Interdomain Routing

Reading: Sections K&R 4.6.3

EE122: Intro to Communication Networks

Fall 2007 (WF 4:00-5:30 in Cory 277)

Guest Lecture by Brighten Godfrey

Instructor: Vern Paxson

TAs: Lisa Fowler, Daniel Killebrew & Jorge Ortiz

http://inst.eecs.berkeley.edu/~ee122/

Materials with thanks to Scott Shenker, Jennifer Rexford, Ion Stoica and colleagues at Princeton and UC Berkeley

Outline

- Why does BGP exist?
  - What is interdomain routing and why do we need it?
  - Why does BGP look the way it does?
- How does BGP work?
  - Boring details
  - Yuck

pay more attention to the "why" than the "how"

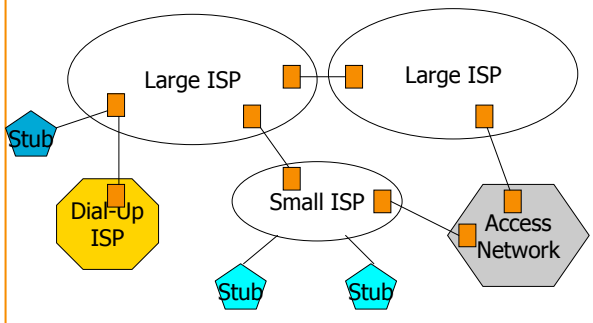
Routing

- Provides paths between networks
- Previous lecture presented two routing designs
  - link-state
  - distance vector
- Previous lecture assumed single domain
  - all routers have same routing metric (shortest path)
  - no privacy issues, no policy issues

Internet is more complicated.....

- Internet not just unstructured collection of networks
- Internet is comprised of a set of "autonomous systems" (ASes)
  - independently run networks, some are commercial ISPs
  - currently around 20,000 ASes
- ASes are sometimes called "domains"
  - hence "interdomain routing"

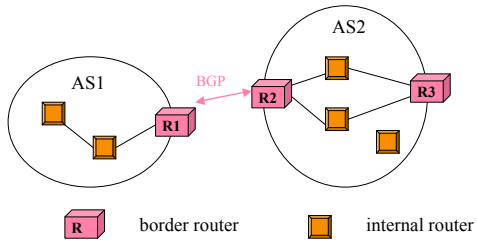
Internet: a large number of ASes



This adds another level in hierarchy

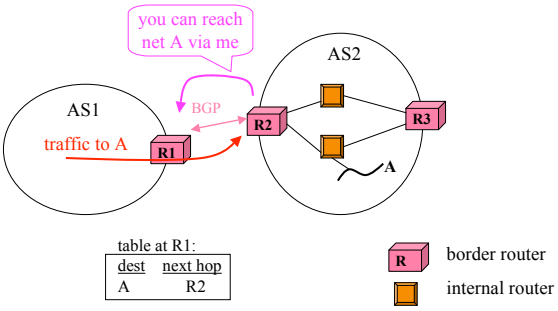
- Three levels in logical routing hierarchy
  - networks: reaches individual hosts
  - intradomain: routes between networks
  - interdomain: routes between ASes
- Need a protocol to route between domains
  - BGP is current standard
  - BGP unifies network organizations

Who speaks BGP?



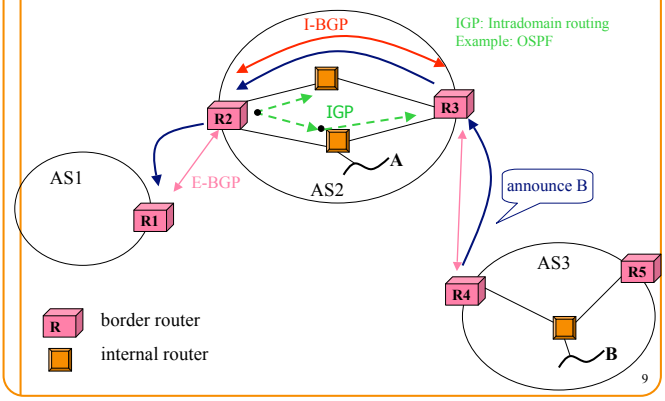
- Two types of routers
  - Border router (Edge), Internal router (Core)

Purpose of BGP

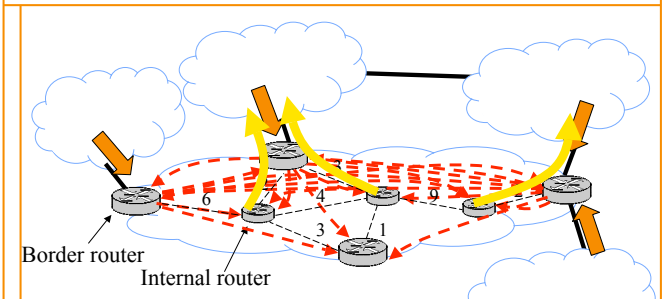


Share connectivity information across ASes

I-BGP and E-BGP



In more detail



1. Provide internal reachability (**IGP**) ------
2. Learn routes to external destinations (**eBGP**) ➡
3. Distribute externally learned routes internally (**iBGP**) - ➡
4. Select closest egress (**IGP**) ------

Rest of lecture...

- Motivate why BGP is the way it is
- Discuss some problems with interdomain routing
- Discuss (briefly!) what a new BGP might look like
- Explain some of BGP's details
  - not fundamental, just series of specific design decisions

Why BGP Is the Way It Is

1. ASes are autonomous

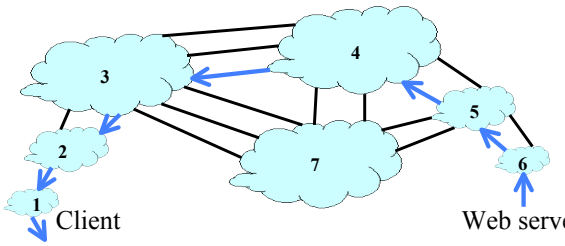
- Want to choose their own internal routing protocol
  - different algorithms and metrics
- Want freedom to route based on policy
  - "my traffic can't be carried over my competitor's network"
  - "I don't want to carry transit traffic through my network"
  - not expressible as Internet-wide "shortest path"!
- Want to keep their connections and policies private
  - would reveal business relationships, network structure

2. ASes have business relationships

- Three kinds of relationships between ASes
  - AS A can be AS B's *customer*
  - AS A can be AS B's *provider*
  - AS A can be AS B's *peer*
- Business implications
  - customer pays provider
  - peers don't pay each other
- Policy implications
  - "When sending traffic, I prefer to route through customers over peers, and peers over providers"
  - "I don't carry traffic from one provider to another provider"

AS-level topology

- Destinations are IP prefixes (e.g., 12.0.0.0/8)
- Nodes are Autonomous Systems (ASes)
  - internals are hidden
- Links are connections & business relationships



What routing algorithm can we use?

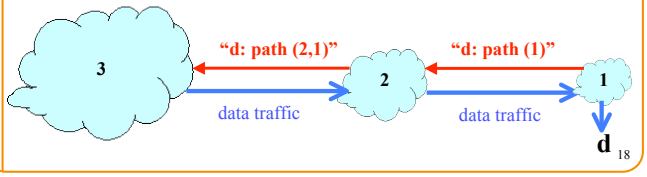
- Key issues are *policy* and *privacy*
- Can't use shortest path
  - domains don't have any shared metric
  - *policy choices might not be shortest path*
- Can't use link state
  - would have to flood policy preferences and topology
  - *would violate privacy*

What about distance vector?

- Does not reveal any connectivity information
- But is designed to compute shortest paths
- Extend distance vector to allow policy choices?

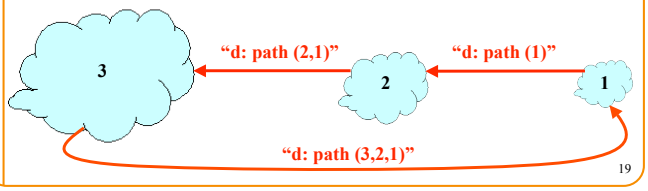
Path-Vector Routing

- Extension of distance-vector routing
  - Support flexible routing policies
  - Faster loop detection (no count-to-infinity)
- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per dest d
  - Path vector: send the *entire path* for each dest d



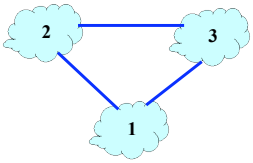
Faster Loop Detection

- Node can easily detect a loop
  - Look for its own node identifier in the path
  - E.g., node 1 sees itself in the path "3, 2, 1"
- Node can simply discard paths with loops
  - E.g., node 1 simply discards the advertisement



Flexible Policies

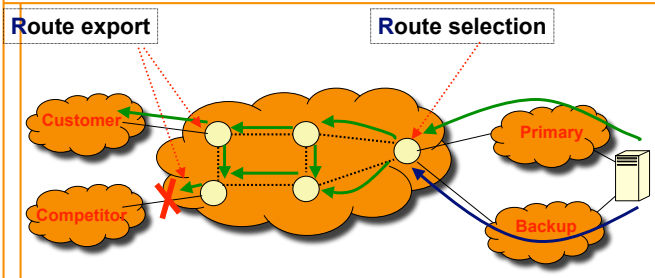
- Each node can apply local policies
  - Path selection: Which path to use?
  - Path export: Which paths to advertise?
- Examples
  - Node 2 may prefer the path "2, 3, 1" over "2, 1"
  - Node 1 may not let node 3 hear the path "1, 2"



Selection vs Export

- Selection policies
  - determines which paths I want my traffic to take
- Export policies
  - determines whose traffic I am willing to carry
- Notes:
  - any traffic I carry will follow the same path my traffic takes, so there is a connection between the two
  - from a protocol perspective, decisions can be *arbitrary*
    - can depend on entire path (advantage of PV approach)

Illustration



Selection: controls traffic out of the network

Export: controls traffic into the network

Examples of Standard Policies

- Transit network:
  - Selection: prefer customer to peer to provider
  - Export:
    - Let customers use any of your routes
    - Let anyone route through you to your customer
    - *Block everything else*
- Multihomed (nontransit) network:
  - Export: Don't export routes for other domains
  - Selection: pick primary over backup

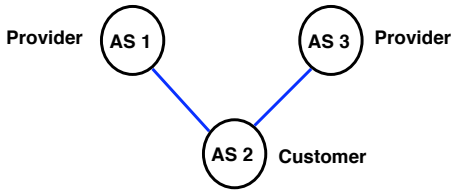
Any Questions?

Issues with Path-Vector Policy Routing

- Reachability
- Security
- Performance
- Lack of isolation
- Policy oscillations

Reachability

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



Security

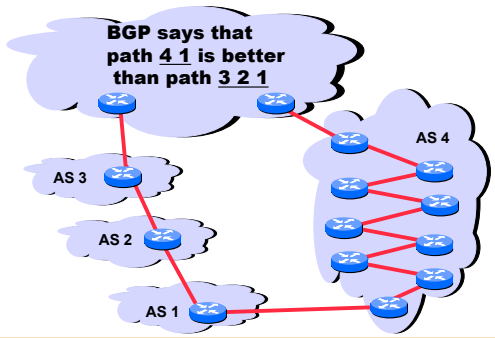
- An AS can claim to serve a prefix that they actually don't have a route to (blackholing traffic)
  - problem not specific to policy or path vector
  - important because of AS autonomy
- Fixable: make ASes "prove" they have a path

Performance

- BGP designed for policy not performance
- "Hot Potato" routing common but suboptimal
  - AS wants to hand off the packet as soon as possible
- Even BGP "shortest paths" are not shortest
  - Fewest AS's != Fewest number of routers
- 20% of paths inflated by at least 5 router hops
- Not clear this is a significant problem

Performance (example)

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



Lack of Isolation: dynamics

- If there is a change in the path, the path must be re-advertised to every node upstream of the change
- "Route Flap Damping" supposed to help here, (but ends up causing more problems)

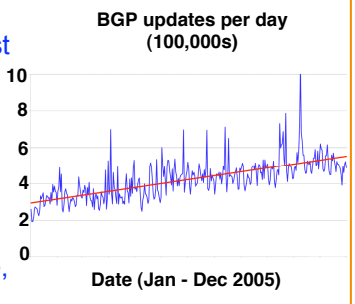


Fig. from [Huston & Armitage 2006]

Lack of isolation: routing table size

- Each BGP router must know path to every other IP prefix
  - but router memory is expensive and thus constrained
- Number of prefixes growing more than linearly
- Subject of current research

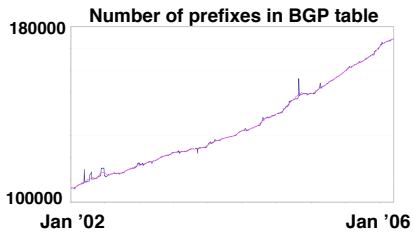
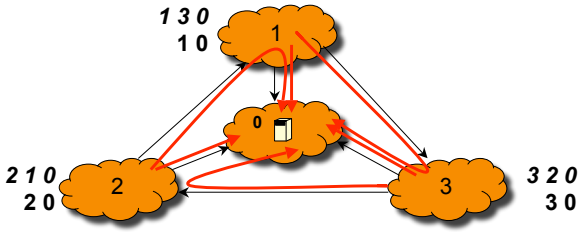


Fig. from [Huston & Armitage 2006]

Persistent Oscillations due to Policies

Depends on the interactions of policies



We are back to where we started!

Policy Oscillations (cont'd)

- Policy autonomy vs network stability
  - focus of much recent research
- Not an easy problem
  - PSPACE-complete to decide whether given policies will eventually converge! [FP08]
- However, if policies follow normal business practices, stability is guaranteed

Redesigning BGP

- If we keep all the current constraints, not many alternative design options (at high-level)
  - Which constraints might we lift?
- Are most policies really private?
  - could use link-state for some of the routing
- Do ASes really need to see the entire path?
  - could hide some of the path, reducing updates
- Can AS structure be integrated into addressing?

Any Questions?

Rest of lecture....

- BGP details
- Stay awake as long as you can.....

Border Gateway Protocol (BGP)

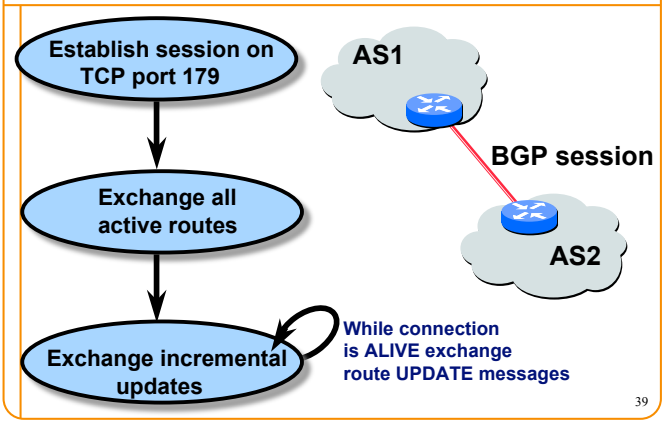
- Interdomain routing protocol for the Internet
  - –Prefix-based path-vector protocol
  - –Policy-based routing based on AS Paths
  - –Evolved during the past 15 years

- **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 Routing Table

```
ner-routes>show ip bgp
BGP table version is 6128791, local router ID is 4.2.34.165
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete
   Network          Next Hop            Metric LocPrf Weight Path
* i3.0.0.0         4.0.6.142             1000     50      0 701 80 i
* i4.0.0.0         4.24.1.35                0    100      0 i
* i12.3.21.0/23    192.205.32.153           0     50      0 7018 4264 6468 ?
* e128.32.0.0/16   192.205.32.153           0     50      0 7018 4264 6468 25 e
```

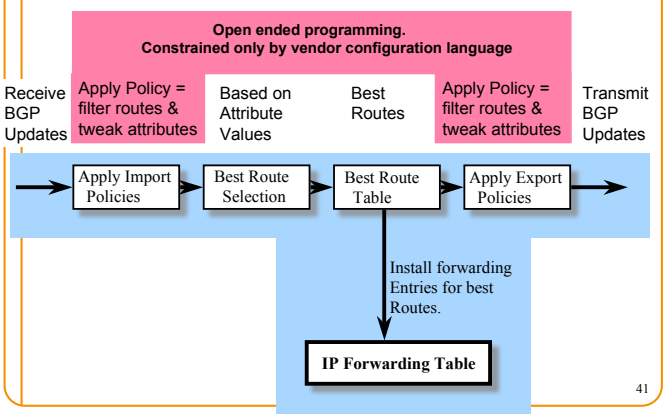

BGP Operations



Incremental Protocol

- A node learns multiple paths to destination
  - –Stores all of the routes in a routing table
  - –Applies policy to select a single active route
  - –… and may advertise the route to its neighbors
- Incremental updates
  - –Announcement
    - Upon selecting a new active route, add node id to path
    - … and (optionally) advertise to each neighbor
  - –Withdrawal
    - If the active route is no longer available
    - … send a withdrawal message to the neighbors

BGP Route Processing



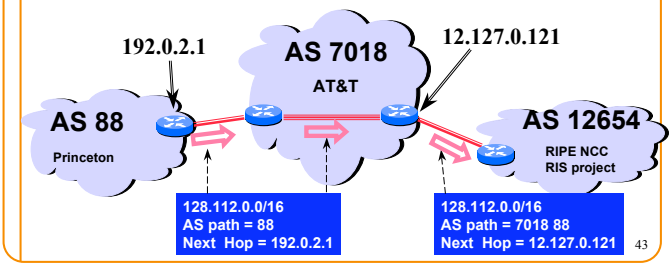
Selecting the best route

- Attributes of routes set/modified according to operator instructions
- Routes compared based on attributes using (mostly) standardized rules

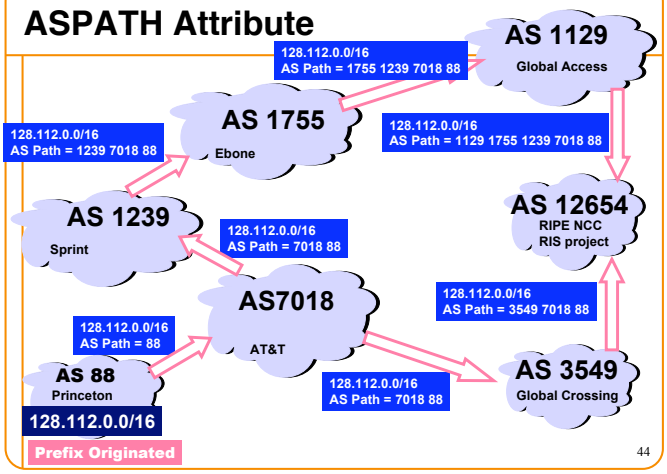
1. Highest local preference (all equal by default…
2. Shortest AS path length …so default = shortest paths)
3. Lowest origin type (IGP < EGP < incomplete)
4. Lowest MED
5. eBGP- over iBGP-learned
6. Lowest IGP cost
7. Lowest next-hop router ID

Attributes

- Destination prefix (e.g,. 128.112.0.0/16)
- Routes have attributes, including
  - AS path (e.g., "7018 88")
  - Next-hop IP address (e.g., 12.127.0.121)



ASPATH Attribute

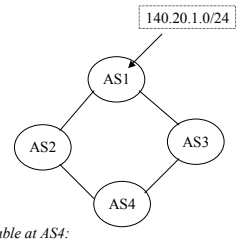


Local Preference attribute

Policy choice between different AS paths

The higher the value the more preferred

Carried by IBGP, local to the AS.

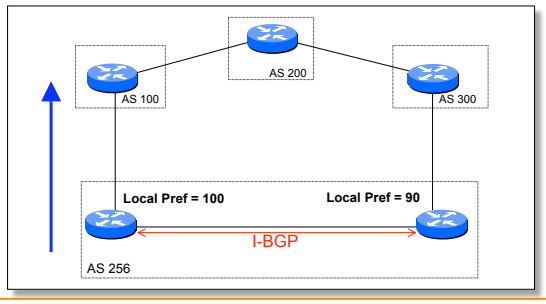


BGP table at AS4:

Destination	AS Path	Local Pref
140.20.1.0/24	AS3 AS1	300
140.20.1.0/24	AS2 AS1	100

Internal BGP and Local Preference

- Example
  - Both routers prefer the path through AS 100 on the left
  - … even though the right router learns an external path

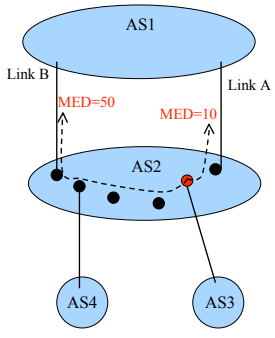


Origin attribute

- Who originated the announcement?
- Where was a prefix *injected* into BGP?
- IGP, BGP or Incomplete (often used for static routes)

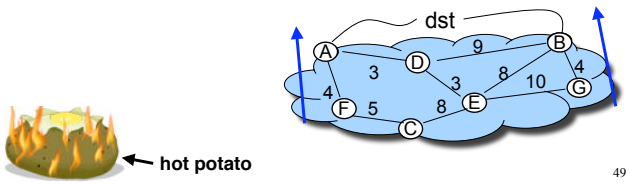
Multi-Exit Discriminator (MED) attr.

- When ASes interconnected via 2 or more links
- AS announcing prefix sets MED (AS2 in picture)
- AS receiving prefix uses MED to select link
- A way to specify how close a prefix is to the link it is announced on



IGP cost attribute

- Used in BGP for hot-potato routing
  - Each router selects the closest egress point
  - … based on the path cost in intradomain protocol
- Somewhat in conflict with MED

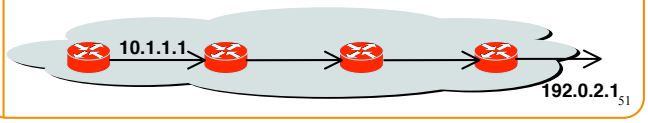


Lowest Router ID

- Last step in route selection decision process
- "Arbitrary" tiebreaking
- But we do sometimes reach this step, so how ties are broken matters

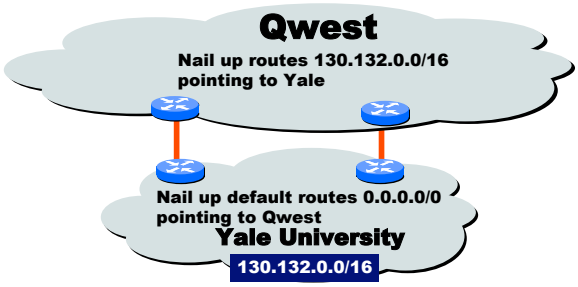
Joining BGP and IGP Information

- Border Gateway Protocol (BGP)
  - Announces reachability to external destinations
  - Maps a destination prefix to an egress point
    - 128.112.0.0/16 reached via 192.0.2.1
- Interior Gateway Protocol (IGP)
  - Used to compute paths within the AS
  - Maps an egress point to an outgoing link
    - 192.0.2.1 reached via 10.1.1.1



Some Routers Don't Need BGP

- Customer that connects to a single upstream ISP
  - The ISP can introduce the prefixes into BGP
  - … and the customer can simply default-route to the ISP



Summary

- BGP is essential to the Internet
  - ties different organizations together
- Poses fundamental challenges....
  - leads to use of path vector approach
- ...and myriad details