# Let's go cisco live! #CiscoLive



# Border Gateway Protocol Fundamentals

Gustavo Sibaja - Systems Architect @GustavoSibaja4

Peter Paluch - Automation Boot Camp Delivery @Peter\_Paluch

BRKENT-1179

cisco Live!

#### Cisco Webex App

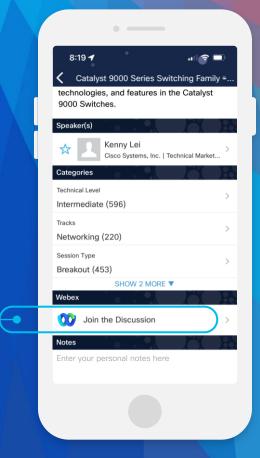
#### Questions?

Use Cisco Webex App to chat with the speaker after the session

#### How

- 1 Find this session in the Cisco Live Mobile App
- 2 Click "Join the Discussion"
- 3 Install the Webex App or go directly to the Webex space
- 4 Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 9, 2023.



https://ciscolive.ciscoevents.com/ciscolivebot/#BRKENT-1179

# Agenda

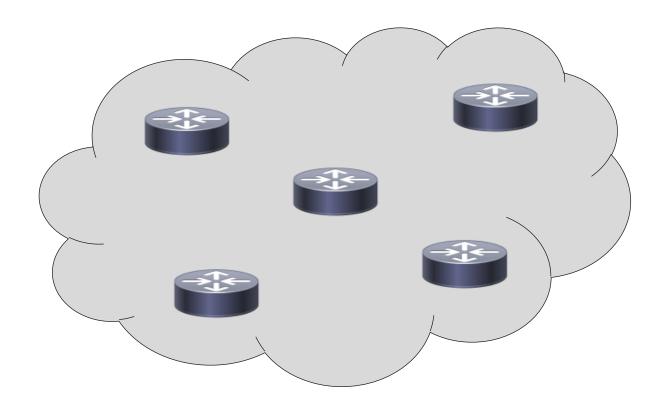
- Why BGP?
- Messages and States
- Attributes
- Internal vs External BGP
- Best Path Selection Algorithm
- Policy Control and Traffic Engineering

## Why BGP?



#### Autonomous System

A group of one or more IP prefixes (lists of IP addresses accessible on a network) run by one or more network operators that maintain a single, clearly-defined routing policy.





## BGP 2-byte ASN RFC1105 (1989)

- Created to address 2-byte ASN depletion
- Interoperable with 2-byte ASNs (range includes 2-byte ASNs)
- 32-bit number
- 0 to 4294967295
- Additional private range 4200000000 to 4294967294





- 16-bit number
- 0 to 65535
- Private range
   64512 to 65534

BGP 4-byte ASN RFC4893 (2007)



#### Border Gateway Protocol

- Border Gateway Protocol
  - Large scale, robust and stable routing protocol designed to operate between autonomous systems
  - Based on TCP, listens on port 179
  - Fundamentally a distance vector protocol
  - Does not have the concept of a simple metric
  - Instead, uses multiple characteristics called attributes
  - Allows for strong control over advertised routes and their attributes
  - Assumes that the routing inside the autonomous system is already fully taken care of by an IGP (EIGRP, OSPF, IS-IS)



Inter-AS Routing Policies **BGP** IS-IS **EIGRP** BGP BGP OSPF



BRKENT-1179

#### Route Scale & Control Plane Stability

```
route-views> show bgp ipv4 unicast summary | ex never|Active|Idl
BGP router identifier 128.223.51.103, local AS number 6447
BGP table version is 2813468887, main routing table version 2813468887
Path RPKI states: 7708169 valid, 10262796 not found, 14390 invalid
962240 network entries using 238635520 bytes of memory
<Output omitted...>
BGP using 3360540172 total bytes of memory
BGP activity 71481761/70340753 prefixes, 3122764440/3095983216 paths, scan
interval 60 secs
Neighbor V AS MsgRcvd MsgSent TblVer InQ OutQ Up/Down State/PfxRcd
4.68.4.46 4
                   3356 5325851 83510 2813468729 12
                                                        0 3w5d
                                                                     905347
12.0.1.63 4
                   7018 5117677 19250 2813468729 124
                                                        0 1w5d
                                                                     906235
<Output omitted...>
```



#### **BGP Stability Considerations**

- Events in networks often occur in bursts
- There is always a challenge how to react
  - Reacting fast improves convergence time but may introduce churn
  - Reacting with a delay improves stability but delays convergence
- BGP favors stability
  - It delays sending updates to smoothen out the churn and to collect possibly multiple changes for a single update
  - It only advertises changes (incremental updates)



#### Route Scale & Control Plane Stability



#### Network Layer Reachability Information Exchange

## Prefix/Length

```
R1(config) # router bgp 64512
R1(config-router) # address-family ?
  ipv4   Address family
  ipv6   Address family
  12vpn   Address family
  nsap   Address family
  vpnv4   Address family
  vpnv4   Address family
  vpnv6   Address family
```

```
route-views> show bgp all neighbors 4.68.4.46 | i family

For address family: IPv4 Unicast

Address family IPv4 Unicast: advertised and received

Address family: IPv4 Multicast: advertised and received

For address family: IPv4 Unicast

For address family: IPv4 Multicast

Address family IPv4 Unicast: advertised and received

Address family: IPv4 Multicast: advertised and received

For address family: L2VPN E-VPN

For address family: MVPNv4 Unicast
```

## Messages



#### BGP Message Types

- BGP runs on top of TCP transport layer protocol which is...
  - Byte stream-oriented
  - Unicast only
  - Connection-oriented and reliable
  - Providing flow and congestion control
- The format of BGP messages partly accommodates TCP specifics
  - Message markers
  - Length indications (lots of!)



## BGP Message Types

- BGPv4 uses (only) 5 message types
  - OPEN
  - UPDATE
  - NOTIFICATION
  - KEEPALIVE
  - ROUTE-REFRESH (not part of initial BGPv4 RFC 1654 specification, brought in through RFC 2918 and nearly universally supported)

#CiscoLive



#### BGP OPEN Message

- BGP speakers use OPEN to advertise their configuration and capabilities once their TCP session comes up
  - Version advertisement
  - Autonomous System Number advertisement
  - Hold Time advertisement/negotiation
  - BGP Router ID advertisement
  - Optional Capabilities advertisement/negotiation
- If the peer advertises an incompatible configuration, the peering is terminated, and the TCP session closed



#### BGP OPEN Message

```
Border Gateway Protocol - OPEN Message
 Length: 57
Type: OPEN Message (1)
Version: 4
 My AS: 64512
 Hold Time: 180
 BGP Identifier: 10.255.255.1
 Optional Parameters Length: 28
-Optional Parameters
 Optional Parameter: Capability
 Optional Parameter: Capability
 Optional Parameter: Capability
 Optional Parameter: Capability
 - Optional Parameter: Capability
   Parameter Type: Capability (2)
   Parameter Length: 6
  Capability: Support for 4-octet AS number capability
```

#CiscoLive





## Select BGP Capability Codes for OPEN Message

https://www.iana.org/assignments/capability-codes/capability-codes.xhtml#capability-codes-2

Value	Name	RFC
1	Multiprotocol Extensions for BGP-4	<u>2858</u>
2	Route Refresh Capability for BGP-4	<u>2918</u>
3	Outbound Route Filtering Capability	<u>5291</u>
5	Extended Next Hop Encoding	<u>8950</u>
6	BGP Extended Message	8654
7	BGPsec Capability	<u>8205</u>
8	Multiple Labels Capability	8277
9	BGP Role	9234
64	Graceful Restart Capability	<u>4724</u>
65	Support for 4-octet AS number capability	<u>6793</u>
69	ADD-PATH Capability	<u>7911</u>
70	Enhanced Route Refresh Capability	<u>7313</u>



#### **BGP NOTIFICATION Message**

- The NOTIFICATION message is sent out by a peer who detected an unrecoverable condition and needs to terminate the peering
- After sending out a NOTIFICATION, the sender closes the session
- The NOTIFICATION contents are useful for diagnostics

#### Border Gateway Protocol - NOTIFICATION Message

Length: 21

Type: NOTIFICATION Message (3)

Major error Code: Cease (6)

Minor error Code (Cease): Administratively Shutdown (2)





#### Select BGP Error Codes for NOTIFICATION

https://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-3

Value	Name	RFC
1	Message Header Error	<u>4271</u>
2	OPEN Message Error	<u>4271</u>
3	UPDATE Message Error	<u>4271</u>
4	Hold Timer Expired	<u>4271</u>
5	Finite State Machine Error	<u>4271</u>
6	Cease	<u>4271</u>
7	ROUTE-REFRESH Message Error	<u>7313</u>





#### Select BGP Error Subcodes

https://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-3

Message Header Error Subcodes		
Value	Name	RFC
0	Unspecific	EID 4493
1	Connection Not Synchronized	<u>4271</u>
2	Bad Message Length	<u>4271</u>
3	Bad Message Type	<u>4271</u>

OPEN Message Error Subcodes		
Value	Name	RFC
0	Unspecific	EID 4493
1	Unsupported Version Number	4271
2	Bad Peer AS	<u>4271</u>
3	Bad BGP Identifier	4271
4	Unsupported Optional Parameter	4271
6	Unacceptable Hold Time	4271
7	Unsupported Capability	4271
11	Role Mismatch	9234





#### Select BGP Error Subcodes

https://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-3

UPDATE Message Error Subcodes		
Value	Name	RFC
0	Unspecific	EID 4493
1	Malformed Attribute List	4271
2	Unrecognized Well-known Attribute	<u>4271</u>
3	Missing Well-known Attribute	<u>4271</u>
4	Attribute Flags Error	<u>4271</u>
5	Attribute Length Error	4271
6	Invalid ORIGIN Attribute	4271
8	Invalid NEXT_HOP Attribute	4271
9	Optional Attribute Error	4271
10	Invalid Network Field	4271
11	Malformed AS_PATH Attribute	<u>4271</u>

Finite State Machine Error Subcodes			
Value	Name	RFC	
0	Unspecified Error	6608	
1	Receive Unexpected Message in OpenSent State	6608	
2	Receive Unexpected Message in OpenCofirm State	6608	
3	Receive Unexpected Message in Established State	6608	



#### Select BGP Error Subcodes

https://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-3

Cease Subcodes		
Value	Name	RFC
1	Max Number of Prefixes Reached	4486
2	Administrative Shutdown	4486
3	Peer De-configured	4486
4	Administrative Reset	4486
5	Connection Rejected	4486
6	Other Configuration Change	4486
7	Connection Collision Resolution	4486
8	Out of Resources	4486
9	Hard Reset	<u>8538</u>
10	BFD Down	9384

ROUTE-REFRESH Message Error Subcodes		
Value	Name	RFC
0	Reserved	7313
1	Invalid Message Length	<u>7313</u>



BRKENT-1179

#### BGP KEEPALIVE Message

- Instead of relying on TCP keepalives, BGP uses the KEEPALIVE message to periodically announce a speaker's liveliness
- KEEPALIVE is sent...
  - Immediately after receiving an agreeable OPEN message from peer
  - Periodically, with the period being one third of Hold Time by default

```
Border Gateway Protocol - KEEPALIVE Message
```

Length: 19

Type: KEEPALIVE Message (4)



#### BGP ROUTE-REFRESH Message

- Original BGP specification did not specify means to ask a peer to resend prefixes to us
  - This is necessary when the inbound route policy changes
  - Vendors worked around this deficiency by storing aside a separate unfiltered copy of all routes from the peer ("Soft Reconfiguration")
- RFC 2918 brought the ROUTE-REFRESH message allowing to ask a peer to resend all routes of any particular address family



#### BGP UPDATE Message

- The UPDATE message is the workhorse of BGP
  - Advertises reachable NLRIs along with their attributes
  - Withdraws unreachable NLRIs
- The format of the UPDATE message targets maximum efficiency
  - The path attributes are included only once, followed by the list of all NLRIs that share them
  - Every NLRI contains only the network prefix (and padding bits to a whole octet if needed)



#### BGP Update Message - New/Updated Routes

```
Border Gateway Protocol - UPDATE Message
 Length: 67
 Type: UPDATE Message (2)
 Withdrawn Routes Length: 0
 Total Path Attribute Length: 28
 Path attributes
 Path Attribute - ORIGIN: IGP
 Path Attribute - AS_PATH: empty
 Path Attribute - NEXT_HOP: 10.255.255.1
 Path Attribute - MULTI EXIT DISC: 1234
 Path Attribute - LOCAL_PREF: 100
 Network Layer Reachability Information (NLRI)
 · 192.168.0.0/24
 \rightarrow 192.168.1.0/24
 · 192.168.2.0/24
 192.168.3.0/24
```

#CiscoLive



© 2023 Cisco and/or its affiliates. All rights reserved. Cisco Public

## BGP Update Message - Withdrawn Routes



## States





#### Active, Idle, Connect

```
1174 17:39:50.701103 10.1.1.2
                                                                                                                          60 52565 → 179 [SYN] Seq=0 Win=16384
                                                                   10.1.1.1
 1175 17:39:50.724304 10.1.1.1
                                                                                                                          60 179 → 52565 [SYN, ACK] Seq=0 Ack=
                                                                   10.1.1.2
                                                                                                               TCP
                                                                                                                          60 52565 → 179 [ACK] Seg=1 Ack=1 Win=16384
 1176 17:39:50.761402 10.1.1.2
                                                                                                               TCP
                                                                   10.1.1.1
Frame 1174: 60 bytes on wire (480 bits), 60 bytes captured (480 bits) on interface -, id 0
Ethernet II, Src: ca:02:9f:44:00:00 (ca:02:9f:44:00:00), Dst: ca:01:5f:3a:00:00 (ca:01:5f:3a:00:00)
Internet Protocol Version 4, Src: 10.1.1.2, Dst: 10.1.1.1
Transmission Control Protocol, Src Port: 52565, Dst Port: 179, Seq: 0, Len: 0
  Source Port: 52565
  Destination Port: 179
   [Stream index: 4]
  [Conversation completeness: Incomplete, DATA (15)]
  [TCP Segment Len: 0]
  Sequence Number: 0
                        (relative sequence number)
  Sequence Number (raw): 3742606240
                              (relative sequence number)]
  [Next Sequence Number: 1
  Acknowledgment Number: 0
  Acknowledgment number (raw): 0
  0110 .... = Header Length: 24 bytes (6)
> Flags: 0x002 (SYN)
  Window: 16384
  [Calculated window size: 16384]
  Checksum: 0xf164 [unverified]
  [Checksum Status: Unverified]
  Urgent Pointer: 0
> Options: (4 bytes), Maximum segment size
> [Timestamps]
```



#### Open Sent & Open Confirmed

```
1177 17:39:50.792164 10.1.1.2
                                                                 10.1.1.1
                                                                                                             BGP
                                                                                                                      112 OPEN Message
 1178 17:39:50.815234 10.1.1.1
                                                                 10.1.1.2
                                                                                                             BGP
                                                                                                                      112 OPEN Message
                                                                 10.1.1.2
                                                                                                                       73 KEEPALIVE Message
 1179 17:39:50.815297 10.1.1.1
                                                                                                             BGP
 1180 17:39:50.873147 10.1.1.2
                                                                 10.1.1.1
                                                                                                             BGP
                                                                                                                       73 KEEPALIVE Message
Frame 1177: 112 bytes on wire (896 bits), 112 bytes captured (896 bits) on interface -, id 0
Ethernet II, Src: ca:02:9f:44:00:00 (ca:02:9f:44:00:00), Dst: ca:01:5f:3a:00:00 (ca:01:5f:3a:00:00)
Internet Protocol Version 4, Src: 10.1.1.2, Dst: 10.1.1.1
Transmission Control Protocol, Src Port: 52565, Dst Port: 179, Seq: 1, Ack: 1, Len: 58
Border Gateway Protocol - OPEN Message
  Length: 58
  Type: OPEN Message (1)
  Version: 4
  My AS: 65500
  Hold Time: 180
  BGP Identifier: 10.1.1.2
  Optional Parameters Length: 29
v Optional Parameters

∨ Optional Parameter: Capability

       Parameter Type: Capability (2)
       Parameter Length: 6
     > Capability: Multiprotocol extensions capability
  v Optional Parameter: Capability
       Parameter Type: Capability (2)
       Parameter Length: 2
     > Capability: Route refresh capability (Cisco)

∨ Optional Parameter: Capability

       Parameter Type: Capability (2)
       Parameter Length: 2
     > Capability: Route refresh capability
```



#### Established

```
R1# show bgp ipv4 unicast summary
BGP router identifier 10.1.1.1, local AS number 64512
BGP table version is 1, main routing table version 1

Neighbor V AS MsgRcvd MsgSent TblVer InQ OutQ Up/Down State/PfxRcd
10.1.1.2 4 65500 15 15 1 0 0 00:10:49 0
```







#### **BGP** Attributes

- An attribute is an additional piece of information accompanying an advertised NLRI
- BGP uses attributes in multiple ways
  - Prevents routing loops
  - Performs best path selection
  - Filters or sorts routes
  - ... and many more
- Basic BGP specification recognizes only a handful of attributes
  - Several new have been added over time for various applications and uses



#### **BGP Attribute Types**

- Well-known: Every BGP implementation must support it
  - Well-known mandatory: Must always be included with a NLRI
  - Well-known discretionary: May be included with a NLRI as needed

- Optional: BGP implementations do not need to support it
  - Optional transitive: When advertising a learned NLRI, keep the attribute with the NLRI even if not recognized
  - Optional non-transitive: When advertising a learned NLRI, remove the attribute from the NLRI if not recognized

Note: All well-known attributes are transitive



#### **BGP** Attributes

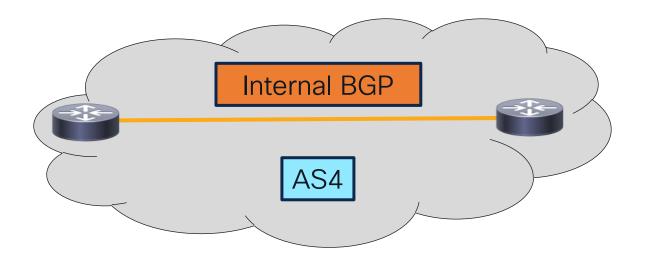
- Well-known mandatory:
  - · AS\_PATH
  - NEXT\_HOP
  - ORIGIN
- Well-known discretionary
  - · LOCAL\_PREF
  - ATOMIC\_AGGREGATE

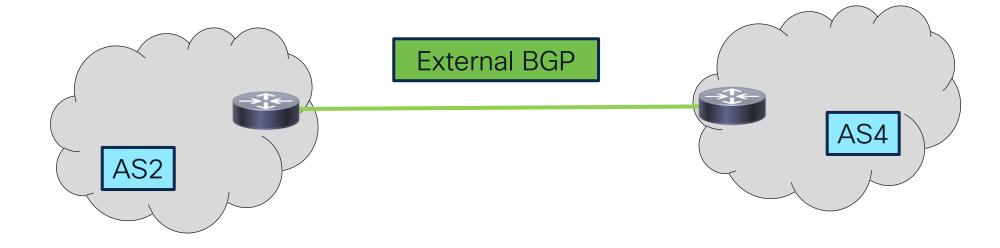
- Optional transitive
  - AGGREGATOR
  - · COMMUNITIES
  - EXTENDED\_COMMUNITIES
- Optional non-transitive
  - MULTI\_EXIT\_DISC
  - · CLUSTER\_LIST



#### Internal vs External BGP

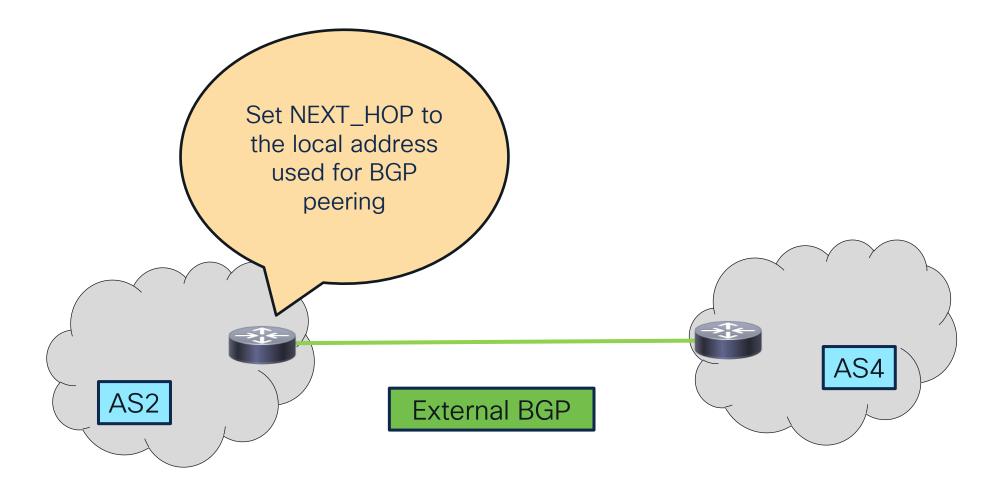






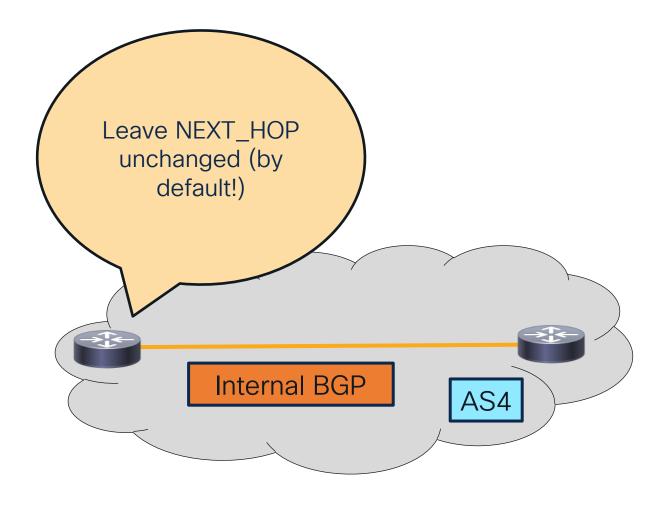


#### NEXT\_HOP in eBGP



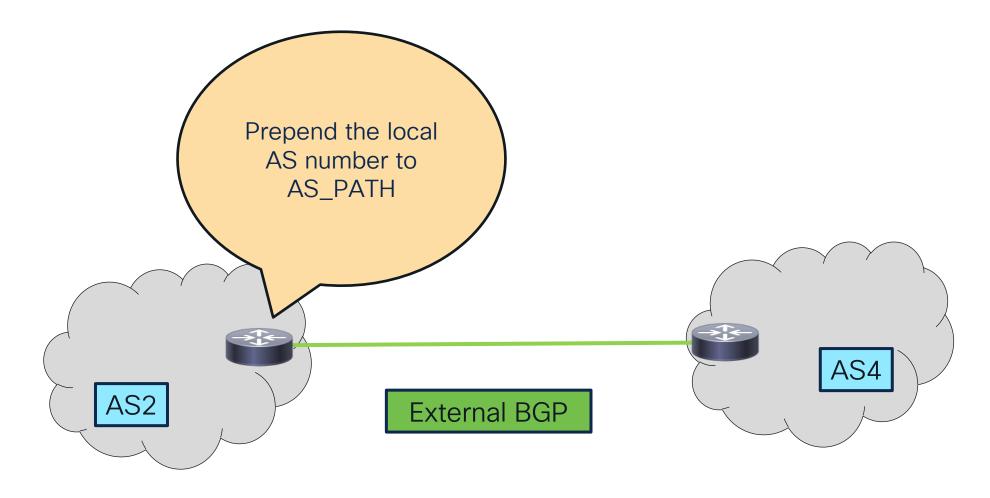


#### NEXT\_HOP in iBGP





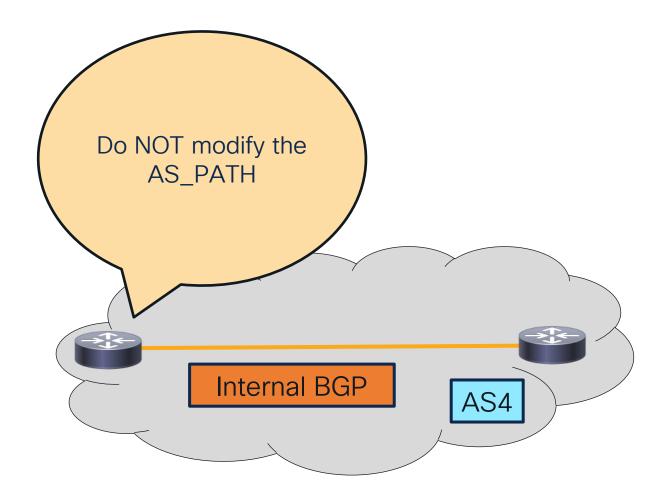
#### AS\_PATH in eBGP



#CiscoLive

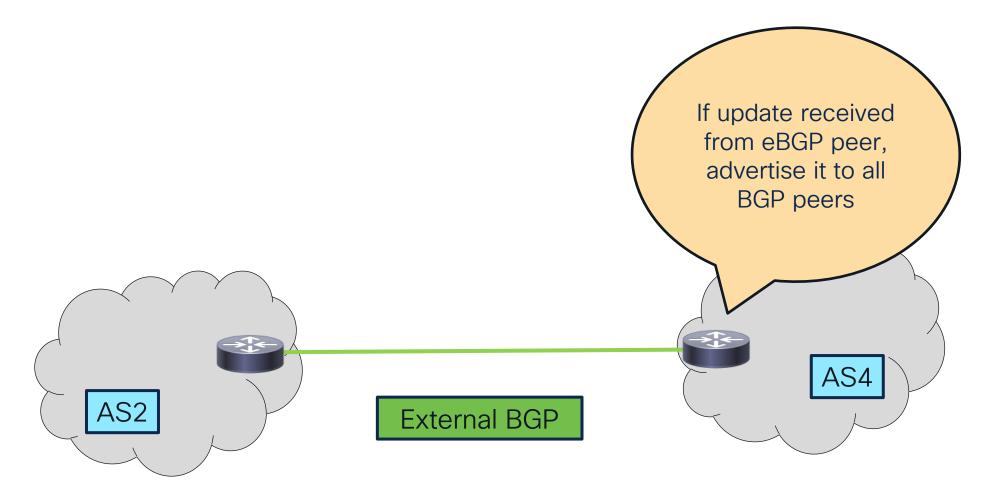


#### AS\_PATH in iBGP



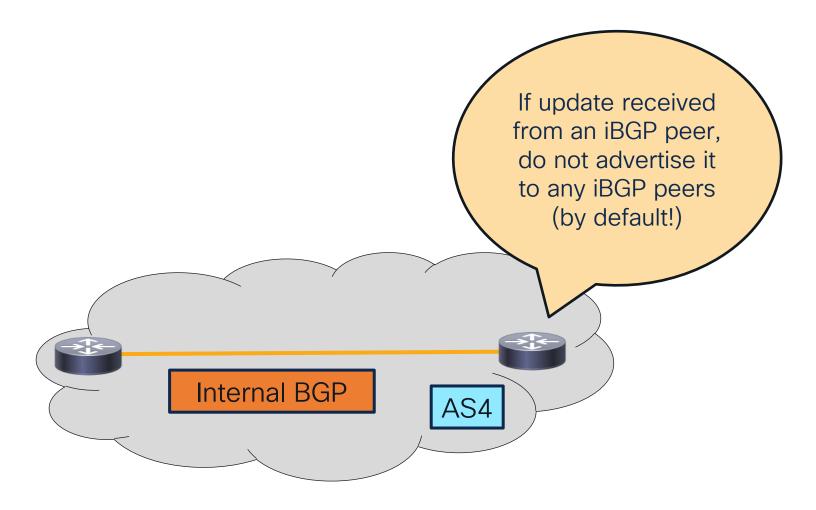


#### Updates in eBGP





#### Updates in iBGP







#### eBGP/iBGP Attribute Handling Summary

	Internal	External
Scope	Within AS	Between AS's
NEXT_HOP	Unchanged (default)	Changed to local address used to establish eBGP session
AS_PATH	Unchanged	Local AS inserted at the beginning of the AS_PATH list
Update	From iBGP not sent to iBGP (default)	From any sent to eBGP From eBGP sent to iBGP



#### Best Path Selection Algorithm



#### Best Path Selection in BGP

- BGP observes multiple constraints when selecting the best path
  - Loop-free
  - Shortest in terms of the number of ASes to traverse
  - Most likely to hit the true destination
  - Leaving the local autonomous system as quickly as possible (if needed)
  - ... but not updated unnecessarily if changed only insignificantly
  - ... and always with the possibility for the admin to override the selection
- BGP rules for best path selection aim to select exactly one best path per NLRI – to be both used locally and advertised to peers



#### 10,000 Miles View at BGP Best Path Selection

- BGP speaker needs to decide on the best path for an NLRI if it learns about multiple options (variants) on how to reach it
- For every learned NLRI:
  - Select the first variant in the database as the best one
  - If there are more variants, go through them one by one, and compare the current best one to the entry at hand, and pick the new best one
    - · The new best one may be the current one (no change), or the entry at hand
  - Finish when all variants have been inspected



#### Before Getting to the Nitty Gritty...

- Before diving into the BGP best path algorithm steps, keep in mind:
  - For every NLRI, BGP always compares the current best path to a variant in its database (injected locally or learned from peers)
  - The best path is both used locally (if it makes sense) and advertised to other peers, subject to the advertisement rules and policies
- In upcoming slides, the diagrams demonstrate typical scenarios but are not exhaustive



Router performing the best path selection



Generic router (unspecific)



The resulting next hop

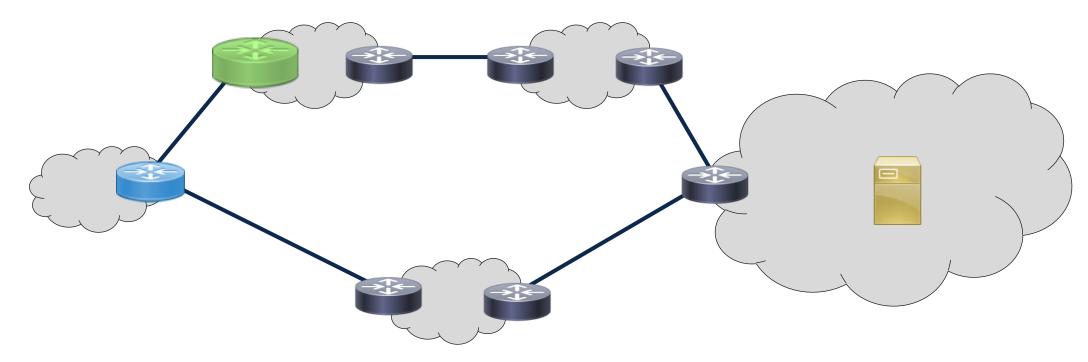


Destination (NLRI)



#### The logic behind BGP best path selection (1)

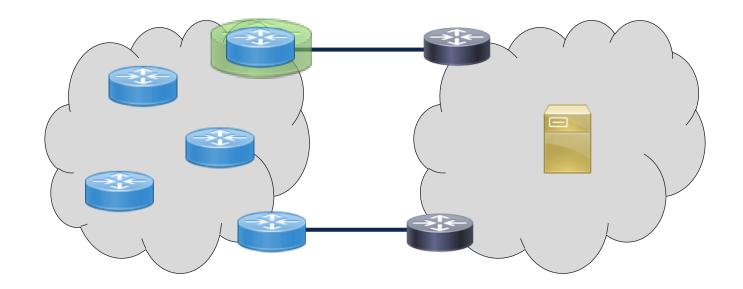
- Step 1: Prefer the path with the higher WEIGHT
  - Rationale: Always have means to override the path choice locally
  - Note: This is an override rule





#### The logic behind BGP best path selection (2)

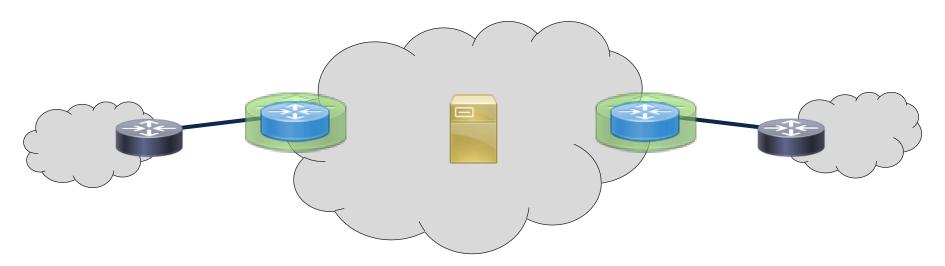
- Step 2: Prefer the path with the higher LOCAL\_PREF
  - Rationale: Have means to override the best path for the entire AS from a single exit point
  - Note: This is an override rule





#### The logic behind BGP best path selection (3)

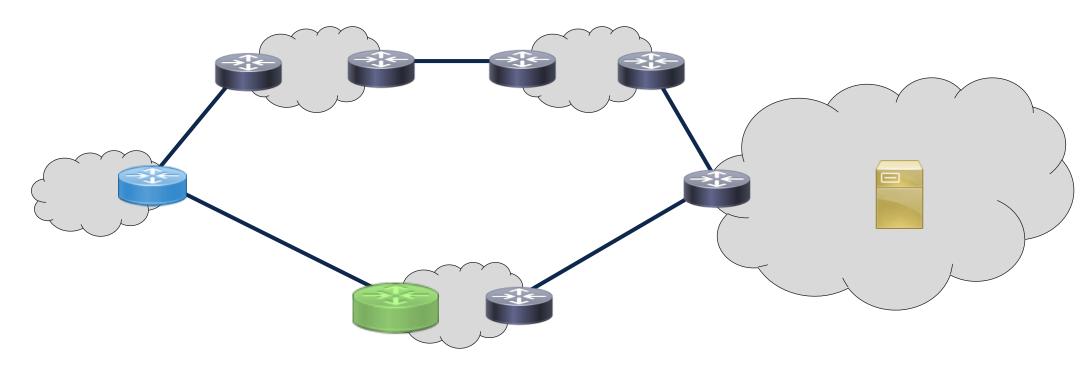
- Step 3: Prefer the locally originated path (network, redistribution, aggregation)
  - Rationale: I get a chance to speak on behalf of my own local AS
  - Note: The best route is not just for me to use but also to advertise to others so that they know





#### The logic behind BGP best path selection (4)

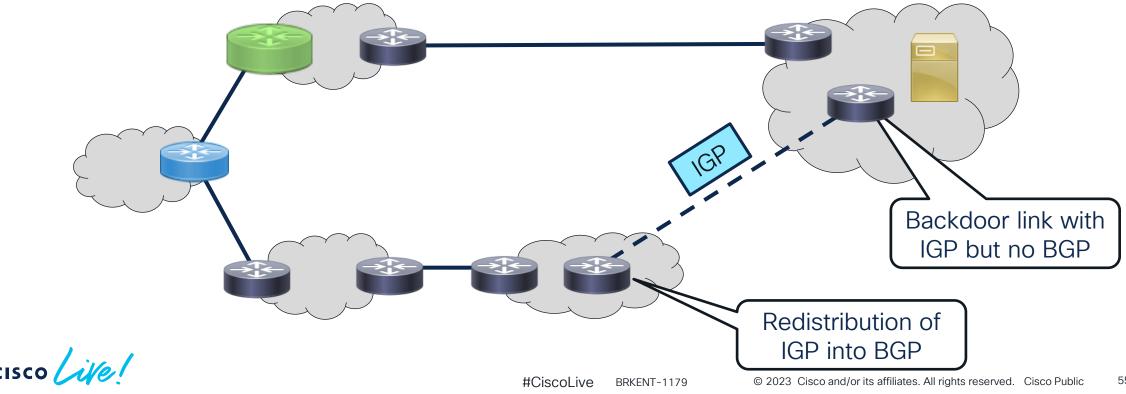
- Step 4: Prefer the path with the shortest AS\_PATH / AS4\_PATH
  - Rationale: Traverse the least amount of autonomous systems





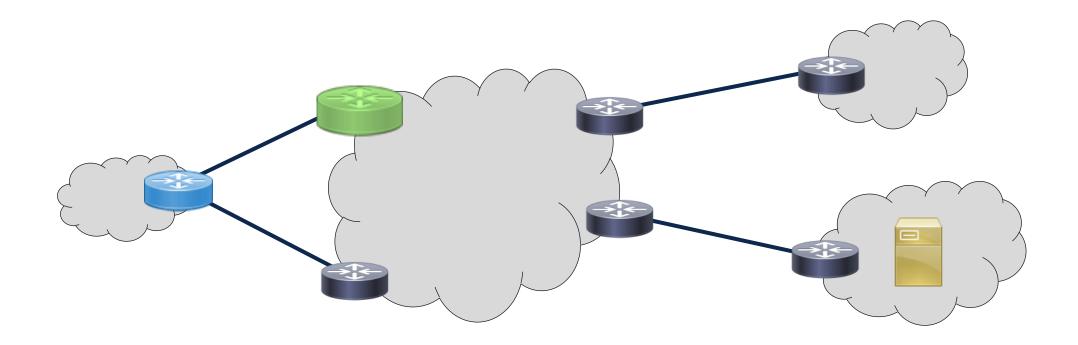
#### The logic behind BGP best path selection (5)

- Step 5: Prefer the path with the lower ORIGIN code
  - Rationale: Take the most trustworthy path
  - Note: IGP is lower than EGP, EGP is lower than Incomplete



#### The logic behind BGP best path selection (6)

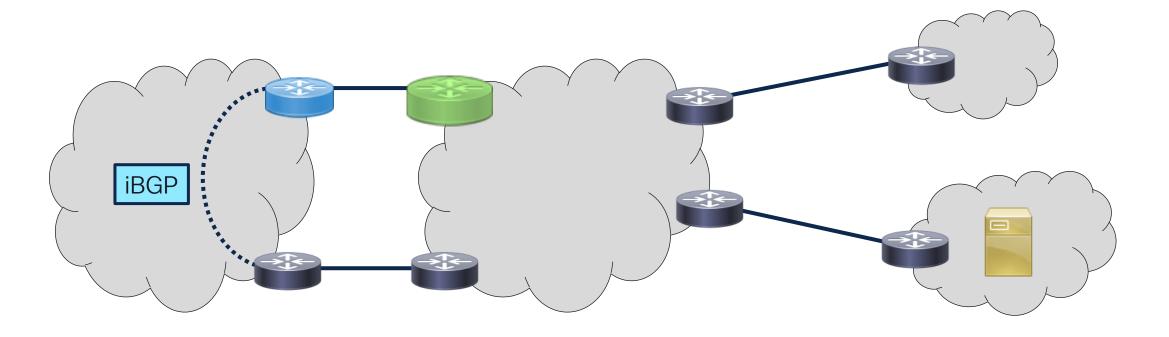
- Step 6: Prefer the path with the lower MULTI\_EXIT\_DISC
  - Rationale: Respect the preferred path hint indicated by the neighbor AS





#### The logic behind BGP best path selection (7)

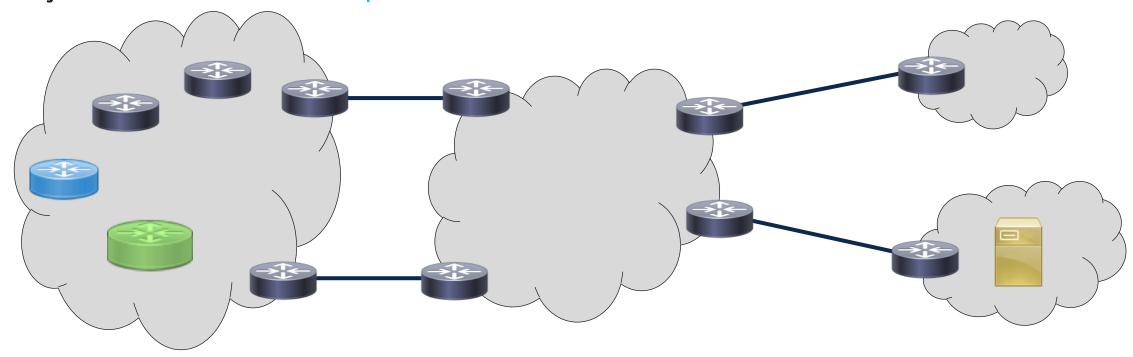
- Step 7: Prefer eBGP-learned path over iBGP-learned one
  - Rationale: If you need to leave the local AS, leave right away if you can





#### The logic behind BGP best path selection (8)

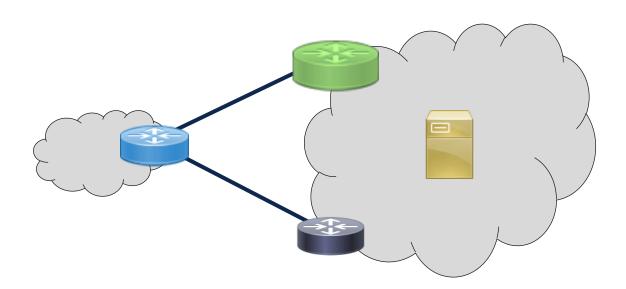
- Step 8: Prefer the path with the lower IGP metric to the next hop
  - Rationale: If you need to traverse the local AS but can't leave right away, just take the shortest path toward the exit





#### The logic behind BGP best path selection (9)

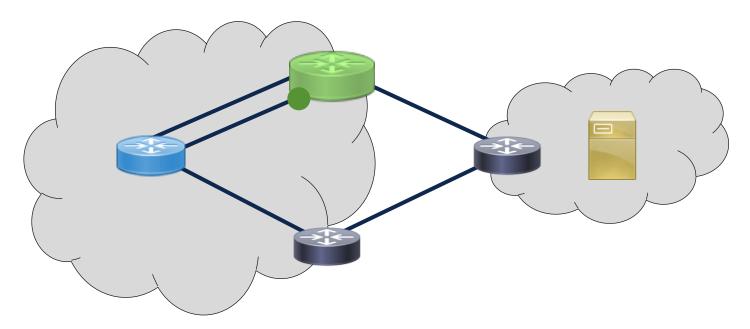
- Step 9: If both paths are learned via eBGP, prefer the older one
  - Rationale: The eBGP paths are, by this point, effectively equal so don't bother updating anything





#### The logic behind BGP best path selection (10-12)

- Steps 10-12: Prefer the path learned from the BGP peer with the lower Router ID, then with the shorter CLUSTER\_LIST, then from the BGP peer with the lower peering IP address
  - Rationale: Technical tiebreakers to arrive at exactly one best path







#### BGP Best Path Selection Algorithm

https://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html

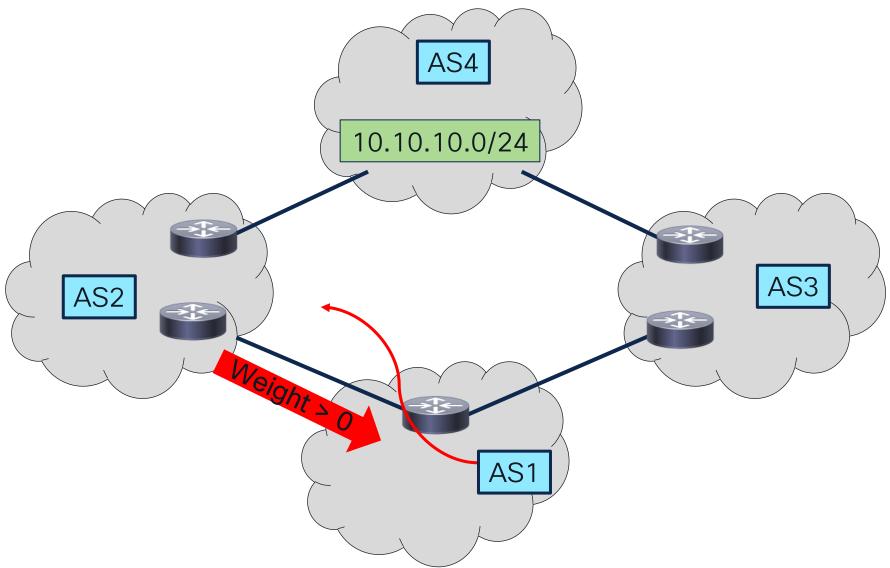
BGP Best Path Selection Algorithm as Implemented on Cisco Devices		
1. Highest WEIGHT	7. eBGP-learned route over iBGP-learned one	
2. Highest LOCAL_PREF	8. Lowest IGP metric to the next hop	
3. Locally originated (injected) route	9. Oldest eBGP-learned route	
4. Shortest AS_PATH / AS4_PATH	10. Lowest BGP peer's Router ID	
5. Lowest ORIGIN code	11. Shortest CLUSTER_LIST	
6. Lowest MULTI_EXIT_DISC	12. Lowest BGP peer's address	



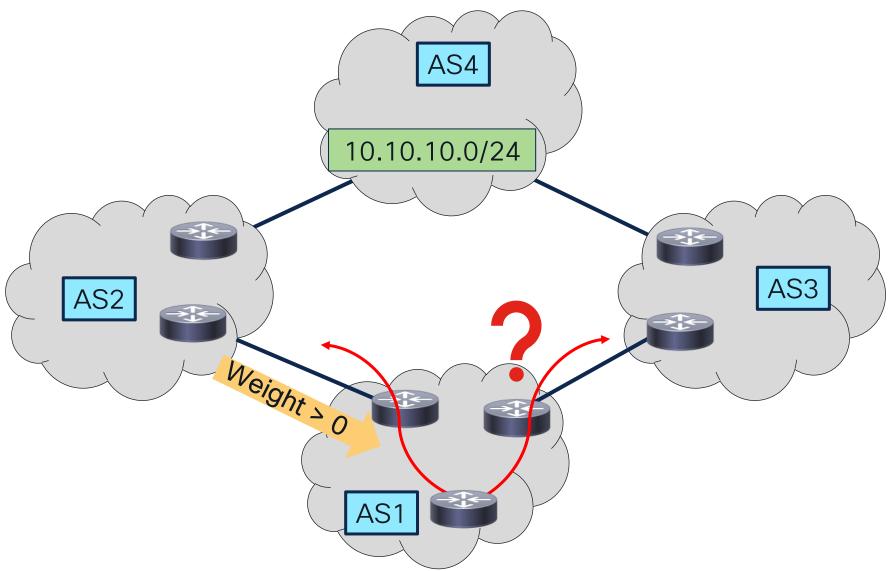
### Policy Control and Traffic Engineering



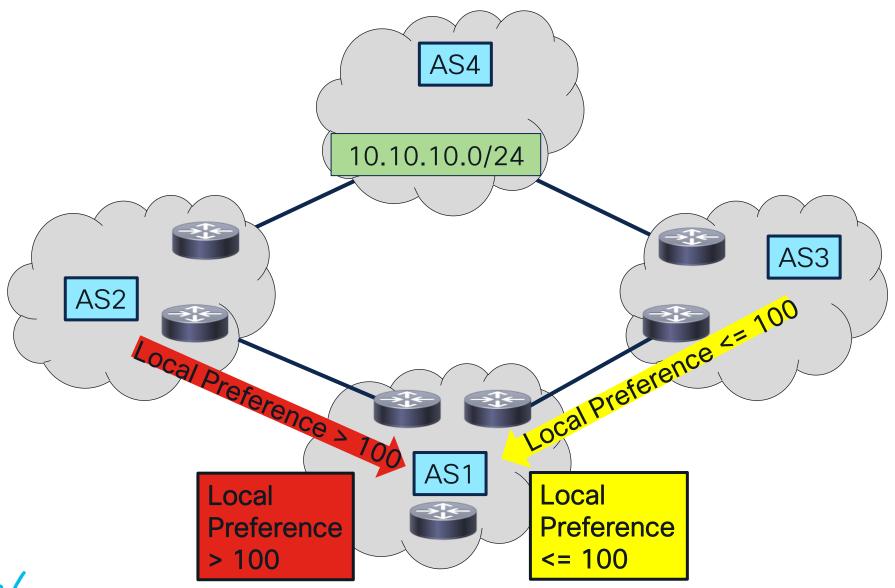
#### Outbound Traffic - Weight



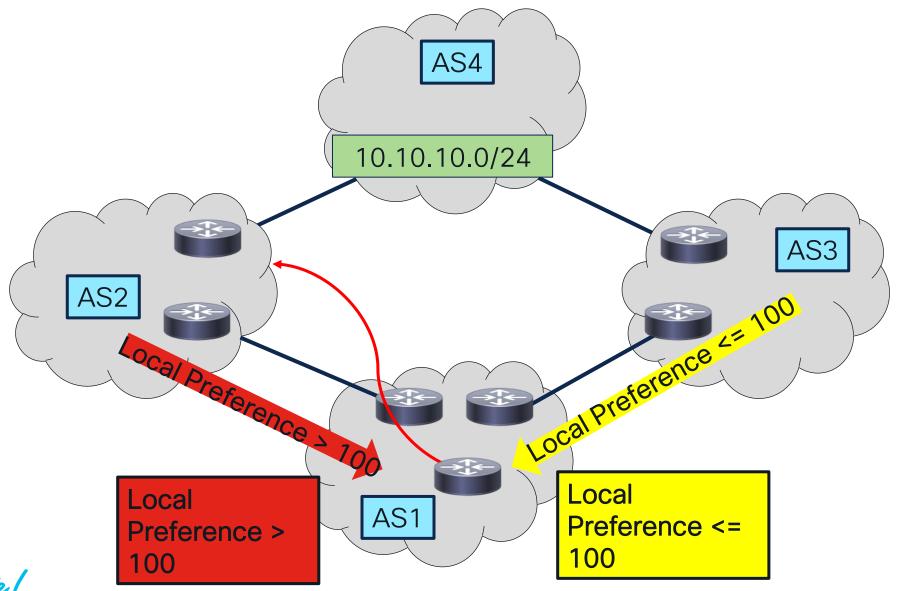
#### Outbound Traffic - Weight



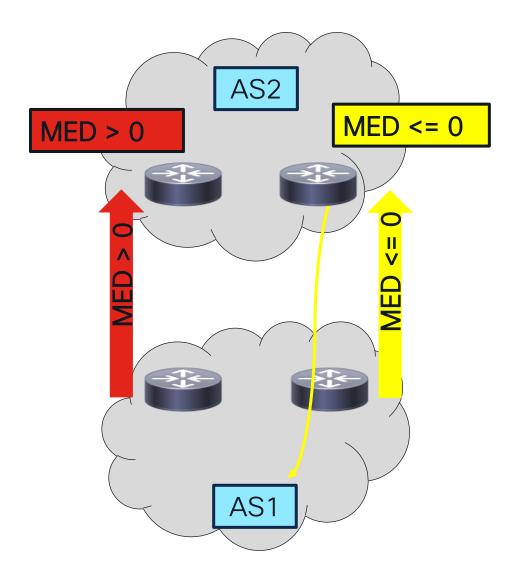
#### Outbound Traffic - LOCAL\_PREF



#### Outbound Traffic - LOCAL\_PREF

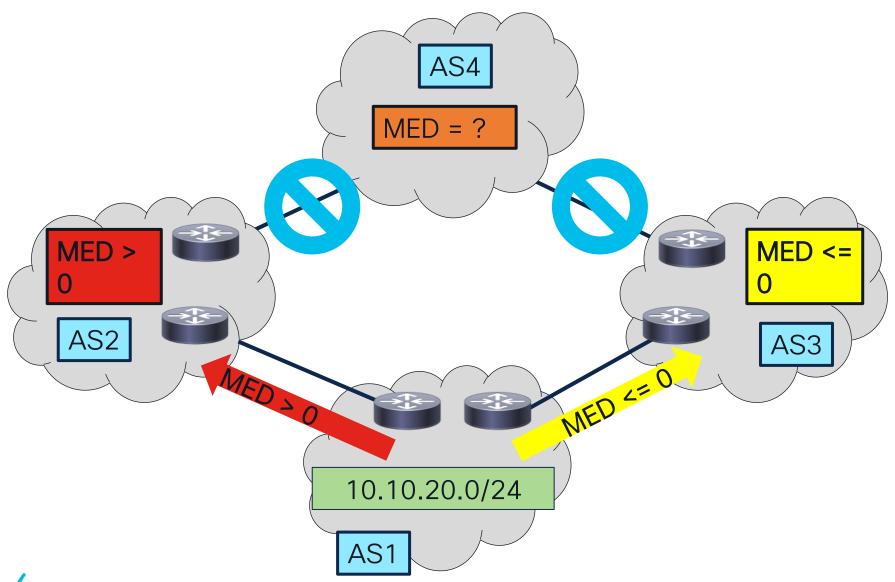


#### Inbound Traffic - MULTI\_EXIT\_DISC



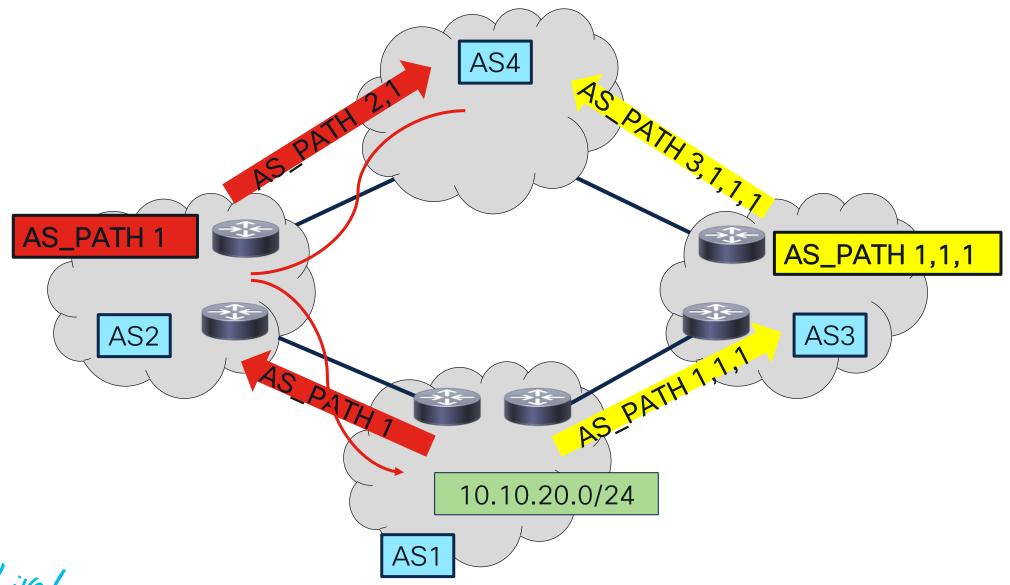


#### Inbound Traffic - MULTI\_EXIT\_DISC





#### Inbound Traffic - AS\_PATH



#### Concluding Remarks

- BGPv4 is ~30 years old but its core is still the same
  - A credit to its well-thought design
- BGP is a world on its own where to learn more?
  - Cisco Press textbooks
  - Cisco Communities, Cisco Learning Network
  - IETF RFCs
  - Wireshark
  - Hands-on experience is key



#Ciscol ive

#### "Like a fine wine, BGP only gets better with age."

Anonymous



#CiscoLive

#### Fill out your session surveys!



Attendees who fill out a minimum of four session surveys and the overall event survey will get Cisco Live-branded socks (while supplies last)!



Attendees will also earn 100 points in the Cisco Live Game for every survey completed.



These points help you get on the leaderboard and increase your chances of winning daily and grand prizes

#Ciscol ive



# Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand



#### Thank you



#### Cisco Live Challenge

Gamify your Cisco Live experience! Get points for attending this session!

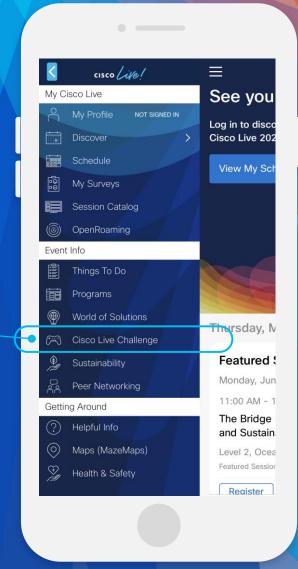
#### How:

- Open the Cisco Events App.
- Click on 'Cisco Live Challenge' in the side menu.
- Click on View Your Badges at the top.
- Click the + at the bottom of the screen and scan the QR code:









## Let's go cisco live! #CiscoLive