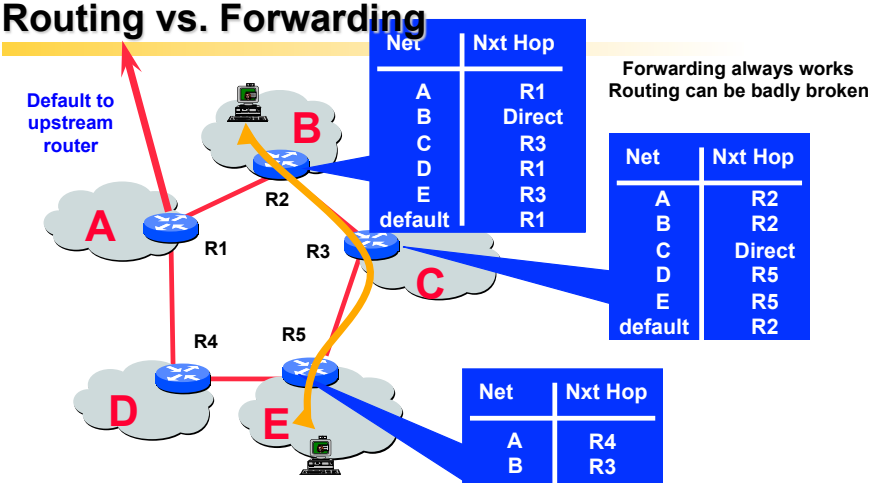
Interdomain Routing
BGP

Prométhée Spathis
promethee.spathis@lip6.fr
Thème NPA, LIP6
Paris, FRANCE

Goals of Today's Lecture

- Challenges of interdomain routing
  - Scale, privacy, and policy
  - Limitations of link-state and distance-vector routing
- Path-vector routing
  - Faster loop detection than distance-vector routing
  - More flexibility than shortest-path routing
- Border Gateway Protocol (BGP)
  - Incremental, prefix-based, path-vector protocol
  - Programmable import and export policies
  - Multi-step decision process for selecting "best" route
- Multiple routers within an AS
- BGP convergence delay

Routing vs. Forwarding



Forwarding always works
Routing can be badly broken

Forwarding: determine next hop
Routing: establish end-to-end paths

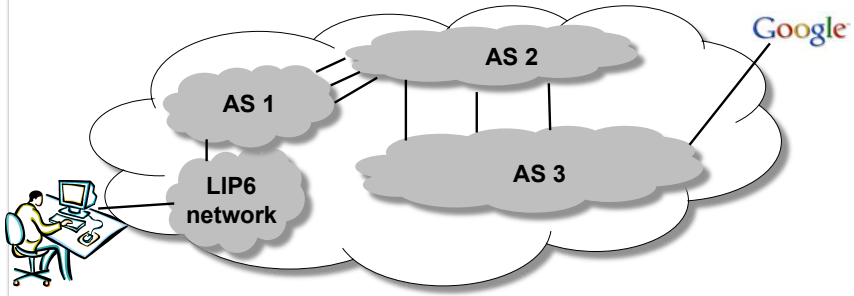
Internet Routing Architecture

- Divided into Autonomous Systems
  - Distinct regions of administrative control
  - Routers/links managed by a single "institution"
  - Service provider, company, university, …
- Hierarchy of Autonomous Systems
  - Large, tier-1 provider with a nationwide backbone
  - Medium-sized regional provider with smaller backbone
  - Small network run by a single company or university
- Interaction between Autonomous Systems
  - Internal topology is not shared between ASes
  - … but, neighboring ASes interact to coordinate routing

Two-Tiered Internet Routing Architecture

- Goal: distributed management of resources
  - Internetworking of multiple networks
  - Networks under separate administrative control
- Solution: two-tiered routing architecture
  - Intradomain: inside a region of control
    - Okay for routers to share topology information
    - Routers configured to achieve a common goal
  - Interdomain: between regions of control
    - Not okay to share complete information
    - Networks may have different/conflicting goals
- Led to the use of different protocols…

Autonomy: network of networks



- Internet = interconnection of Autonomous Systems (AS)
  - Distinct regions of administrative control
  - Routers/links managed by a single “institution”
  - Service provider, company, university, etc.

AS Numbers (ASNs)

ASNs are 16 bit values.
64512 through 65535 are “private”

Currently around 20,000 in use.

- Level 3: 1
- MIT: 3
- Harvard: 11
- Jussieu: 1307
- AT&T: 7018, 6341, 5074, …
- UUNET: 701, 702, 284, 12199, …
- Sprint: 1239, 1240, 6211, 6242, …
- …

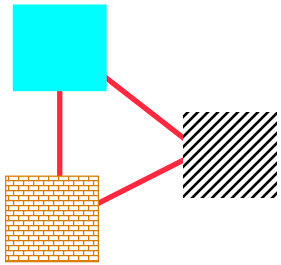
ASNs represent units of routing policy

AS ≠ Institution

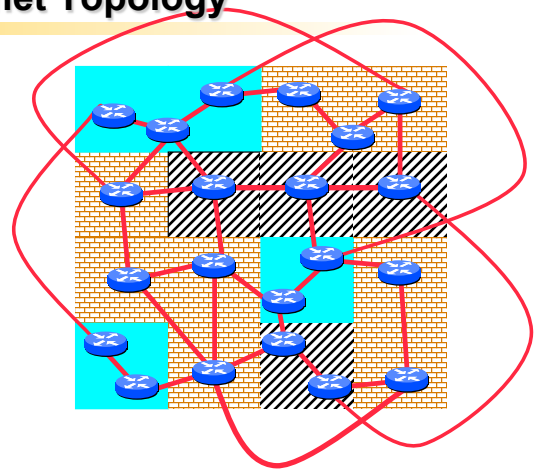
- Not equivalent to an AS
  - Many institutions span multiple autonomous systems
  - Some institutions do not have their own AS number
  - Ownership of an AS may be hard to pinpoint (whois)
- Not equivalent to a block of IP addresses (prefix)
  - Many institutions have multiple (non-contiguous) prefixes
  - 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)

AS Graph != Internet Topology

BGP was designed to throw away information!



The AS graph may look like this.

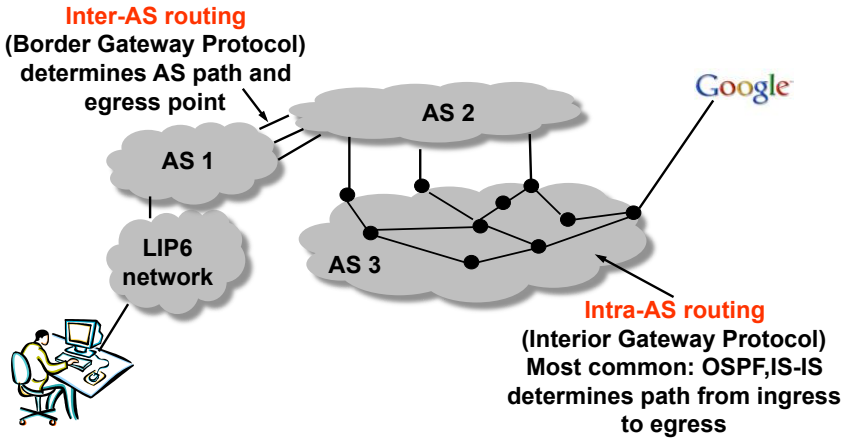


Reality may be closer to this...

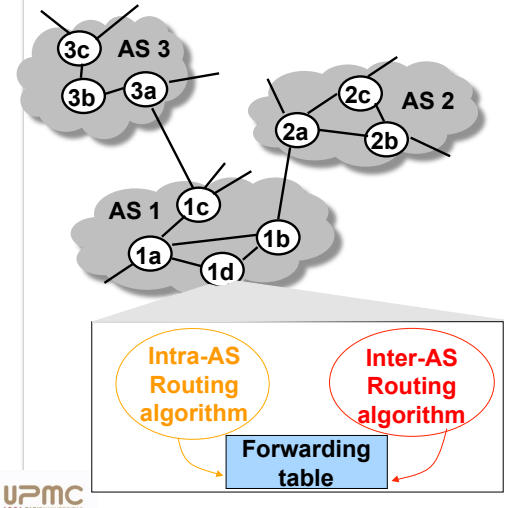
Characterizations of AS Topology

- Tier-1: small number of tier-1 ASes
  - A near-clique of ~15 ASes with no providers
  - AT&T, Sprint, UUNET, …
- Transit core: peer with tier-1s and each other
  - Around 100-200 large ASes
  - UUNET Europe, KDDI, and Singapore Telecom
- Regional ISPs: non-stubs near the edge
  - Around 2000 medium-sized ASes
  - Minnesota Regional Network, US West
- Stub ASes: no peer or customer neighbors
  - Jussieu, MIT, AT&T Research, …

Hierarchical routing



Interconnected ASes

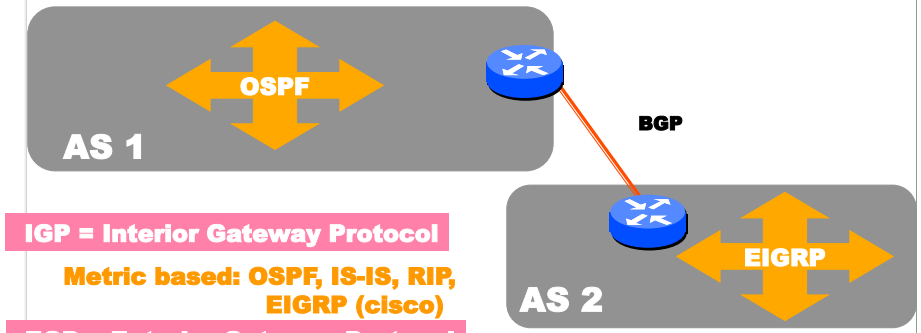


- Forwarding table is configured by both intra- and inter-AS routing algorithm
  - Intra-AS sets entries for internal dests
  - Inter-AS & Intra-As sets entries for external dests

Two-Tiered Internet Routing System

- Interdomain routing: between ASes
  - Routing policies based on *business relationships*
  - No common metrics, and limited cooperation
  - BGP: policy-based, path-vector routing protocol
- Intradomain routing: within an AS
  - Shortest-path routing based on *link metrics*
  - Routers all managed by a single institution
  - OSPF and IS-IS: link-state routing protocol
  - RIP and EIGRP: distance-vector routing protocol

Architecture of Dynamic Routing



IGP = Interior Gateway Protocol

Metric based: OSPF, IS-IS, RIP, EIGRP (cisco)

EGP = Exterior Gateway Protocol

Policy based: BGP

The Routing Domain of BGP is the entire Internet

Technology of Distributed Routing

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

Intradomain Routing Today

- Link-state routing with static link weights
  - Static weights: avoid stability problems
  - Link state: faster reaction to topology changes
- Most common protocols in backbones
  - OSPF: Open Shortest Path First
  - IS-IS: Intermediate System–Intermediate System
- Some use of distance vector in enterprises
  - RIP: Routing Information Protocol
  - EIGRP: Enhanced Interior Gateway Routing Protocol
- Growing use of Multi-Protocol Label Switching

Link-State Routing is Problematic

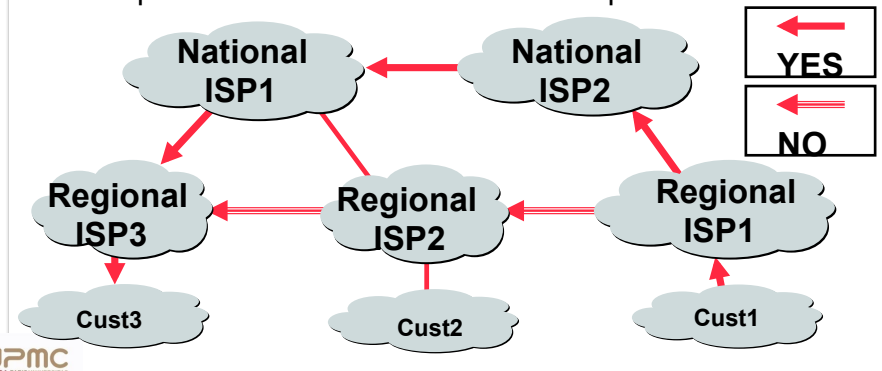
- Topology information is flooded
  - High bandwidth and storage overhead
  - Forces nodes to divulge sensitive information
- Entire path computed locally per node
  - High processing overhead in a large network
- Minimizes some notion of total distance
  - Works only if policy is shared and uniform
- Typically used only inside an AS
  - E.g., OSPF and IS-IS

Challenges for Interdomain Routing

- Scale
  - Prefixes: 150,000-200,000, and growing
  - ASes: 20,000 visible ones, and growing
  - AS paths and routers: at least in the millions…
- Privacy
  - ASes don't want to divulge internal topologies
  - … or their business relationships with neighbors
- Policy
  - No Internet-wide notion of a link cost metric
  - Need control over where you send traffic
  - … and who can send traffic through you

Shortest-Path Routing is Restrictive

- All traffic must travel on shortest paths
- All nodes need common notion of link costs
- Incompatible with commercial relationships

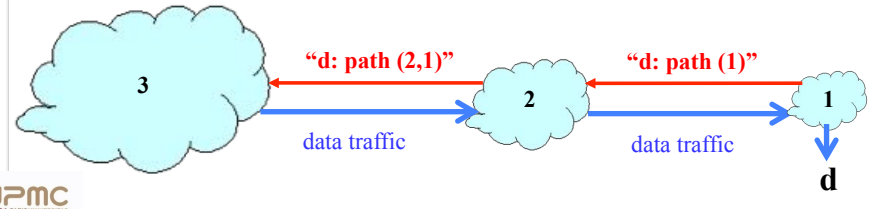


Distance Vector is on the Right Track

- Advantages
  - Hides details of the network topology
  - Nodes determine only "next hop" toward the dest
- Disadvantages
  - Minimizes some notion of total distance, which is difficult in an interdomain setting
  - Slow convergence due to the counting-to-infinity problem ("bad news travels slowly")
- Idea: extend the notion of a distance vector

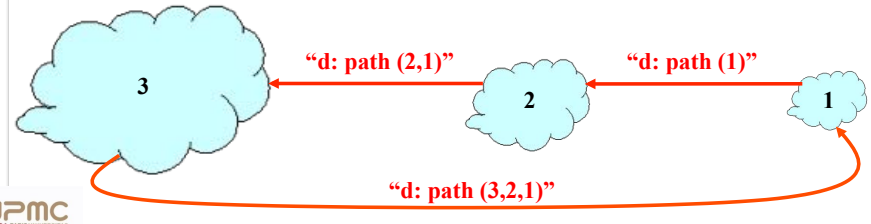
Path-Vector Routing

- Extension of distance-vector routing
  - Support flexible routing policies
  - Avoid count-to-infinity problem
- 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



Routing Protocols

	Link State	Distance Vector
Dissem-ination	Flood link state advertisements to all routers	Update distances from neighbors' distances
Algorithm	Dijsktra's shortest path	Bellman-Ford shortest path
Converge	Fast due to flooding	Slow, due to count-to-infinity
Protocols	OSPF, IS-IS	RIP, EIGRP

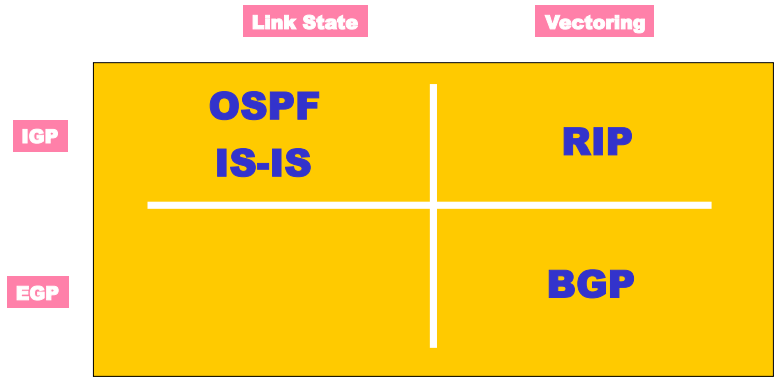
Routing Protocols

	Link State	Distance Vector	Path Vector
Dissem-ination	Flood link state advertisements to all routers	Update distances from neighbors' distances	Update paths based on neighbors' paths
Algorithm	Dijsktra's shortest path	Bellman-Ford shortest path	Local policy to rank paths
Converge	Fast due to flooding	Slow, due to count-to-infinity	Slow, due to path exploration
Protocols	OSPF, IS-IS	RIP, EIGRP	BGP

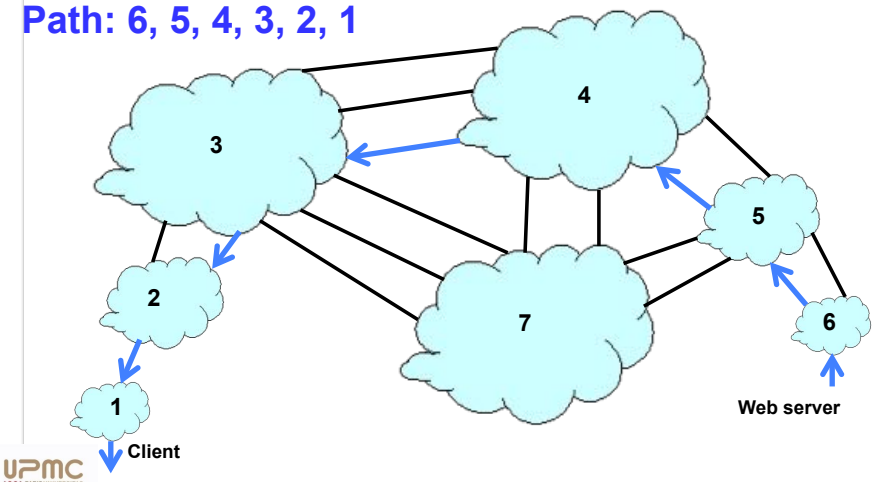
Two-Tiered Internet Routing System

- Intradomain routing: within an AS
  - Shortest-path routing based on *link metrics*
  - Routers all managed by a single institution
  - OSPF and IS-IS: link-state routing protocol
  - RIP and EIGRP: distance-vector routing protocol
- Interdomain routing: between ASes
  - Routing policies based on *business relationships*
  - No common metrics, and limited cooperation
  - BGP: policy-based, path-vector routing protocol

The Gang of Four

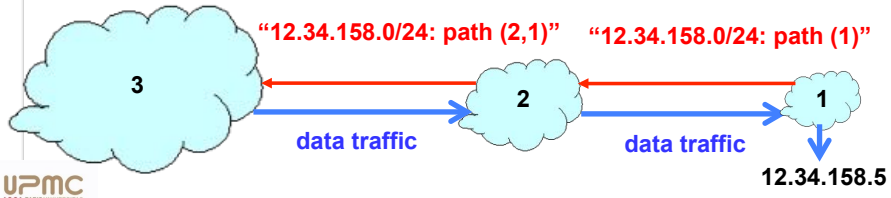


Interdomain Routing (Between ASes)



Interdomain Routing: Border Gateway Protocol

- 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?



Border Gateway Protocol

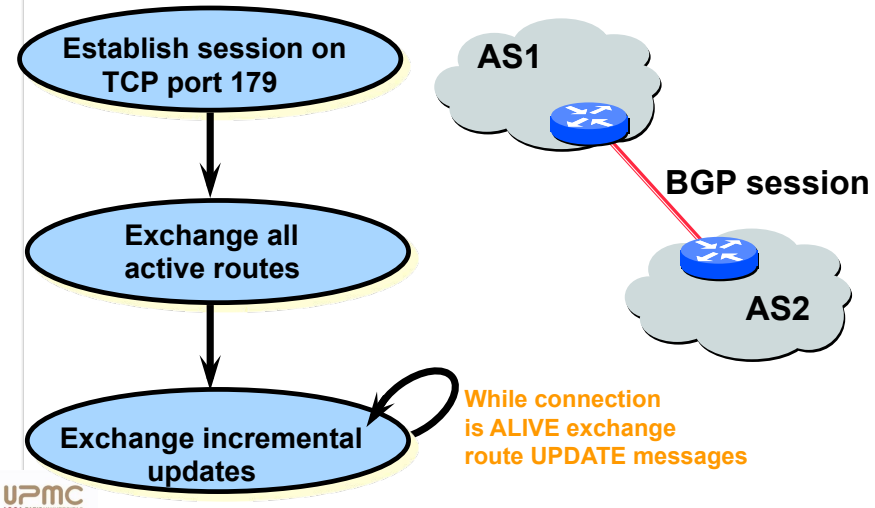
- Interdomain routing protocol for the Internet
  - Prefix-based path-vector protocol
  - Policy-based routing based on AS Paths
  - Evolved during the past 25 years

- **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)**

Components of BGP

- BGP protocol
  - Definition of how two BGP neighbors communicate
  - Message formats, state machine, route attributes, etc.
  - Standardized by the IETF
- Policy specification
  - Flexible language for filtering and manipulating routes
  - Indirectly affects the selection of the best route
  - Varies across vendors, though constructs are similar
- BGP decision process
  - Complex sequence of rules for selecting the best route
  - De facto standard applied by router vendors
  - Being codified in a new RFC for BGP coming soon

BGP Operations



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

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

Update Messages

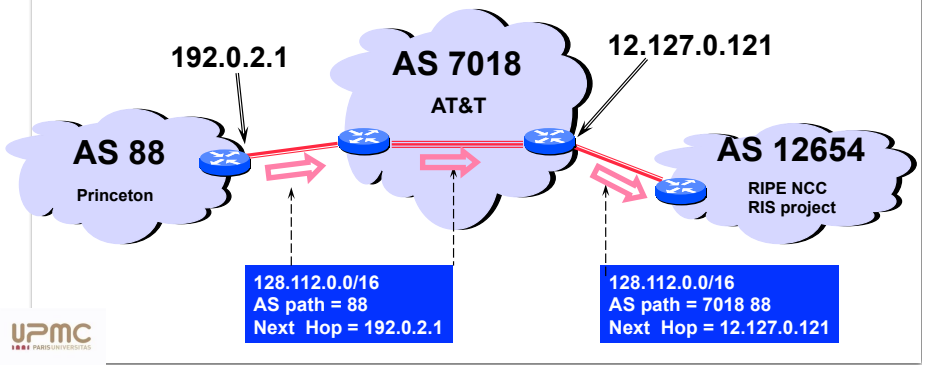
- Update messages
  - Advertisement
    - New route for the prefix (e.g., 12.34.158.0/24)
    - Attributes such as the AS path (e.g., “2 1”)
  - Withdrawal
    - Announcing that the route is no longer available
- Numerous BGP attributes
  - AS path
  - Next-hop IP address
  - Local preference
  - Multiple-Exit Discriminator
  - …

Advertising a prefix

- When a router advertises a prefix to one of its BGP neighbors:
  - information is valid until first router explicitly advertises that the information is no longer valid
  - BGP does not require routing information to be refreshed
  - if node A advertises a path for a prefix to node B, then node B can be sure node A is using that path itself to reach the destination.

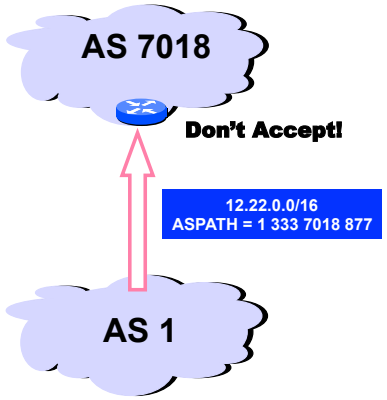
BGP Route

- Destination prefix (e.g,. 128.112.0.0/16)
- Route attributes, including
  - AS path (e.g., “7018 88”)
  - Next-hop IP address (e.g., 12.127.0.121)

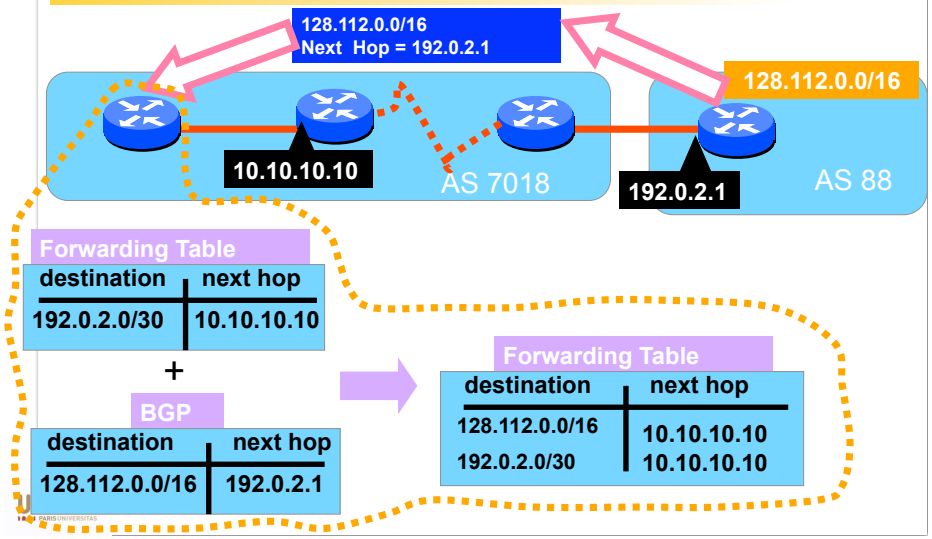


Interdomain Loop Prevention

BGP at AS YYY will never accept a route with ASPATH containing YYY.

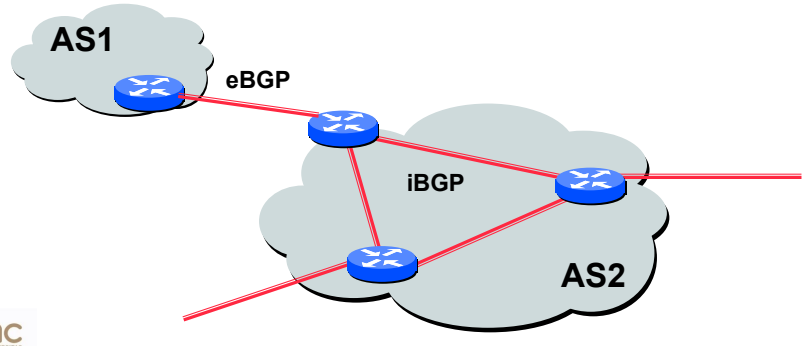


Joining BGP and IGP Information



An AS is Not a Single Node

- Multiple routers in an AS
  - Need to distribute BGP information within the AS
  - Internal BGP (iBGP) sessions between routers



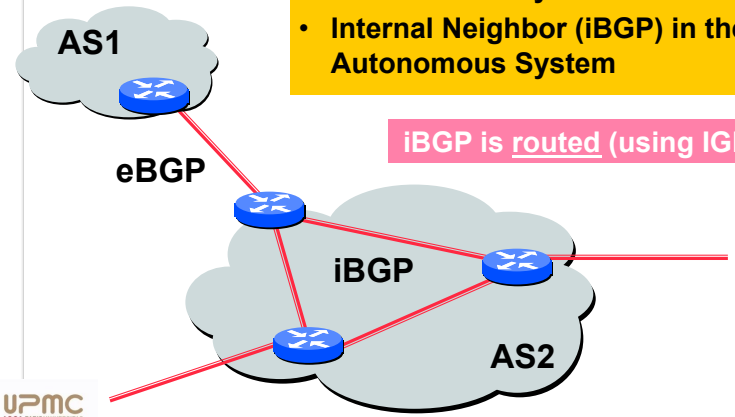
Internal BGP (I-BGP)

- Used to distribute routes learned via EBGP to all the routers within an AS
- I-BGP and E-BGP are same protocol in that
  - same message types used
  - same attributes used
  - same state machine
  - BUT use different rules for readvertising prefixes
- Rule #1: prefixes learned from an E-BGP neighbor *can* be readvertised to an I-BGP neighbor, and vice versa
- Rule #2: prefixes learned from an I-BGP neighbor *cannot* be readvertised to another I-BGP neighbor

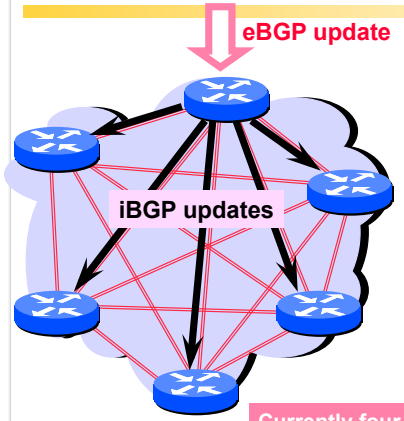
Two Types of BGP Neighbor Relationships

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

iBGP is routed (using IGP!)



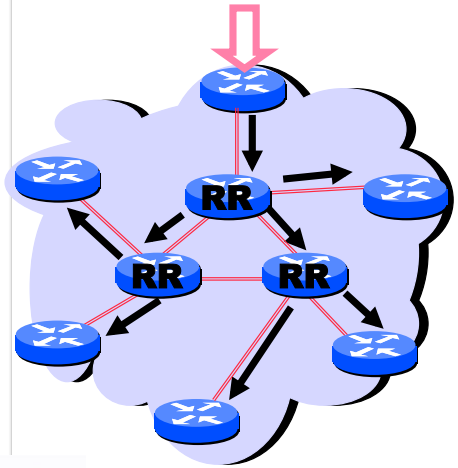
iBGP Mesh Does Not Scale



- N border routers means N(N-1)/2 peering sessions
- Each router must have N-1 iBGP sessions configured
- The addition a single iBGP speaker requires configuration changes to all other iBGP speakers
- Size of iBGP routing table can be order N larger than number of best routes (remember alternate routes!)
- Each router has to listen to update noise

Currently four solutions:
(0) Buy bigger routers!
(1) Break AS into smaller ASes
(2) BGP Route reflectors
(3) BGP confederations

Route Reflectors

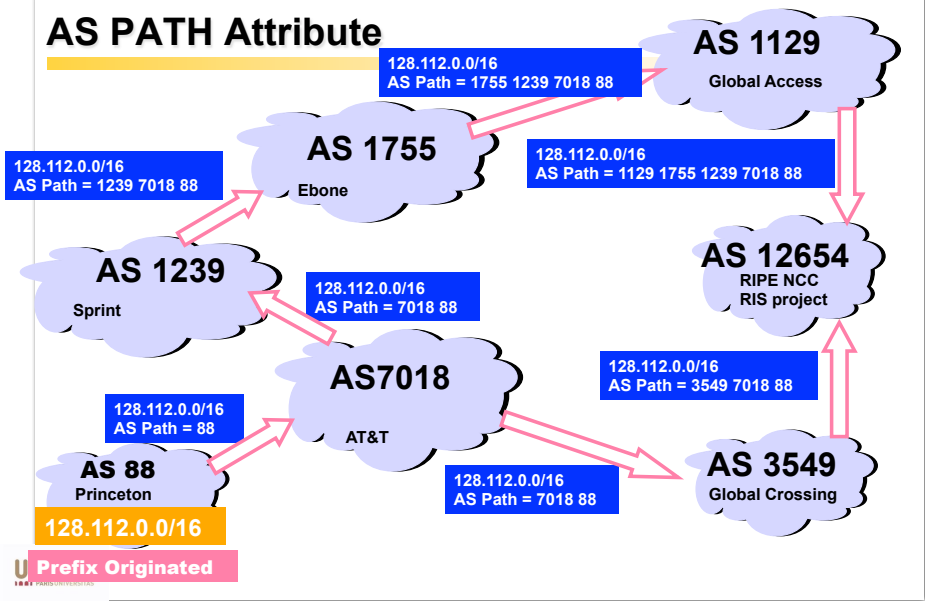


- Route reflectors can pass on iBGP updates to clients
- Each RR passes along ONLY best routes
- ORIGINATOR_ID and CLUSTER_LIST attributes are needed to avoid loops

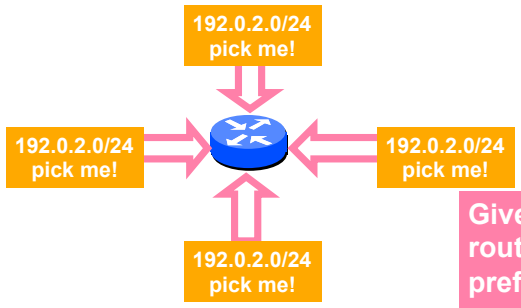
Constructing the Forwarding Table

- 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

AS PATH Attribute



Attributes are Used to Select Best Routes

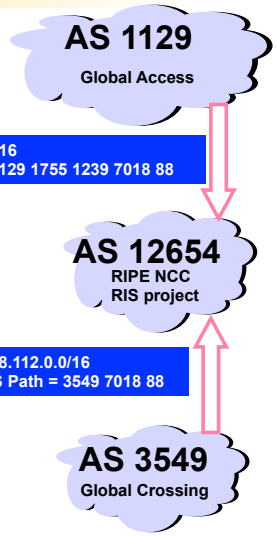


Given multiple routes to the same prefix, a BGP speaker must pick at most one best route

(Note: it could reject them all!)

BGP Path Selection

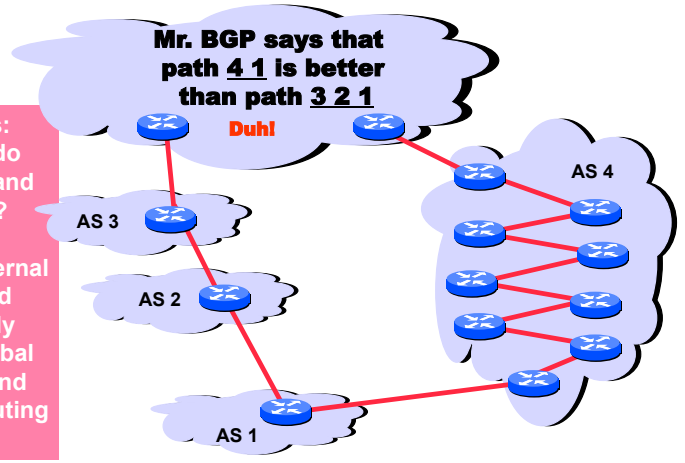
- Simplest case
  - Shortest AS path
  - Arbitrary tie break
- Example
  - Four-hop AS path preferred over a three-hop AS path
  - AS 12654 prefers path through Global Crossing
- But, BGP is not limited to shortest-path routing
  - Policy-based routing



Shorter Doesn't Always Mean Shorter

In fairness: could you do this "right" and still scale?

Exporting internal state would dramatically increase global instability and amount of routing state



BGP Attributes

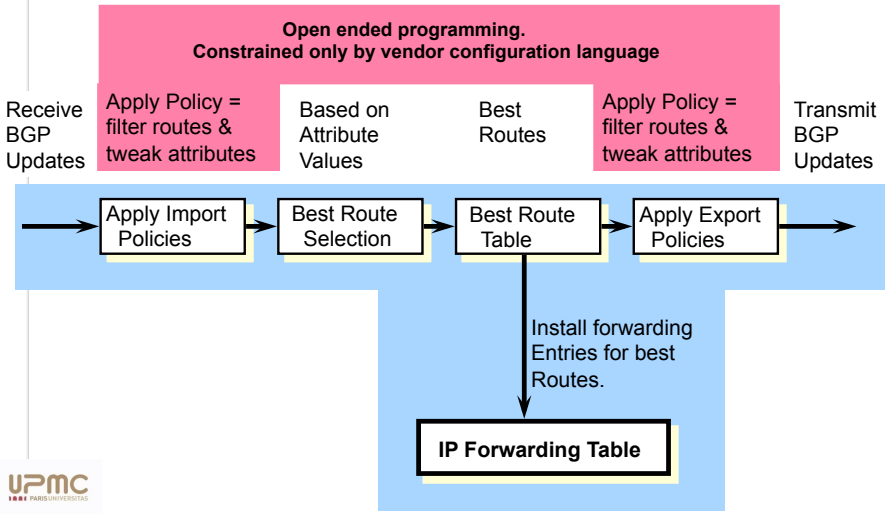
Value	Code	Reference
1	ORIGIN	RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH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

Most important attributes

Not all attributes need to be present in every announcement

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

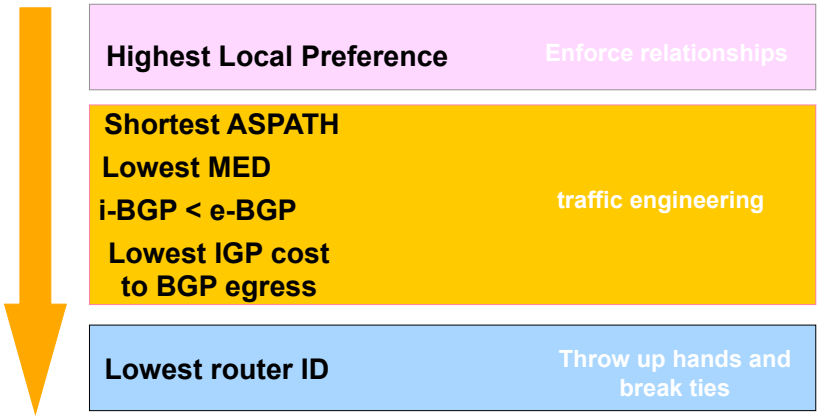
BGP Policy: Influencing Decisions



BGP Decision Process: Path Selection on a Router

- Routing Information Base
  - Store all BGP routes for each destination prefix
  - Withdrawal message: remove the route entry
  - Advertisement message: update the route entry
- Selecting the best route
  - Consider all BGP routes for the prefix
  - Apply rules for comparing the routes
  - Select the one best route
    - Use this route in the forwarding table
    - Send (optionally) this route to neighbors

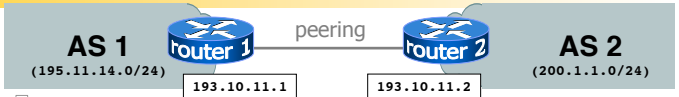
Route Selection Summary



BGP Policy: Applying Policy to Routes

- Import policy
  - Filter unwanted routes from neighbor
    - E.g. prefix that your customer doesn't own
  - Manipulate attributes to influence path selection
    - E.g., assign local preference to favored routes
- Export policy
  - Filter routes you don't want to tell your neighbor
    - E.g., don't tell a peer a route learned from other peer
  - Manipulate attributes to control what they see
    - E.g., make a path look artificially longer than it is

announcement example




```
! router 1 configuration file
router bgp 1
network 195.11.14.0/24
neighbor 193.10.11.2 remote-as 2
```

```
! router 2 configuration file
router bgp 2
network 200.1.1.0/24
neighbor 193.10.11.1 remote-as 1
```

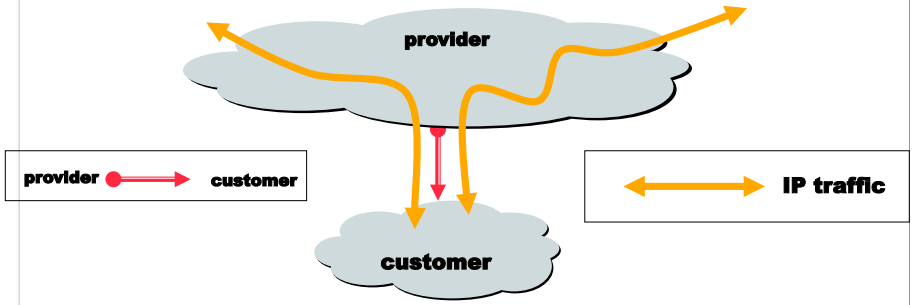
prefix filtering: example

zebra configuration file

```
router bgp 1
network 195.11.14.0/24
network 195.11.15.0/24
neighbor 193.10.11.2 remote-as 2
neighbor 193.10.11.2 description Router 2 of AS2
neighbor 193.10.11.2 prefix-list partialOut out
neighbor 193.10.11.2 prefix-list partialIn in
!
ip prefix-list partialOut permit 195.11.14.0/24
!
ip prefix-list partialIn deny 200.1.1.0/24
ip prefix-list partialIn permit any
```

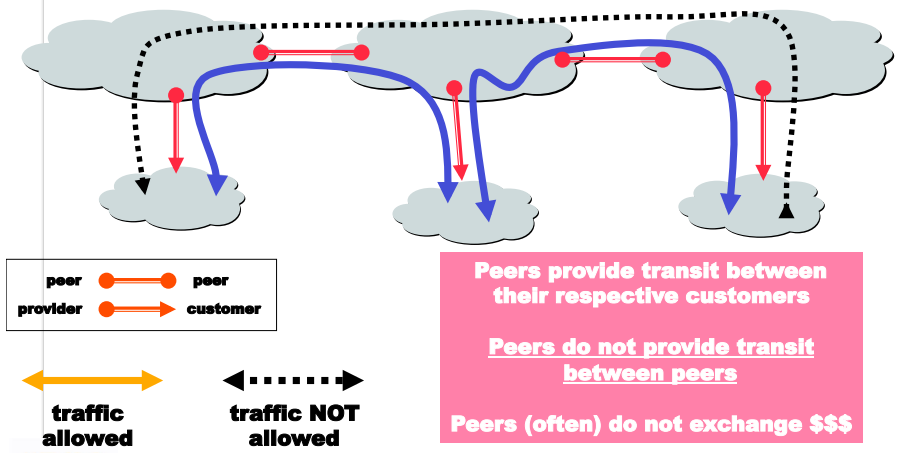
only `195.11.14.0/24` is announced to neighbor `193.10.11.2`
all with the exception of `200.1.1.0/24` is accepted from `193.10.11.2`

Customers and Providers

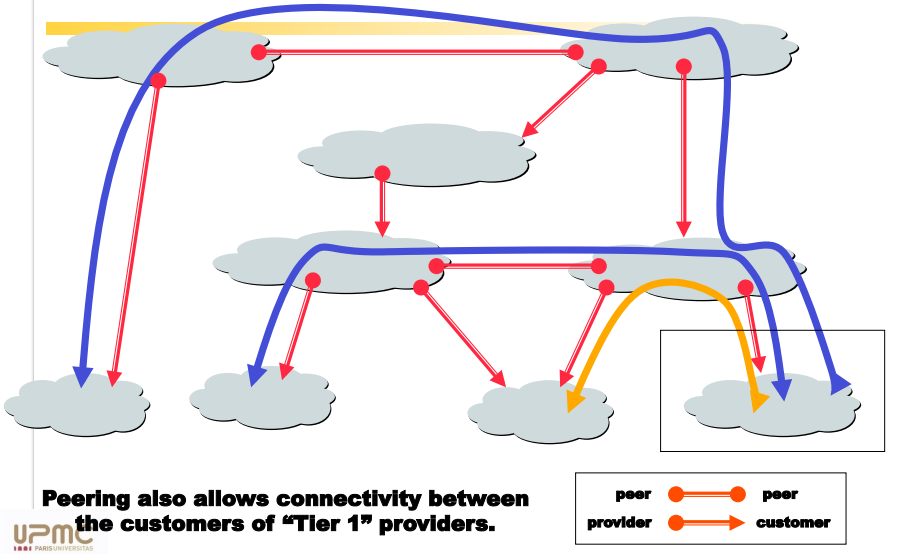


Customer pays provider for access to the Internet

The "Peering" Relationship

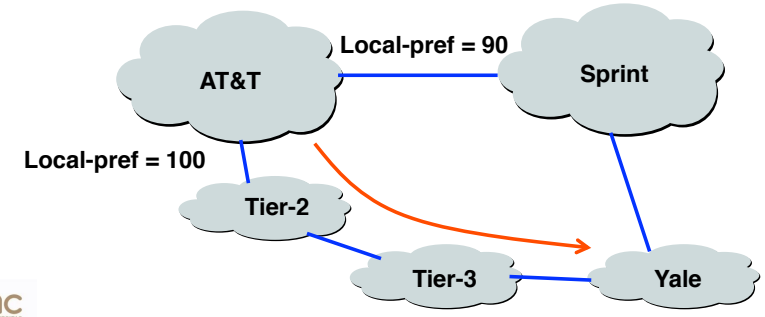


Peering Provides Shortcuts



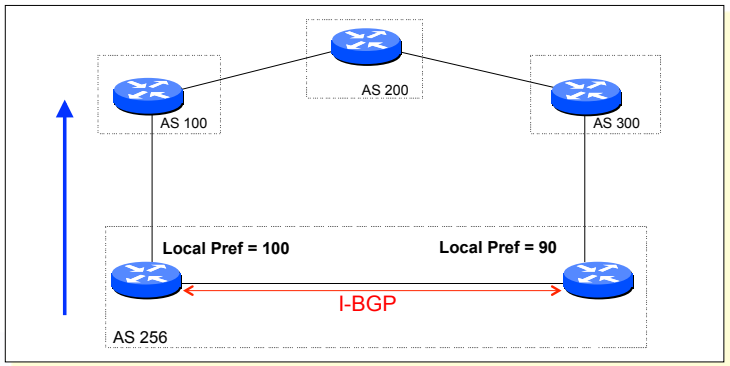
Import Policy: Local Preference

- Favor one path over another
  - Override the influence of AS path length
  - Apply local policies to prefer a path
- Example: prefer customer over peer



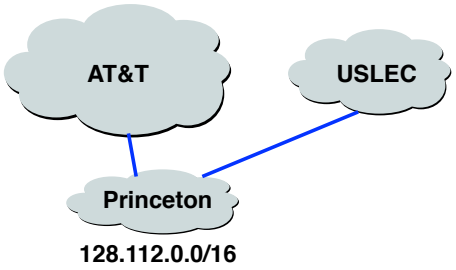
Internal BGP and Local Preference

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



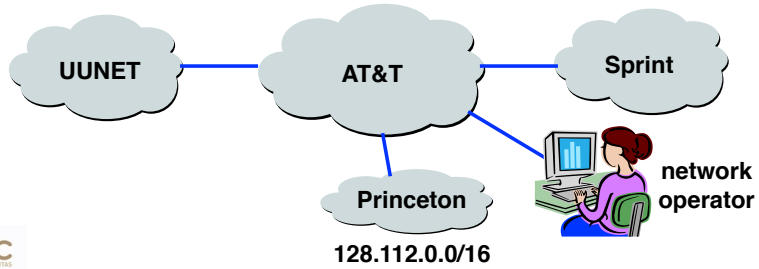
Import Policy: Filtering

- Discard some route announcements
  - Detect configuration mistakes and attacks
- Examples on session to a customer
  - Discard route if prefix not owned by the customer
  - Discard route that contains other large ISP in AS path



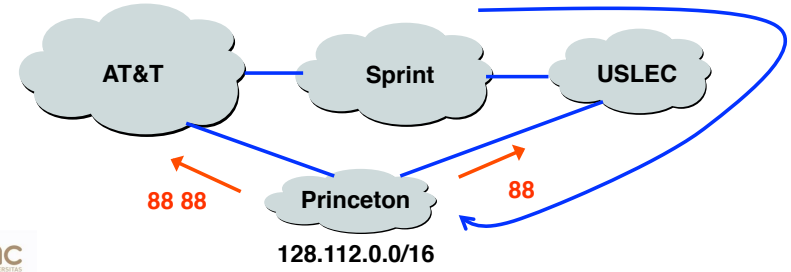
Export Policy: Filtering

- Discard some route announcements
  - Limit propagation of routing information
- Examples
  - Don't announce routes from one peer to another
  - Don't announce routes for network-management hosts



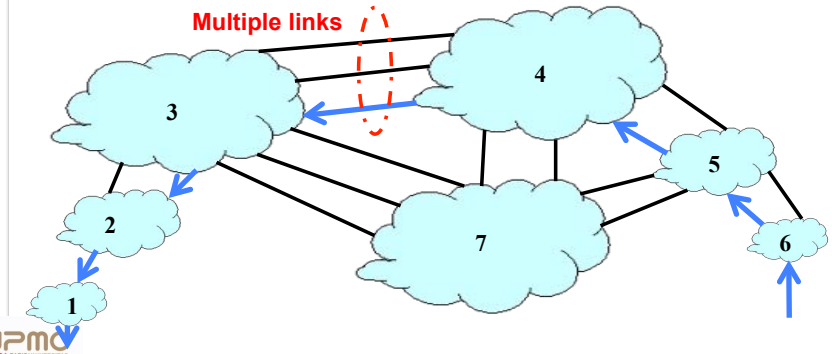
Export Policy: Attribute Manipulation

- Modify attributes of the active route
  - To influence the way other ASes behave
- Example: AS prepending
  - Artificially inflate the AS path length seen by others
  - To convince some ASes to send traffic another way



An AS is Not a Single Node

- Multiple connections to neighboring ASes
  - Multiple border routers may learn good routes
  - … with the same local-pref and AS path length

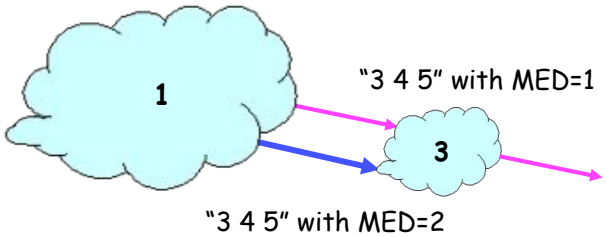


Multiple Exit Discriminator Attribute (MED)

- when AS's interconnected via 2 or more links
- AS announcing prefix sets MED
- enables AS(3) to indicate its preference
- AS(1) receiving prefix uses MED to select link
- a way to specify how close a prefix is to the link it is announced on

Multiple Exit Discriminator Attribute (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



BGP Policy Configuration

- Routing policy languages are vendor-specific
  - Not part of the BGP protocol specification
  - Different languages for Cisco, Juniper, etc.
- Still, all languages have some key features
  - Policy as a list of clauses
  - Each clause matches on route attributes
  - … and either discards or modifies the matching routes
- Configuration done by human operators
  - Implementing the policies of their AS
  - Business relationships, traffic engineering, security, …

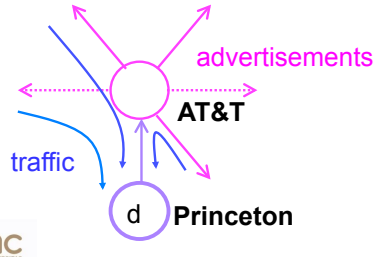
Policies in Practice : Business Relationships

- Common relationships
  - Customer-provider
  - Peer-peer
  - Backup, sibling, …
- Implementing in BGP
  - Import policy
    - Ranking customer routes over peer routes
  - Export policy
    - Export only customer routes to peers and providers

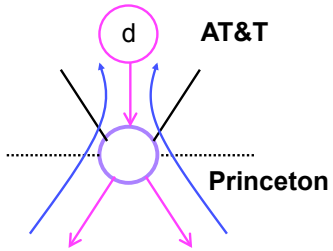
Customer-Provider Relationship

- Customer pays provider for access to Internet
  - Customer needs to be reachable from everyone
  - Provider exports customer's routes to everybody
  - Customer exports provider's routes to customers
- Customer does not want to provide transit service
  - Customer does not export from one provider to another

Traffic **to** the customer



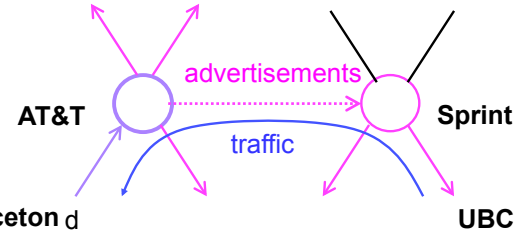
Traffic **from** the customer



Peer-Peer Relationship

- Peers exchange traffic between customers
  - AS exports *only* customer routes to a peer
  - AS exports a peer's routes *only* to its customers

Traffic to/from the peer and its customers



How Peering Decisions are Made?

Peer

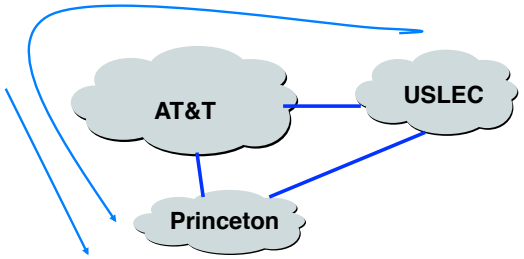
- Reduces upstream transit costs
- Can increase end-to-end performance
- May be the only way to connect your customers to some part of the Internet ("Tier 1")

Don't Peer

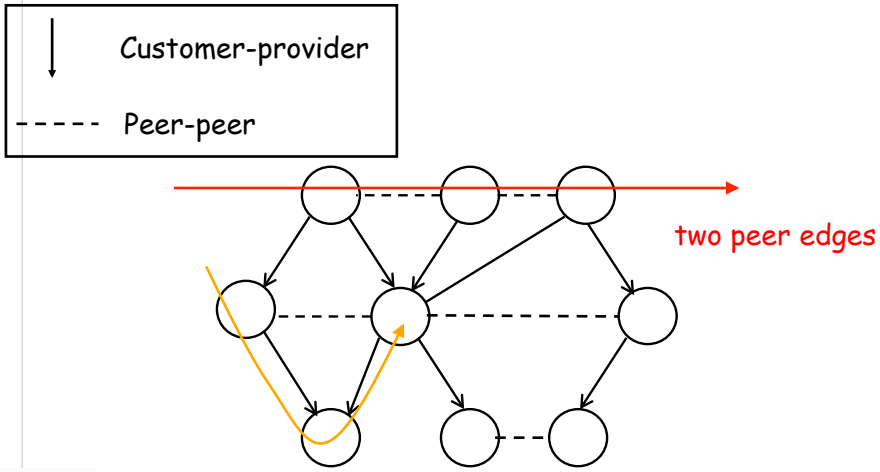
- You would rather have customers
- Peers are usually your competition
- Peering relationships may require periodic renegotiation

Backup Relationship

- Backup provider
  - Only used if the primary link fails
  - Routes through other paths

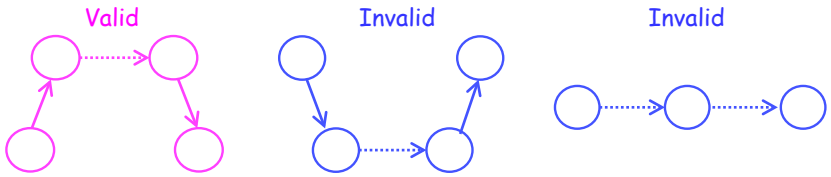


Paths You Should Never See ("Invalid")



Valid and Invalid Paths

- AS relationships limit the kinds of valid paths
  - Uphill portion: customer-provider relationships
  - Plateau: zero or one peer-peer edge
  - Downhill portion: provider-customer relationships

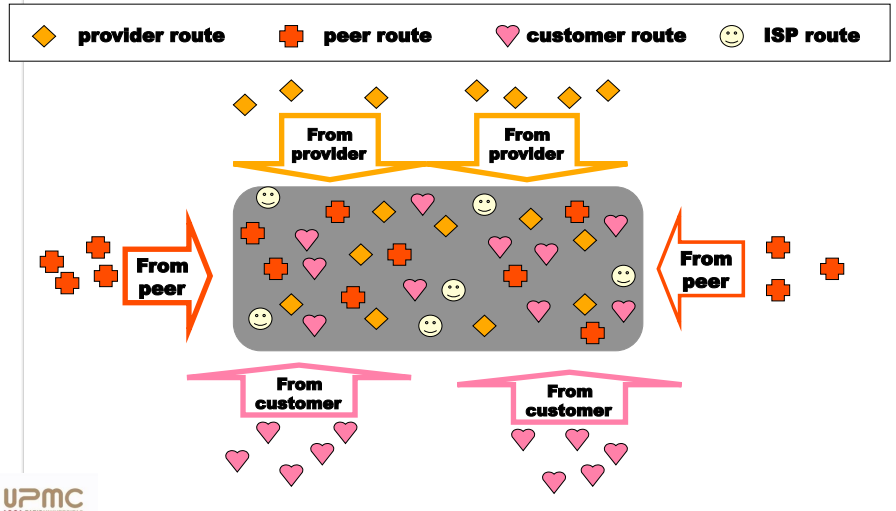


Implementing Customer/Provider and Peer/Peer relationships

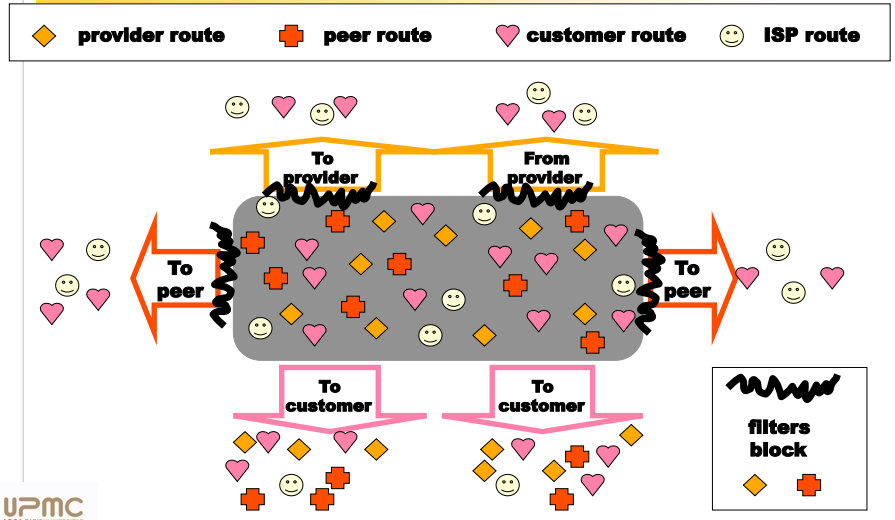
Two parts:

- Enforce transit relationships
  - Outbound route filtering
- Enforce order of route preference
  - provider < peer < customer

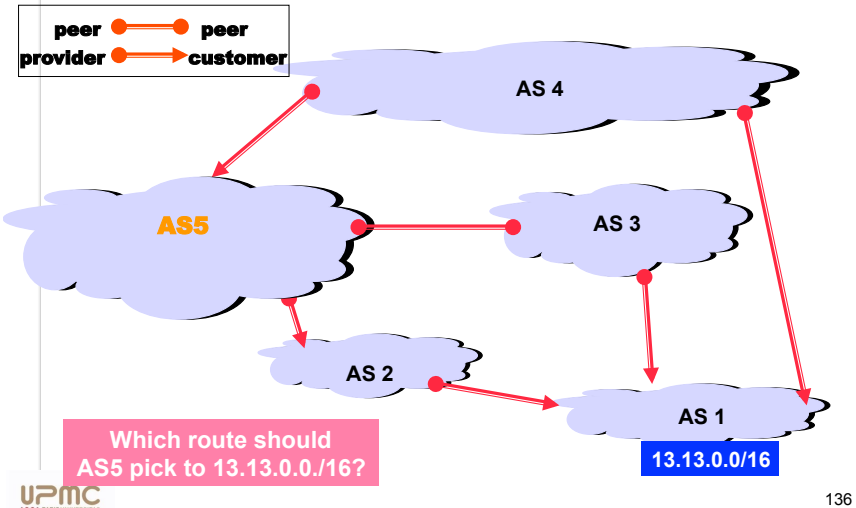
Import Routes



Export Routes



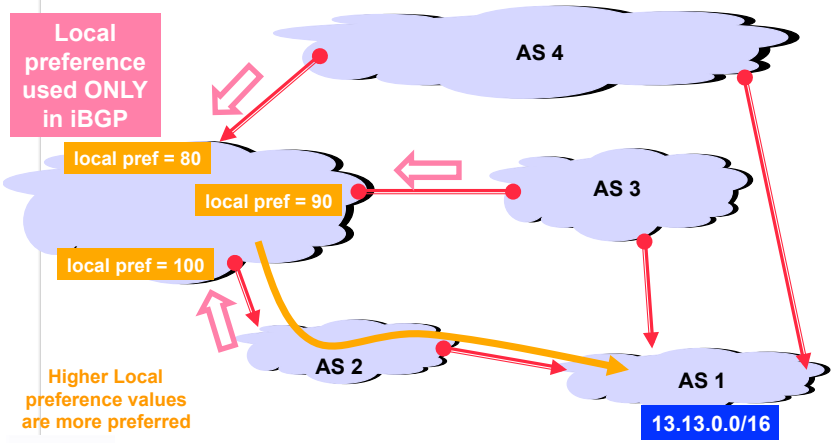
So Many Choices



Which route should AS5 pick to 13.13.0.0./16?

LOCAL PREFERENCE

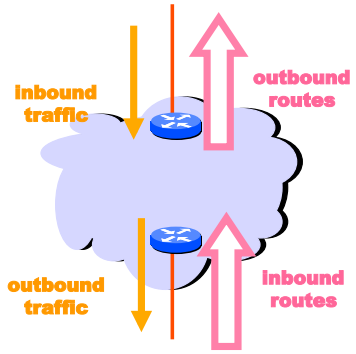
Local preference used ONLY in iBGP



Higher Local preference values are more preferred

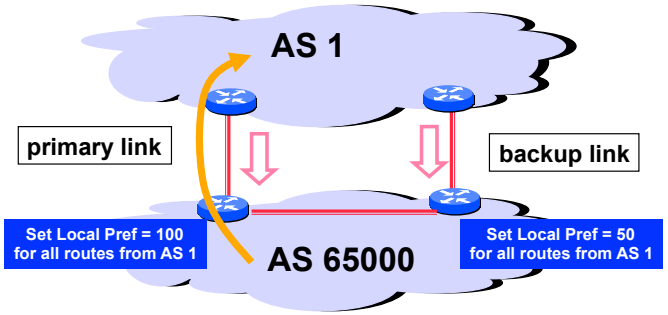
Tweak Tweak Tweak

- 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



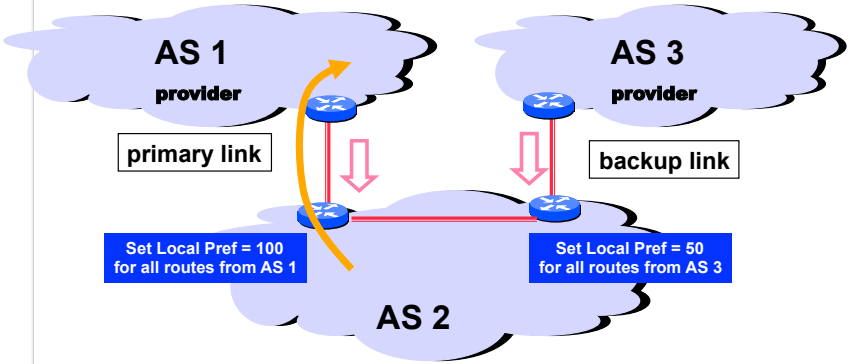
In general, an AS has more control over outbound traffic

Implementing Backup Links with Local Preference (Outbound Traffic)



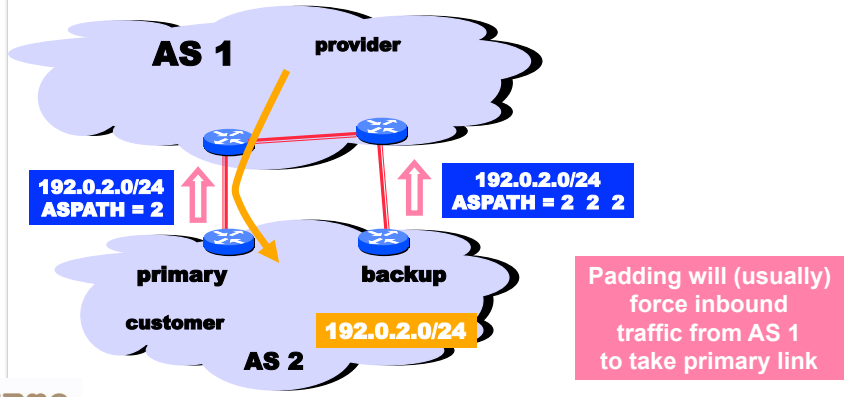
Forces outbound traffic to take primary link, unless link is down.

Multihomed Backups (Outbound Traffic)



Forces outbound traffic to take primary link, unless link is down.

Shedding Inbound Traffic with ASPATH Padding. Yes, this is a Glorious Hack …



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

Traffic Engineering

- Load balancing
  - Making good use of network resources
  - Alleviating network congestion
- End-to-end performance
  - Avoiding paths with downstream congestion
  - By moving traffic to alternate paths
- Mechanisms
  - Preferring some paths over other paths
  - E.g., by setting local-preference attribute
  - Among routes within the same business class

Route Stability

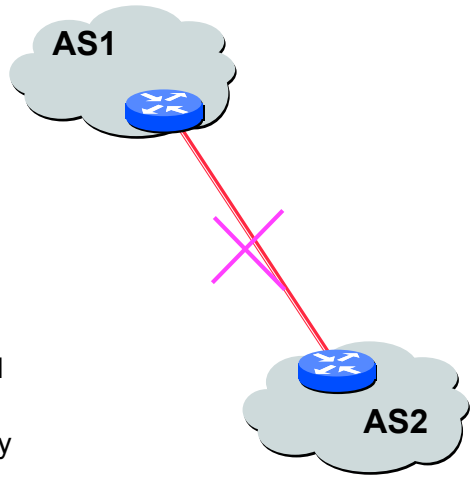
- Routing instability: rapid fluctuation of network reachability information
- route flapping: when a route is withdrawn and re-announced repeatedly in a short period of time
  - happens via UPDATE messages
- because messages propagate to global Internet, route flapping behavior can cascade and deteriorate routing performance in many places
- Effects: increased packet loss, increased network latency, CPU overhead, loss of connectivity

Causes of BGP Routing Changes

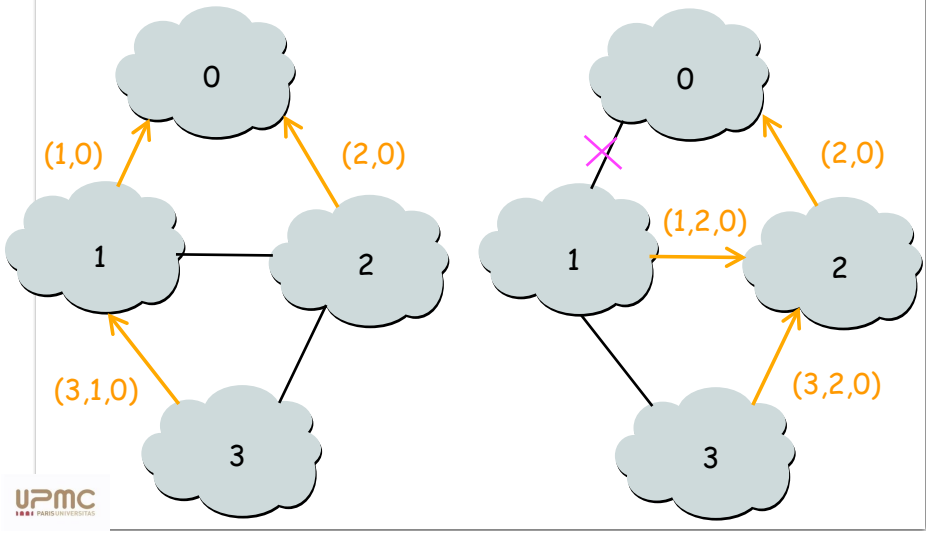
- Topology changes
  - Equipment going up or down
  - Deployment of new routers or sessions
- BGP session failures
  - Due to equipment failures, maintenance, etc.
  - Or, due to congestion on the physical path
- Changes in routing policy
  - Reconfiguration of preferences
  - Reconfiguration of route filters
- Persistent protocol oscillation
  - Conflicts between policies in different ASes

BGP Session Failure

- BGP runs over TCP
  - BGP only sends updates when changes occur
  - TCP doesn't detect lost connectivity on its own
- Detecting a failure
  - Keep-alive: 60 seconds
  - Hold timer: 180 seconds
- Reacting to a failure
  - Discard all routes learned from the neighbor
  - Send new updates for any routes that change

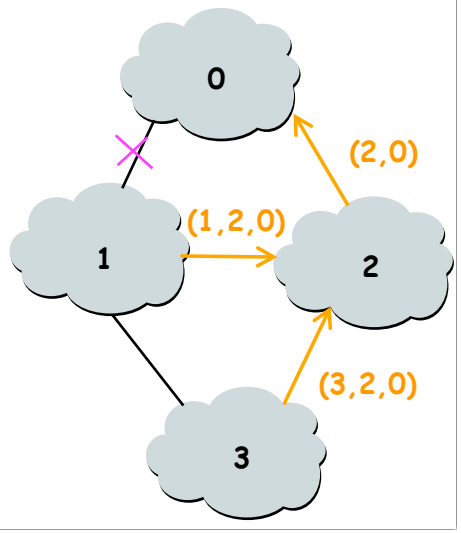


Routing Change: Before and After



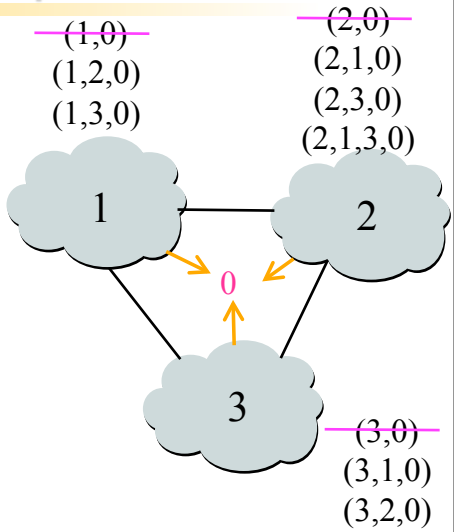
Routing Change: Path Exploration

- AS 1
  - Delete the route (1,0)
  - Switch to next route (1,2,0)
  - Send route (1,2,0) to AS 3
- AS 3
  - Sees (1,2,0) replace (1,0)
  - Compares to route (2,0)
  - Switches to using AS 2



Routing Change: Path Exploration

- Initial situation
  - Destination 0 is alive
  - All ASes use direct path
- When destination dies
  - All ASes lose direct path
  - All switch to longer paths
  - Eventually withdrawn
- E.g., AS 2
  - (2,0) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null



Time Between Steps in Path Exploration

- Minimum route advertisement interval (MRAI)
  - Minimum spacing between announcements
  - For a particular (prefix, peer) pair
- Advantages
  - Provides a rate limit on BGP updates
  - Allows grouping of updates within the interval
- Disadvantages
  - Adds delay to the convergence process
  - E.g., 30 seconds for each step

BGP Converges Slowly, if at All

- Path vector avoids count-to-infinity
  - But, ASes still must explore many alternate paths
  - … to find the highest-ranked path that is still available
- Fortunately, in practice
  - Most popular destinations have very stable BGP routes
  - And most instability lies in a few unpopular destinations
- Still, lower BGP convergence delay is a goal
  - Can be tens of seconds to tens of minutes
  - High for important interactive applications
  - … or even conventional application, like Web browsing

Conclusions

- BGP is solving a hard problem
  - Routing protocol operating at a global scale
  - With tens of thousands of independent networks
  - That each have their own policy goals
  - And all want fast convergence
- Key features of BGP
  - Prefix-based path-vector protocol
  - Incremental updates (announcements and withdrawals)
  - Policies applied at import and export of routes
  - Internal BGP to distribute information within an AS
  - Interaction with the IGP to compute forwarding tables