15-441 Computer Networking

Inter-Domain Routing

BGP (Border Gateway Protocol)

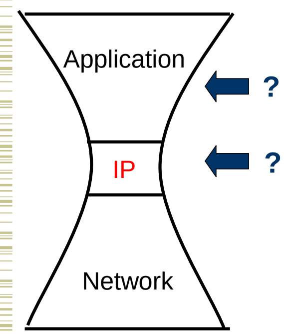
Review

- Overlay Multicast

Failure of IP Multicast

- Not widely deployed even after 15 years!
  - Use carefully – e.g., on LAN or campus, rarely over WAN
- Various failings
  - Scalability of routing protocols
  - Hard to manage
  - Hard to implement TCP equivalent
  - Hard to get applications to use IP Multicast without existing wide deployment
  - Hard to get router vendors to support functionality and hard to get ISPs to configure routers to enable

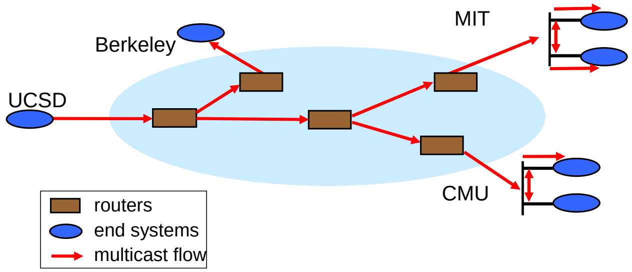
Supporting Multicast on the Internet



Internet architecture

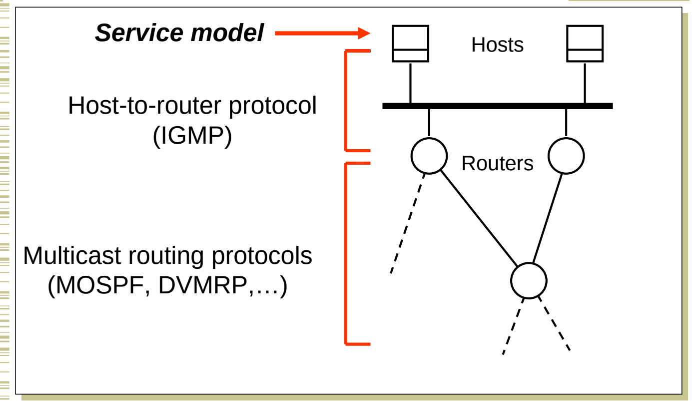
At which layer should multicast be implemented?

IP Multicast

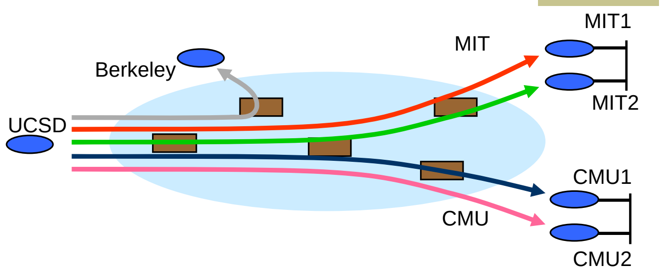


- Highly efficient
- Good delay

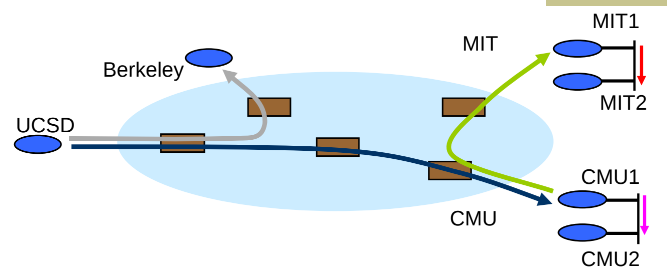
IP Multicast Architecture



Naïve Overlay Multicast

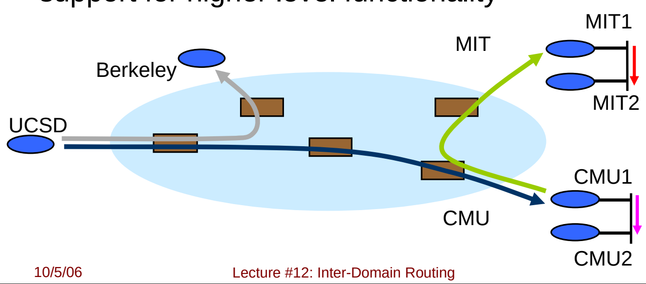


Smart Overlay Multicast



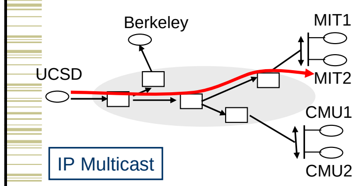
Benefits Over IP Multicast

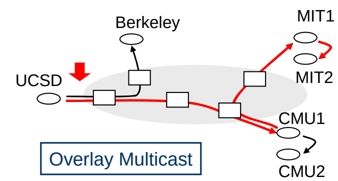
- Quick deployment
- All multicast state in end systems
- Computation at forwarding points simplifies support for higher level functionality



Concerns with Overlay Multicast

- Self-organize recipients into multicast delivery overlay tree
  - Must be closely matched to real network topology to be efficient
- Performance concerns compared to IP Multicast
  - Increase in delay
  - Bandwidth waste (packet duplication)





Important Multicast Concepts

- Multicast provides support for efficient data delivery to multiple recipients
- Requirements for IP Multicast routing
  - Keeping track of interested parties
  - Building distribution tree
  - Broadcast/suppression technique
- Difficult to deploy new IP-layer functionality
- End system-based techniques can provide similar efficiency
  - Easier to deploy

Routing Review

- 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

Outline

- Routing hierarchy
- Internet structure
- External BGP (E-BGP)

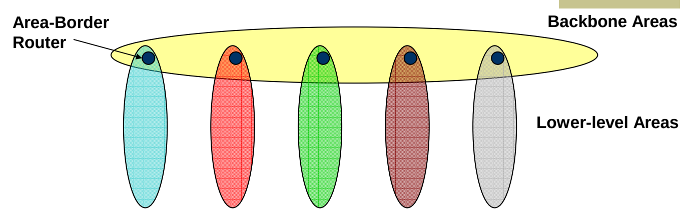
Routing Hierarchies

- Flat routing doesn't scale
  - Storage → Each node cannot be expected to store routes to every destination (or destination network)
  - Convergence times increase
  - Communication → Total message count increases
- Key observation
  - Need less information with increasing distance to destination
  - Need lower diameters networks
- Solution: area hierarchy

Areas

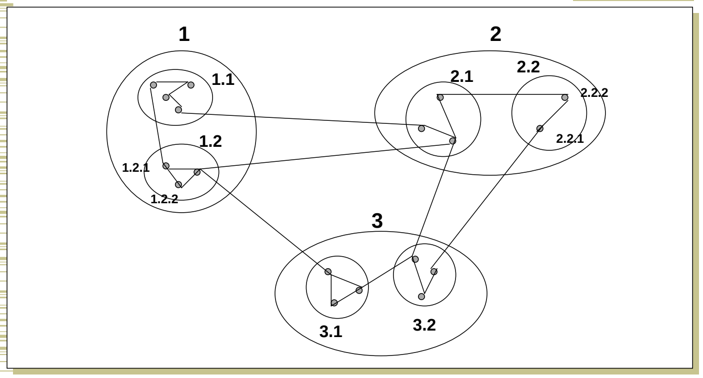
- Divide network into areas
  - Areas can have nested sub-areas
- Hierarchically address nodes in a network
  - Sequentially number top-level areas
  - Sub-areas of area are labeled relative to that area
  - Nodes are numbered relative to the smallest containing area

Routing Hierarchy



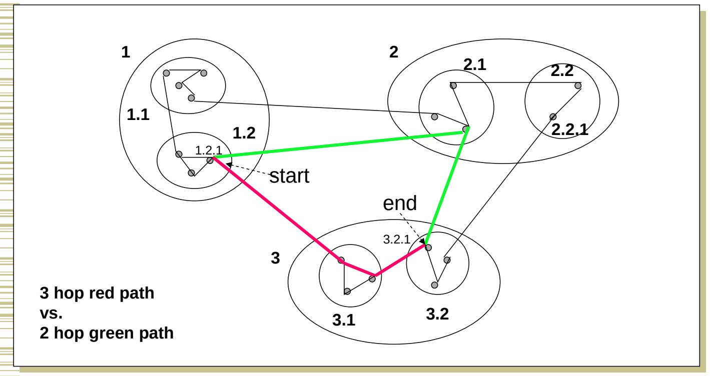
- Partition Network into "Areas"
  - Within area
    - Each node has routes to every other node
  - Outside area
    - Each node has routes for other top-level areas only
    - Inter-area packets are routed to nearest appropriate border router
- Constraint: no path between two sub-areas of an area can exit that area

Area Hierarchy Addressing



Path Sub-optimality

- Can result in sub-optimal paths

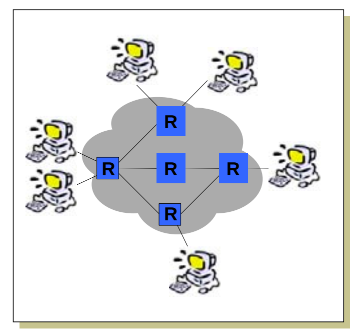


Outline

- Routing hierarchy
- Internet structure
- External BGP (E-BGP)

A Logical View of the Internet?

- After looking at RIP/OSPF descriptions
  - End-hosts connected to routers
  - Routers exchange messages to determine connectivity
- NOT TRUE!



Internet's Area Hierarchy

- What is an Autonomous System (AS)?
  - A set of routers under a single technical administration, using an *interior gateway protocol (IGP)* and common metrics to route packets within the AS and using an *exterior gateway protocol (EGP)* to route packets to other AS's
- Each AS assigned unique ID
- AS's peer at network exchanges

AS Numbers (ASNs)

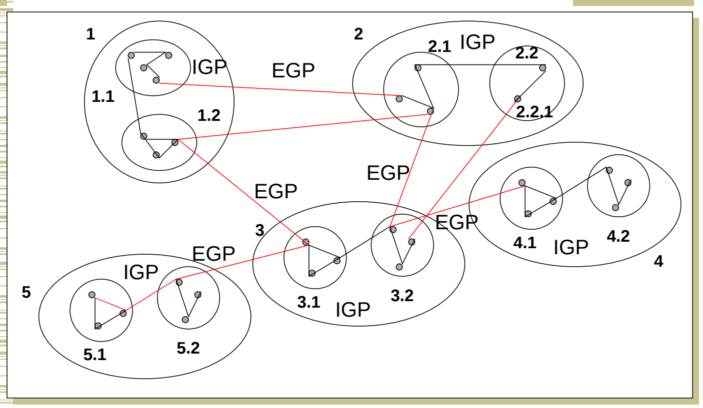
ASNs are 16 bit values 64512 through 65535 are "private"

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, …
- …

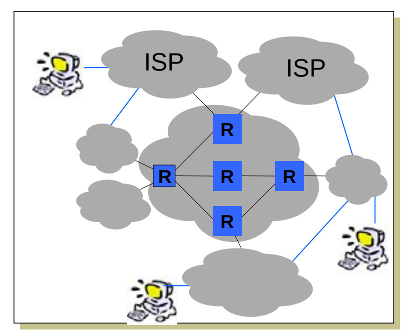
ASNs represent units of routing policy

Example



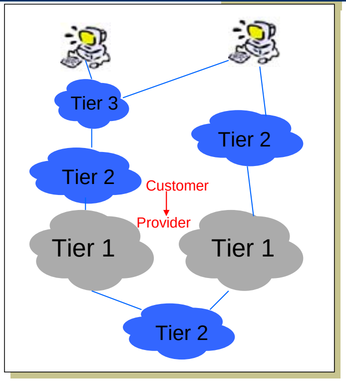
A Logical View of the Internet?

- RIP/OSPF not very scalable → area hierarchies
- NOT TRUE EITHER!
- ISP's aren't equal
  - Size
  - Connectivity

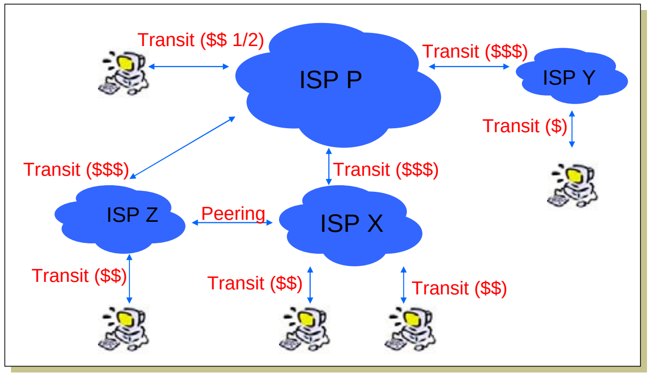


A Logical View of the Internet

- Tier 1 ISP
  - "Default-free" with global reachability info
- Tier 2 ISP
  - Regional or country-wide
- Tier 3 ISP
  - Local



Transit vs. Peering



Policy Impact

- "Valley-free" routing
  - Number links as (+1, 0, -1) for provider, peer and customer
  - In any 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

Outline

- Routing hierarchy
- Internet structure
- External BGP (E-BGP)

Choices

- Link state or distance vector?
  - No universal metric – policy decisions
- Problems with distance-vector:
  - Bellman-Ford algorithm may not converge
- Problems with link state:
  - Metric used by routers not the same – loops
  - LS database too large – entire Internet
  - May expose policies to other AS's

Solution: Distance Vector with Path

- Each routing update carries the entire path
- 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
- Advantage:
  - Metrics are local - AS chooses path, protocol ensures no loops

Interconnecting BGP Peers

- BGP uses TCP to connect peers
- Advantages:
  - Simplifies BGP
  - No need for periodic refresh - routes are valid until withdrawn, or the connection is lost
  - Incremental updates
- Disadvantages
  - Congestion control on a routing protocol?
  - Poor interaction during high load

Hop-by-hop Model

- BGP advertises to neighbors only those routes that it uses
  - Consistent with the hop-by-hop Internet paradigm
  - e.g., AS1 cannot tell AS2 to route to other AS's in a manner different than what AS2 has chosen (need source routing for that)
- BGP enforces policies by choosing paths from multiple alternatives and controlling advertisement to other AS's

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

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

BGP UPDATE Message

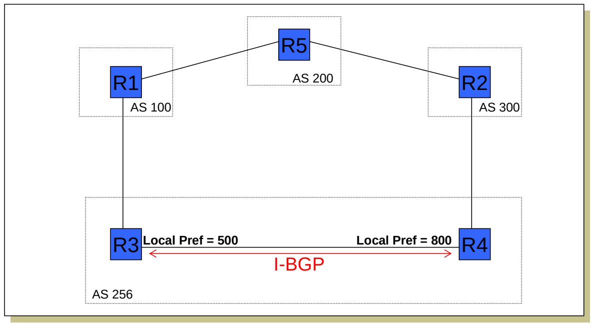
- List of withdrawn routes
- Network layer reachability information
  - List of reachable prefixes
- Path attributes
  - Origin
  - Path
  - Metrics
- All prefixes advertised in message have same path attributes

Path Selection Criteria

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

LOCAL PREF

- Local (within an AS) mechanism to provide relative priority among BGP routers (e.g. R3 over R4)

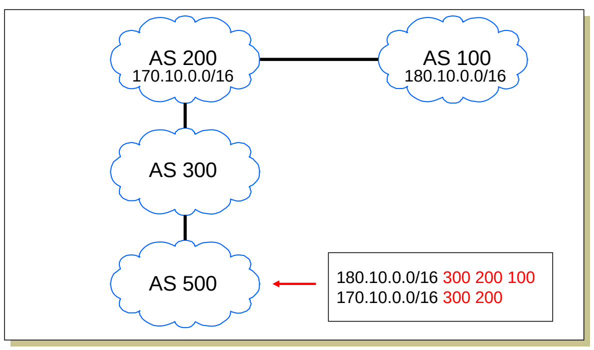


LOCAL PREF – Common Uses

- Peering vs. transit
  - Prefer to use peering connection, why?
- In general, customer > peer > provider
  - Use LOCAL PREF to ensure this

AS_PATH

- List of traversed AS's

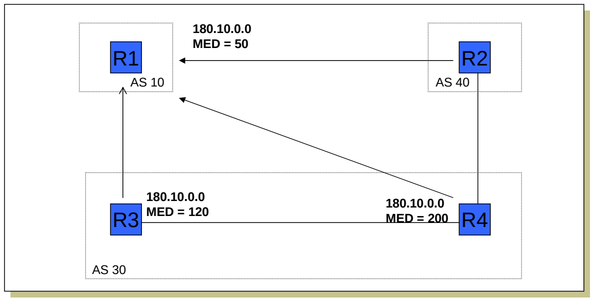


Multi-Exit Discriminator (MED)

- Hint to external neighbors about the preferred path into an AS
  - Non-transitive attribute
    - Different AS choose different scales
- Used when two AS's connect to each other in more than one place

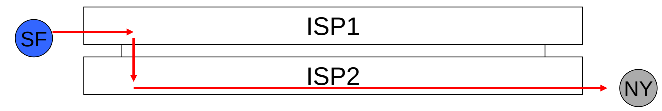
MED

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



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

Decision Process

- Processing order of attributes:
  - Select route with highest LOCAL-PREF
  - Select route with shortest AS-PATH
  - Apply MED (if routes learned from same neighbor)

Important Concepts

- Wide area Internet structure and routing driven by economic considerations
  - Customer, providers and peers
- BGP designed to:
  - Provide hierarchy that allows scalability
  - Allow enforcement of policies related to structure
- Mechanisms
  - Path vector – scalable, hides structure from neighbors, detects loops quickly

Next Lecture: DNS

- How to resolve names like www.google.com into IP addresses

EXTRA SLIDES

The rest of the slides are FYI

History

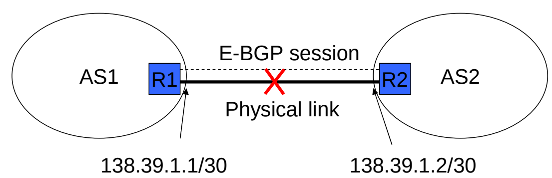
- Mid-80s: EGP
  - Reachability protocol (no shortest path)
  - Did not accommodate cycles (tree topology)
  - Evolved when all networks connected to NSF backbone
- Result: BGP introduced as routing protocol
  - Latest version = BGP 4
  - BGP-4 supports CIDR
  - Primary objective: connectivity not performance

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?

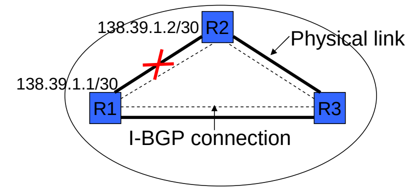
Failure on an E-BGP Link

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

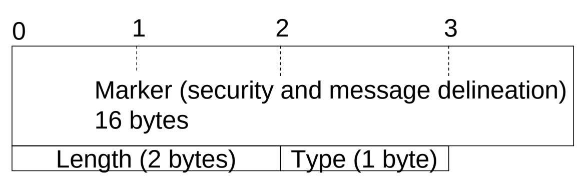


Failure on an I-BGP Link

- 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

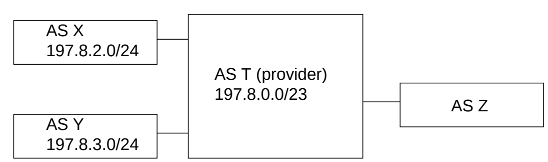


BGP Common Header



Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

CIDR and BGP



What should T announce to Z?

Options

- Advertise all paths:
  - Path 1: through T can reach 197.8.0.0/23
  - Path 2: through T can reach 197.8.2.0/24
  - Path 3: through T can reach 197.8.3.0/24
- But this does not reduce routing tables! We would like to advertise:
  - Path 1: through T can reach 197.8.0.0/22

Sets and Sequences

- Problem: what do we list in the route?
  - List T: omitting information not acceptable, may lead to loops
  - List T, X, Y: misleading, appears as 3-hop path
- Solution: restructure AS Path attribute as:
  - Path: (Sequence (T), Set (X, Y))
  - If Z wants to advertise path:
    - Path: (Sequence (Z, T), Set (X, Y))
  - In practice used only if paths in set have same attributes

Other Attributes

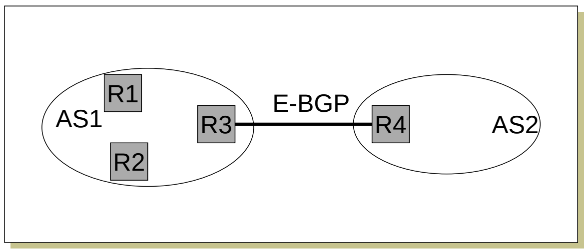
- ORIGIN
  - Source of route (IGP, EGP, other)
- NEXT_HOP
  - Address of next hop router to use
- Check out http://www.cisco.com for full explanation

Outline

- Routing hierarchy
- Internet structure
- External BGP (E-BGP)
- Internal BGP (I-BGP)

Internal vs. External BGP

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



Internal BGP (I-BGP)

- Same messages as E-BGP
- Different rules about re-advertising prefixes:
  - Prefix learned from E-BGP can be advertised to I-BGP neighbor and vice-versa, but
  - Prefix learned from one I-BGP neighbor cannot be advertised to another I-BGP neighbor
  - Reason: no AS PATH within the same AS and thus danger of looping.

Internal BGP (I-BGP)

•R3 can tell R1 and R2 prefixes from R4
•R3 can tell R4 prefixes from R1 and R2
•R3 cannot tell R2 prefixes from R1

R2 can only find these prefixes through a *direct connection* to R1
Result: I-BGP routers must be fully connected (via TCP)!
•contrast with E-BGP sessions that map to physical links