BGP

CS168, Fall 2014
Sylvia Ratnasamy

http://inst.eecs.berkeley.edu/~cs168/fa14/
Announcement

- Canceling my office hours this week (09/25)
- Instead, additional office hours
  - Monday (09/29): 1-2pm
  - Tuesday (09/30): 1-2pm
“Autonomous System (AS)” or “Domain”
Region of a network under a single administrative entity

“Border Routers”

“Interior Routers”
Topology and routes shaped by the business relationships between ASes

- Three basic relationships between two ASes
  - A is a customer of B
  - A is a provider of B
  - A and B are peers

- Business implications
  - customer pays provider
  - peers don’t pay each other
Routing Follows the Money!

traffic allowed

traffic not allowed
Interdomain Routing: Setup

- Destinations are IP prefixes (12.0.0.0/8)
- Nodes are Autonomous Systems (ASes)
  - Internals of each AS are hidden
- Links represent both physical links and business relationships
- BGP (Border Gateway Protocol) is the Interdomain routing protocol
  - Implemented by AS border routers
BGP: Basic Idea

An AS advertises its best routes to one or more IP prefixes.

Each AS selects the “best” route it hears advertised for a prefix.

You’ve heard this story before!
BGP inspired by Distance Vector

- Per-destination route advertisements
- No global sharing of network topology information
- Iterative and distributed convergence on paths
- With four crucial differences!
Differences between BGP and DV

1) **not picking shortest path routes**

- BGP selects the best route based on policy, not shortest distance (least cost)

Node 2 may prefer “2, 3, 1” over “2, 1”

- How do we avoid loops?
Differences between BGP and DV (2) path-vector routing

- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per destination
  - Path vector: send the *entire path* for each destination
Differences between BGP and DV (2) path-vector routing

- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per destination
  - Path vector: send the *entire path* for each destination

- Benefits
  - loop avoidance is easy
For policy reasons, an AS may choose not to advertise a route to a destination. Hence, reachability is not guaranteed even if the graph is connected.

Example: AS#2 does not want to carry traffic between AS#1 and AS#3.
Differences between BGP and DV

(4) BGP may aggregate routes

- For scalability, BGP may aggregate routes for different prefixes

- a.*.*.* is this way
- a.b.*.* is this way
- a.c.*.* is this way
- a.d.*.* is this way

- LBL a.b.0.0/16
- AT&T a.0.0.0/8
- UCB a.c.0.0.0/16
- foo.com a.d.0.0.0/16
BGP: Outline

- BGP policy
  - typical policies, how they’re implemented

- BGP protocol details

- Issues with BGP
Policy imposed in how routes are selected and exported

- **Selection**: Which path to use?
  - controls whether/how traffic leaves the network
- **Export**: Which path to advertise?
  - controls whether/how traffic enters the network
Typical Selection Policy

- In decreasing order of priority
  - make/save money (send to customer > peer > provider)
  - maximize performance (smallest AS path length)
  - minimize use of my network bandwidth ("hot potato")
  - ...
  - ...
Typical Export Policy

<table>
<thead>
<tr>
<th>Destination prefix advertised by...</th>
<th>Export route to...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer</td>
<td>Everyone (providers, peers, other customers)</td>
</tr>
<tr>
<td>Peer</td>
<td>Customers</td>
</tr>
<tr>
<td>Provider</td>
<td>Customers</td>
</tr>
</tbody>
</table>

We’ll refer to these as the “Gao-Rexford” rules (capture common -- but not required! -- practice!)
With Gao-Rexford, the AS policy graph is a DAG (directed acyclic graph) and routes are “valley free”
BGP: Today

- BGP policy
  - typical policies, how they’re implemented

- BGP protocol details
  - stay awake as long as you can…

- BGP issues
Who speaks BGP?

Border routers at an Autonomous System
What does “speak BGP” mean?

- Implement the BGP protocol standard

- Specifies what messages to exchange with other BGP “speakers”
  - message types (e.g., route advertisements, updates)
  - message syntax

- And how to process these messages
  - e.g., “when you receive a BGP update, do…
  - follows BGP state machine in the protocol spec + policy decisions, etc.
A border router speaks BGP with border routers in other ASes
BGP “sessions”

A border router speaks BGP with other (interior and border) routers in its own AS
eBGP, iBGP, IGP

- **eBGP**: BGP sessions between border routers in **different** ASes
  - Learn routes to external destinations

- **iBGP**: BGP sessions between border routers and other routers within the **same** AS
  - Distribute externally learned routes internally

- **IGP**: “Interior Gateway Protocol” = Intradomain routing protocol
  - Provide internal reachability
  - e.g., OSPF, RIP
Some Border Routers Don’t Need BGP

- Customer that connects to a single upstream ISP
  - The ISP can advertise prefixes into BGP on behalf of customer
  - ... and the customer can simply default-route to the ISP
1. Provide internal reachability (IGP)
2. Learn routes to external destinations (eBGP)
3. Distribute externally learned routes internally (iBGP)
4. Travel shortest path to egress (IGP)
Basic Messages in BGP

- **Open**
  - Establishes BGP session
  - BGP uses TCP *[will make sense in 1-2 weeks]*

- **Notification**
  - Report unusual conditions

- **Update**
  - Inform neighbor of new routes
  - Inform neighbor of old routes that become inactive

- **Keepalive**
  - Inform neighbor that connection is still viable
Route Updates

- Format *<IP prefix: route attributes>*
  - attributes describe properties of the route

- Two kinds of updates
  - announcements: new routes or changes to existing routes
  - withdrawal: remove routes that no longer exist
Route Attributes

- Routes are described using attributes
  - Used in route selection/export decisions
- Some attributes are local
  - i.e., private within an AS, not included in announcements
- Some attributes are propagated with eBGP route announcements
- There are many standardized attributes in BGP
  - We will discuss a few
Attributes (1): **ASPATH**

- Carried in route announcements
- Vector that lists all the ASes a route advertisement has traversed (in reverse order)

**Example:**
- AS path = 7018 88
- IP prefix = 128.112.0.0/16
- AS path = 88
- 128.112.0.0/16
- AS path = 7018 88
Attributes (2): **LOCAL PREF**

- “Local Preference”
- Used to choose between different AS paths
- The higher the value the more preferred
- Local to an AS; carried only in iBGP messages

**BGP table at AS4:**

<table>
<thead>
<tr>
<th>Destination</th>
<th>AS Path</th>
<th>Local Pref</th>
</tr>
</thead>
<tbody>
<tr>
<td>140.20.1.0/24</td>
<td>AS3 AS1</td>
<td>300</td>
</tr>
<tr>
<td>140.20.1.0/24</td>
<td>AS2 AS1</td>
<td>100</td>
</tr>
</tbody>
</table>
Example: iBGP and LOCAL PREF

- Both routers prefer the path through AS 2 on the left
Attributes (3) : MED

- “Multi-Exit Discriminator”

- Used when ASes are interconnected via 2 or more links to specify how close a prefix is to the link it is announced on

- Lower is better

- AS announcing prefix sets MED

- AS receiving prefix (optionally!) uses MED to select link
Attributes (4): IGP cost

- Used for hot-potato routing
  - Each router selects the closest egress point based on the path cost in intra-domain protocol
IGP may conflict with MED
Typical Selection Policy

- In decreasing order of priority
  - make/save money (send to customer > peer > provider)
  - maximize performance (smallest AS path length)
  - minimize use of my network bandwidth (“hot potato”)
  - ...
  - ...
Using Attributes

- Rules for route selection in priority order

<table>
<thead>
<tr>
<th>Priority</th>
<th>Rule</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LOCAL PREF</td>
<td>Pick highest LOCAL PREF</td>
</tr>
<tr>
<td>2</td>
<td>ASPATH</td>
<td>Pick shortest ASPATH length</td>
</tr>
<tr>
<td>3</td>
<td>MED</td>
<td>Lowest MED preferred</td>
</tr>
<tr>
<td>4</td>
<td>eBGP &gt; iBGP</td>
<td>Did AS learn route via eBGP (preferred) or iBGP?</td>
</tr>
<tr>
<td>5</td>
<td>iBGP path</td>
<td>Lowest IGP cost to next hop (egress router)</td>
</tr>
<tr>
<td>6</td>
<td>Router ID</td>
<td>Smallest next-hop router’s IP address as tie-breaker</td>
</tr>
</tbody>
</table>
BGP UPDATE Processing

Open ended programming. Constrained only by vendor configuration language.

Control plane

BGP Updates

Apply Import Policies → Best Route Selection → Best Route Table → Apply Export Policies

Data plane

Data packets → forwarding Entries → IP Forwarding Table → Data packets

BGP Updates
BGP: Today

- BGP policy
  - typical policies, how they’re implemented

- BGP protocol details

- BGP issues
Issues with BGP

- Reachability
- Security
- Convergence
- Performance
- Anomalies
Reachability

- In normal routing, if graph is connected then reachability is assured.
- With policy routing, this does not always hold.
Security

- An AS can claim to serve a prefix that they actually don’t have a route to (blackholing traffic)
  - Problem not specific to policy or path vector
  - Important because of AS autonomy
  - *Fixable: make ASes “prove” they have a path*

- Note: AS may forward packets along a route different from what is advertised
  - Tell customers about fictitious short path…
  - Much harder to fix!
Convergence

- Result: If all AS policies follow “Gao-Rexford” rules, BGP is guaranteed to converge (safety)

- For arbitrary policies, BGP may fail to converge!
Example of Policy Oscillation

“1” prefers “1 3 0” over “1 0” to reach “0”
Step-by-Step of Policy Oscillation

Initially: nodes 1, 2, 3 know only shortest path to 0
1 advertises its path 1 0 to 2
Step-by-Step of Policy Oscillation
Step-by-Step of Policy Oscillation

3 advertises its path 3 0 to 1
Step-by-Step of Policy Oscillation
1 withdraws its path 1 0 from 2
Step-by-Step of Policy Oscillation
Step-by-Step of Policy Oscillation

2 advertises its path 2 0 to 3
Step-by-Step of Policy Oscillation
3 withdraws its path 3 0 from 1
Step-by-Step of Policy Oscillation
Step-by-Step of Policy Oscillation

1 advertises its path 1 0 to 2
Step-by-Step of Policy Oscillation
Step-by-Step of Policy Oscillation

2 withdraws its path 2 0 from 3
Step-by-Step of Policy Oscillation

We are back to where we started!
Convergence

- Result: If all AS policies follow “Gao-Rexford” rules, BGP is guaranteed to converge (safety)

- For arbitrary policies, BGP may fail to converge!

- Why should this trouble us?
Performance Nonissues

- Internal routing (non)
  - Domains typically use “hot potato” routing
  - Not always optimal, but economically expedient

- Policy not about performance (non)
  - So policy-chosen paths aren’t shortest

- AS path length can be misleading (non)
  - 20% of paths inflated by at least 5 router hops
Performance (example)

- AS path length can be misleading
  - An AS may have many router-level hops

BGP says that path 4 1 is better than path 3 2 1
Real Performance Issue: Slow convergence

- BGP outages are biggest source of Internet problems

- Labovitz et al. *SIGCOMM’97*
  - 10% of routes available less than 95% of time
  - Less than 35% of routes available 99.99% of the time

- Labovitz et al. *SIGCOMM 2000*
  - 40% of path outages take 30+ minutes to repair

- But most popular paths are very stable
BGP Misconfigurations

- BGP protocol is both bloated and underspecified
  - lots of attributes
  - lots of leeway in how to set and interpret attributes
  - necessary to allow autonomy, diverse policies
  - but also gives operators plenty of rope

- Much of this configuration is manual and *ad hoc*

- And the core abstraction is fundamentally flawed
  - disjoint per-router configuration to effect AS-wide policy
  - now strong industry interest in changing this! [later: SDN]
BGP: How did we get here?

- BGP was designed for a different time
  - before commercial ISPs and their needs
  - before address aggregation
  - before multi-homing

- We don't get a second chance: 'clean slate' designs virtually impossible to deploy

- Thought experiment: how would you design a policy-driven interdomain routing solution? How would you deploy it?

- 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)
Next Time.

- Wrap up the network layer!
  - the IPv4 header
  - IP routers