# TCOM 515

# Lecture 5: BGP

# Objectives

- BGP Operation
- BGP Message Types
- BGP Attributes
- iBGP vs eBGP
- BGP Best Path Selection
- BGP Path Selection Examples

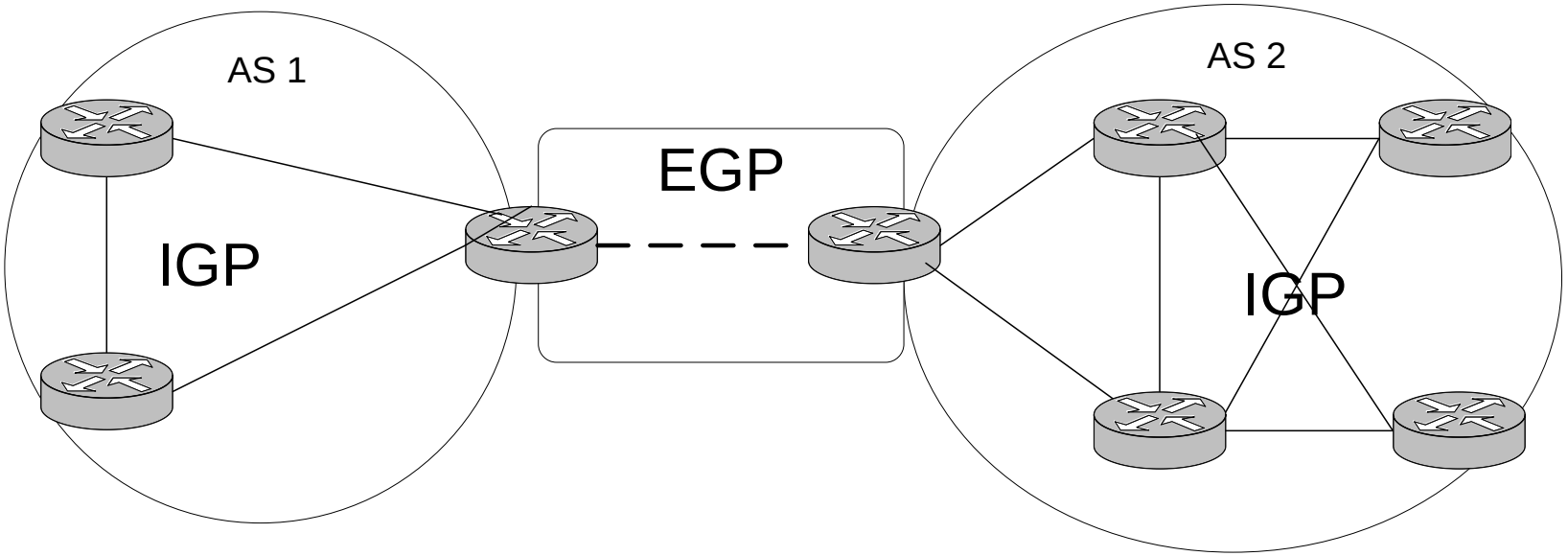
# BGP Basics

- BGP - Border Gateway Protocol
- BGP is an inter-AS routing protocol
- It is primarily used as an External Gateway Protocol (EGP).
- BGP is the EGP of choice for the Internet.
- Path Vector Protocol
  - BGP carries routing information that includes a next hop and a set of AS numbers, describing a “AS path” that the route has traversed,
- Uses TCP as transport protocol. TCP port number is 179
- BGP version 4 was introduced to allow for classless routing.
- Current RFCs - 4271, 1772, 2858, 2918

# Autonomous System

- Set of routers having a single routing policy under a single set of technical administrators.
- Single IGP domain or collection of IGP Domains
- Necessary when connecting to another group of routers
- Autonomous System number range: 1 - 65535
- Assigned by International Internet Registry
- 64512 - 65534 designated for private use
- Non-Transit AS
    - Only traffic originating or destined for the Local AS goes through it
- Transit AS
    - Traffic originating from another AS and destined for yet a third AS may also transit through the Local AS, additionally.

# IGP and EGP




**IGP: OSPF, EIGRP, ISIS, RIP**
**EGP: BGP**

# BGP Operation

- BGP involves two routers forming a BGP session after establishing neighbor relationship (peering). The routers will then pass routing information, each router send 31updates about the state of their BGP routes
- BGP updates include prefix and attributes
  - A BGP Prefix is a subnet
    - 10.10.0.0/16
    - 192.168.1.0/26
  - Attributes are used to describe the prefix
- BGP prefixes can be advertised
  - 1. Local AS (network command)
  - 2. Route aggregation
  - 3. Forwarded from other AS

# BGP Neighbors

- BGP neighbors are known as BGP Peers
- A Peer is defined by both the Neighbor AS and IP address
- ISPs peer with each other to exchange routes on the Internet.  Peerings can be private or public.
- Two types of peers
  - Internal Peer - a neighbor within your own AS
  - External Peer - a neighbor outside your own AS
- Both neighbors must identify the other as a Neighbor to establish a peering
- Once neighbors are configured on both sides, neighbor establishment begins

# BGP Neighbor States

- Idle-before the neighbors attempt TCP connection.
- Connect-one of the neighbor attempting to establish TCP connection
- Active-Attempt for TCP connection timed out, will periodically try to re-connect
- OpenSent- Identication packet sent to neighbor
- OpenConfirm- OpenSent message is received from neighbor, neighbor will decide to accept/refuse BGP session.
- Established-BGP session fully active, peers start exchanging update packets with peer

# BGP Message Types

- Open
- Notification
- Update
- Keepalive

# BGP Open Message

- The Open message is the first sent.  It is used for identification and agreement of protocol parameters.
- Open Message fields:
  - Version - BGP Version number of sender
  - ASN - Autonomous System Number - the AS of the sending router, compared to BGP neighbor configured AS
  - Hold Time - number of seconds that the sender proposes to use as a hold timer, max time it will wait for keepalive from neighbor- once exceeded, neighbor is marked as dead, negotiated to lower of the two neighbors
  - BGP Identifier - value chosen by sender to identify the BGP speaker.

# BGP Notification Message

- The Notification message is used to identify an error in the underlying TCP connection before it closes the connection.
- Used by Network Administrators to troubleshoot
- Notification Message fields:
  - Error Code - identifies the type of error that occurred
    1. Message Header Error
    2. Open Message Error (Bad Peer AS or Identifier)
    3. Update Message Error (Invalid Next_Hop, Invalid Origin)
    4. Hold timer Error
    5. Finite State Machine Error
    6. Cease - no other code applies
  - Error Subcode - narrows down more specific the type of error, applicable to Codes 1-3
  - Data - only present for specific error code and error subcode combinations

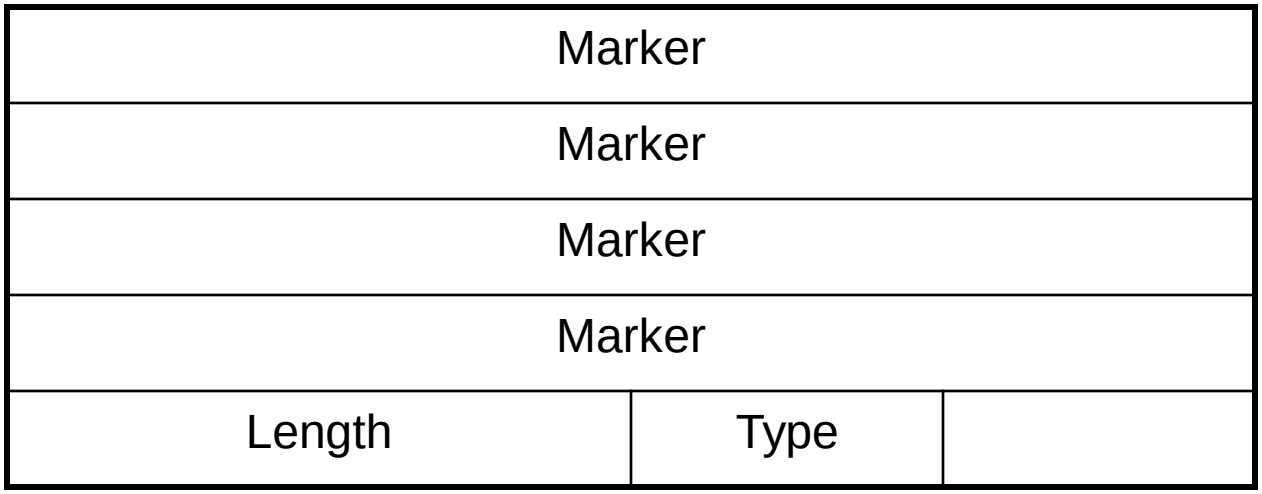
# BGP Update Message

- The Update message is used for most communications between two BGP peers. Used to advertise or withdraw a prefix
- Update Message fields:
  - **Network Layer Reachability Information** - NLRI - list of prefixes advertised and associated with the Path Attributes, all prefixes are described by all path attributes
  - **Path Attributes** - list of BGP attributes that describe the prefixes in the next field - Attribute type, length and value, for each attribute.
  - **Withdrawn Routes** - list of IP prefixes for which the sender no longer wishes to forward packets

# BGP Keepalive Message

- The Keepalive message is used by BGP Neighbors to maintain that the connection between them is active.
- The hold timer is negotiated at the beginning of the session and used to determine the maximum amount of time between keepalives before a neighbor is considered dead.
- Recommended keepalive interval is 1/3 the hold timer
- Either a Keepalive message or an Update message will reset the hold timer.
- A Keepalive message has only the BGP header with no other data contained within it.

# BGP Header




- Marker - used for synchronization and security, based on message type and options
- Length - length of entire BGP message including header
- Message Type - 1 - Open, 2 - Update, 3 - Notification, 4 - Keepalive

# BGP RIBs

- **RIB** - Routing Information Base - BGP-4's term for the routing table
- There are a few types of RIBs:
  - **Adj-RIB-In** - the location where prefixes from specific neighbors are stored. Each peer has an Adj-RIB-In.
  - **Loc-RIB** - all the prefixes in the different Adj-RIB-In are processed. The chose paths for each individual prefix is stored in the Loc-RIB. Each system has one Loc-RIB.
  - **Adj-RIB-Out** - the location where the prefixes to be advertised to a specific peer are stored. Each peer has its own Adj-RIB-Out.

# BGP Attributes

- BGP uses Path Attributes in the Update packets to give information about the prefixes advertised.
- Attributes belong to the following categories,
  - Well-known mandatory-attribute must be supported and included in update
  - Well-known discretionary-attribute is recognized by all BGP implementations, but does not have to be included in updates.
  - Optional transitive-attribute not required to be supported, but will be passed to other BGP speakers.
  - Optional non-transitive-attribute not required to be supported, but will not be passed to other BGP speakers.

- Some important attributes are:
  - Origin
  - AS-Path
  - Next-Hop
  - Multi-Exit Discriminator
  - Local-Pref
  - Atomic-Aggregate
  - Aggregator
  - Communities

We will discuss these attributes, but there are more that won't be covered. Please see the reading for more information and the RFCs.

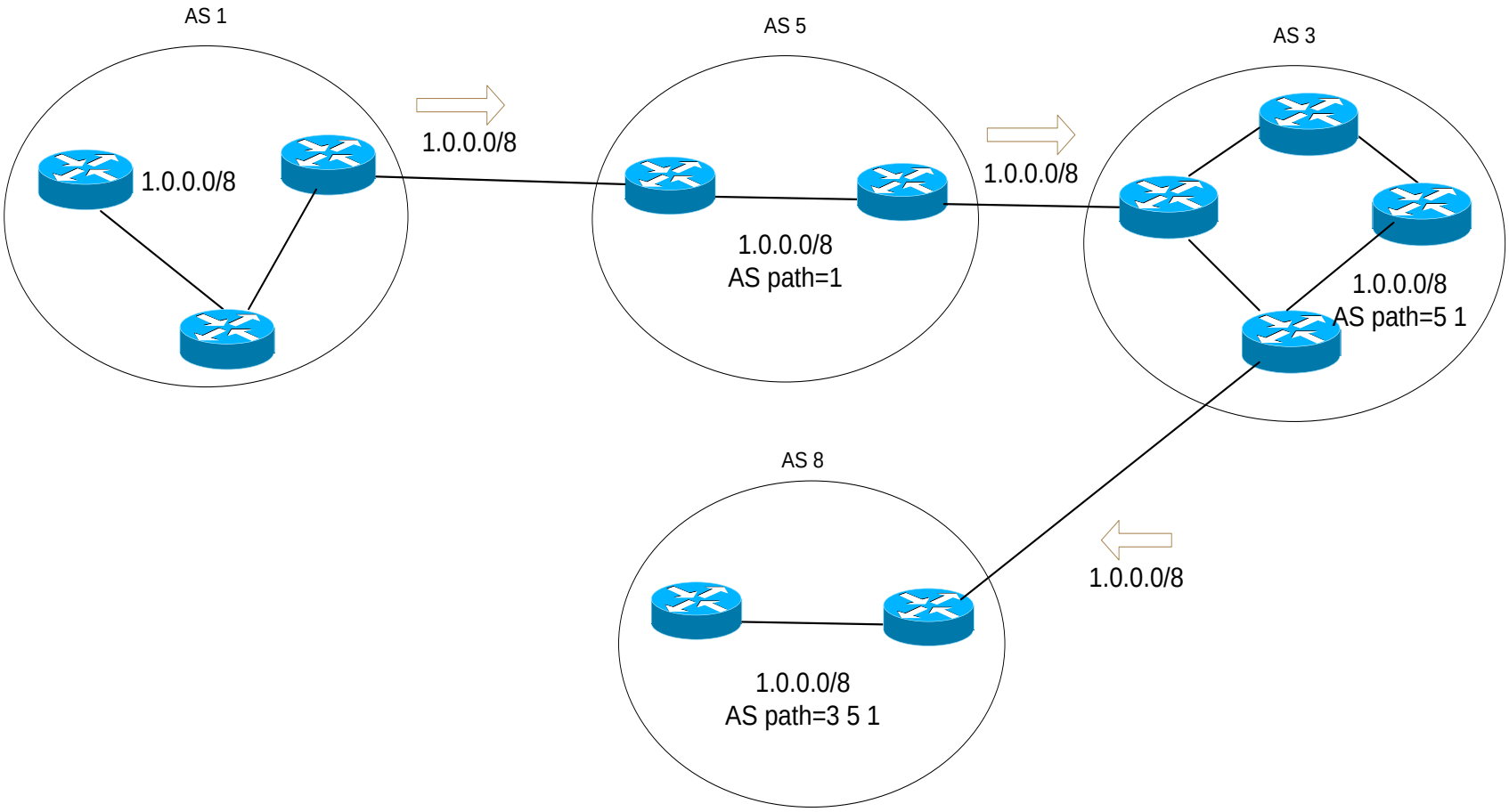
# BGP Origin

- The BGP Origin Path Attribute describes how the advertised prefix came into BGP table at the originating AS.
- The Origin attribute is a well-known mandatory path attribute.
- The prefix can come from directly attached interfaces, static routes, or other routing protocols.
- The possible values of the attribute are:
  - 1 - IGP - prefix learned from an IGP
  - 2 - EGP - prefix learned from EGP
  - 3 - Incomplete - learned through method other than IGP or EGP, most often redistribution.

# BGP AS-Path

- The BGP AS-Path attribute contains the Autonomous System numbers for each AS that the announcement for the prefix passed through.
- AS-Path attribute is a well-known mandatory attribute.
- The first AS in this attribute is the originating AS, with each subsequent AS appending their numbers as it leaves the AS.
- Used to detect and prevent routing loops.

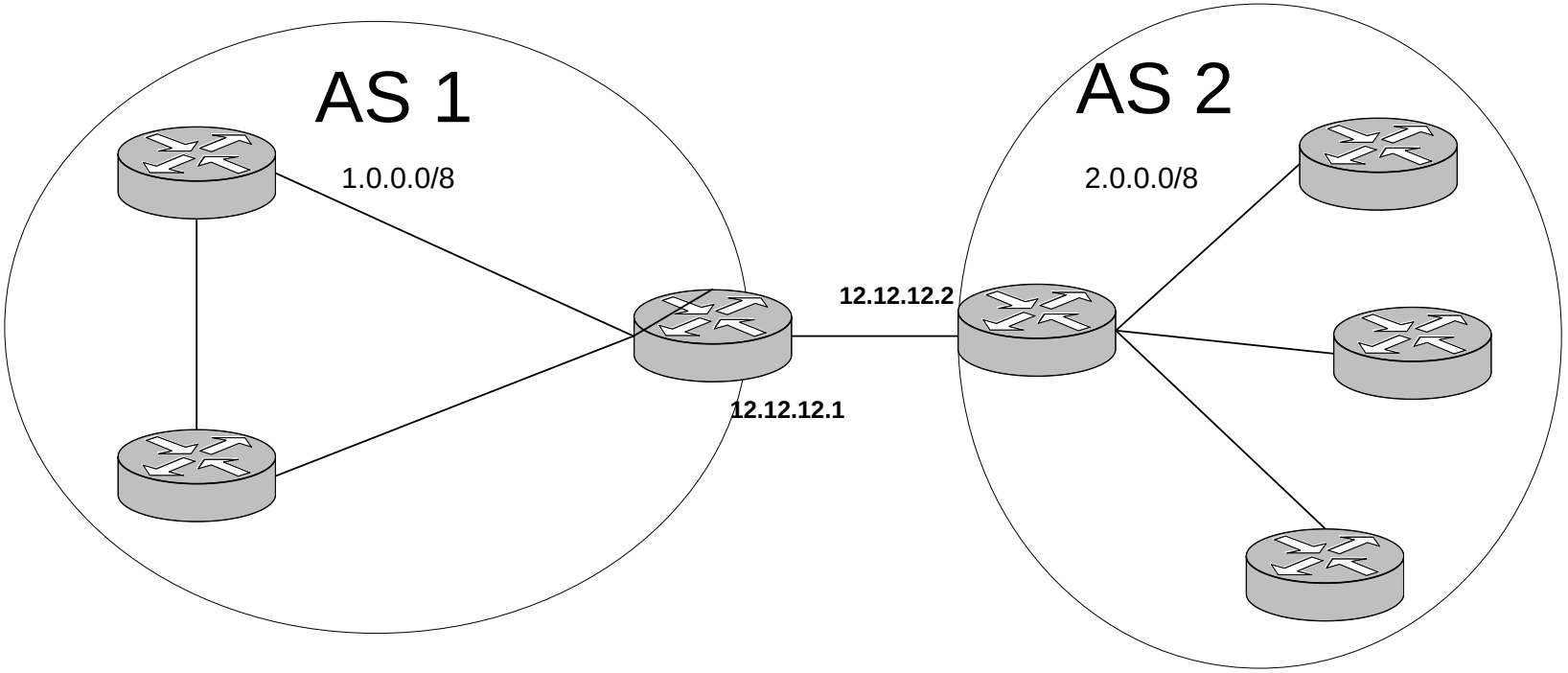
# BGP AS Path


AS 1
1.0.0.0/8
1.0.0.0/8
AS 5
1.0.0.0/8
AS path=1
1.0.0.0/8
AS 3
1.0.0.0/8
AS path=5 1
AS 8
1.0.0.0/8
AS path=3 5 1
1.0.0.0/8

# BGP Next-Hop

- The BGP Next-Hop attribute tells the receiver of the advertisement the node to send packets to in order to reach the advertised prefix destination.
- Next-Hop attribute is a well-known mandatory.
- The Next-Hop maybe a BGP speaker from the advertising system but it does not have to be.
- This Next-Hop is used in building the routing table for BGP.
- The Next-Hop is encoded as an IP address.
- BGP Next-Hop reachability issues

# BGP next hop




BGP next hop for 1.0.0.0/8 is 12.12.12.1
BGP next hop for 2.0.0.0/8 is 12.12.12.2

# BGP Next Hop Reachability

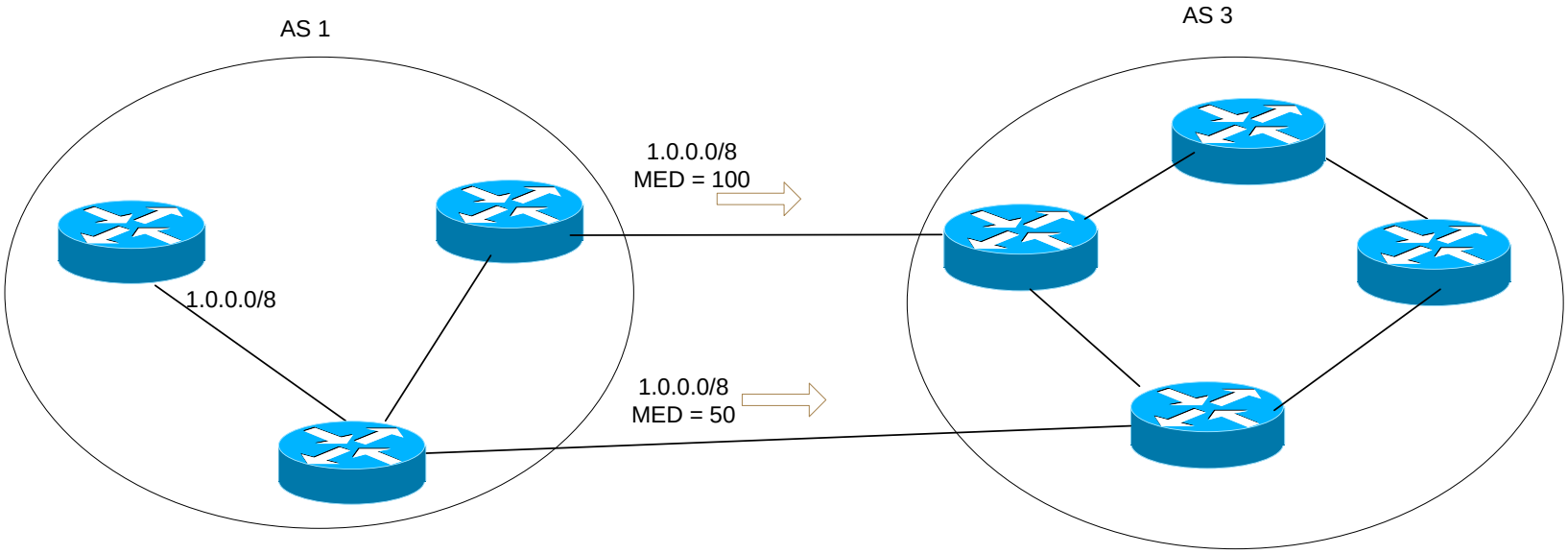
BGP next hop have to be reachable In order for the BGP route to be added to the BGP table. The following methods can be used to ensure BGP next hop reachability,

1. Static route to next hop
2. BGP Next hop self - this changes the default next hop from your peering neighbor IP address to your own IP address.
3. Advertise next hop into IGP
   - passive interface – link that connects next hop will be in IGP, but no neighbor relationship will be formed, therefore preventing routing information from being leaked out.
   - Redistribution of connected interface – link that connects next hop will be in IGP, but no neighbors will form on this link.

# BGP Multi-Exit-Discriminator

- The BGP Multi-Exit-Discriminator attribute carries a value expressing preference when there are two or more paths to one AS.
- This attribute is optional and nontransitive.
- The sender of this attribute is informing the peer AS on which link it would prefer to receive traffic from the receiving AS.
- The lower the value of the Multi-Exit-Discriminator the more preferred the path.
- This attribute is configured by Network Admin.
- Only apply to AS that has dual peering with another AS

# BGP MED

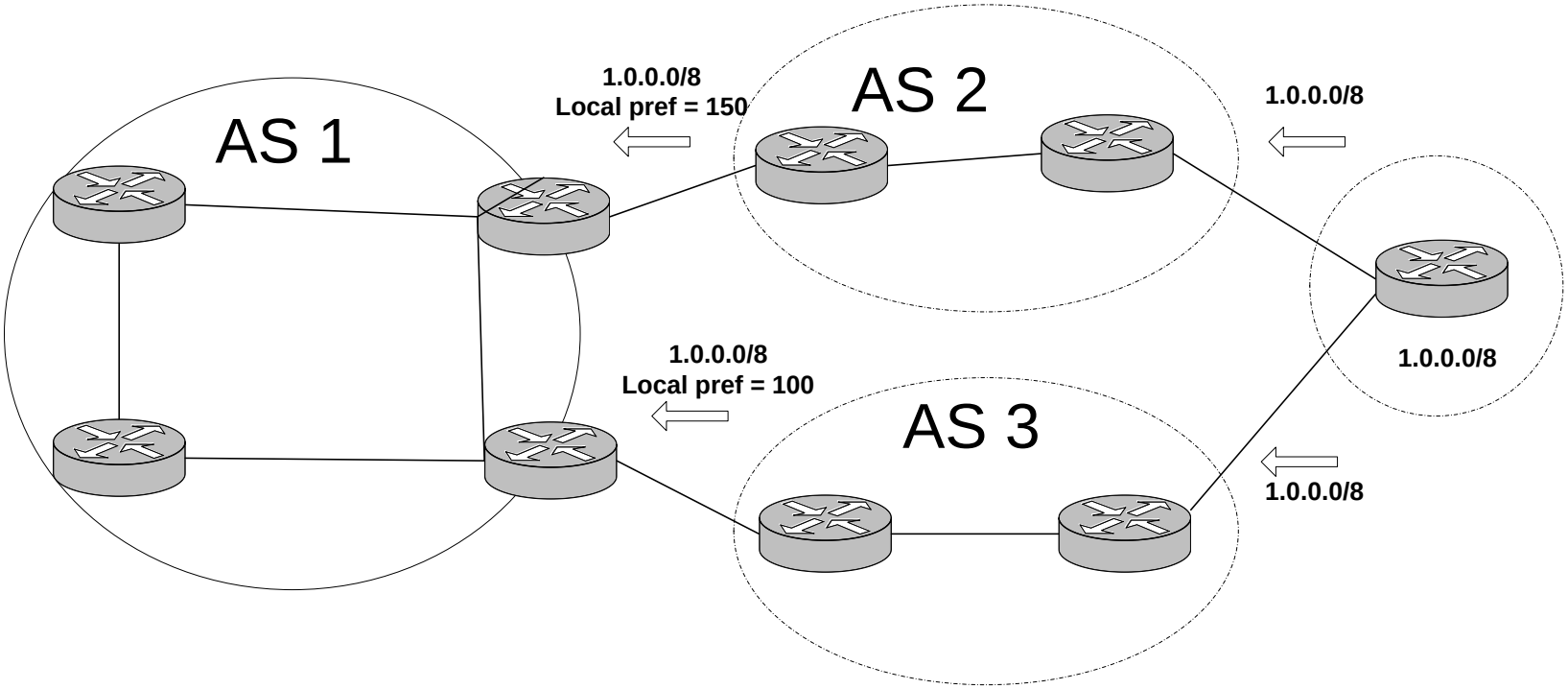



Traffic destined to 1.0.0.0/8 from AS3 will
use the bottom path because of lower MED

# BGP Local-Pref

- The BGP Local-Pref goes beyond Mutli-Exit-Discriminator for use by large ISPs with multiple connections and paths to other Autonomous Systems.
- This attribute is a well-known and discretionary attribute.
- The Local-Pref attribute is configured for the local AS and is assigned based on the advertising AS. The remote AS does not assign the value.
- The higher the Local-Pref value the better the route.  It allows the local Network Administrator to choose the preferred paths and change the preference dynamically based on new info, etc.

# Local preference




Traffic destined to 1.0/8 from AS 1 will go to AS 2 due to higher local pref

# BGP Communities

- The BGP Community attribute is used to simplify configuration of routing policies for BGP. It allows the network administrator to define policies for types of neighbors rather than for each. A route is identified as being within a certain category.
- The BGP Community attribute is transitive and optional. The value is a list of 32 bit community values. These values are only significant within the Autonomous System. Each route can have multiple community values.
- Well-known Community values:
  - 0xFFFFFF01 - No-Export
  - 0xFFFFFF02 - No-Advertise
  - 0xFFFFFF03 – No-Export-Subconfederation

# Other BGP Attributes

- Atomic Aggregate
  - Well-known discretionary attribute
  - Indicate loss of information due to aggregation
- Aggregator
  - Optional transitive
  - Identifies AS and BGP speaker that performed aggregation of some of the routes

# iBGP vs eBGP

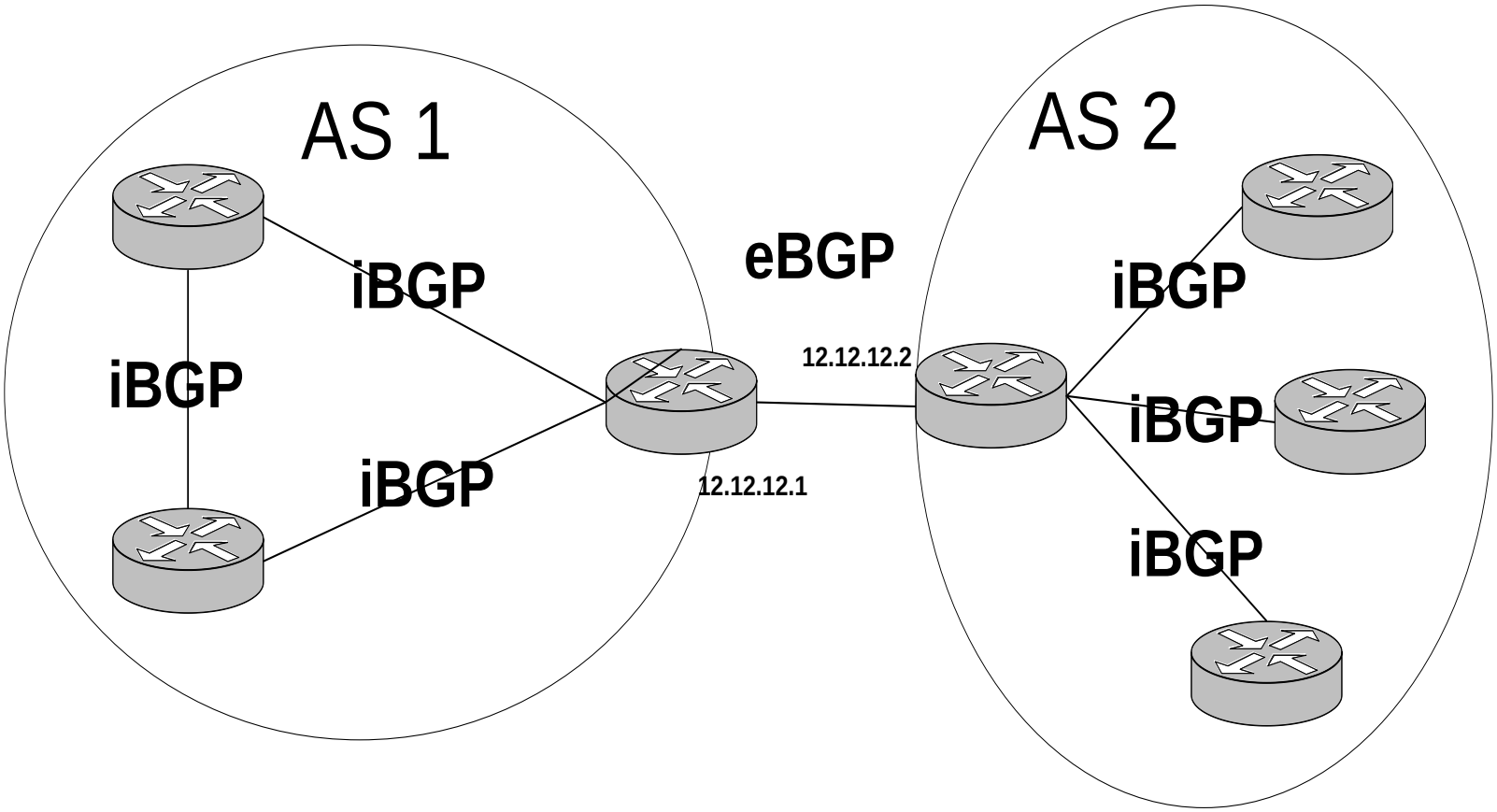
- eBGP - External BGP - between routers in two separate ASes – Admin distance of 20
- iBGP - Internal BGP - between routers within a single AS- Admin distance of 200
- eBGP most often has an IGP as a source of its routes. Some prefer to use iBGP because of the common attributes and metrics. Importing other IGPs such as OSPF or IS-IS there can be loss of metrics used to make routing decisions.
- Most networks running eBGP do not use iBGP as their primary IGP but use instead either OSPF, IS-IS, or EIGRP.

# iBGP vs eBGP (2)

Four main differences between iBGP and EBGP

1. Routes learned via iBGP Peers are not advertised to other iBGP Peers to prevent routing loops
2. iBGP path attributes are not manipulated to affect path selection
3. AS Path is not changed when advertising to iBGP Peer, ie the local AS is not added
4. BGP Next hop is not changed on the route learned from an iBGP Peer

# eBGP and iBGP



# iBGP Scaling

- iBGP routers have to be fully meshed. This means a lot of BGP sessions in AS with a lot of iBGP routers.
- Route reflection
  - Router reflector is a router that can re-advertise a route learned from an iBGP neighbor
  - Route reflector client is a router that receives that reflected routes from a route reflector
- Confederations
  - Divide an confederation into smaller sub-AS

# BGP Route Selection

- BGP has a very specific process for choosing a route from overlapping routes.
- Multiple routes to the same prefix mean that the prefix must be of exactly the same subnet length to be considered the same. Routes with more specific subnets or longer subnet lengths are preferred over less specific.
- BGP uses the Attributes we discussed already to help it make decisions when there are equal length prefixes with more than one possible route.
- Paths are compared in pairs, the better path is compared to the next possible path until 1 path is determined as Best.
- Important consideration in Route Selection is that a Network Administrator should predict or choose the route that will be preferred.

# BGP Route Selection

1. Highest Local-Pref value.
2. Shortest AS-Path list.
3. Lowest origin code.
4. Lowest MED value if there are multiple links to the same neighbor AS.
5. eBGP routes are preferred over iBGP routes.
6. Lowest IGP metric to the BGP Next-Hop.
7. BGP neighbor with the lowest BGP identifier preferred.
8. Neighbor with lowest IP address preferred.

# BGP Decision List (Cisco)

1. Only consider paths with reachable NEXT_HOPs
2. Do not consider iBGP path if not synchronized
3. Highest WEIGHT
4. Highest LOCAL_PREF
5. Prefer locally originated route
6. Shortest AS_PATH
7. Lowest ORIGIN code
   IGP < EGP < incomplete
8. Lowest Multi-Exit Discriminator (MED)
   - 8a IF bgp deterministic-med, order the paths before comparing
   - 8b IF bgp always-compare-med, then compare it for all paths
   - 8c Considered only if paths are from the same neighbor AS
9. Prefer an *External* path over an *Internal* one
10. Lowest IGP metric to the NEXT_HOP
11. IF multipath is enabled, the router may install up to N parallel paths in the routing table
12. For eBGP paths, select the "oldest"
    To minimize route-flap
13. Lowest Router-ID
    Originator-ID is considered for reflected routes
14. Shortest Cluster-List
    Client must be aware of RR attributes!
15. Lowest neighbor IP address

# BGP Example 1

*BigSky#show ip bgp 192.168.3.3*
*BGP routing table entry for 192.168.3.3/32, version 10*
*Paths: (2 available, best #2)*
*6 3*
*192.168.0.26 from 192.168.0.26 (192.168.6.6)*
*Origin IGP, localpref 100, valid, external*
*2 3*
*192.168.0.5 from 192.168.0.5 (192.168.2.2)*
*Origin IGP, localpref 150, valid, external, best*

# BGP Example 2

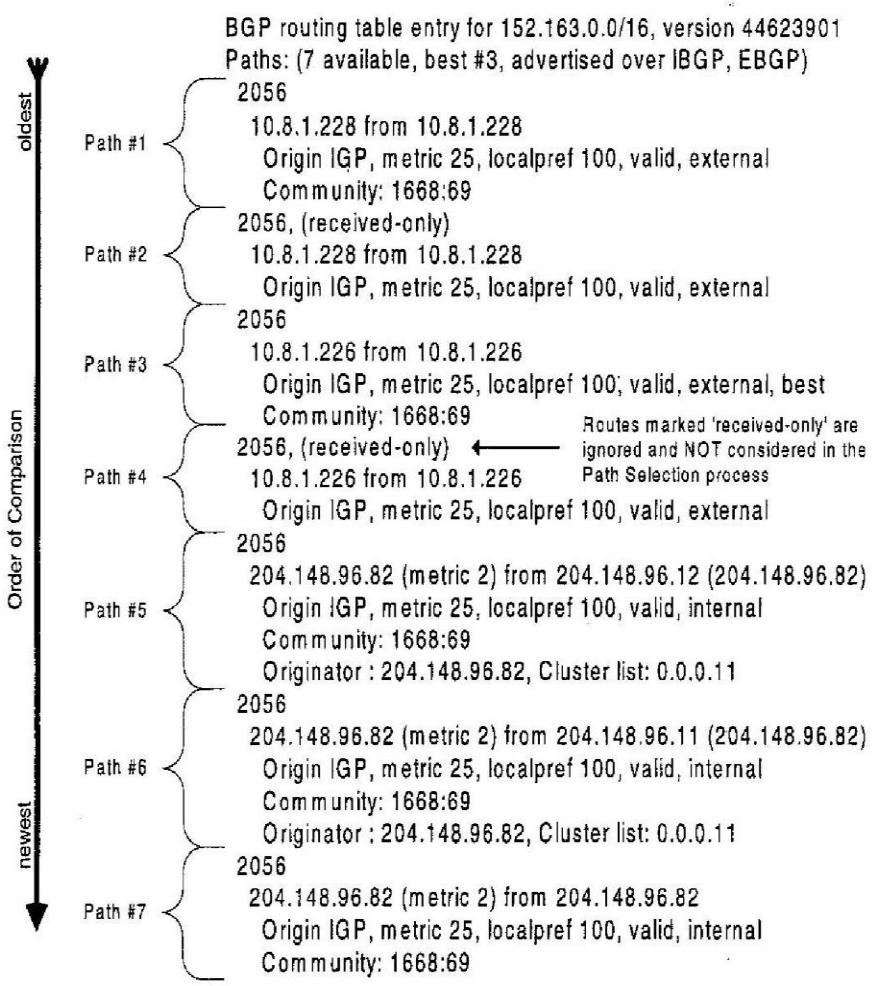
*Seattle#show ip bgp 192.168.1.0*
*BGP routing table entry for 192.168.1.0/26, version 5*
*Paths: (2 available, best #1)*
*6 3*
*192.168.0.15 from 192.168.0.26 (192.168.6.6)*
*Origin IGP, localpref 100, valid, external, best*
*2 4 3*
*192.168.0.1 from 192.168.0.5 (192.168.2.2)*
*Origin IGP, localpref 100, valid, external*

# BGP Example 3

## Why is path #3 selected as 'best'

```
BGP routing table entry for 64.12.128.0/18, version 6502563
Paths: (3 available, best #3)
  Advertised to peer-groups:
    rr-pop
  6461 6461 1668 14853
    198.32.136.31 from 198.32.136.31 (207.126.96.1)
      Origin IGP, metric 10020, localpref 100, valid, external
      Community: 2548:201 2548:666
  6461 1668 14853
    165.117.51.94 (metric 41) from 165.117.1.128 (165.117.1.128)
      Origin IGP, metric 4294967294, localpref 100, valid, internal
      Community: 2548:184 2548:207 2548:666 3706:127
  6461 1668 14853
    209.133.31.53 (metric 10) from 165.117.1.145 (165.117.1.145)
      Origin IGP, metric 4294967294, localpref 100, valid, internal, best
      Community: 2548:184 2548:208 2548:666 3706:153
```

# BGP Example 4



# BGP Example Answers

- Example 1 – path #2 is best because its local-pref is higher than that of path #1 (66)
- Example 2 – path #1 is best because its AS-path (6 3) is shorter than path #2 (2 3 4)
- Example 3 – path #3 is best because its IGP metric to next hop (10) is lower than IGP metric of path #2 (41).
- Example 4 – path #3 is best because its BGP ID(10.8.1.226) is lower than path #1 BGP ID (10.8.1.228)

# BGP Commands

- *Show ip bgp summary*
- *Show ip bgp neighbor <neighbor> advertised-routes*
- *Show ip bgp neighbor <neighbor> received-routes*
- *Show ip bgp neighbor <neighbor> routes*

# BGP Information

*Pop1#show ip bgp summary*

| *Neighbor* | *V* | *AS* | *MsgRcvd* | *MsgSent* | *TblVer* | *InQ* | *OutQ* | *Up/Down* | *State/PfxRcd* |
|---|---|---|---|---|---|---|---|---|---|
| *10.185.10.1* | *4* | *64888* | *98516* | *7626* | *5049* | *0* | *0* | *1w0d* | *6536* |
| *10. 85.28.153* | *4* | *65024* | *8480* | *8642* | *6305* | *0* | *0* | *3w4d* | *1536* |
| *10.120.8.155* | *4* | *64999* | *7494* | *7611* | *6496* | *0* | *0* | *3w4d* | *2* |
| *192.168.4.5* | *4* | *65096* | *4116* | *4793* | *1500* | *0* | *0* | *8w4d* | *Idle (Admin)* |

*pop1#show ip bgp neigh 10.120.8.155 adv*

| *Network* | *Next Hop* | *Metric* | | *LocPrf* | *Weight* | *Path* |
|---|---|---|---|---|---|---|
| **> 10.12.144.0/24* | *0.0.0.0* | *0* | *200* | *32768* | *i* | |
| **> 192.168.144.0/22* | *0.0.0.0* | *0* | *300* | *32768* | *i* | |
| **> 10.10.0.0/16* | *0.0.0.0* | *0* | *300* | *32768* | *i* | |

*pop1#show ip bgp neigh 10.120.8.155 routes*

| *Network* | *Next Hop* | *Metric* | *LocPrf* | *Weight* | *Path* |
|---|---|---|---|---|---|
| **>i172.16.0.0/24* | *10.120.8.155* | *0* | *300* | *0* | *i* |
| **>i.172.28.0.0/24* | *10.120.8.155* | *0* | *300* | *0* | *i* |

*Total number of prefixes 2*

# Summary

- BGP Operation
- BGP Message Types
- BGP Attributes
- BGP Best Path Selection
- iBGP vs eBGP
- BGP Path Selection Examples
- Reading Assignment: BGP4 Inter-Domain Routing in the Internet by John W. Stewart
  - section 1.4 p. 17 -18
  - section 1.7 to section 4.4 p. 29-109