CS 640: Introduction to Computer Networks

Aditya Akella

Lecture 11 -
Inter-Domain Routing -
BGP (Border Gateway Protocol)

Intra-domain routing

- The Story So Far…
  - Routing protocols generate the forwarding table
  - Two styles: distance vector, link state
  - Scalability issues:
    - Distance vector protocols suffer from count-to-infinity
    - Link state protocols must flood information through network
- Today's lecture
  - How to make routing protocols support large networks
  - How to make routing protocols support business policies

2

Inter-domain Routing: Hierarchy

- "Flat" routing not suited for the Internet
  - Doesn't scale with network size
    - Storage → Each node cannot be expected to store routes to every destination (or destination network)
    - Convergence times increase
    - Communication → Total message count increases
  - Administrative autonomy
    - Each internetwork may want to run its network independently
      - E.g hide topology information from competitors
- Solution: Hierarchy via autonomous systems

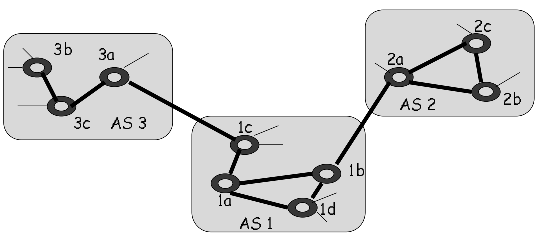
3

Internet's Hierarchy

- What is an Autonomous System (AS)?
  - A set of routers under a single technical administration
    - Use an *interior gateway protocol (IGP)* and common metrics to route packets within the AS
    - Connect to other ASes using *gateway routers*
    - Use an *exterior gateway protocol (EGP)* to route packets to other AS's
  - IGP: OSPF, RIP (last class)
  - Today's EGP: BGP version 4

4

An example



Intra-AS routing algorithm + Inter-AS routing algorithm → Forwarding table

5

The Problem

- Easy when only one link leading to outside AS
- Much harder when two or more links to outside ASes
  - Which destinations reachable via a neighbor?
  - Propagate this information to other internal routers
  - Select a "good route" from multiple choices
  - Inter-AS routing protocol
    - Communication between distinct ASes
    - Must be the same protocol!

6

History

- Mid-80s: EGP
  - Reachability protocol (no shortest path)
  - Did not accommodate cycles (tree topology)
  - Evolved when all networks connected to NSF backbone
- Result: BGP introduced as routing protocol
  - Latest version = BGP 4
  - BGP-4 supports CIDR
  - Primary objective: connectivity *not* performance

7

BGP Preliminaries

- Pairs of routers exchange routing info over TCP connections (port 179)
  - One TCP connection for every pair of neighboring gateway routers
  - Routers called "BGP peers"
  - BGP peers exchange routing info as messages
  - TCP connection + messages → BGP session
- Neighbor ASes exchange info on which CIDR prefixes are reachable via them

8

Choices for Routing

- How to propagate routing information?
- Link state or distance vector?
  - No universal metric – policy decisions
  - Problems with distance-vector:
    - Very slow convergence
  - Problems with link state:
    - Metric used by ISPs not the same → loops
    - LS database too large – entire Internet
- BGP: Path vector

9

AS Numbers (ASNs)

ASNs are 16 bit values 64512 through 65535 are "private"

Currently over 15,000 in use

- Genuity: 1
- MIT: 3
- CMU: 9
- 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

10

Distance Vector with Path

- Each routing update carries the entire AS-level path so far
  - "AS_Path attribute"
- Loops are detected as follows:
  - When AS gets route, check if AS already in path
    - If yes, reject route
    - If no, add self and (possibly) advertise route further
      - Advertisement depends on metrics/cost/preference etc.
- Advantage:
  - Metrics are local - AS chooses path, protocol ensures no loops

11

Hop-by-hop Model

- BGP advertises to neighbors only those routes that it uses
  - Consistent with the hop-by-hop Internet paradigm
  - Consequence: hear only one route from neighbor
    - (although neighbor may have chosen this from a large set of choices)
    - Could impact view into availability of paths

12

Policy with BGP

- BGP provides capability for enforcing various policies
- Policies are **not** part of BGP: they are provided to BGP as configuration information
- **Enforces** policies by
  - *Choosing appropriate paths* from multiple alternatives
  - *Controlling advertisement* to other AS's

13

Examples of BGP Policies

- A multi-homed AS refuses to act as transit
  - Limit path advertisement
- A multi-homed AS can become transit for some AS's
  - Only advertise paths to some AS's
- An AS can favor or disfavor certain AS's for traffic transit from itself

14

BGP Messages

- Open
  - Announces AS ID
  - Determines hold timer – interval between keep_alive or update messages, zero interval implies no keep_alive
- Keep_alive
  - Sent periodically (but before hold timer expires) to peers to ensure connectivity.
  - Sent in place of an UPDATE message
- Notification
  - Used for error notification
  - TCP connection is closed *immediately* after notification

15

BGP UPDATE Message

- List of withdrawn routes
- Network layer reachability information
  - List of reachable prefixes
- Path attributes
  - Origin
  - Path
  - Local_pref
  - MED
  - Metrics
- All prefixes advertised in message have same path attributes

16

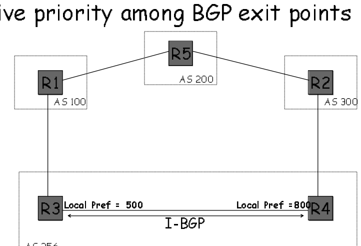
Path Selection Criteria

- Attributes + external (policy) information
- Examples:
  - Policy considerations
    - Preference for AS
    - Presence or absence of certain AS
  - Hop count
  - Path origin

17

LOCAL PREF

- Local (within an AS) mechanism to provide relative priority among BGP exit points

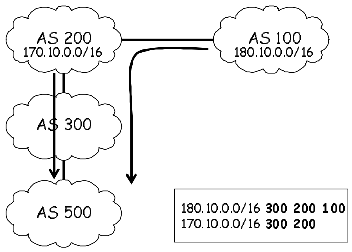


- Prefer routers announced by one AS over another or general preference over routes

18

AS_PATH

- List of traversed AS's



19

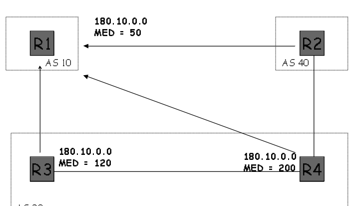
Multi-Exit Discriminator (MED)

- Hint to external neighbors about the preferred path *into* an AS
  - Different AS choose different scales
- Used when two AS's connect to each other in more than one place
  - More useful in a customer provider setting
  - Not honored in other settings
    - Will see later why

20

MED

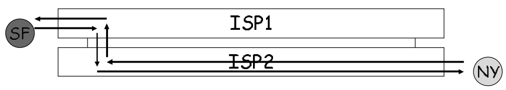
- Hint to R1 to use R3 over R4 link
- Cannot compare AS40's values to AS30's



21

MED

- MED is typically used in provider/subscriber scenarios
- It can lead to unfairness if used between ISP because it may force one ISP to carry more traffic:



- ISP1 ignores MED from ISP2
- ISP2 obeys MED from ISP1
- ISP2 ends up carrying traffic most of the way

22

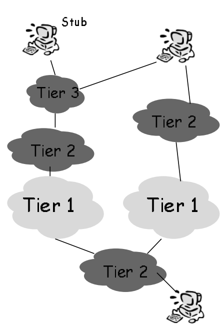
Decision Process (First cut)

- Rough processing order of attributes:
  - Select route with highest LOCAL-PREF
  - Select route with shortest AS-PATH
  - Apply MED (to routes learned from same neighbor)
- How to set the attributes?
  - Especially local_pref?
  - Policies in action

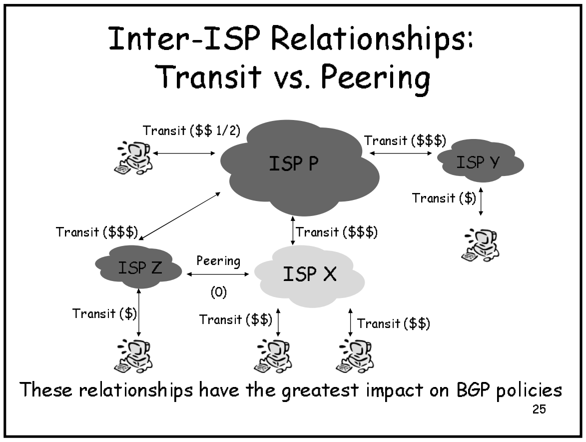
23

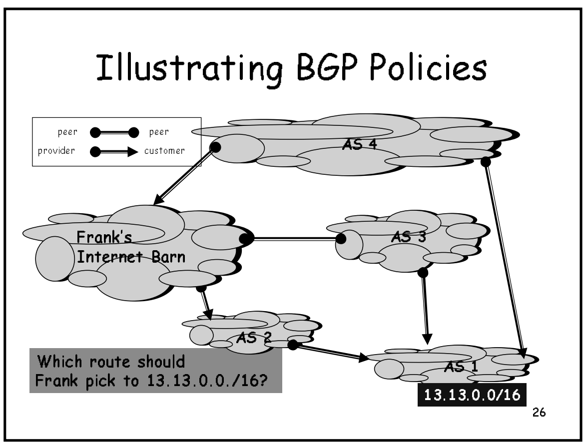
A Logical View of the Internet

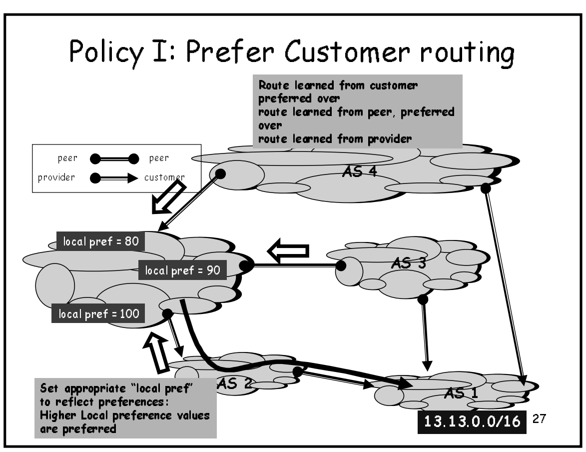
- Tier 1 ISP
  - "Default-free" with global reachability info
- Tier 2 ISP
  - Regional or country-wide
  - Typically route through tier-1
    - Customer
- Tier 3/4 ISPs
  - Local
  - Route through higher tiers
- Stub AS
  - End network such as IBM or UW-Madison



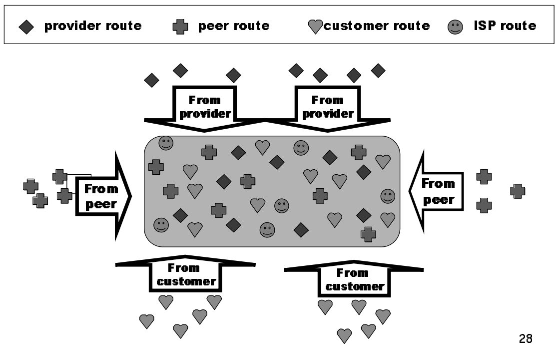
24


Inter-ISP Relationships:
Transit vs. Peering
Transit ($$ 1/2)
ISP P
Transit ($$$)
ISP Y
Transit ($)
Transit ($$$)
Transit ($$$)
ISP Z
Peering
(0)
ISP X
Transit ($)
Transit ($$)
Transit ($$)
These relationships have the greatest impact on BGP policies
25


Illustrating BGP Policies
peer
peer
provider
customer
AS 4
Frank's
Internet Barn
AS 3
AS 2
AS 1
Which route should
Frank pick to 13.13.0.0./16?
13.13.0.0/16
26

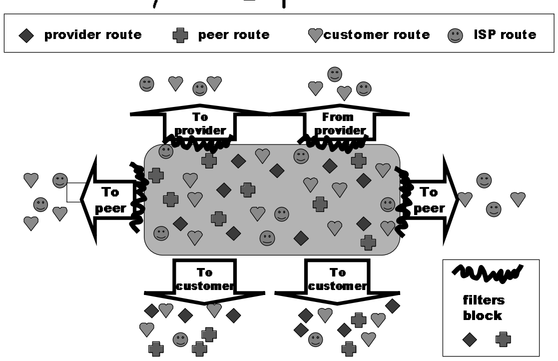

Policy I: Prefer Customer routing
Route learned from customer
preferred over
route learned from peer, preferred
over
route learned from provider
peer
peer
provider
customer
AS 4
local pref = 80
local pref = 90
AS 3
local pref = 100
AS 2
AS 1
Set appropriate "local pref"
to reflect preferences:
Higher Local preference values
are preferred
13.13.0.0/16
27

Policy II: Import Routes



28

Policy II: Export Routes



Policy II: Valley-Free Routes

- "Valley-free" routing
  - Number links as (+1, 0, -1) for provider, peer and customer
  - In any *valid* path should only see sequence of +1, followed by at most one 0, followed by sequence of -1
  - Why?
    - Consider the economics of the situation
- How to make these choices?
  - Prefer-customer routing: LOCAL_PREF
  - Valley-free routes: control route advertisements (see previous slide)

30

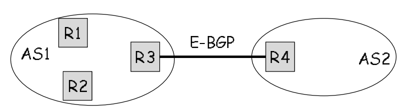
BGP Route Selection Summary

Highest Local Preference	Enforce relationships E.g. prefer customer routes over peer routes
Shortest ASPATH Lowest MED i-BGP < e-BGP Lowest IGP cost to BGP egress	traffic engineering
Lowest router ID	Throw up hands and break ties

31

Internal vs. External BGP

- BGP can be used by R3 and R4 to learn routes
- How do R1 and R2 learn best routes?



- Use I-BGP
- Create a full mesh
  - TCP connections
- Use this to exchanged BGP route information

32

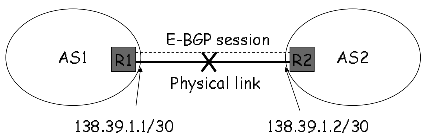
Link Failures

- Two types of link failures:
  - Failure on an E-BGP link
  - Failure on an I-BGP Link
- These failures are treated completely different in BGP
- Why?

33

Failure on an E-BGP Link

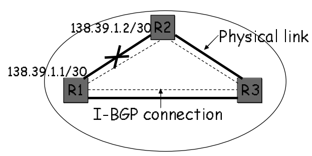
- If the link R1-R2 goes down
  - The TCP connection breaks
  - BGP routes are removed
- This is the *desired* behavior



34

Failure on an I-BGP Link

- If link R1-R2 goes down, R1 and R2 should still be able to exchange traffic
- The indirect path through R3 must be used
- Thus, E-BGP and I-BGP must use *different conventions* with respect to TCP endpoints



35

Next Class

- Multicast
  - Service model
  - IGMP
  - IP Multicast routing protocols
  - Overlay-based multicast

36