Agenda

- BGP refresher
  - BGP protocol and attributes
  - IBGP and EBGP
  - Route damping
- BGP routing policy
- Scaling the IBGP full mesh
  - Route reflection
  - Confederations
- New features in BGP
- Other BGP Interests
- Resource Location

Agenda

- **BGP refresher**
  - **BGP protocol and attributes**
  - **IBGP and EBGP**
  - **Route damping**
- **BGP routing policy**
- **Scaling the IBGP full mesh**
  - **Route reflection**
  - **Confederations**
- **New features in BGP**
- **Other BGP Interests**
- **Resource Location**

BGP Overview

- BGP runs over TCP
  - TCP port 179
  - Why re-invent the wheel; for reliable delivery, re-transmission, sequencing, etc.
  - Some hackers are now targeting BGP, attacking port 179
- JUNOS™ Internet software supports BGP Version 4 and many extensions to the protocol
  - RFCs 1771 and 1772 - BGP4
  - RFC 1965 & 3065 - Autonomous System Confederations
  - RFC 1966 & 2796 - Route Reflection
  - RFC 1997 - Communities
  - RFC 2283 - MBGP
  - RFC 2385 - BGP MD5 Authentication
  - RFC 2439 - Route Damping
  - RFC 2842 - Capabilities Negotiation
  - Other Internet drafts

BGP Routes

- Consist of
  - Destination, described as an IP address prefix
  - Information that describes path to the destination
    - BGP route attributes
- BGP peers advertise NLRI to each other in update messages

BGP Routes

- BGP stores routes in the JUNOS software routing table
  - **RIB**
    - Inet.0 = IPv4 unicast routing table
  - **FIB**
    - IP forwarding table
- Routing table stores
  - Routing information learned from update messages
  - Local routing information selected by applying local policies to routes received in update messages
  - Information selected to advertise to BGP peers

Looking at BGP Routes

- Look at specific entries in the routing table

```
user@host> show route 172.16.1/24 extensive

inet.0: 11 destinations, 12 routes (11 active, 0 holddown, 0 hidden)
172.16.1.0/24 (1 entry, 1 announced)
TSI:
KRT in-kernel 172.16.1.0/24 -> {indirect(58)}
        *BGP    Preference: 170/-101
                Source: 10.0.0.8
                Next hop: via so-0/2/1.0, selected
                Protocol next hop: 10.0.0.8 Indirect next hop: 84d0b28 58
                State: <Active Int Ext>
                Local AS:      1 Peer AS:      1
                Age: 23          Metric: 0        Metric2: 1
                Task: BGP_1.10.0.0.8+4371
                Announcement bits (2): 0-KRT 4-Resolve inet.0
```

Looking at BGP Routes (cont)

- Look at specific entries in the routing table

```
AS path: I
Localpref: 100
Router ID: 10.0.0.8
Indirect next hops: 1
        Protocol next hop: 10.0.0.8 Metric: 1 Indirect next hop: 84d0b28 58
        Indirect path forwarding next hops: 1
Next hop:         via so-0/2/1.0
        10.0.0.8/32 Originating RIB: inet.0
          Metric: 1                          Node path count: 1
          Forwarding nexthops: 1
                Nexthop: via so-0/2/1.0
```

BGP Messages

- Four (or five?) types of messages
  - Type 1 - Open
  - Type 2 - Update
  - Type 3 - Notification
  - Type 4 - Keepalive
  - (Type 5 – Route Refresh Message)

BGP Messages

- Open messages
  - After a TCP connection is established, BGP peers exchange open messages to negotiate a BGP connection
  - Upon successful BGP connection, the peers exchange other BGP messages and data, such as routing information

BGP Messages

- Update messages
  - Used to exchange network reachability information
- BGP systems use this information to construct a type of graph describing relationships among all known autonomous systems
  - How this is constructed is Vendor implementation specific
  - Autonomous system numbers that are associated with each IP prefix prevent routing loops
  - No Dijkstra type of algorithm run

BGP Messages

- Notification messages
  - BGP systems send notification messages when an error condition is detected
  - After the notification message is sent, the BGP session and the TCP connection between the BGP peers is closed by the sender
  - Notification messages consist of
    - BGP header
    - Error code
    - Error Subcode
    - Data that describes the error

BGP Notification Messages

- Notification error codes
  - 1 – Message header error
  - 2 – Open message error
  - 3 – Update message error
  - 4 – Hold timer expired
  - 5 – Finite state machine error
  - 6 – Cease

BGP Notification Messages

- Notification error subcodes
  - Message header error codes
    - 1 – Connection not synchronized
    - 2 – Bad message length
    - 3 – Bad message type
  - Open message error codes
    - 1 – Unsupported version number
    - 2 – Bad peer autonomous system (AS)
    - 3 – Bad BGP identifier
    - 4 – Unsupported optional parameter
    - 5 – Authentication failure
    - 6 – Unacceptable hold time
    - 7 – Unsupported Capability
  - Update message error codes
    - 1 – Malformed attribute list
    - 2 – Unrecognized well-known attribute
    - 3 – Missing well-known attribute
    - 4 – Attribute flags error
    - 5 – Attribute length error
    - 6 – Invalid origin attribute
    - 8 – Invalid next-hop attribute
    - 9 – Optional attribute error
    - 10 – Invalid network field
    - 11 – Malformed AS-path

BGP Messages

- Keepalive messages
  - BGP systems exchange keepalive messages to determine whether a peer has failed or is no longer available
  - Exchanged often enough so that the hold timer does not expire
  - 30-second keepalive interval with a 90-second hold timer are the JUNOS software defaults
    - Hold timer is negotiated between peers
    - Highest hold timer is used
  - Consist of only the BGP header (19 bytes)

BGP Neighbor States

- Idle
  - BGP always begins in the Idle state
  - In this state, all BGP connections are refused
  - Basically, BGP hasn't started
- Connect
  - TCP connection is being attempted
- Active
  - BGP is trying to initiate a TCP connection
  - Likely there was a problem with the first TCP connection attempt

BGP Neighbor States

- OpenSent
  - Open message has been sent
  - BGP waits for an Open message from its peer
- OpenConfirm
  - Open message received from peer
  - Waiting for a Keepalive message
- Established
  - BGP is complete and Updates can now be exchanged

Show BGP Neighbor

```
user@host> show bgp neighbor
Peer: 10.0.0.8+4371   AS 1     Local: 10.0.0.7+179    AS 1
  Type: Internal    State: Established    Flags: <ImportEval>
  Last State: OpenConfirm   Last Event: RecvKeepAlive
  Last Error: None
  Options: <Preference LocalAddress HoldTime PeerAS Refresh>
  Local Address: 10.0.0.7 Holdtime: 90 Preference: 170
  Number of flaps: 0
  Peer ID: 10.0.0.8         Local ID: 10.0.0.7         Active Holdtime: 90
  Keepalive Interval: 30
  NLRI advertised by peer: inet-unicast
  NLRI for this session: inet-unicast
```

Important BGP Attributes

- AS-path
  - Lists the ASNs that a route has traversed
  - Used for
    - Loop detection
    - BGP decision process (shorter AS-path is preferred)
- Local preference
  - Used to influence routing (higher LP is preferred)
  - Assists with choosing which exit to take out of the local AS
- Multi exit discriminator (MED)
  - Used to influence routing (lower MED is preferred)
  - Assists with choosing which entrance to take into the local AS

Important BGP Attributes

- BGP next hop
  - The IP address of the router that is used to reach a BGP destination
- Origin
  - Provides a slight clue about how the route was originated
- Community
  - Used to identify or classify routes into groups
  - Typically used as basis for applying routing policy

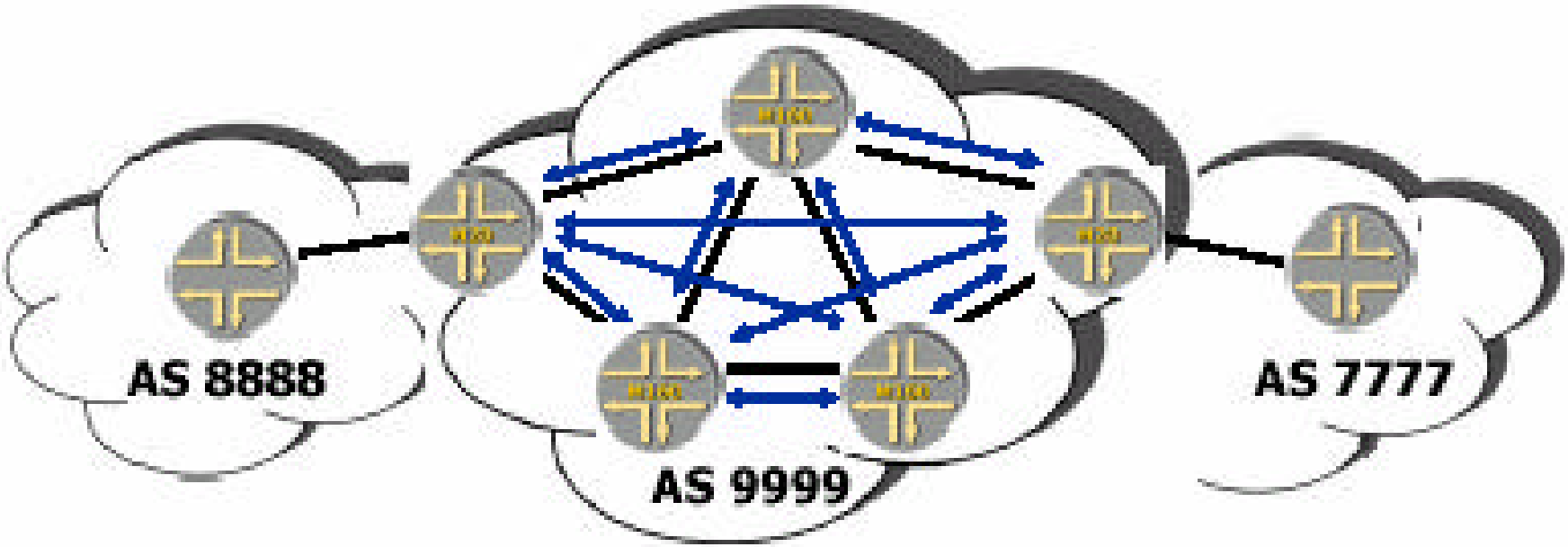
Autonomous Systems

- What is an AS?
  - Group of routers
    - Administered with a common routing policy
    - Running under a single technical administration
    - Viewed externally as a single, coherent interior routing domain
    - Could be running more than one IGP
  - 16-bit integer (1-65535)
  - 64512-65535 are private ASs
  - AS-path is checked upon receipt to determine routing loops

Interior BGP

- IBGP used inside an AS
- Typically implemented as full IBGP mesh
- Why do you need a full mesh?
  - AS-path check not applicable to IBGP
  - IBGP speakers, by default, cannot forward IBGP learned routes to other IBGP speakers
- BGP next hop not reset when re-sending routes
- Local preference often used for internal routing policy
  - Selecting preferred exit point from AS
  - Common practice to set local preference based on community matching
- IBGP peers almost always peer between loopback addresses
  - Provides extra redundancy

Interior BGP



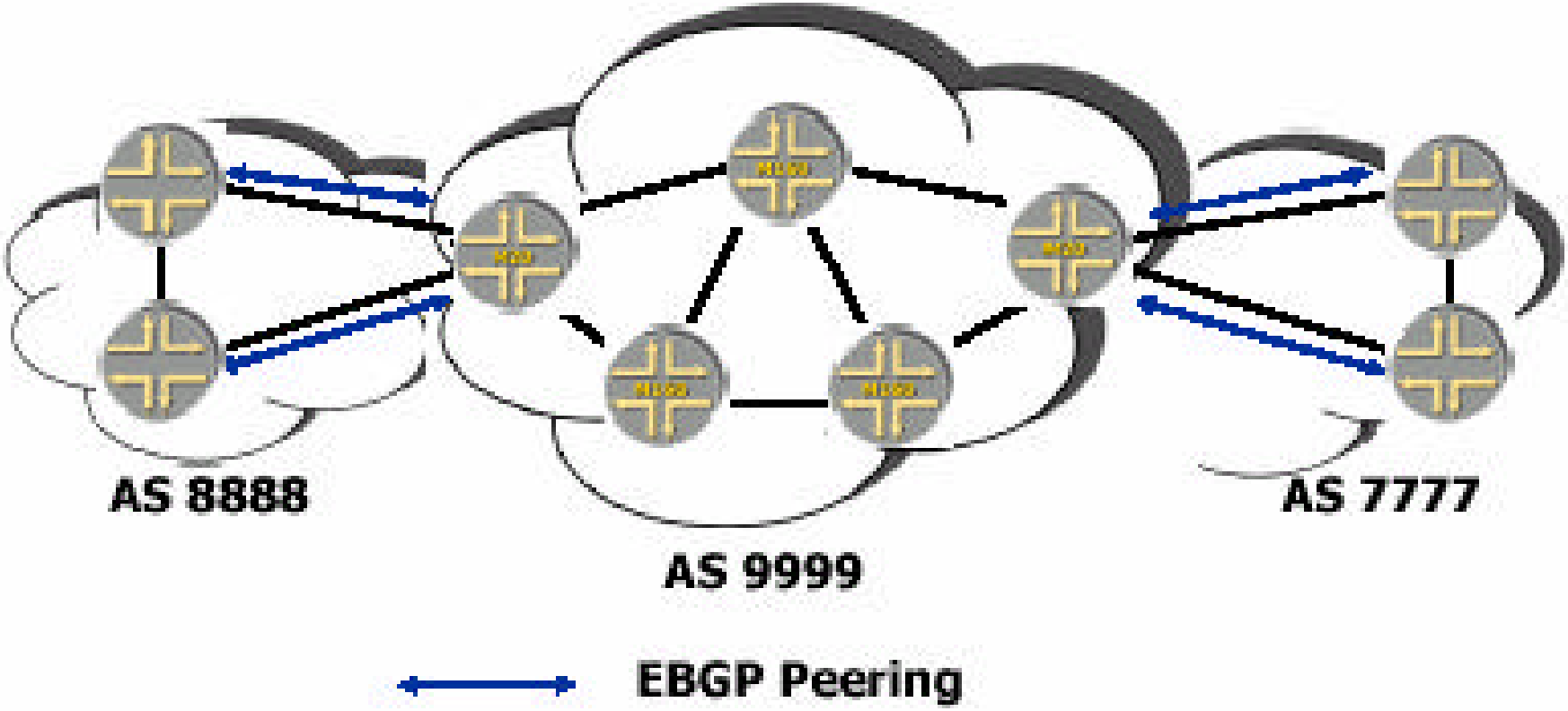
Interior BGP

- What is synchronization?
  - BGP speakers should not advertise routes unless IGP knows how to reach them
  - Not needed if running full IBGP mesh
  - Historically an Enterprise, not an ISP, issue
- Sync permanently disabled in JUNOS software

Exterior BGP

- Used for passing routes between autonomous systems
- Differences with IBGP
  - BGP next hop is reset when re-sending routes
    - Nexthop address used will be in the IGP
  - Local AS number is pre-pended to AS-path
  - MED often used for routing policy
    - Selecting preferred entry point to local AS
    - Common practice to set MED based on IGP metric
  - EBGP peers typically peer between physical interface addresses
    - Re-routing EBGP sessions not desirable
    - If interface goes down, EBGP session should drop
    - Exception is the use of multihop for load balancing over parallel links

Exterior BGP



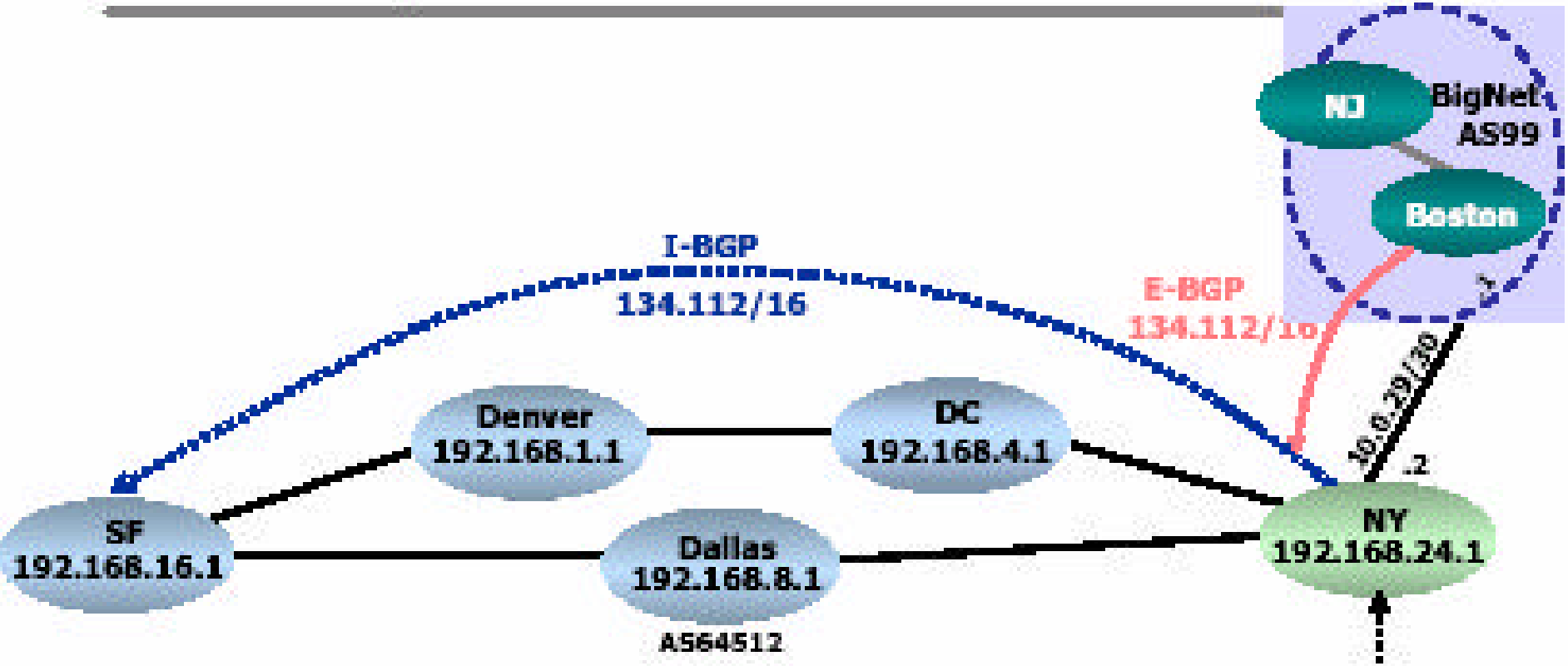
Why Do You Need an IGP?

- Isn't IBGP enough?
- What is IGP used for?
  - Loopback addresses
    - Which are typically the BGP nexthop addresses
  - Interior links
  - NMS/server subnets
- Is it feasible to not have an IGP?
  - Possible, yes, but … optimal?

Resolving EBGP Next Hops

- Two common methods
  - **Next-hop self**
    - Resets BGP nexthop address to local loopback address when exporting routes to IBGP peers
    - Local loopback address should already be in IGP
  - **Passive interface**
    - Adds external link subnet to IGP
    - But no IGP adjacencies can be formed over external link
    - Also allows external peer router to be pinged from internal network
  - **Implications for MPLS FEC**
    - Next-hop self works better if using MPLS for traffic engineering to external BGP destinations
    - If using passive interface, MPLS knobs are needed

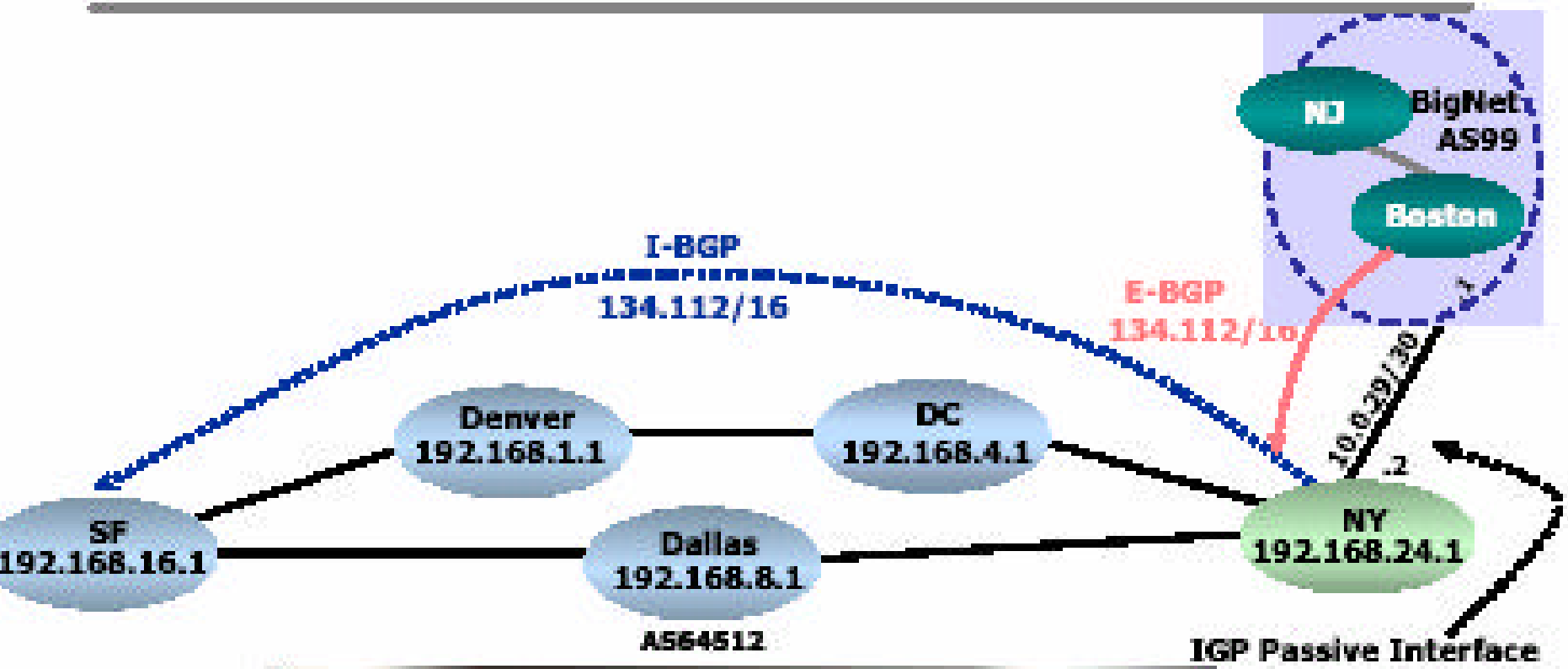
Next-hop Self




```
lab@SF> show route resolution 134.112/16
Table inet.0 Nodes 12
134.112.0.0/16 Originating RIB: inet.0
  Metric: 20  Node path count: 1
  Indirect nexthops: 1
        Protocol Nexthop: 192.168.24.1 Metric: 20
        Indirect nexthop: 84d0b28 58
        Indirect path forwarding nexthops: 1
                Nexthop: via so-0/2/1.0
```

Configure Next-hop Self Policy

Passive Interface




```
lab@SF> show route resolution 134.112/16
Table inet.0 Nodes 12
134.112.0.0/16 Originating RIB: inet.0
  Metric: 30  Node path count: 1
  Indirect nexthops: 1
        Protocol Nexthop: 10.0.29.1 Metric: 30
        Indirect nexthop: 84d0b28 58
        Indirect path forwarding nexthops: 1
              Nexthop: via so-0/2/1.0
```

JUNOS Route Preference

- Next hop is reachable?
  -1 = Not reachable
- Lower route preference
  0 = Directly connected
  5 = Static routes
  7 = RSVP
  9 = LDP
  10 = OSPF internal
  15 = IS-IS L1 internal
  18 = IS-IS L2 internal
  100 = RIP
  130 = Aggregate or generated
  150 = OSPF external
  160 = IS-IS L1 external
  165 = IS-IS L2 external
  170 = BGP

JUNOS BGP Route Selection

- Lower route preference
- Higher local preference
- Shortest AS-path
- Lower origin (IGP < EGP < incomplete)
- Lower MED
- External over confederation over internal
- Lower IGP metric
- Shorter cluster list
- Lower router-id

JUNOS BGP Route Advertisement

- JUNOS software default BGP advertisement rules
  - Send all active BGP routes
    - All BGP learned routes (except IBGP rule)
    - Advertise-inactive knob available
  - Export policies needed to
    - Advertise static routes
    - Advertise aggregate routes
    - Redistribute/export other protocol routes to BGP
    - Originate routes into BGP

Route Damping

- Reduce the route update load without limiting convergence time for well behaved routes
- Applied to EBGP routes
  - Can also be used with confederations
- Configured by creating a named set of damping parameters that you apply as a damping policy action
- Some prefixes should never be damped
  - DNS root servers
- BGP damping off by default
- Historically used to protect legacy routers
- How many service providers still use damping?
- RFC 2439

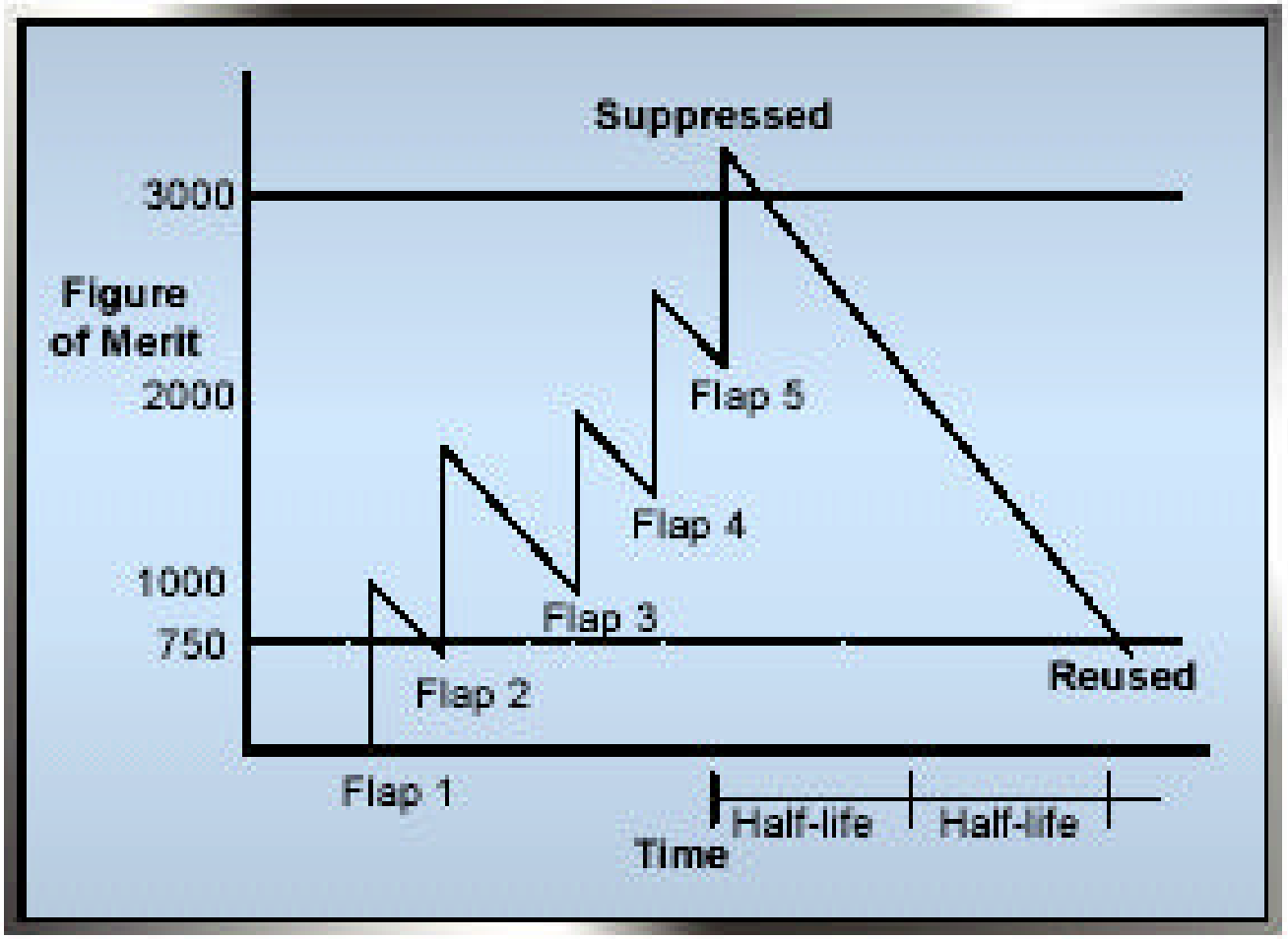
Damping Figure of Merit

- New route given a figure of merit of 0
- Figure of merit increases with each incident
  - Withdrawn route—1,000
  - Path attribute change—500
- Route is suppressed when figure of merit exceeds suppress threshold
  - Default suppress threshold is 3,000
- Route is reused when figure of merit drops below reuse threshold
  - Default reuse threshold is 750

Damping Figure of Merit

- Exponential decay
  - Reduces figure of merit over time
  - Default 15 minute half-life
- Maximum suppression time limit
  - Default is 60 minutes
- Maximum figure of merit
  - Stops increasing when ceiling is reached
  - Determined by formula
  - Not explicitly configurable

Route Damping



Damping Configuration

- Defining damping parameters are similar to defining a community

```
policy-options {
    damping name {
        half-life minutes;
        max-suppress minutes;
        reuse number;
        suppress number;
    }
}
```

Damping Example

```
policy-options {
     policy-statement damp {
          from {
               route-filter 11/8 exact damping high;
               route-filter 15/8 exact damping medium;
               route-filter 0/0 upto /24 damping none;
          }
          then accept;
     }
     damping high {
          half-life 15;
          suppress 3000;
          reuse 2500;
          max-suppress 50;
     }
     damping medium {
          half-life 3;
          max-suppress 4;
     }
     damping none {
          disable;
     }
}
```

Show BGP Summary

- View basic information about all BGP neighbors

```
user@host> show bgp summary
Groups: 1 Peers: 1 Down peers: 0
Table            Tot Paths  Act Paths Suppressed    History Damp State    Pending
inet.0                   2          2          0          0          0          0
Peer                     AS      InPkt     OutPkt    OutQ   Flaps Last Up/Dwn
State|#Active/Received/Damped...
10.0.0.8                  1         18         19       0       0     8:02 2/2/0
10.0.0.9                  1         20         21       0       0     8:05 3/3/0
```

Agenda

- BGP refresher
  - BGP protocol and attributes
  - IBGP and EBGP
  - Route damping
- **BGP routing policy**
- Scaling the IBGP full mesh
  - Route reflection
  - Confederations
- New features in BGP
- Other BGP Interests
- Resource Location

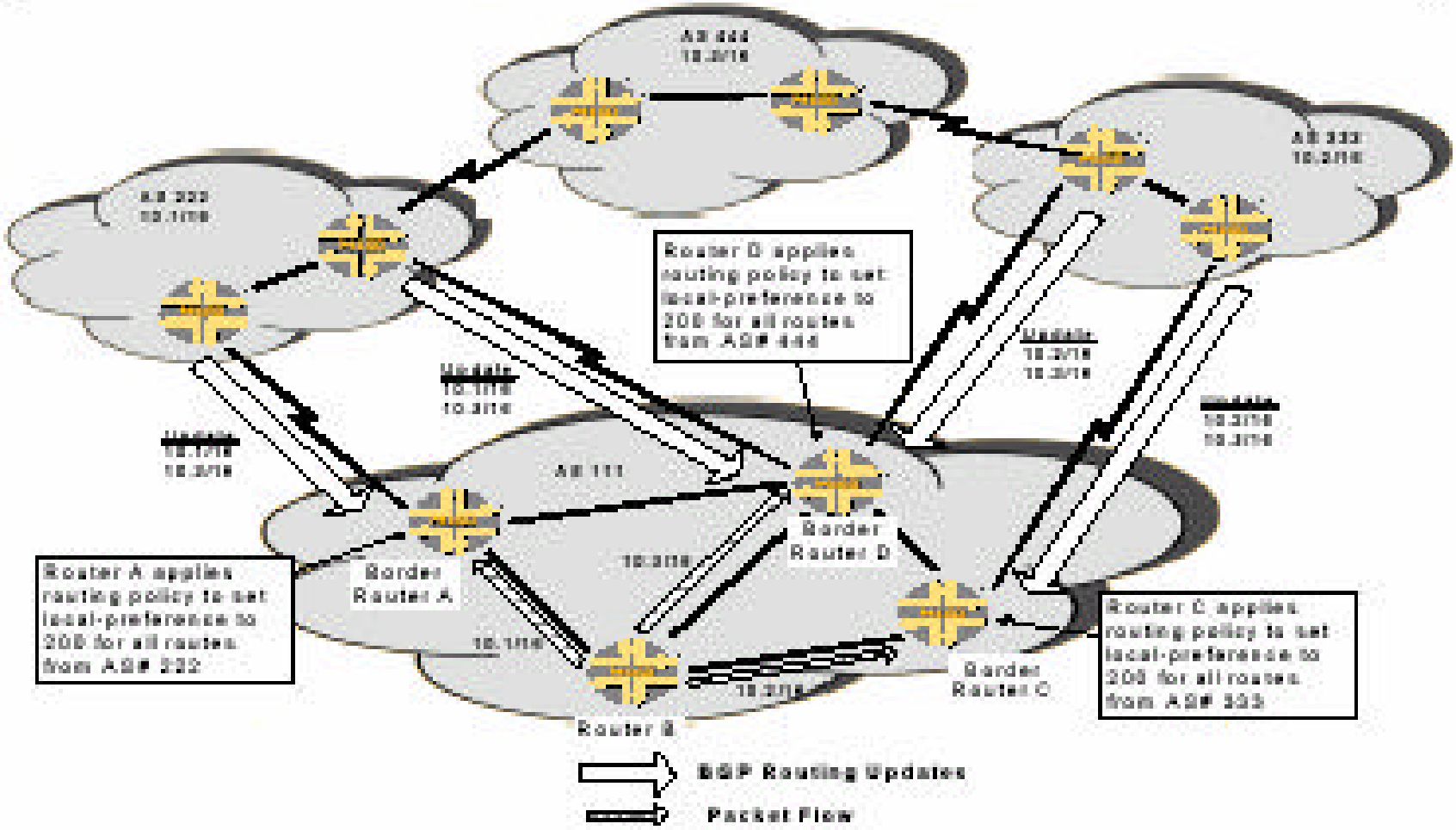
Routing Policy

- Controls routing information transferred between routing table and each routing protocol
  - Incoming routing information can be ignored, accepted or changed
  - Outgoing routing information can be suppressed, accepted or changed

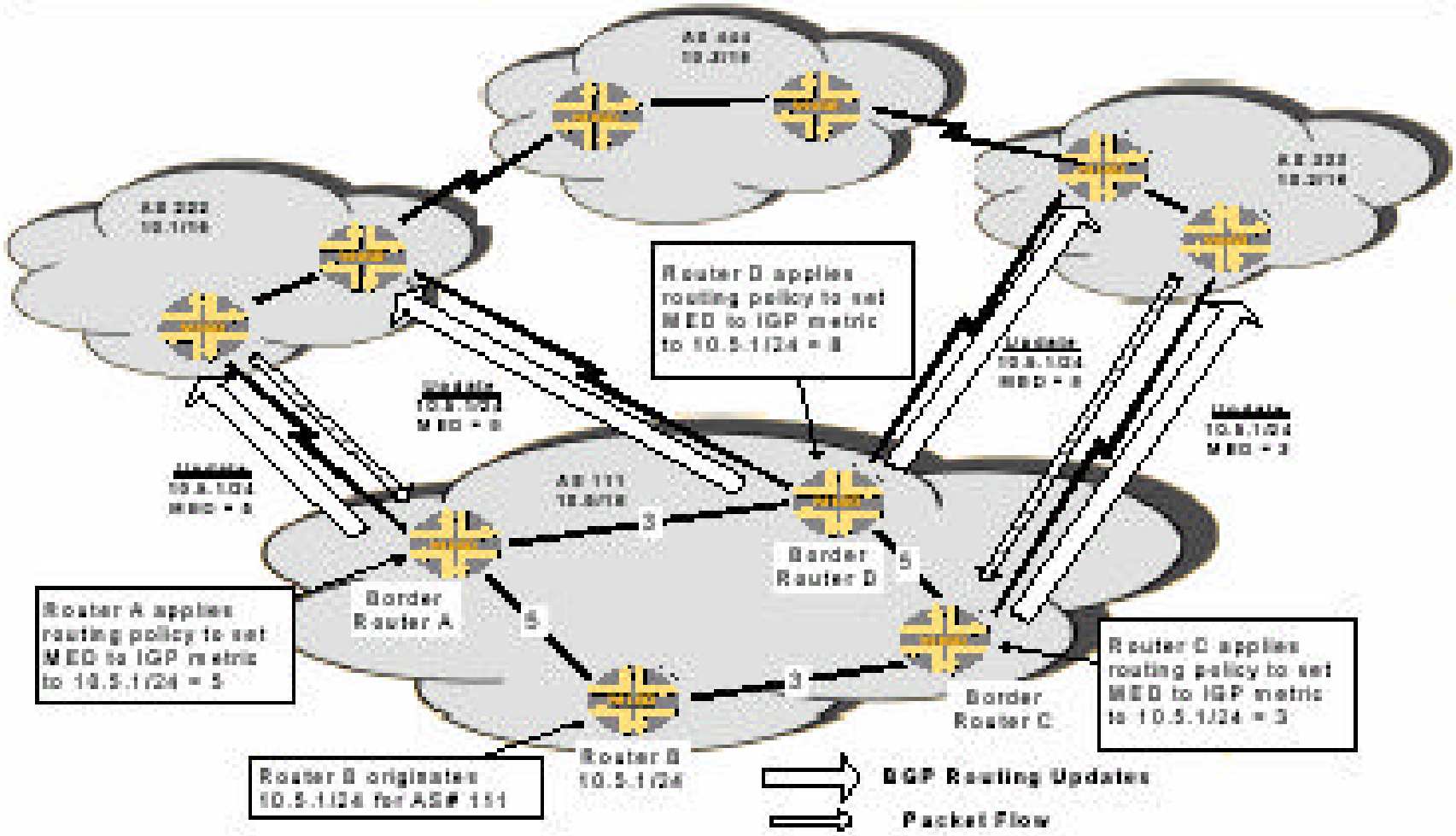
When to Apply Policy

- You do not want to import all learned routes into the routing table
- You do not want to export all learned routes to neighboring routers
- You want to export locally defined routes
- You want one protocol to receive routes from another protocol
- You want to modify information (attributes) associated with a route
- You want to control routing decisions

Local Preference Example



MED Example


Router D applies routing policy to set MED to IGP metric to 10.5.1/24 = 8
Router A applies routing policy to set MED to IGP metric to 10.5.1/24 = 5
Router C applies routing policy to set MED to IGP metric to 10.5.1/24 = 3
Router B originates 10.5.1/24 for AS# 111
Border Router A
Border Router D
Border Router C
Router B 10.5.1/24
3
5
5
3
BGP Routing Updates
Packet Flow

Show BGP Routes

- Look at routes received from a specific peer before policy is applied

```
user@host> show route receive-protocol bgp 11.1.1.1
inet.0: 6 destinations, 6 routes (5 active, 0 holddown, 1 hidden)
Prefix                  Nexthop      MED     Lclpref    AS path
* 10.0.0.0/8            11.1.1.1     0       100        I
* 172.16.0.0/12         11.1.1.1     0       100        I
```

- Look at routes advertised to a specific peer after policy is applied

```
user@host> show route advertising-protocol bgp 11.1.1.2
inet.0: 10 destinations, 10 routes (8 active, 0 holddown, 2
hidden)
Prefix                  Nexthop      MED     Lclpref    AS path
* 10.0.0.0/8            Self         0       100        I
* 172.16.0.0/12         Self         0       100        I
```

Best Practice Routing Policies

- Recommended and often used policies
  - Martian route-filter
  - Prefix length filters
  - Send aggregate, yet suppress specifics
  - Prefer customer routes over all other routes
  - Prefer peer routes over transit routes
  - Mark routes with communities

Communities

- Group of prefixes that share a common property
- Routing decisions can more easily be based on the identity of the community group
  - Rather than on each IP prefix
- Facilitates and simplifies the control of routing information
- Incoming routes should be tagged by ingress border router
- Can tag locally defined routes with community
- RFCs 1997 and 1998

Configuration Example

```
policy-options {
    policy-statement TRANSIT-IN {
        term DENY-PREFIX-LENGTH {
            from policy PREFIX-LENGTH;
            then reject;
        }
        term DENY-MARTIANS {
            from policy MARTIANS;
            then reject;
        }
        term PERMIT-REST {
            then {
                community set TRANSIT-
    ROUTES;
                local-preference 80;
                accept;
            }
        }
    }
    community TRANSIT-ROUTES members
    6666:80;
```

Route Aggregation

- Summarizes group of routes with common prefixes
- Reduces size of routing table, routing updates, and route flapping
- Four common methods
  - "Nail it up" with passive knob
  - Allow it to disappear if no contributing route present
  - Use generate route
  - Use static route

```
routing-options {
        aggregate {
            route 8.8.0.0/16; {
                                passive;
                        }
        }
```

Route Aggregation

- Must explicitly suppress contributing routes

```
policy-options {
    policy-statement SUPPRESS-SPECIFICS {
        from route-filter 8.8/16 longer reject;
        }
```

- RFC 2519 = Inter-domain Route Aggregation
- CIDR Report sent weekly to Nanog list on who is aggregating and who is being bad

Agenda

- BGP refresher
  - BGP protocol and attributes
  - IBGP and EBGP
  - Route damping
- BGP routing policy
- **Scaling the IBGP full mesh**
  - **Route reflection**
  - **Confederations**
- New features in BGP
- Other BGP Interests
- Resource Location

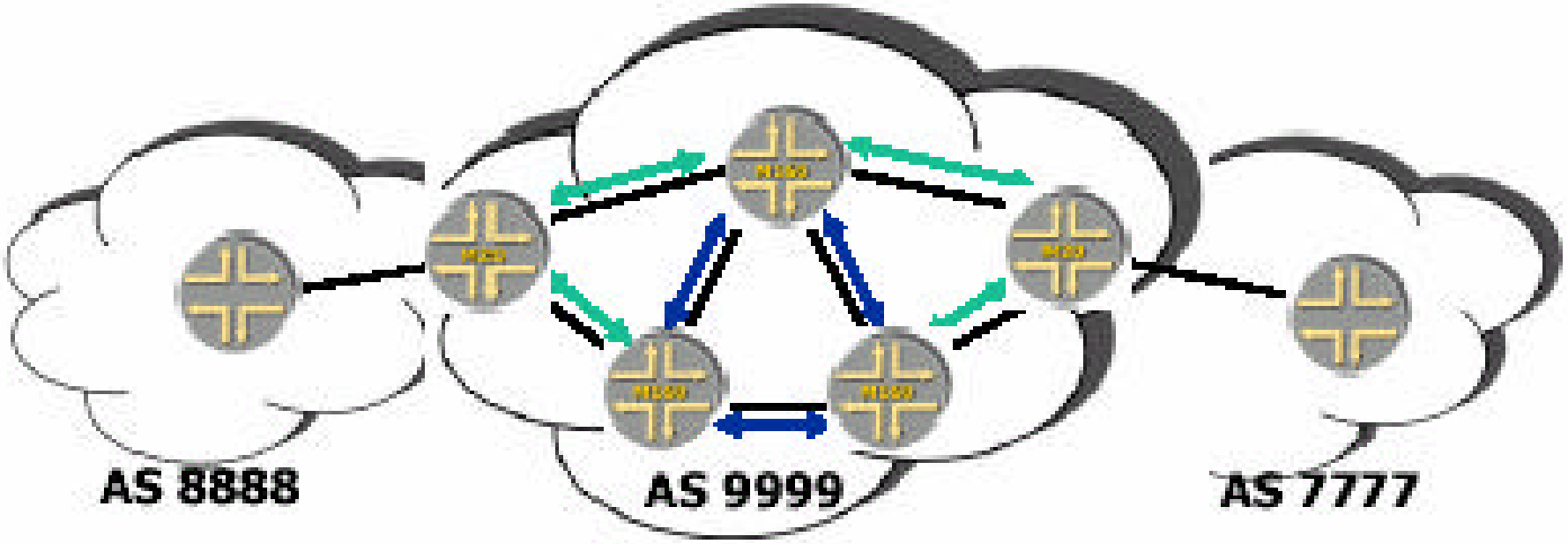
IBGP Full-mesh Scaling

- N-squared problem
  - Add one new router to mesh, one must
    - Peer to all IBGP speakers
    - Add new router to all IBGP speaker configurations
- Increases TCP processing overhead
- Increases router CPU processing
- Increases router memory requirements
- Increases size of routing tables
- Two common methods to scale IBGP
  - Route Reflection (RFC 2796)
  - Confederations (RFC 3065)

Route Reflection

- Allows an IBGP speaker to export an IBGP learned route to another IBGP speaker
- Relaxes IBGP rule
- Reduces IBGP meshing requirement
- Route reflector (RR) only reflects its best path
  - "Hides" alternate paths
  - May result in sub-optimal routing
  - May result in oscillating routes (more on that later)
- RR, by default, does not change IBGP attributes

Route Reflection



IBGP Full Mesh Peering

RR Peering

Route Reflection

- Since IBGP rule is being relaxed, doesn't this create possible routing loops?
- New attributes added
  - Cluster-id
    - Identifies the route reflection cluster ID
    - RR adds this attribute
  - Cluster-list
    - Sequence of cluster-ids that an update has traversed
    - Similar to AS-path list
  - Originator-id
    - Identifies the router that originated the route into the AS
    - RR adds this attribute

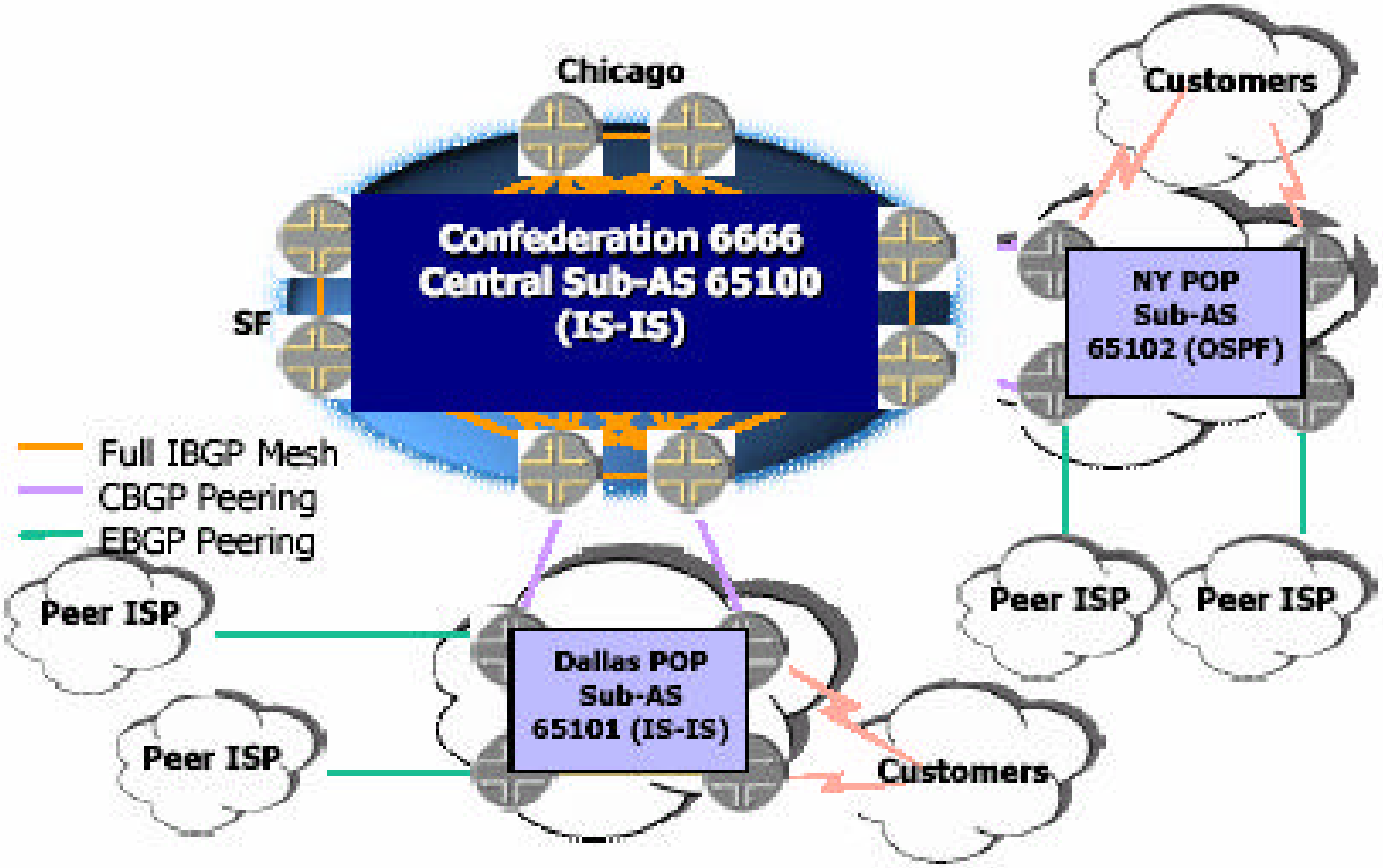
Configuration Example

```
routing-options {
     autonomous-system 6666;
}
protocols {
     bgp {
          damping;
        group ibgp-mesh {
            export [ nexthopself send-connected ];
            local-address 8.8.254.253;
            peer-as 6666;
            neighbor 1.2.3.4;
            neighbor 2.3.4.5;
            neighbor 3.4.5.6;
        }
        group rr-cluster {
              cluster 1.1.1.1;
              export [ nexthopself send-connected
   ];
              local-address 8.8.254.253;
              peer-as 6666;
              neighbor 4.5.6.7;
```

Confederations

- Second method of reducing full IBGP mesh
- Break AS into multiple sub-autonomous systems (sub-AS)
- Sub-AS
  - Can use private AS numbers
  - Full IBGP mesh still needed inside sub-AS or use RR
  - Hidden from external ASs
- AS still viewed externally as single AS
- Sub-AS not counted as AS-path hop for route selection

Confederations



Confederation BGP

- CBGP (or E-IBGP)
  - Is this EBGP or IBGP?
- BGP next hop
- AS-path
- Local preference
- MED

Agenda

- BGP refresher
  - BGP protocol and attributes
  - IBGP and EBGP
  - Route damping
- BGP routing policy
- Scaling the IBGP full mesh
  - Route reflection
  - Confederations
- **New features in BGP**
- Other BGP Interests
- Resource Location

Capabilities Negotiation

- Allows capabilities negotiation between BGP speakers
- RFC 1771 says
  - If Open message received with unsupported capability
  - Send Notification error subcode 4 "Unsupported Optional Parameter"
  - Terminate peering session
  - Did not facilitate introduction of new capabilities in BGP
- RFC 2842

Extended Communities

- Two important enhancements
  - Provides extended range (4 octet value to 8 octet value)
  - Adds Type field (2 octets)
- Route target community
  - Identifies the destination to which the route is going
- Route origin community
  - Identifies where the route originated
- Used to control route distribution in MPLS L3 VPNs
- draft-ietf-idr-bgp-ext-communities-05.txt

Extended Communities

```
[edit]
 policy-options {
          community test-a members [target:9999:70];
          community test-b members [target:1.1.1.1:90];
          community test-c members [origin:6666:110];
 }
```

Route Refresh Capability

- Dynamically and non-disruptively request a re-advertisement of routes from a peer

```
user@host> clear bgp neighbor 11.1.1.1 soft-inbound
```

- JUNOS software stores unmodified copy of all valid routes in RIB-In

```
user@host> show route receive-protocol bgp 11.1.1.1
```

- Creates a new BGP message type
- RFC 2918

Outbound Route Filter (ORF)

- Address Prefix Outbound Route Filter
- Used to perform address prefix based route filtering
- A BGP router can send to its peer a list of prefixes that can be used to create an outbound route filter on the peer
- Useful in Layer 3 VPNs to control routing

BGP Multipath

- Installs multiple, non-multihop BGP nexthops as active
  - Distributes prefixes across those active nexthops
- Works with EBGP and IBGP
- Eliminates the tie-break "lower router-id" decision
  - Which would normally result in all prefixes pointing to single nexthop
- Where might this be applicable?

MBGP

- Extensions to enable BGP to carry information for different network layers and address families
- Used for multicast

```
user@host# set nlri [ multicast | unicast | any ]
```

- Used for MPLS VPNs
  - MP_REACH_NLRI
    - VPN-IPv4 + Label
- Used for IPv6

BGP Graceful Restart

- Optional capability, negotiated in Open message
- Allows BGP router to continue forwarding while BGP is restarting
  - Based on FIB state before restart
- Once BGP restart is complete, then forwarding table is updated with new routing information
- Similar in functionality to OSPF and ISIS Graceful Restart