Inter-domain Routing

- Autonomous Systems (AS)
- Border Gateway Protocol (BGP)
Autonomous System (AS)

• **Definition of an AS**
  – An autonomous system is an autonomous routing domain
  – the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it.

• Internet consists of thousands of ASes.
  – Our objective is to understand routing between the Autonomous Systems
  – First describe the types of ASes and peering relationships

• Remaining Slides Modified and Re-ordered from first half of Tim’s Tutorial
Nontransit vs. Transit ASes

Internet Service providers (often) have transit networks

Nontransit AS might be a corporate or campus network. Could be a “content provider”

Traffic NEVER flows from ISP 1 through NET A to ISP 2 (At least not intentionally!)

ISP 1  
NET A  
ISP 2
Selective Transit

NET A provides transit between NET B and NET C and between NET D and NET C.

NET A DOES NOT provide transit between NET D and NET B.

Most transit networks transit in a selective manner...
Customers and Providers

Customer pays provider for access to the Internet
Customers Don’t Always Need BGP

Static routing is the most common way of connecting an autonomous routing domain to the Internet. This helps explain why BGP is a mystery to many ...

provider

Nail up routes 192.0.2.0/24 pointing to customer

customer

Nail up default routes 0.0.0.0/0 pointing to provider.

192.0.2.0/24
Customer-Provider Hierarchy

provider → customer

IP traffic
The Peering Relationship

Peers provide transit between their respective customers

Peers do not provide transit between peers

Peers (often) do not exchange $$$

peers

traffic allowed

traffic NOT allowed
Peering also allows connectivity between the customers of “Tier 1” providers.
BGP-4

- BGP = Border Gateway Protocol
- Is a Policy-Based routing protocol
- Is the de facto EGP of today’s global Internet
- Relatively simple protocol, but configuration is complex and the entire world can see, and be impacted by, your mistakes.

- 1989 : BGP-1 [RFC 1105]
  - Replacement for EGP (1984, RFC 904)
- 1990 : BGP-2 [RFC 1163]
- 1991 : BGP-3 [RFC 1267]
- 1995 : BGP-4 [RFC 1771]
  - Support for Classless Interdomain Routing (CIDR)
BGP Operations

Establish session on TCP port 179

Exchange all active routes

Exchange incremental updates

While connection is ALIVE exchange route UPDATE messages
Two Types of BGP Neighbor Relationships

- External Neighbor (eBGP) in a different Autonomous Systems
- Internal Neighbor (iBGP) in the same Autonomous System

iBGP is routed (using IGP!)
iBGP Peers Fully Meshed

- iBGP is needed to avoid routing loops within an AS
- Injecting external routes into IGP does not scale and causes BGP policy information to be lost
- BGP does not provide “shortest path” routing
- Is iBGP an IGP? NO!

iBGP neighbors do not announce routes received via iBGP to other iBGP neighbors.
Four Types of BGP Messages

• **Open**: Establish a peering session.
• **Keep Alive**: Handshake at regular intervals.
• **Notification**: Shuts down a peering session.
• **Update**: Announcing new routes or withdrawing previously announced routes.

announcement = prefix + attributes values
### BGP Attributes

<table>
<thead>
<tr>
<th>Value</th>
<th>Code</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ORIGIN</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>2</td>
<td>AS_PATH</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>3</td>
<td>NEXT_HOP</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>4</td>
<td>MULTI_EXIT_DISC</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>5</td>
<td>LOCAL_PREF</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>6</td>
<td>ATOMIC_AGGREGATE</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>7</td>
<td>AGGREGATOR</td>
<td>[RFC1771]</td>
</tr>
<tr>
<td>8</td>
<td>COMMUNITY</td>
<td>[RFC1997]</td>
</tr>
<tr>
<td>9</td>
<td>ORIGINATOR_ID</td>
<td>[RFC2796]</td>
</tr>
<tr>
<td>10</td>
<td>CLUSTER_LIST</td>
<td>[RFC2796]</td>
</tr>
<tr>
<td>11</td>
<td>DPA</td>
<td>[Chen]</td>
</tr>
<tr>
<td>12</td>
<td>ADVERTISER</td>
<td>[RFC1863]</td>
</tr>
<tr>
<td>13</td>
<td>RCID_PATH / CLUSTER_ID</td>
<td>[RFC1863]</td>
</tr>
<tr>
<td>14</td>
<td>MP_REACH_NLRI</td>
<td>[RFC2283]</td>
</tr>
<tr>
<td>15</td>
<td>MP_UNREACH_NLRI</td>
<td>[RFC2283]</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>reserved for development</td>
<td></td>
</tr>
</tbody>
</table>

Not all attributes need to be present in every announcement.

From IANA: [http://www.iana.org/assignments/bgp-parameters](http://www.iana.org/assignments/bgp-parameters)

Tim’s tutorial will cover these attributes.
Attributes are Used to Select Best Routes

Given multiple routes to the same prefix, a BGP speaker must pick at most one best route (Note: it could reject them all!)
So Many Choices

Which route should Frank pick to 13.13.0.0/16?
BGP Route Processing

Receive BGP Updates
- Apply Import Policies
- Based on Attribute Values
  - Best Routes
  - Apply Policy = filter routes & tweak attributes
- Best Route Selection
- Best Route Table
- Install forwarding Entries for best Routes.
- IP Forwarding Table
- Transmit BGP Updates
- Apply Export Policies

Open ended programming. Constrained only by vendor configuration language

Apply Policy = filter routes & tweak attributes

Based on Attribute Values

Best Routes

Apply Policy = filter routes & tweak attributes
Route Selection Process

- Highest Local Preference
- Shortest ASPATH
- Lowest MED
- i-BGP < e-BGP
- Lowest IGP cost to BGP egress
- Lowest router ID

Enforce relationships
traffic engineering
Throw up hands and break ties
Prefix Originated
BGP at AS YYY will never accept a route with ASPATH containing YYY.
Traffic Often Follows ASPATH

IP Packet
Dest = 135.207.44.66
... But It Might Not

AS 2 filters all subnets with masks longer than /24.

From AS 4, it may look like this packet will take path 3 2 1, but it actually takes path 3 2 5.
In fairness: could you do this “right” and still scale?

Exporting internal state would dramatically increase global instability and amount of routing state
Shedding Inbound Traffic with ASPATH Padding Hack

Padding will (usually) force inbound traffic from AS 1 to take primary link

192.0.2.0/24 ASPATH = 2

192.0.2.0/24 ASPATH = 2 2 2

Padding will (usually) force inbound traffic from AS 1 to take primary link
Padding May Not Shut Off All Traffic

AS 1
provider

customer
AS 2

AS 3
provider

192.0.2.0/24
ASPATH = 2

192.0.2.0/24
ASPATH = 2 2 2 2 2 2 2 2 2 2 2 2 2 2

AS 3 will send traffic on “backup” link because it prefers customer routes and local preference is considered before ASPATH length!

Padding in this way is often used as a form of load balancing.
Route Selection Process

- Highest Local Preference
- Shortest AS PATH
- Lowest MED
- i-BGP < e-BGP
- Lowest IGP cost to BGP egress
- Lowest router ID

Enforce relationships
traffic engineering
Throw up hands and break ties
COMMUNITY Attribute to the Rescue!

Customer import policy at AS 3:
- If 3:90 in COMMUNITY then set local preference to 90
- If 3:80 in COMMUNITY then set local preference to 80
- If 3:70 in COMMUNITY then set local preference to 70

AS 3: normal customer local pref is 100, peer local pref is 90
Back to Frank ...

Local preference only used in iBGP

Higher Local preference values are more preferred