Exterior Gateway Protocols: EGP, BGP-4, CIDR

Shivkumar Kalyanaraman
Rensselaer Polytechnic Institute
shivkuma@ecse.rpi.edu
http://www.ecse.rpi.edu/Homepages/shivkuma

Based in part upon slides of Tim Griffin (AT&T), Ion Stoica (UCB), J. Kurose (U Mass), Noel Chiappa (MIT), Jennifer Rexford (Princeton)
Cores, Peers, and the limit of default routes
- Autonomous systems & EGP
- BGP4
- CIDR: reducing router table sizes

Refs: Chap 10, 14, 15. Books: “Routing in Internet” by Huitema, “Interconnections” by Perlman, “BGP4” by Stewart, Sam Halabi, Danny McPherson, Internet Routing Architectures

Reading: Geoff Huston, Commentary on Inter-domain Routing in the Internet
Reference: BGP-4 Standards Document: In TXT

Reading: Norton, Internet Service Providers and Peering
Reading: Labovitz et al, Delayed Internet Routing Convergence
Reference: Paxson, End-to-End Routing Behavior in the Internet,
Reading: Interdomain Routing: Additional Notes: In PDF | In MS Word
Reference Site: Griffin, Interdomain Routing Links
Intra-AS and Inter-AS routing

Gateways:
- perform inter-AS routing amongst themselves
- perform intra-AS routing with other routers in their AS

inter-AS, intra-AS routing in gateway A.c
History: Default Routes: limits

- Default routes => partial information
- Routers/hosts w/ default routes rely on other routers to complete the picture.
- In general routing “signposts” should be:
  - Consistent, i.e., if packet is sent off in one direction then another direction should not be more optimal.
  - Complete, i.e., should be able to reach all destinations
Core

- A small set of routers that have consistent & complete information about all destinations.
- Outlying routers can have partial information provided they point default routes to the core.
- Partial info allows site administrators to make local routing changes independently.
Peer Backbones

- Initially NSFNET had only one connection to ARPANET (router in Pittsburg) => only one route between the two.
- Addition of multiple interconnections => multiple possible routes => need for dynamic routing
- Single core replaced by a network of peer backbones => more scalable
  - Today there are over 30 backbones!
- Routing protocol at cores/peers: GGP -> EGP-> BGP-4
Today’s Big Picture

Large ISP

Large ISP

Small ISP

Large number of diverse networks
Internet AS Map: caida.org
Purpose of EGP

Share connectivity information across ASes
Who speaks Inter-AS routing?

- Two types of routers
  - Border router (Edge), Internal router (Core)
  - Two border routers of different ASes will have a BGP session
Intra-AS vs Inter-AS

- An AS is a routing domain
- Within an AS:
  - Can run a link-state routing protocol
  - Trust other routers
  - Scale of network is relatively small
- Between ASes:
  - Lack of information about other AS’s network (Link-state not possible)
  - Crossing trust boundaries
  - Link-state protocol will not scale
  - Routing protocol based on route propagation
Requirements for Inter-AS Routing

- Should **scale** for the size of the global Internet.
  - Focus on **reachability**, not optimality
  - Use **address aggregation** techniques to minimize core routing table sizes and associated control traffic
  - At the same time, it should allow **flexibility in topological structure** (eg: don’t restrict to trees etc)

- Allow **policy-based routing** between autonomous systems
  - Policy refers to **arbitrary preference among a menu of available routes** (based upon routes’ **attributes**)
  - Fully distributed routing (as opposed to a signaled approach) is the only possibility.
  - Extensible to meet the demands for newer policies.
Autonomous System (AS)

- Internet is not a single network
  - Collection of networks controlled by different administrations
- An autonomous system is a network under a single administrative control
- An AS owns an IP prefix
- Every AS has a unique AS number
- ASes need to inter-network themselves to form a single virtual global network
  - Need a common protocol for communication
Autonomous Systems (ASes)

- An autonomous system is an autonomous routing domain that has been assigned an Autonomous System Number (ASN).
- All parts within an AS remain connected.

… 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.

RFC 1930: Guidelines for creation, selection, and registration of an Autonomous System
IP Address Allocation and Assignment: Internet Registries

Allocate to National and local registries and ISPs
Addresses assigned to customers by ISPs

IANA
www.iana.org

ARIN
www.arin.org

RIPE
www.ripe.org

APNIC
www.apnic.org

RFC 2050 - Internet Registry IP Allocation Guidelines
RFC 1918 - Address Allocation for Private Internets
RFC 1518 - An Architecture for IP Address Allocation with CIDR
Shivkumar Kalyanaraman
Rensselaer Polytechnic Institute

AS Numbers (ASNs)

ASNs are 16 bit values.
64512 through 65535 are “private”
Currently over 11,000 in use.

- Genuity: 1
- MIT: 3
- Harvard: 11
- UC San Diego: 7377
- AT&T: 7018, 6341, 5074, ...
- UUNET: 701, 702, 284, 12199, ...
- Sprint: 1239, 1240, 6211, 6242, ...
- ...

ASNs represent units of routing policy
AS != Institution

- Not equivalent to an AS
  - Many institutions span multiple autonomous systems
  - Some institutions do not have their own AS number
  - Ownership of an AS may be hard to pinpoint (whois)
- Not equivalent to a block of IP addresses (prefix)
  - Many institutions have multiple (non-contiguous) prefixes
  - Some institutions are a small part of a larger address block
  - Ownership of a prefix may be hard to pinpoint (whois)
- Not equivalent to a domain name (att.com)
  - Some sites may be hosted by other institutions
  - Some institutions have multiple domain names (att.net)
Characteristics of the AS Graph

- AS graph structure
  - High variability in node degree ("power law")
  - A few very highly-connected ASes
  - Many ASes have only a few connections

Graph:
- CCDF
- AS degree
- All ASes have degree $\geq 1$
- Very few have degree $\geq 100$
Where to Get BGP Routes: Public Servers

BGP sessions
Nontransit vs. Transit ASes

ISP 1

NET A

ISP 2

Internet Service providers (ISPs) have transit networks.

Traffic NEVER flows from ISP 1 through NET A to ISP 2.

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

Shivkumar Kalyanaraman
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 ASes allow only selective transit: impact of commercialization.
Customers and Providers

Customer pays provider for access to the Internet
Customer-Provider Hierarchy
The Peering Relationship

- Peers provide transit between their respective customers
- Peers do not provide transit between peers
- Peers (often) do not exchange $$$

Traffic allowed
Traffic NOT allowed

peer peer
provider customer
AOL’s Settlement-Free Interconnection Policy

- Operational requirements on a peer network
  - Handle a single-node outage w/o traffic impact
  - Single AS number
  - Network Operations Center staffed at all times
- Backbone capacity
  - At least 10 gigabits/sec between 8 or more cities
  - Minimum peering link speed of 622 megabits/sec
- Peering locations (in U.S.)
  - At least four locations
  - Must include D.C. area, middle of country, Bay area, and NYC or Atlanta
AOL Routing Requirements

- Consistent advertisements
  - All customer routes
  - At all peering points
  - With the same AS path length

- Address blocks
  - Routes aggregated as much as possible
  - No address blocks smaller than /24
  - Address blocks are registered (e.g., with ARIN)

- No default routing
  - Only send traffic to destinations AOL advertises
Peering Wars

Peer
- Reduces upstream transit costs
- Can increase end-to-end performance
- May be the only way to connect your customers to some part of the Internet (“Tier 1”)

Don’t Peer
- You would rather have customers
- Peers are usually your competition
- Peering relationships may require periodic renegotiation

Peering struggles are by far the most contentious issues in the ISP world!

Peering agreements are often confidential.

Rensselaer Polytechnic Institute
Shivkumar Kalyanaraman
Recall: Distributed Routing Techniques

**Link State**
- Topology information is flooded within the routing domain
- Best end-to-end paths are computed locally at each router.
- Best end-to-end paths determine next-hops.
- Based on minimizing some notion of distance
- Works only if policy is shared and uniform
- Examples: OSPF, IS-IS

**Vectoring**
- Each router knows little about network topology
- Only best next-hops are chosen by each router for each destination network.
- Best end-to-end paths result from composition of all next-hop choices
- Does not require any notion of distance
- Does not require uniform policies at all routers
- Examples: RIP, BGP
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)
Border Gateway Protocol (BGP) Model

- ASes exchange info about who they can reach
  - IP prefix: block of destination IP addresses
  - AS path: sequence of ASes along the path
- Policies configured by the AS’s operator
- Path selection: which of the paths to use?
- Path export: which neighbors to tell?

```
12.34.158.0/24: path (2,1)
```

```
12.34.158.0/24: path (1)
```

Data traffic flow:
- From AS 3 to AS 2
- From AS 2 to AS 1

Shivkumar Kalyanaraman
BGP Operations (Simplified)

1. Establish session on TCP port 179
2. Exchange all active routes
3. Exchange incremental updates

While connection is ALIVE exchange route UPDATE messages
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
Border Gateway Protocol (BGP)

- Allows multiple cores and arbitrary topologies of AS interconnection.
  - Uses a path-vector concept which enables loop prevention in complex topologies.
- In AS-level, shortest path may not be preferred for policy, security, cost reasons.
  - Different routers have different preferences (policy) => as packet goes thru network it will encounter different policies.
- Bellman-Ford/Dijkstra don’t work!
- BGP allows attributes for AS and paths which could include policies (policy-based routing).
BGP (Cont’d)

- When a BGP Speaker A advertises a prefix to its B that it has a path to IP prefix C, B can be certain that A is actively using that AS-path to reach that destination.
- BGP uses TCP between 2 peers (reliability)
  - Exchange entire BGP table first (50K+ routes!)
  - Later exchanges only incremental updates
- Application (BGP)-level keepalive messages
- Hold-down timer (at least 3 sec) locally config
- Interior and exterior peers: need to exchange reachability information among interior peers before updating intra-AS forwarding table.
Border routers

- Border router
  - Learns BGP route from neighbor AS
  - Creates forwarding-table entry for prefix
- But, how do the other routers get there?

Border router: 12.34.158.0/24
How do Other Routers Learn the BGP Route?

- Internal BGP
  - iBGP sessions between the routers
  - Allows other routers to get the big picture
  - Simplest case: “full mesh” of iBGP sessions
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!)
How To Get to the Egress Router?

- Interior Gateway Protocol (OSPF/IS-IS)
  - Routers flood information to learn topology
  - Routers determine “next hop” to other routers…
  - Compute shortest paths based on the link weights
  - Link weights configured by the operator

“Use Serial0/0.1 to get to the red router”
Constructing the Forwarding Table

- Three protocols
  - External BGP: learn the external route
  - Internal BGP: propagate inside the AS
  - IGP: learn outgoing link on path to other router
- Router joins the data
  - Prefix 12.34.158.0/24 reached through red router
  - Red router reached via link Serial0/0.1
  - Forwarding entry: 12.34.158.0/24 → Serial0/0.1
- Router forwards packets
  - Lookup destination 12.34.158.5 in table
  - Forward packet out link Serial0/0.1
What if There are Multiple Choices?

Hot-potato routing

This router has two BGP routes to 192.44.78.0/24.

Hot potato: get traffic off of your network as soon as possible. Go for egress 1!
Routing is Not Symmetric

Web request and TCP ACKs

Web response
Revisit: I-BGP

- Why is IGP (OSPF, ISIS) not used?
  - In large ASs full route table is very large (100K routes!)
  - Rate of change of routes is frequent
  - Tremendous amount of control traffic
  - Not to mention Dijkstra computation being evoked for any change...
  - BGP policy information may be lost

- I-BGP: Within an AS
  - Same protocol/state machines as EBGP
  - But different rules about advertising prefixes
  - Prefix learned from an I-BGP neighbor cannot be advertised to another I-BGP neighbor to avoid looping
    => need full IBGP mesh!
  - AS-PATH cannot be used internally. Why?
iBGP Peers: Fully Meshed

- iBGP is needed to avoid routing loops within an AS
- Full Mesh =>
  - Independent of physical connectivity.
  - Single link may see same update multiple times!
  - iBGP neighbors do not announce routes received via iBGP to other iBGP neighbors.
- Is iBGP an IGP? NO!
- Set of neighbor relationships to transfer BGP info
IBGP Scaling: *Route Reflection*

- Add hierarchy to I-BGP
- **Route reflector:** A router whose BGP implementation supports the re-advertisement of routes between I-BGP neighbors
- **Route reflector client:** A router which depends on route reflector to re-advertise its routes to entire AS and learn routes from the route reflector
Route Reflection

AS1

128.23.0.0/16

RR-C1
RR-C2
RR1
RR2
RR3
RR-C3
RR-C4

ER

AS2

10.0.0.0/24

EBGP
IBGP

Shivkumar Kalyanaraman
Rensselaer Polytechnic Institute
AS Confederations

- Divide and conquer: Divides a large AS into sub-ASs
CIDR

- **Shortage of class Bs** => give out a set of class Cs instead of one class B address
  - Problem: every class C n/w needs a routing entry!

- Solution: *Classless* Inter-domain Routing (CIDR).
  - Also called “supernetting”
  - Key: allocate addresses such that they can be summarized, i.e., contiguously.
  - Share same higher order bits (i.e. prefix)
  - Routing tables and protocols must be capable of carrying a subnet mask. **Notation: 128.13.0/23**

- When an **IP address matches multiple entries** (eg 194.0.22.1), choose the one which had the longest mask (“longest-prefix match”)
RFC 1519: Classless Inter-Domain Routing (CIDR)

Pre-CIDR: Network ID ended on 8-, 16, 24- bit boundary
CIDR: Network ID can end at any bit boundary

IP Address : 12.4.0.0       IP Mask: 255.254.0.0

Usually written as 12.4.0.0/15, a.k.a “supernetting”
Understanding Prefixes and Masks (Recap)

12.5.9.16 is covered by prefix 12.4.0.0/15

12.5.9.16
00001100 00000101 00001001 00010000

12.4.0.0/15
00001100 00000100 00000000 00000000
11111111 11111110 00000000 00000000

12.7.9.16
00001100 00000111 00001001 00010000

12.7.9.16 is not covered by prefix 12.4.0.0/15
Inter-domain Routing Without CIDR

204.71.0.0
204.71.1.0
204.71.2.0
.............
204.71.255.0

Service Provider

Global Internet Routing Mesh

.............
204.71.255.0

Inter-domain Routing With CIDR

204.71.0.0
204.71.1.0
204.71.2.0
.............
204.71.255.0

Service Provider

204.71.0.0/16

Global Internet Routing Mesh

Shivkumar Kalyanaraman
Longest Prefix Match (Classless) Forwarding

Destination = 12.5.9.16

payload

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Next Hop</th>
<th>Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.0.0/0</td>
<td>10.14.11.33</td>
<td>ATM 5/0/9</td>
</tr>
<tr>
<td>12.0.0.0/8</td>
<td>10.14.22.19</td>
<td>ATM 5/0/8</td>
</tr>
<tr>
<td>12.4.0.0/15</td>
<td>10.1.3.77</td>
<td>Ethernet 0/1/3</td>
</tr>
<tr>
<td>12.5.8.0/23</td>
<td>attached</td>
<td>Serial 1/0/7</td>
</tr>
</tbody>
</table>
What is Routing Policy

- Policy refers to *arbitrary preference among a menu of available routes* (based upon routes’ *attributes*)
  - Public description of the relationship between external BGP peers
  - Can also describe internal BGP peer relationship
- Eg: Who are my BGP peers
- What routes are
  - Originated by a peer
  - Imported from each peer
  - Exported to each peer
  - Preferred when multiple routes exist
- What to do if no route exists?
Attributes are Used to Select Best Routes

Given multiple routes to the same IP prefix, a BGP speaker must pick at most one best route

(Note: it could reject them all! Or have arbitrary preference based upon route attributes)
BGP Policy Knob: 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>[RFC1771]</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>RCVD_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>16</td>
<td>EXTENDED COMMUNITIES</td>
<td>[Rosen]</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>reserved for development</td>
<td></td>
</tr>
</tbody>
</table>

From IANA: http://www.iana.org/assignments/bgp-parameters

We will cover a subset of these attributes.

Not all attributes need to be present in every announcement.

Shivkumar Kalyanaraman
BGP Route Processing

- **Receive BGP Updates**
  - Apply Policy = filter routes & tweak attributes

- **Based on Attribute Values**
  - Best Routes

- **Best Route Selection**

- **Best Route Table**

- **Apply Policy = filter routes & tweak attributes**
  - Transmit BGP Updates

- **IP Forwarding Table**
  - Install forwarding Entries for best Routes.

- **Apply Import Policies**

- **Apply Export Policies**

- **Receive BGP Updates**

- **Apply Policy = filter routes & tweak attributes**

- **Based on Attribute Values**

- **Best Routes**

- **Apply Policy = filter routes & tweak attributes**

- **Transmit BGP Updates**

- **Apply Import Policies**

- **Apply Export Policies**
Import and Export Policies

- For **inbound** traffic
  - Filter outbound routes
  - Tweak attributes on outbound routes in the hope of influencing your neighbor’s best route selection

- For **outbound** traffic
  - Filter inbound routes
  - Tweak attributes on inbound routes to influence best route selection

**In general, an AS has more control over outbound traffic**
Import and Export Policies

- Inbound filtering controls outbound traffic
  - filters route updates received from other peers
  - filtering based on IP prefixes, AS_PATH, community
- Outbound Filtering controls inbound traffic
  - forwarding a route means others may choose to reach the prefix through you
  - not forwarding a route means others must use another router to reach the prefix
- Attribute Manipulation
  - Import: LOCAL_PREF (manipulate trust)
  - Export: AS_PATH and MEDs
Policy Implementation Flow

Incomming

Adj RIB
In

Main BGP RIB

Outgoing

Adj RIB
Out

IGPs

Main RIB/FIB

Static & HW Info

Rensselaer Polytechnic Institute

Shivkumar Kalyanaraman
Conceptual Model of BGP Operation

- **RIB**: Routing Information Base
- **Adj-RIB-In**: Prefixes learned from neighbors. As many Adj-RIB-In as there are peers
- **Loc-RIB**: Prefixes selected for local use after analyzing Adj-RIB-Ins. This RIB is advertised internally.
- **Adj-RIB-Out**: Stores prefixes advertised to a particular neighbor. As many Adj-RIB-Out as there are neighbors
### UPDATE message in BGP

- Primary message between two BGP speakers.
- Used to **advertise/withdraw** IP prefixes (NLRI).
- **Path attributes** field: unique to BGP.
  - Apply to all prefixes specified in NLRI field.
  - Optional vs **Well-known**; Transitive vs Non-transitive.

<table>
<thead>
<tr>
<th><strong>2 octets</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Withdrawn Routes Length</td>
</tr>
<tr>
<td>Withdrawn Routes (variable length)</td>
</tr>
<tr>
<td>Total Path Attributes Length</td>
</tr>
<tr>
<td>Path Attributes (variable length)</td>
</tr>
<tr>
<td>Network Layer Reachability Info. (NLRI: variable length)</td>
</tr>
</tbody>
</table>
BGP *Route Selection Process*

*Series of tie-breaker decisions...*

- IfNEXTHOPis inaccessible do not consider the route.
- Prefer largest LOCAL-PREF
- If same LOCAL-PREF prefer the shortest AS-PATH.
- If all paths are external prefer the lowest ORIGIN code (IGP<EGP<INCOMPLETE).
- If ORIGIN codes are the same prefer the lowest MED.
- If MED is same, prefer min-cost NEXT-HOP
- If routes learned from EBGP or IBGP, prefer paths learnt from EBGP
- Final tie-break: Prefer the route with I-BGP ID (IP address)
Route Selection Summary

- **Highest Local Preference**: Enforce relationships
- **Shortest AS PATH**
- **Lowest MED**
- **i-BGP < e-BGP**
- **Lowest IGP cost to BGP egress**: traffic engineering
- **Lowest router ID**: Throw up hands and break ties
Path Attributes: ORIGIN

- ORIGIN:
  - Describes how a prefix came to BGP at the origin AS
  - Prefixes are learned from a source and "injected" into BGP:
    - Directly connected interfaces, manually configured static routes, dynamic IGP or EGP
  - Values:
    - IGP (EGP): Prefix learnt from IGP (EGP)
    - INCOMPLETE: Static routes
Path Attributes: AS-PATH

- List of **ASs** thru which the prefix announcement has passed. AS on path adds ASN to AS-PATH.
- Eg: 138.39.0.0/16 originates at AS1 and is advertised to AS3 via AS2.
- Eg: AS-SEQUENCE: “100 200”
- Used for loop detection and path selection.
Traffic Often Follows ASPATH

AS 1
135.207.0.0/16

AS 2

AS 3

AS 4

135.207.0.0/16
ASPATH = 3 2 1

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

AS 2 filters all subnets with masks longer than /24

IP Packet
Dest = 135.207.44.66

From AS 4, it may look like this packet will take path 3 2 1, but it actually takes path 3 2 5
Shorter AS-PATH Doesn’t Mean Shorter # Hops

BGP says that path 4 1 is better than path 3 2 1

Duh!
Path Attributes: NEXT-HOP

- Next-hop: node to which packets must be sent for the IP prefixes. **May not be same as peer.**
- **UPDATE** for 180.20.0.0, **NEXT-HOP** = 170.10.20.3
If routes (prefix) are learnt thru iBGP, NEXT-HOP is the iBGP router which originated the route.

- Note: iBGP peer might be several IP-level hops away as determined by the IGP
- Hence BGP NEXT-HOP is not the same as IP next-hop
- BGP therefore checks if the “NEXT-HOP” is reachable through its IGP.
- If so, it installs the IGP next-hop for the prefix
- This process is known as “recursive lookup” – the lookup is done in the control-plane (not data-plane) before populating the forwarding table.
- Example in next slide
Join EGP with IGP For Connectivity

**Forwarding Table**

<table>
<thead>
<tr>
<th>Destination</th>
<th>Next Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.0.2.0/30</td>
<td>10.10.10.10</td>
</tr>
<tr>
<td>135.207.0.0/16</td>
<td>192.0.2.1</td>
</tr>
</tbody>
</table>

**EGP**

<table>
<thead>
<tr>
<th>Destination</th>
<th>Next Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>135.207.0.0/16</td>
<td>192.0.2.1</td>
</tr>
</tbody>
</table>

**192.0.2.1**

**Next Hop = 192.0.2.1**

**135.207.0.0/16**

**135.207.0.0/16**

**192.0.2.0/30**
Real World: Multiple Links Between Domains

Client

Web server

Multiple links

Middle of path
Hot-Potato Routing

- All traffic from customer to peers
- All traffic to customer prefixes with multiple connections

multiple egress points

San Francisco

New York

Dallas

dest

ISP network
BGP Decision Process

- Highest local preference
- **Lowest AS path length**
- Lowest origin type
- Lowest MED (with same next hop AS)
- **Lowest IGP cost to next hop**
- Lowest router ID of BGP speaker

“Equally good”
Motivations for Hot-Potato Routing

- Simple computation for the routers
  - IGP path costs are already computed
  - Easy to make a direct comparison
- Ensures consistent forwarding paths
  - Next router in path picks same egress point
- Reduces resource consumption
  - Get traffic out as early as possible
  - (But, what does IGP distance really mean???)
Hot-Potato Routing Change

- failure
- planned maintenance
- traffic engineering

Consequences:
- Transient forwarding instability
- Traffic shift
- Interdomain routing changes

Routes to thousands of destinations switch egress points!!!

Shivkumar Kalyanaraman
Load-Balancing Knobs in BGP

- **LOCAL-PREF**: outbound traffic, local preference (box-level knob)
- **MED**: Inbound-traffic, typically from the same ISP (link-level knob) AS1
Path Attribute: LOCAL-PREF

- Locally configured indication about which path is preferred to exit the AS in order to reach a certain network. Default value = 100. Higher is better.
Why Inbound Traffic is Hard to Manage

- Other ASes decide how to send to you
  - Destination-based routing
- Other ASes decide which path to take
- Based on their own policies

AS 2 doesn’t know how AS 1 will send traffic toward p.
AS Prepending

- Artificial inflate AS path length
  - Prepend your own AS in the path
  - E.g., turn “3 4 5” into “3 3 3 4 5”
  - Hope to make the path less attractive
ASPATH Padding: Shed inbound traffic

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

192.0.2.0/24
ASPATH = 2

primary

192.0.2.0/24
ASPATH = 2 2 2

backup

192.0.2.0/24

customer

provider
Padding May Not Shut Off All Traffic

AS 1
provider

192.0.2.0/24
ASPATH = 2

AS 3
provider

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

AS 2

primary

backup

customer

192.0.2.0/24

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.
Multiple Exit Discriminator (MED)

- Tell your neighbor what you want
  - MED attribute to indicate receiver preference
  - Decision process picks route with smallest MED
  - Can use MED for “cold potato” routing
  - But, have to get your neighbor to accept MEDs

```
1
```

“3 4 5” with MED=2

```
3
```

“3 4 5” with MED=1
This Router has two BGP routes to 192.44.78.0/24.

Hot potato: get traffic off of your network as Soon as possible. Go for egress 1!
Getting Burned by the Hot Potato

Many customers want their provider to carry the bits!

Shivkumar Kalyanaraman
Rensselaer Polytechnic Institute

High bandwidth Provider backbone

Low b/w customer backbone

Heavy Content Web Farm

San Diego

SFF

NYC

2865

17

15

56

tiny http request

huge http reply
Cold Potato Routing with MEDs (Multi-Exit Discriminator Attribute)

This means that MEDs must be considered BEFORE IGP distance!

Note1: some providers will not listen to MEDs

Note2: MEDs need not be tied to IGP distance
MEDs Can Export Internal Instability

Shivkumar Kalyanaraman
If AS 1 does not announce the more specific prefix, then most traffic to AS 2 will go through AS 3 because it is a longer match.

AS 2 is "punching a hole" in the CIDR block of AS 1 => subverts CIDR.
CIDR at Work, No load balancing

Table at ISP3

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Next Hop</th>
<th>ORIGIN AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>128.32/11</td>
<td>ISP1</td>
<td>ISP1</td>
</tr>
<tr>
<td>140.64/10</td>
<td>ISP2</td>
<td>ISP2</td>
</tr>
</tbody>
</table>

Diagram:
- AS1: 128.40/16, 140.127/16
- ISP1: 128.32/11
- ISP2: 140.64/10
- ISP3: 140.127/16

Links:
- Link A: 128.40/16
- Link B: 140.127/16
CIDR Subverted for Load Balancing

Table at ISP3

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Next Hop</th>
<th>ORIGIN AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>128.32/11</td>
<td>ISP1</td>
<td>ISP1</td>
</tr>
<tr>
<td>140.64/10</td>
<td>ISP2</td>
<td>ISP2</td>
</tr>
<tr>
<td>140.255.20/24</td>
<td>ISP1</td>
<td>AS1</td>
</tr>
<tr>
<td>128.42.10/24</td>
<td>ISP2</td>
<td>AS1</td>
</tr>
</tbody>
</table>

Diagram:
- AS1
  - 128.40/16
  - 140.127/16
- ISP1
  - 128.32/11
  - 140.255.20/24
  - 128.42.10/24
- ISP2
  - 140.64/10
  - 128.40/16
  - 140.127/16
- ISP3
  - 140.255.20/24
Inter-AS Negotiation

- Better to cooperate?
  - Negotiate where to send
  - Inbound and outbound
  - Mutual benefits
- But, how to do it?
  - What info to exchange?
  - How to prioritize the many choices?
  - How prevent cheating?
- Open research territory
How Can Routes be Colored?

BGP Communities

A community value is 32 bits

By convention, first 16 bits is ASN indicating who is giving it an interpretation

Community Attribute = a list of community values. (So one route can belong to multiple communities)

Two reserved communities

RFC 1997 (August 1996)

- no_export = 0xFFFFFFFF01: don’t export out of AS
- no_advertise 0xFFFFFFFF02: don’t pass to BGP neighbors
Communities Example

- 1:100
  - Customer routes
- 1:200
  - Peer routes
- 1:300
  - Provider Routes

- To Customers
  - 1:100, 1:200, 1:300
- To Peers
  - 1:100
- To Providers
  - 1:100

Import

Export

AS 1

Shivkumar Kalyanaraman

Rensselaer Polytechnic Institute
Inbound Traffic: RFC 1998 on BGP Communities

- Provider and customer agree on a “tag”
  - One tag means “primary” and the other “backup”
  - Customer includes tags in BGP advertisements
  - Provider sets local preference based on tags
- BGP community attribute
  - Opaque attribute with no real meaning
    - Two numbers: usually AS number and arbitrary number
  - Sprint example (http://www.sprint.net/policy/bgp.html)
    - 1239:70 means “assign local pref of 70”
    - ...
    - 1239:110 means “assign local pref of 110”
Example: Tier-1 ISP Setting Local-Preference

- **Customers**
  - 110: Primary path
  - 100: Secondary path
  - 80: Primary backup path
  - 70: Secondary backup path

- **Peers**
  - 81-99: In between
  - Range for traffic engineering
Route Selection Summary

Highest Local Preference
- Enforce relationships

Shortest AS PATH

Lowest MED

i-BGP < e-BGP

Lowest IGP cost to BGP egress

Lowest router ID
- Throw up hands and break ties
BGP Route Selection Process

Series of tie-breaker decisions...

- If NEXTHOP is inaccessible do not consider the route.
- Prefer largest LOCAL-PREF
- If same LOCAL-PREF prefer the shortest AS-PATH.
- If all paths are external prefer the lowest ORIGIN code (IGP<EGP<INCOMPLETE).
- If ORIGIN codes are the same prefer the lowest MED.
- If MED is same, prefer min-cost NEXT-HOP
- If routes learned from EBGP or IBGP, prefer paths learnt from EBGP
- Final tie-break: Prefer the route with I-BGP ID (IP address)
Caveat

- **BGP is not guaranteed** to converge on a stable routing. Policy interactions could lead to “livelock” protocol oscillations.

  See “Persistent Route Oscillations in Inter-domain Routing” by K. Varadhan, R. Govindan, and D. Estrin. ISI report, 1996

- **Corollary:** BGP is not guaranteed to recover from network failures.
Large BGP Tables Considered Harmful

• Routing tables must store best routes and alternate routes
• Burden can be large for routers with many alternate routes (route reflectors for example)
• Routers have been known to die
• Increases CPU load, especially during session reset
Dealing with ASN growth…

- Make ASNs larger than 16 bits
  - How about 32 bits?
  - See Internet Draft: “BGP support for four-octet AS number space” (draft-ietf-idr-as4bytes-03.txt)
  - Requires protocol change and wide deployment

- Change the way ASNs are used
  - Allow multihomed, non-transit networks to use private ASNs
  - Uses ASE (AS number Substitution on Egress )
  - See Internet Draft: “Autonomous System Number Substitution on Egress” (draft-jhaas-ase-00.txt)
  - Works at edge, requires protocol change (for loop prevention)
Daily Update Count

Prefixes announced + Prefixes withdrawn, June 2001 (data source = RIPE NCC)
A Few Bad Apples ...

Typically, 80% of the updates are for less than 5% of the prefixes.

Most prefixes are stable most of the time. On this day, about 83% of the prefixes were not updated.

Data source: RIPE NCC

Thanks to Madanlal Musuvathi for this plot.
Squashing Updates

- **Rate limiting** on sending updates
  - Send batch of updates every MinRouteAdvertisementInterval seconds (± random fuzz)
  - Default value is 30 seconds
  - A router can change its mind about best routes many times within this interval without telling neighbors

- **Route Flap Dampening**
  - Punish routes for misbehaving

Effective in dampening oscillations inherent in the vectoring approach

Must be turned on with configuration
Routes are given a penalty for changing. If penalty exceeds suppress limit, the route is dampened. When the route is not changing, its penalty decays exponentially. If the penalty goes below reuse limit, then it is announced again.

- Can dramatically reduce the number of BGP updates
- Requires additional router resources
- Applied on eBGP inbound only
Persistent Routing Changes

- **Causes**
  - Link with intermittent connectivity
  - Congestion causing repeated session resets
  - Persistent oscillation due to policy conflicts

- **Effects**
  - Lots of BGP update messages
  - Disruptions to data traffic
  - High overhead on routers

- **Solution**
  - Suppress paths that go up/down repeatedly
  - ... to avoid updates and prefer stable paths
Route Flap Dampening Example

Route Flap Dampening (halflife = 15 minutes) -- 6 flaps, one every 10 minutes

penalty
suppress limit
reuse limit

route dampened for nearly 1 hour

penalty for each flap = 1000
Route Flap Damping

- BGP-speaking router
  - One or more BGP neighbors
  - Keep an “RIB-in” per neighbor
  - Select single best route per destination prefix
- Route-flap damping
  - Penalty counter per (peer, prefix) pair
  - Increment penalty when peer changes route
  - Decrease penalty over time when route is stable
- Design and deployed in the mid 1990s
  - Widely viewed as helping improve stability
How Long Does BGP Take to Adapt to Changes?

From: Abha Ahuja and Craig Labovitz
Rensselaer Polytechnic Institute
Two Main Factors in Delayed Convergence

- Rate limiting timer slows everything down
- BGP can explore many alternate paths before giving up or arriving at a new path
  - No global knowledge in vectoring protocols
Implementation Does Matter!

![Graph showing number of updates (thousands) over time with labels for stateless and stateful withdraws.]

Shivkumar Kalyanaraman

Rensselaer Polytechnic Institute
Summary

- BGP is a fairly simple protocol …
- … but it is not easy to configure
- BGP is running on more than 100K routers making it one of world’s largest and most visible distributed systems
- Global dynamics and scaling principles are still not well understood
- Traffic Engineering hacked in as an afterthought…