSR considerations for convergence**
  - ✓ **`hold-queue 1000 in` (on all intfs)**

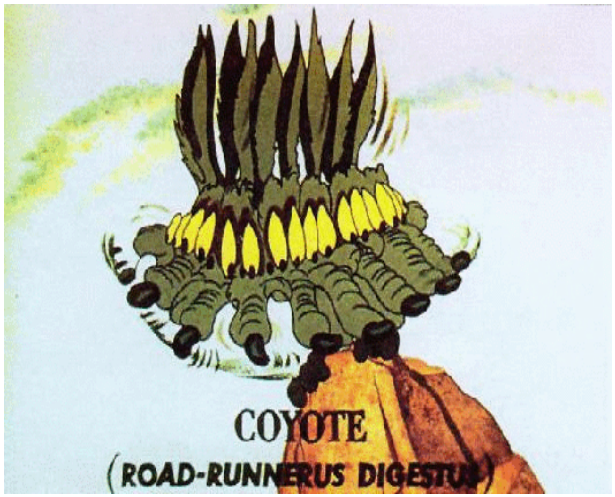
    This reduces the risk of loosing TCP acks during heavy BGP activity. PMTU discovery also helps.
  - ✓ **`spd headroom 1000` (Global cmd)**
  - ✓ **`bgp update-delay 300` (BGP cmd)**

    Used to minimize resource contention between CEF and BGP. The update delay is a maximum time bgp will wait after the first peer comes up before selecting the bestpath and pushing updates to the LCs. If all peers come up before 300s, then bestpath will be computed at that time.

BGP Convergence Optimizations

- **7500 considerations for convergence**
  - ✓ **`hold-queue 750 in` (on all intfs)**
  - ✓ **`spd headroom 1000`**
  - ✓ **BGP update-delay normally does not need to be adjusted on the 7500.**

Issues w/ Static peer-groups



Issues w/ Static peer-groups

- **Peer-groups: Two roles…**
  - ✓ **Update grouping efficiency**
  - ✓ **Configuration efficiency**
- **Disadvantages:**
  - ✓ **Static configuration**
  - ✓ **Any given neighbor must be in same peer-group for all AFs.**
- **Result: Suboptimal BGP update replication**

Peer Template Feature

- **Peer templates are the “next gen” version of peer-groups**
  - ✓ **Retains configuration benefits**
  - ✓ **Supports attribute inheritance between templates**
  - ✓ **Dynamic update-group feature groups neighbors based on outbound policies for optimal update packing**
- **CSCdp42769 - BGP:peer-template implementation**
  - ✓ **Integrated: 12.0(24)S**

Peer Template Feature

- **Two kinds of peer-templates**
  - ✓ **peer-session templates**
  - ✓ **peer-policy templates**

Peer Template Feature

- **peer-session templates**
  - ✓ **Configures the "BGP session parameters" in a template form**
  - ✓ **peer-session templates are configured per-neighbor with attributes common to all AFs**
  - ✓ **peer-session templates can inherit attributes from up to 7 other peer-session templates**
  - ✓ **peer-groups and peer templates cannot be mixed with the same neighbor**

Peer Template Feature

- **Peer-session templates are for general session commands**
  - **description**
  - **ebgp-multihop**
  - **inherit peer-session**
  - **password**
  - **shutdown**
  - **translate-update**
  - **version**
  - **disable-connected-check**
  - **exit-peer-session**
  - **local-as**
  - **remote-as**
  - **timers**
  - **update-source**

Peer Template Feature

- **peer-policy templates**
  - ✓ **Configures the “BGP policies” in a template form**
  - ✓ **peer-policy templates are configured per-neighbor inside individual AFs**
  - ✓ **peer-policy templates can inherit attributes from up to 7 other peer-policy templates**

Peer Template Feature

- **Peer-Policy Templates**

BGP Policy commands that are configured for specific AFs or NLRI configuration modes are configured in a peer-policy template.

- advertisement-interval
- as-override
- default-originate
- exit-peer-policy
- inherit-peer-policy
- next-hop-self
- prefix-list
- route-map
- route-reflector-client
- send-label
- unsuppress-map
- allowas-in
- capability
- distribute-list
- dmzlink-bw
- filter-list
- maximum-prefix
- next-hop-unchnaged
- remove-private-as
- send-community
- soft-configuration
- weight

Peer Template Feature

- **Peer Template inheritance priorities**
  - ✓ **Inherited templates have sequence numbers; higher sequence numbers supercede parameters inherited from templates with lower sequence numbers**
  - ✓ **Configuration within a template supercedes all inherited parameters in the template**
  - ✓ **Neighbor-level configuration supercedes all inherited parameters**
  - ✓ **See example on the next slide**

Peer Template Feature - Configuration

Cisco.com

```
router bgp 100
 template peer-policy SET_MAXPREFIX_LOW
  prefix-list SEND_MARTIANS_ONLY out
  advertisement-interval 2
  maximum-prefix 100
 exit-peer-policy
 !
 template peer-policy SET_MAXPREFIX_HIGHER
  route-map BOGUS out
  prefix-list DENY_MARTIANS out
  maximum-prefix 2147483647
 exit-peer-policy
 !
 template peer-policy SET_LOCALPREF
  route-map LOCALPREF out
  inherit peer-policy SET_MAXPREFIX_HIGHER 20
  inherit peer-policy SET_MAXPREFIX_LOW 10
 exit-peer-policy
 !
 no synchronization
 bgp log-neighbor-changes
 bgp deterministic-med
 bgp update-delay 300
 neighbor 1.1.1.2 remote-as 200
 neighbor 1.1.1.2 send-community
 neighbor 1.1.1.2 advertisement-interval 20
 neighbor 1.1.1.2 inherit peer-policy SET_LOCALPREF
```

SET_MAXPREFIX_LOW sets the advertisement interval, an outbound prefix-list, and maximum prefix

SET_MAXPREFIX_HIGHER sets an outbound route-map, an outbound prefix-list, and maximum prefix

SET_LOCALPREF inherits from SET_MAXPREFIX_LOW and SET_MAXPREFIX_HIGHER, then sets an outbound route-map

Peer Template Feature - Monitoring

Cisco.com

```
GSR1#show ip bgp neighbor 1.1.1.2
BGP neighbor is 1.1.1.2,  remote AS 200, external link
[snip]
  Index 2, Offset 0, Mask 0x4
  Member of update-group 2
  Inherits from template SET_LOCALPREF
  Community attribute sent to this neighbor
  Outbound path policy configured
  Outgoing update prefix filter list is DENY_MARTIANS
  Route map for outgoing advertisements is LOCALPREF
[snip]
  Maximum prefixes allowed 2147483647
  Threshold for warning message 75%
  Number of NLRIs in the update sent: max 0, min 0
  Minimum time between advertisement runs is 20 seconds
```

Dynamic update-groups

- **Idea: Calculate update-groups dynamically per AF**
- **Dynamic update-groups solve the MP-BGP update grouping problem**
  - ✓ **Calculate update-groups at configuration time**
  - ✓ **Peer-groups still work, but have no dynamic or per-AF benefits**

Dynamic update-groups

- **Neighbors belong to same update-group, *if*:**
  - ✓ **Same kind of peer: IBGP or EBGP**
  - ✓ **For IBGP: RR-Clients or non-RR-Client**
  - ✓ **Within the same AF**
  - ✓ **Configured with same outbound policy parameters**

Dynamic update-groups - Monitoring

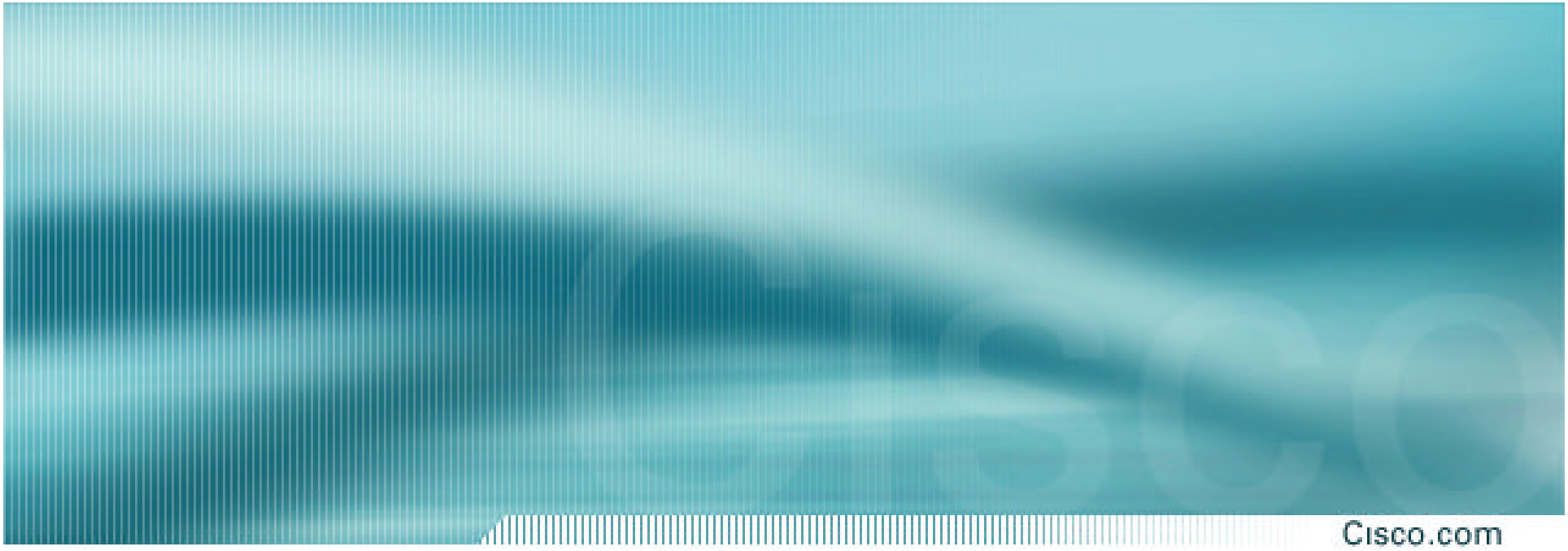
Cisco.com

```
GSR1#show ip bgp update-group summary
Summary for Update-group 1 :
------------------------------
BGP router identifier 10.94.165.85, local AS number 100
BGP table version is 1, main routing table version 1

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.94.165.2     4     1       0       0        0    0    0 never    Active

Summary for Update-group 2 :
------------------------------
BGP router identifier 10.94.165.85, local AS number 100
BGP table version is 1, main routing table version 1

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
1.1.1.2         4   200      55      56        1    0    0 00:52:02        0

GSR1#
```

Dynamic update-groups - Monitoring

Cisco.com

```
GSR1#sh ip bgp update-group 2
BGP version 4 update-group 2, external, Address Family: IPv4 Unicast
  BGP Update version : 0, messages 0/0
  Community attribute sent to this neighbor
  Outgoing update prefix filter list is DENY_MARTIANS
  Route map for outgoing advertisements is LOCALPREF
  Update messages formatted 0, replicated 0
  Number of NLRIs in the update sent: max 0, min 0
  Minimum time between advertisement runs is 20 seconds
  Has 1 member (* indicates the members currently being sent updates):
   1.1.1.2

GSR1#
```

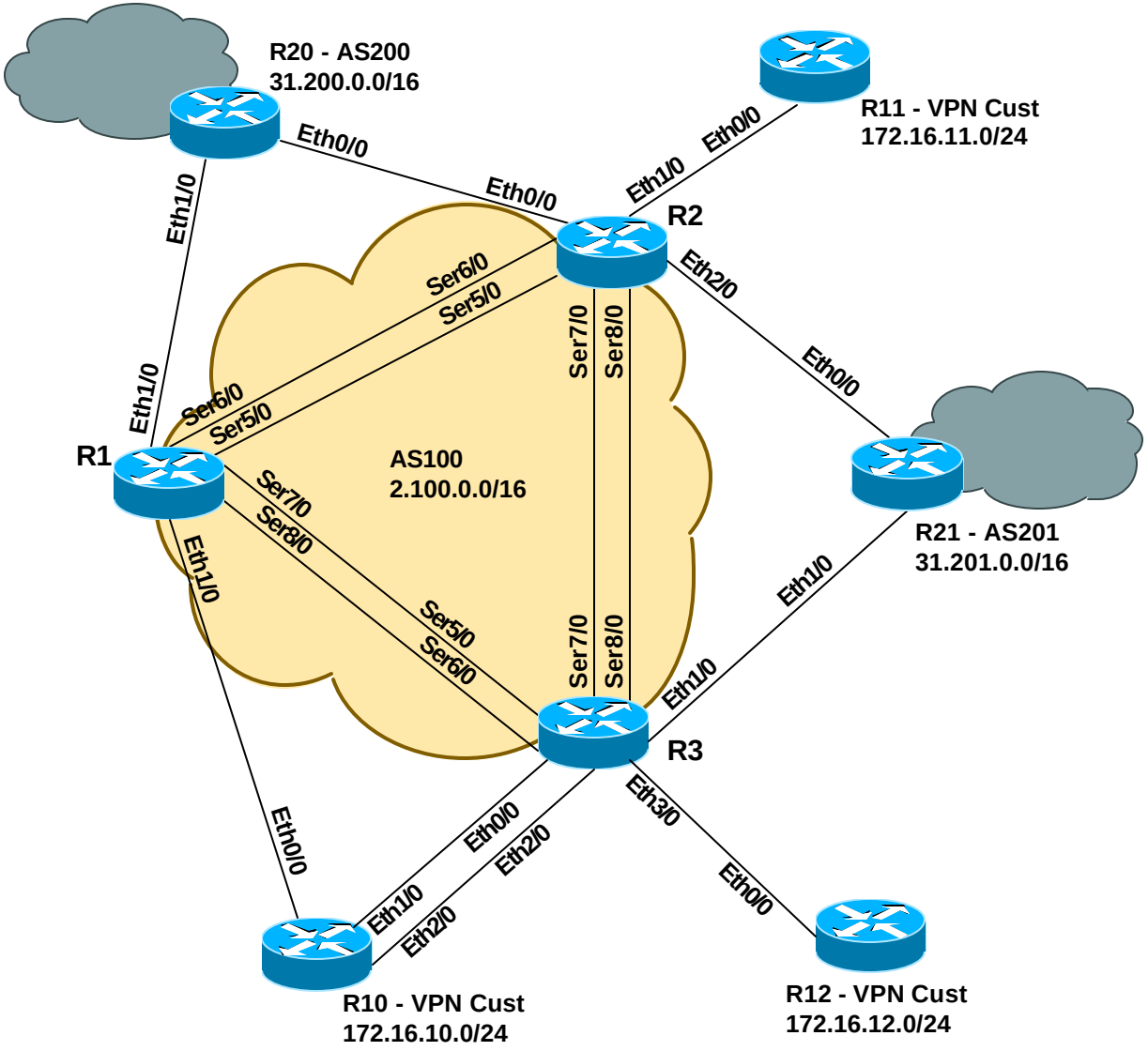


BGP Multipath

BGP Multipath

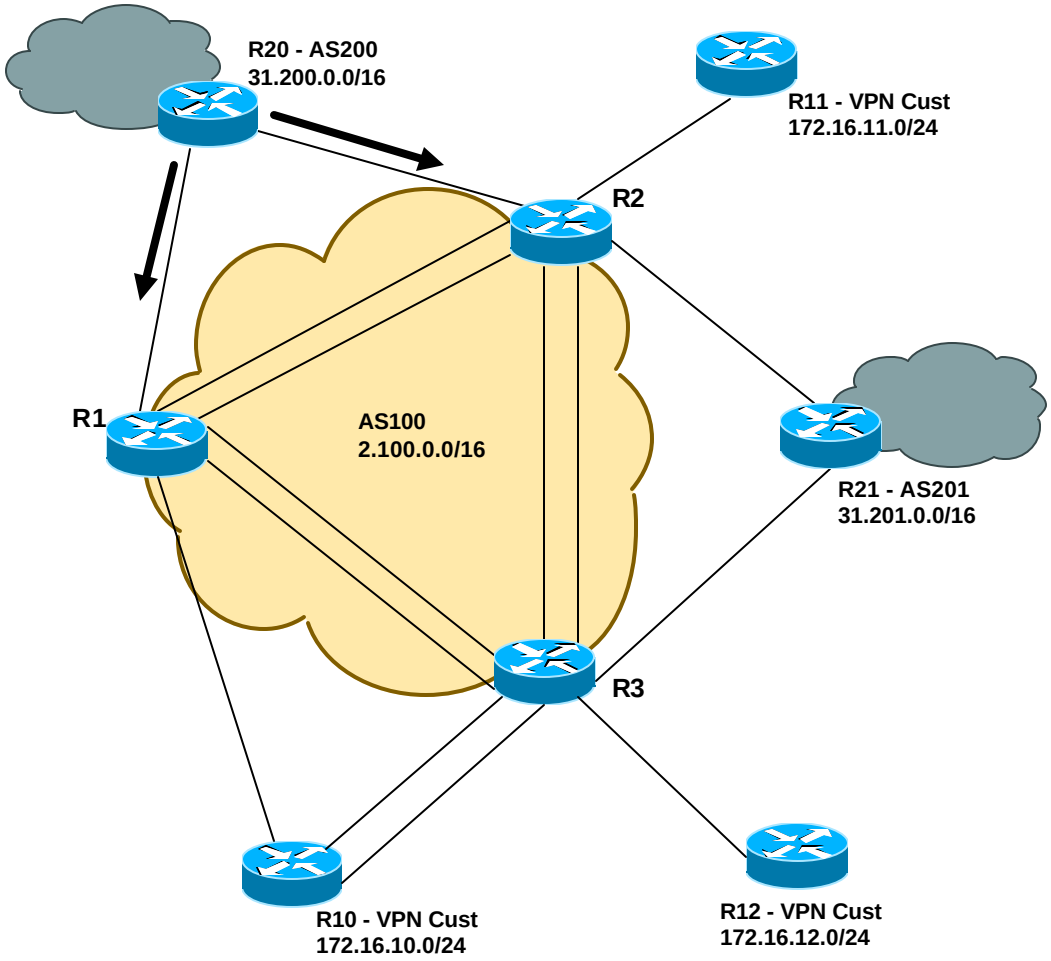
- **eBGP Multipath: Legacy feature**
  - ✓ **Syntax: `maximum-paths 8`**
- **iBGP Multipath: New feature (IPv4 and VPNv4)**
  - ✓ **Syntax: `maximum-paths ibgp 8`**
- **eiBGP Multipath: New feature (VPNv4 only)**
  - ✓ **Syntax: `maximum-paths eibgp 8`**

BGP Multipath - Topology



BGP Multipath – eBGP LB Example

- **Legacy eBGP Load Balancing**
- **R20 load balances across two paths to AS100**



BGP Multipath – eBGP LB Example

Cisco.com

```
R20#sh runn | b router bgp
router bgp 200
 no synchronization
 bgp log-neighbor-changes
 bgp deterministic-med
 bgp bestpath compare-routerid
 bgp maxas-limit 100
 bgp dampening 15 750 3000 45
 network 31.200.0.0 mask 255.255.0.0
 neighbor 2.100.1.1 remote-as 100
 neighbor 2.100.1.1 send-community
 neighbor 2.100.2.1 remote-as 100
 neighbor 2.100.2.1 send-community
 maximum-paths 6
 no auto-summary
```

BGP Multipath – eBGP LB Example

Cisco.com

```
R20#sh ip bgp 2.100.0.0
BGP routing table entry for 2.100.0.0/16, version 8
Paths: (2 available, best #1, table Default-IP-Routing-Table)
Multipath: eBGP
  Advertised to non peer-group peers:
  2.100.2.1
  100
    2.100.1.1 from 2.100.1.1 (2.0.0.1)
      Origin IGP, metric 0, localpref 100, valid, external, multipath, best
  100
    2.100.2.1 from 2.100.2.1 (2.0.0.2)
      Origin IGP, localpref 100, valid, external, multipath
R20#
```

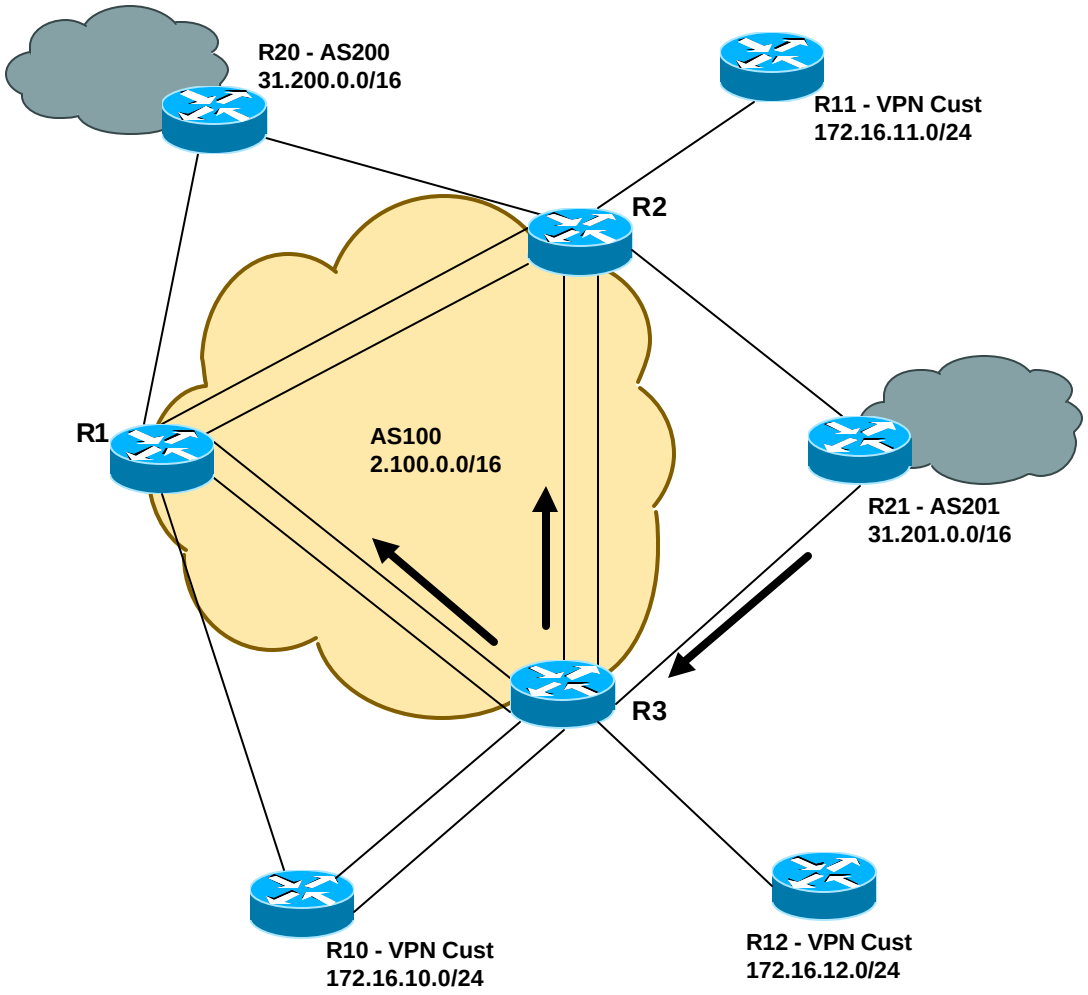
BGP Multipath – eBGP LB Example

Cisco.com

```
R20#sh ip cef 2.100.0.0 internal
2.100.0.0/16, version 19, epoch 0, per-destination sharing
0 packets, 0 bytes
  via 2.100.1.1, 0 dependencies, recursive
    traffic share 1
    next hop 2.100.1.1, Ethernet1/0 via 2.100.1.1/32
    valid adjacency
  via 2.100.2.1, 0 dependencies, recursive
    traffic share 1
    next hop 2.100.2.1, Ethernet0/0 via 2.100.2.1/32
    valid adjacency
[snip]
R20#
```

BGP Multipath – iBGP LB Example

- **iBGP Load Balancing**
- **R3 load balances across two paths to AS200**
- **Balance to both eBGP nexthops for AS200**
- **iBGP LB does NOT support IGP recursion (i.e. only two paths on R3, not four)**



BGP Multipath – iBGP LB Example

Cisco.com

```
R3#show runn | b router bgp
router bgp 100
![snip]
neighbor IBGP peer-group
 neighbor IBGP remote-as 100
 neighbor IBGP update-source Loopback0
 neighbor 2.0.0.1 peer-group IBGP
 neighbor 2.0.0.2 peer-group IBGP
 neighbor 2.100.3.2 remote-as 201
 maximum-paths ibgp  8
 !
 address-family ipv4
 neighbor IBGP activate
 neighbor IBGP send-community
 neighbor 2.0.0.1 peer-group IBGP
 neighbor 2.0.0.2 peer-group IBGP
 neighbor 2.100.0.26 activate
 neighbor 2.100.0.26 send-community both
 neighbor 2.100.0.30 activate
 neighbor 2.100.0.30 send-community both
 neighbor 2.100.3.2 activate
 neighbor 2.100.3.2 send-community
 maximum-paths ibgp  8
![snip]
```

BGP Multipath – iBGP LB Example

Cisco.com

```
R3#sh ip bgp 31.200.0.0
BGP routing table entry for 31.200.0.0/16, version 8
Paths: (2 available, best #1)
Multipath: iBGP
  Advertised to update-groups:
     1           2
  200
    2.100.1.2 (metric 74) from 2.0.0.1 (2.0.0.1)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath, best
  200
    2.100.2.2 (metric 74) from 2.0.0.2 (2.0.0.2)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath
R3#
```

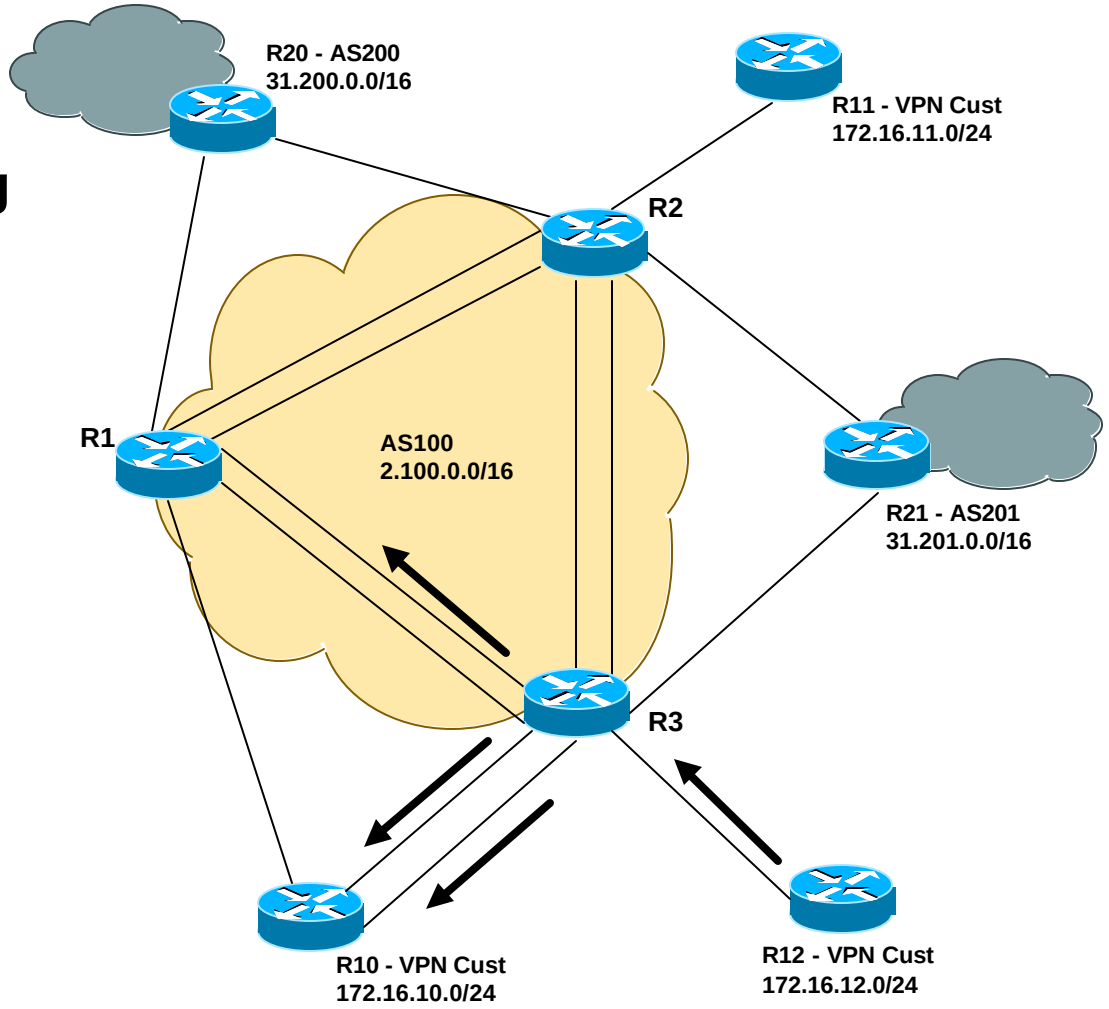
BGP Multipath – iBGP LB Example

Cisco.com

```
R3#sh ip cef 31.200.0.0 internal
31.200.0.0/16, version 72, epoch 0, per-destination sharing
0 packets, 0 bytes
  tag information from 2.100.1.0/24, shared, all rewrites owned
    local tag: 20
  via 2.100.1.2, 0 dependencies, recursive
    traffic share 1
    next hop 2.100.0.13, Serial6/0 via 2.100.1.0/24 (Default)
    valid adjacency (0x00E26518)
    tag rewrite with Se6/0, point2point, tags imposed {}
  via 2.100.2.2, 0 dependencies, recursive
    traffic share 1
    next hop 2.100.0.17, Serial7/0 via 2.100.2.0/24 (Default)
    valid adjacency (0x00E26268)
    tag rewrite with Se5/0, point2point, tags imposed {}
[snip]
R3#
```

BGP Multipath – eiBGP LB Example 1

- **eiBGP Load Balancing**
- **Example of traffic ingress from the CE router, R12**
- **R3 load balances across both iBGP and eBGP paths to MPLS VPN Customer, R10**



BGP Multipath – eiBGP LB Example 1

Cisco.com

```
R3#show runn | b router bgp
router bgp 100
![snip]
 address-family ipv4 vrf VRF01
![snip]
 neighbor 172.16.1.130 remote-as 65000
 neighbor 172.16.1.130 activate
 neighbor 172.16.1.130 send-community
 neighbor 172.16.1.130 as-override
 neighbor 172.16.1.130 route-map SET_SOO_R3 in
 neighbor 172.16.1.194 remote-as 65000
 neighbor 172.16.1.194 activate
 neighbor 172.16.1.194 send-community
 neighbor 172.16.1.194 as-override
 neighbor 172.16.1.194 route-map SET_SOO_R3 in
 maximum-paths eibgp 8
 no auto-summary
 no synchronization
 exit-address-family
```

BGP Multipath – eiBGP LB Example 1

Cisco.com

```
R3#show ip bgp vpnv4 vrf VRF01 172.16.10.0
BGP routing table entry for 100:1:172.16.10.0/24, version 45
Paths: (3 available, best #1, table VRF01)
Multipath: eiBGP
  Advertised to update-groups:
     5
  65000
    172.16.1.130 (via VRF01) from 172.16.1.130 (172.16.1.194)
      Origin IGP, metric 0, localpref 100, valid, external, multipath, best
      Extended Community: SoO:100:2 RT:100:1
  65000
    172.16.1.194 (via VRF01) from 172.16.1.194 (172.16.1.194)
      Origin IGP, metric 0, localpref 100, valid, external, multipath
      Extended Community: SoO:100:2 RT:100:1
  65000
    2.0.0.1 (metric 65) from 2.0.0.1 (2.0.0.1)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath
      Extended Community: RT:100:1
R3#
```

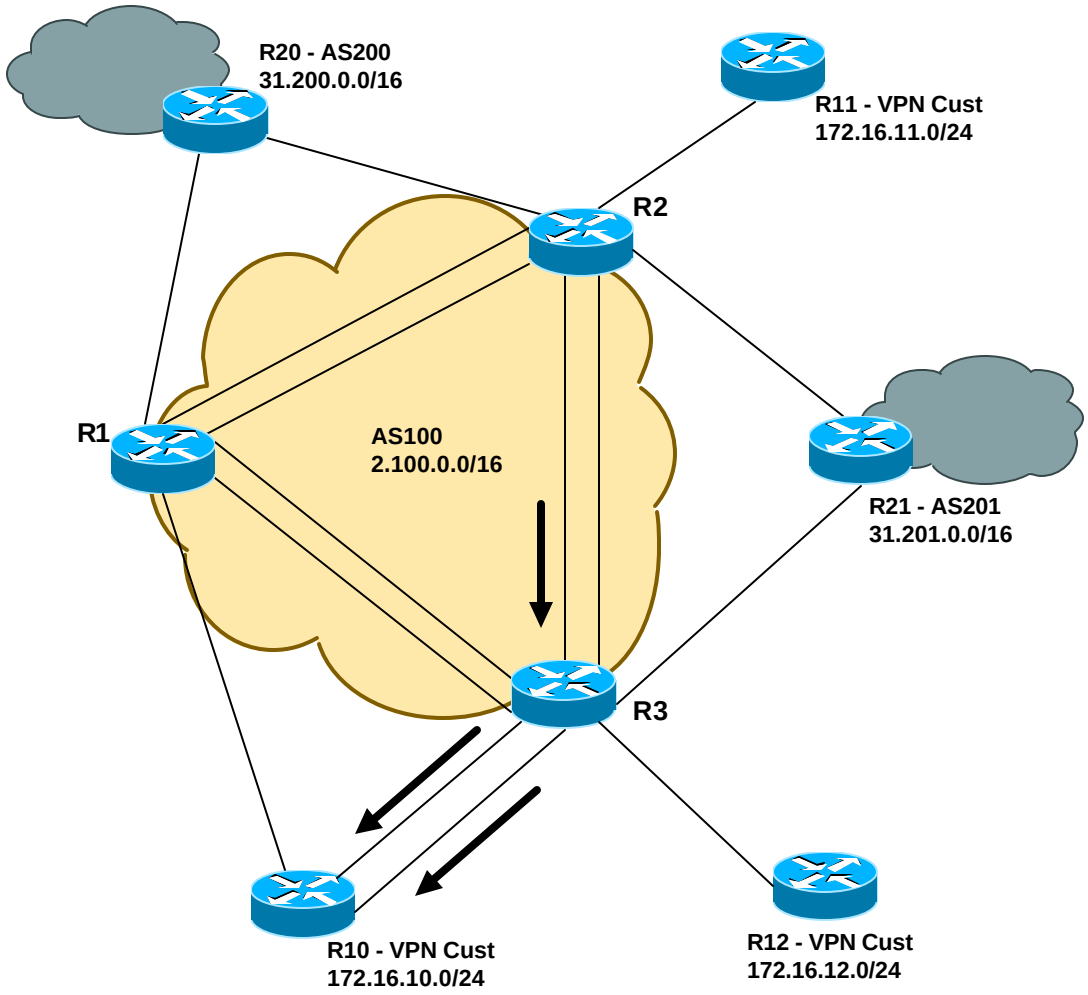
BGP Multipath – eiBGP LB Example 1

Cisco.com

```
R3#sh ip cef vrf VRF01 172.16.10.0 internal
172.16.10.0/24, version 51, epoch 0, per-destination sharing
0 packets, 0 bytes
  tag information set, all rewrites owned
    local tag: 26
  via 172.16.1.130, 0 dependencies, recursive
    traffic share 1
    next hop 172.16.1.130, Ethernet0/0 via 172.16.1.130/32 (VRF01)
    valid adjacency (0x00E25A58)
    tag rewrite with Et0/0, 172.16.1.130, tags imposed {}
  via 172.16.1.194, 0 dependencies, recursive
    traffic share 1
    next hop 172.16.1.194, Ethernet2/0 via 172.16.1.194/32 (VRF01)
    valid adjacency (0x00E25900)
    tag rewrite with Et2/0, 172.16.1.194, tags imposed {}
  via 2.0.0.1, 0 dependencies, recursive
    traffic share 1
    next hop 2.100.0.9, Serial5/0 via 2.0.0.1/32 (Default)
    valid adjacency (0x00E267C8)
    tag rewrite with
        Recursive rewrite via 2.0.0.1/32, tags imposed {28}
[snip]
```

BGP Multipath – eiBGP LB Example 2

- **eiBGP Load Balancing**
- **Example of traffic ingress from R11; therefore the ingress traffic to R3 has a label**
- **R3 load balances across eBGP paths only to MPLS VPN Customer, R10**
- **R3 cannot forward the labeled traffic for R10 to R1 via iBGP. R1 could do the same and cause a loop!**



BGP Multipath – eiBGP LB Example 2

Cisco.com

```
R2#sh ip bgp vpn vrf VRF01 labels
   Network          Next Hop      In label/Out label
Route Distinguisher: 100:1 (VRF01)
   172.16.0.11/32   172.16.2.2      26/nolabel
   172.16.1.0/25    2.0.0.3         nolabel/23
                    2.0.0.1         nolabel/25
   172.16.1.128/26  2.0.0.3         nolabel/25
                    2.0.0.1         nolabel/27
   172.16.1.192/26  2.0.0.3         nolabel/26
                    2.0.0.1         nolabel/28
   172.16.2.0/24    172.16.2.2      24/aggregate(VRF01)
   172.16.10.0/24   2.0.0.3         nolabel/24
                    2.0.0.1         nolabel/26
   172.16.11.0/24   172.16.2.2      25/nolabel

R2#
```

BGP Multipath – eiBGP LB Example 2

Cisco.com

```
R3#sh ip bgp vpn vrf VRF01 172.16.10.0
BGP routing table entry for 100:1:172.16.10.0/24, version 23
Paths: (3 available, best #1, table VRF01)
Multipath: eiBGP
  Advertised to update-groups:
     1          2
  65000
    172.16.1.130 (via VRF01) from 172.16.1.130 (172.16.1.194)
      Origin IGP, metric 0, localpref 100, valid, external, multipath, best
      Extended Community: RT:100:1
  65000
    172.16.1.194 (via VRF01) from 172.16.1.194 (172.16.1.194)
      Origin IGP, metric 0, localpref 100, valid, external, multipath
      Extended Community: RT:100:1
  65000
    2.0.0.1 (metric 65) from 2.0.0.1 (2.0.0.1)
      Origin IGP, metric 0, localpref 100, valid, internal, multipath
      Extended Community: RT:100:1
R3#
```

iBGP path will not be used for tagged traffic

BGP Multipath – eiBGP LB Example 2

Cisco.com

```
R3#sh mpls forwarding-table label 24
Local  Outgoing    Prefix           Bytes tag  Outgoing    Next Hop
tag    tag or VC   or Tunnel Id     switched   interface
24     Untagged    172.16.10.0/24[V] 0         Et0/0       172.16.1.130
       Untagged    172.16.10.0/24[V] 0         Et2/0       172.16.1.194
R3#
```

BGP Multipath

- **Forwarding Restrictions**
  - ✓ **eiBGP multipath only works for MPLS VPN**

    Packets from CE router will receive eiBGP LB (i.e. example 1)

    Packets from another PE or P router receive iBGP LB only (i.e. example 2)
  - ✓ **BGP still advertises ONE AND ONLY ONE bestpath**
  - ✓ **IOS currently does not support iBGP/eiBGP multipath for Route-reflector topologies**

    draft-walton-bgp-add-paths-01.txt addresses this possibility

BGP Multipath

- **Platform Restrictions**
  - ✓ **Requires platform FIB/DFIB support.**
  - ✓ **eiBGP is currently supported on:**
    - **12000 (E0-E3, E4+ supported, E4 not supported)**
    - **7500**
    - **7200**