What’s Included?

1) Routing Reprise
2) BGP Basics
3) BGP Path Selection (many examples)
4) Use of Community Strings
5) Multiprotocol Extensions to BGP
   1) Multicast Routing Covered
   2) Layer-III VPN (aka VRF) Omitted
What’s Missing?

• Confederations
• Dampening
• Minor Extensions
  – 32-bit AS
  – Extended Communities
• Routing Policy Implementation
• BGP State Machine
Part I: Routing Reprise
Cisco Architecture

IP Routing Table

- 192.168.0.0/16
  - Metric:
  - Next-Hop:

- 172.16.0.0/12
  - Metric:
  - Next-Hop:

- 10.0.0.0/8
  - Metric:
  - Next-Hop:

Network Protocols:

- BGP
- EIGRP
- OSPF
- RIP
- Static

Export
Import
Juniper Architecture

- **BGP**
  - Metric 1:
  - Metric 2:
  - Next Hop:
  - Area:
  - AS-Path:
  - Community:
  - Level:

- **IS-IS**

- **OSPF**

- **Static**

- **RIP**

- **Export**

- **Import**
BGP vs IGP

• IGP (OSPF or ISIS)
  – Provides a Next-Hop address and an egress interface to any known destination address

• BGP
  – Only provides a Next-Hop address to a destination prefix
  – This has to be resolved to an egress interface using a second route lookup
Best Route Selection

• Longest prefix always wins regardless of routing protocol

• Source of routing information
  – Connected > Static > eBGP > {IGP} > iBGP

• BGP ignores received prefixes if
  – There is no route to the NEXT_HOP
  – The AS_PATH contains:
    • the local AS number
    • (a private AS number)
  – Not synchronized (Cisco IOS only)
BGP Schematic (IPv4 Unicast)

inet.0

Juniper Networks

172.16.0.0/12

192.168.0.0/16

BGP Path Selection

Routing Policy

Routing Table

BGP Table

172.16.0.0/12 → 192.168.0.0/16
The Autonomous System

- Collection of routers under one administrative control
- Single internal routing protocol
- Identified using an AS Number
AS Numbers

• An ASN is a 16-bit number
  – 1 through 64511 are assigned by RIRs
  – 64512 through 65534 are for private use and should never appear on the Internet
  – Numbers ‘0’ and ‘65535’ are reserved
  – AS 23456 used to represent 4-byte ASN to routers unable to handle the new standard

• All major routing platforms now support 32-bit ASN: http://www.get4byteasrn.info/

• Interesting contrasts between European and US-based R&E networks
Routing Components Depicted

- Circuit 4 (Metric = 10) connects Router A (A.0) to Router D (D.0).
- Circuit 1 (Metric = 10) connects Router A (A.0) to Router D (D.0).
- Circuit 3 (Metric = 50) connects Router C (C.0) to Router D (D.0).
- Circuit 2 (Metric = 10) connects Router C (C.0) to Router B (B.0).
- iBGP Peering between Router B (B.0) and Router C (C.0).
- eBGP Peering between Router B (B.0) and Router K (K.0).
- IS-IS Adjacency between Router A (A.0) and Node A.4.
- IS-IS Adjacency between Router B (B.0) and Node B.1.
- IS-IS Adjacency between Router C (C.0) and Node C.2.
- IS-IS Adjacency between Router D (D.0) and Node D.4.
- IS-IS Adjacency between Router J (J.0) and Node J.5.

Nodes:
- Router A (A.0)
- Router B (B.0)
- Router C (C.0)
- Router D (D.0)
- Router J (J.0)
- Router K (K.0)
Routing Components Explained

• Interior Gateway Protocol (IGP)
• Internal BGP (iBGP)
  – Routes customer prefixes around internal infrastructure
  – Is NOT congruent with physical connectivity
• External BGP (eBGP)
  – Prefix interchange with customers
  – Most routing policy located here
The Golden Rule

• Never redistribute routes from the IGP into BGP
• Never redistribute routes from BGP into the IGP
Part 2: BGP Basics
Introducing Prefixes Into BGP

1. Use *network* statement
   1. With *auto-summary* disabled
   2. With *auto-summary* configured

2. Configure *aggregate* routing

3. Use route maps to redistribute
   1. Prefixes learned from an IGP
   2. Static routes
1.1 The *network* Statement (Cisco)

```plaintext
router bgp 87
  no auto-summary
  network 129.79.0.0
!
ip route 129.79.0.0 255.255.0.0 Null0 200
```

1. There is no mask following the prefix in the ‘network’ statement as we are advertising a classful network

2. The static route serves two important purposes:
   1. The prefix will not be advertised by BGP unless there is an exact match in the IP routing table
   2. Traffic sent to non-existent IP addresses in the range will be silently dropped. This avoids wasting b/w sending ICMP Unreachables and is a valuable defense against scanners and DDoS attacks.
1.2 Using *auto-summary* (Cisco)

```plaintext
router bgp 87
  auto-summary
  network 129.79.0.0
```

1. The prefix will be advertised by BGP providing there is at least one contained prefix in the IP routing table.

2. This IGP-learned prefix can be any length; it does not have to match the classful network that BGP is being asked to advertise.

3. Use this command:

   ```plaintext
   show ip route 129.79.0.0 255.255.0.0 longer-prefixes
   ```
   
   to check whether there is an IGP-learned route. If there are none, then BGP will not advertise the 129.79/16 parent.
3.2 Redistribute Static (Juniper)

```bash
policy-options {
    policy-statement ORIGINATE {
        term Seed {
            from {
                protocol static;
                route-filter 129.79.0.0/16 exact;
            }
            then accept;
        }
    }
}

routing-options {
    static {
        route 129.79.0.0/16 discard;
    }
}
```
Global NOC Recommendation

• Use a ‘network’ statement to originate ARIN allocations to Internet2
• Configure a supporting static route
• Disable the ‘auto-summary’ capability
• Filter more specific contained prefixes using an outbound route-map applied to the peering with Internet2 or NLR
Recommended Routing Policy

- **Should be implemented**
  - Reject any prefix with a private AS in the AS_PATH
  - Reject bogon prefixes (following slide)
- **Consider implementing**
  - Assign higher LOCAL_PREF to Internet2 or NLR prefixes than to commodity.
  - Max prefixes limit on some peers
# BGP Messages

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Open</td>
<td>RFC 4271</td>
</tr>
<tr>
<td>2</td>
<td>Update</td>
<td>RFC 4271</td>
</tr>
<tr>
<td>3</td>
<td>Notification</td>
<td>RFC 4271</td>
</tr>
<tr>
<td>4</td>
<td>Keepalive</td>
<td>RFC 4271</td>
</tr>
<tr>
<td>5</td>
<td>Route-Refresh</td>
<td>RFC 2918</td>
</tr>
</tbody>
</table>
The BGP Update Message

- BGP Header
  - Data
    - NLRI
      - Length
      - Prefix
    - Path Attributes
      - Type
        - Length
        - Value
      - Total Path Attributes Length
        - Length
      - Unfeasible Routes Length
        - Length
        - Prefix
    - Withdrawn Routes
# BGP Path Vector Algorithm

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Which Preferred?</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCAL_PREF</td>
<td>Highest</td>
</tr>
<tr>
<td>AS_PATH</td>
<td>Shortest</td>
</tr>
<tr>
<td>ORIGIN</td>
<td>Lowest</td>
</tr>
<tr>
<td>MED</td>
<td>Lowest</td>
</tr>
<tr>
<td>eBGP &gt; iBGP</td>
<td></td>
</tr>
<tr>
<td>IGP Metric to NEXT_HOP</td>
<td>Lowest</td>
</tr>
<tr>
<td>ROUTER_ID</td>
<td>Lowest</td>
</tr>
<tr>
<td>CLUSTER_LIST</td>
<td>Shortest</td>
</tr>
<tr>
<td>Neighbour IP Address</td>
<td>Lowest</td>
</tr>
</tbody>
</table>
## Attribute Classes

<table>
<thead>
<tr>
<th>Well-Known</th>
<th>Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mandatory</strong></td>
<td><strong>Discretionary</strong></td>
</tr>
<tr>
<td>AS_PATH</td>
<td>LOCAL_PREF</td>
</tr>
<tr>
<td>NEXT_HOP</td>
<td>ATOMIC__AGGREGATE</td>
</tr>
<tr>
<td>ORIGIN</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The AS_PATH Attribute

172.16.0.0/12 200 100

172.16.0.0/12 300 200 100

172.16.0.0/12 100

172.16.0.0/12 400 300 200 100

172.16.0.0/12 500 400 300 200

AS100 detects its own AS in the path from AS500 and ignores the prefix.
Which is the ‘Better’ Path?
Aggregation (With AS_SET)

192.168.0.0/23 [64513 {64512,64514} i]

192.168.0.0/23

192.168.2.0/23
Aggregation (Without AS_SET)
The NEXT_HOP Attribute

Router | NEXT_HOP
-------|---------
A      | —       | —       
B      | A.1     | A.1     
C      | A.1     | B.0     
D      | C.3     | C.3     

172.16.0.0/24
Third Party Next Hop

192.168.17.0/24 → A.1

192.168.17.0/24 → A.1

192.168.17.0/24 → A.1
eBGP and NBMA Networks

192.168.17.0/24 → A.1

Frame Relay
The LOCAL_PREF Attribute

Graphic used with kind permission of Philip Smith, Cisco Systems
Multi-Exit Discriminator

Graphic used with kind permission of Philip Smith, Cisco Systems
MED: Metric Confusion

- MED is non-transitive *and* optional
  - Some implementations send learned MED to iBGP peers by default and others do not
  - Some implementations send MEDs to eBGP peers by default while others do not

- Default metric varies by vendor
  - No explicit metric implies $2^{32}-1$ or $2^{32}-2$
  - No explicit metric implies zero
### eBGP vs iBGP

<table>
<thead>
<tr>
<th>eBGP Features</th>
<th>iBGP Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Uses AS_PATH for loop avoidance</td>
<td>1. Chinese whispers prohibited</td>
</tr>
<tr>
<td>2. NEXT_HOP is modified</td>
<td>2. NEXT_HOP is unchanged</td>
</tr>
<tr>
<td>3. Peers must be directly connected</td>
<td>3. Peers not necessarily directly connected</td>
</tr>
<tr>
<td>4. MED is reset</td>
<td>4. MED is propagated</td>
</tr>
<tr>
<td>5. LOCAL_PREF is never advertised</td>
<td>5. LOCAL_PREF always advertised</td>
</tr>
</tbody>
</table>
Internal Peering Topologies

- Daisy Chain (Wrong)
- Full Mesh (Allowed)
- Route Reflector (Allowed)
How Are Prefixes Passed Around?

• On any given router only the best path for a prefix is passed to other peers
• Best path learned via eBGP
  – Advertised to all other eBGP peers
  – Advertised to iBGP peers
• Best path learned via iBGP
  – Advertised onto eBGP peers
  – **Not** advertised to other iBGP peers
Within Internet2:

1. ‘Next Hop’ for customer prefixes in RIB provided by iBGP and does not change if backbone circuit fails.

2. Entry in FIB provided by IS-IS and depends on backbone connectivity.
Part 3: BGP Path Selection
iBGP Tracks the IGP Metric
NREN and Internet2

Metric 2 = 2932
Metric 2 = 2019

External Peering
Internal Peering
Circuit
Possible Layer-II Connectivity

- Internet2 and NREN routers directly connected
- Connection through a VLAN on a single peering switch
- Connection via VLAN that traverses multiple Layer-II and optical devices
Internet2 Connection to NGIX-W

ESNet

GE

10GE

NGIX-W

VLAN 166

NREN

VLAN 166

SUNN

CI

OC192

SALT

CI

OC192

DENV

CI

OC192

ALBU

CI

10GE

VLAN 166

SALT
Prefer the 10G-Connected Path

• Internet2 router in Seattle
  – Associate a high LOCAL_PREF with prefixes received on external peering with NREN

• Internet2 router in Salt Lake City
  – Receives NREN prefixes over iBGP peering with Seattle with high LOCAL_PREF
  – Now prefers iBGP to SEAT over eBGP via NGIX-W
LOCAL_PREF Gotcha

149.165.128.0/17

149.165.240.64/26

149.165.128.0/17

Indiana Gigapop

Commodity

R&E
Result: Asymmetric Routing!

Fermi $\rightarrow$ UPenn

UPenn $\rightarrow$ Fermi
Cold Potato Routing with MED

FERMI → UPenn

128.175/16 [0]

128.175/16 [1001]

128.175/16 [1000]

128.175/16 [278]

128.175/16 [278]
Steering Inbound Traffic (1/2)
Steering Inbound Traffic (2/2)

**LOCAL_PREF = 100**

- UEN
- FRGP

**LOCAL_PREF = 200**

- Internet2
- NLR

- 129.79/16
- 129.79.9/24
- Indiana
- IU

129.79/16
129.79/16
Part 4: Community Strings
Functions Served by Communities

1. Assign prefixes to pre-defined groups (local significance only)
2. Control how prefix is advertised by peer
3. Control your peer’s LOCAL_PREF for the specific prefix
4. Signal peer to prepend multiple AS numbers to AS_PATH
5. Blackhole all traffic to specific prefix
Expressing A Community

1. A community is just a 32-bit number
2. By convention, the most-significant 16 bits represent an AS number
3. To convert from ‘new’ to ‘old’ formats
   1. Multiply the ‘high’ 16-bits by $2^{16}$
   2. Add the ‘low’ 16-bits to the result

<table>
<thead>
<tr>
<th>JunOS &amp; Cisco ‘New Format’</th>
<th>11537:260</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cisco ‘Old Format’</td>
<td>756089092</td>
</tr>
</tbody>
</table>

JunOS & Cisco ‘New Format’ 11537:260
Cisco ‘Old Format’ 756089092
NO-EXPORT Community

172.16.0.0/12
172.16.1.0/24 +NO-EXPORT
172.16.0.0/16

AS 100

A
B
C

AS 200

D
E
F
G

172.16.0.0/12
172.16.2.0/24 +NO-EXPORT
Using Communities to Create Prefix ‘Sets’

ATLA

128.169.0.0/16
141.225.0.0/16
192.55.208.0/24

Memphis

+14048:10
+14048:10

A
B
C

96.4.0.0/15
128.169.0.0/16
192.55.208.0/24
Routers are optimized for packet forwarding; not packet filtering

Routing to Null saves valuable CPU cycles
Customer-Triggered Blackhole

Inbound Prefix

11537:911

Prefix > /24

Yes

Prefix > /24

Yes

Discard

ISP

No

No

Forward

Customer

Static Route

11537:911

Prefix > /24

Redistribute Into BGP
Customer-Triggered Blackhole (ISP Perspective; Cisco IOS)

interface Null0
   no ip unreachable

ip policy-list BLACKHOLE permit
   match ip address prefix-list 24_TO_32
   match community 10
!
ip community-list 10 permit 756089743
!
ip prefix-list 24_TO_32 seq 5 permit 0.0.0.0/0 ge 24
!
route-map CUSTOMER_IN permit 10
   match policy-list BLACKHOLE
   set community no-export
   set interface Null0
Customer-Triggered Blackhole (Customer Perspective; JunOS)

```text
routing-options {
    static {
        route 192.168.17.15/32 {
            discard;
            community 11537:911;
        }
    }
}

policy-options {
    policy-statement ORIGINATE {
        term BLACKHOLE {
            from {
                protocol static;
                route-filter 0.0.0.0/0 prefix-length-range /24-/32;
                community BLACKHOLE;
            }
            then accept;
        }
        term ... {
            ... 
        }
    }
}
```
Customer-Side Policy: IOS vs JunOS

- **Juniper**
  - Blackhole prefix statically routed to Discard
  - Attach a community tag to the static route
- **Cisco IOS**
  - Blackhole prefix statically routed to Null0
  - Add the new prefix to the ‘blackhole’ prefix list
  - Existing route-map
    - Redistributions ‘blackhole’ prefix list into BGP
    - Attaches the correct community
Part 5: Multiprotocol BGP
Uses of MBGP

- Divergent multicast routing for IPv4
- Layer-III VPNs (aka Virtual Routing and Forwarding)
- Routing Layer-III Protocols Other Than IPv4
Multiprotocol Extensions to BGP

<table>
<thead>
<tr>
<th>AFI</th>
<th>SAFI</th>
<th>Length of Next Hop Address</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Network Address of Next Hop</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NLRI</td>
</tr>
</tbody>
</table>

This diagram has been deliberately simplified with the SNPA fields omitted
Multicast Routing

Unicast traffic AS400 <> AS200
Multicast traffic AS400 <> AS200

Unicast and Multicast IPv4 Prefixes

Unicast IPv4 Prefixes
Source-Specific Multicast

(S,G) Join

IGMPv3 Join

G → R

S → G
hat's All Folks!

Bugs Bunny