Lecture 12
BGP

Peter Steenkiste
Departments of Computer Science and
Electrical and Computer Engineering
Carnegie Mellon University

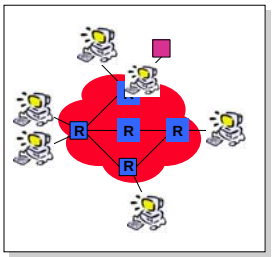
15-441 Networking, Spring 2006
http://www.cs.cmu.edu/~prs/15-441

Outline

- Routing hierarchy
- Internet structure
- Border Gateway Protocol – BGP
  - » External BGP (E-BGP)
  - » Internal BGP (I-BGP)

A Logical View of the Internet?

- After looking a RIP/OSPF descriptions
  - » End-hosts connected to routers
  - » Routers exchange messages to determine connectivity
- Not practical – why?



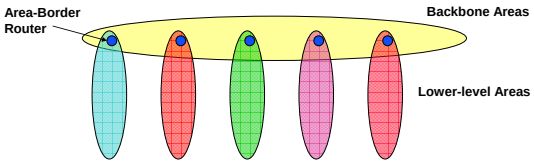
Routing Hierarchies

- Flat routing does not scale
  - » Each node cannot be expected to store routes to every destination or even destination network
  - » Convergence times increase with network diameter
  - » Communication overhead (message count) increases
- Key observation
  - » Need less information with increasing distance to destination
  - » Need lower diameters networks
- Solution: area hierarchy

Areas

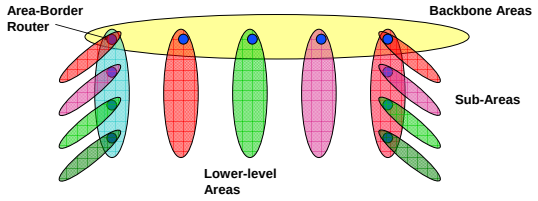
- Divide network into areas
  - » Areas can have nested sub-areas
- Hierarchically address nodes in a network
  - » Sequentially number top-level areas
  - » Sub-areas of area are labeled relative to that area
  - » Nodes are numbered relative to the smallest containing area

Routing Hierarchy

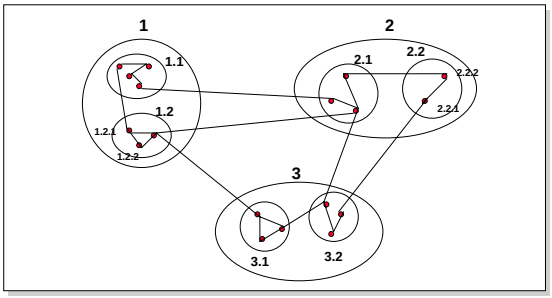


- Partition Network into "Areas"
  - » Within area
    - – Each node has routes to every other node
  - » Outside area
    - – Each node has routes for other top-level areas only
    - – Inter-area packets are routed to nearest appropriate border router
- Constraint: no path between two sub-areas of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

Routing Hierarchy: Multiple Levels

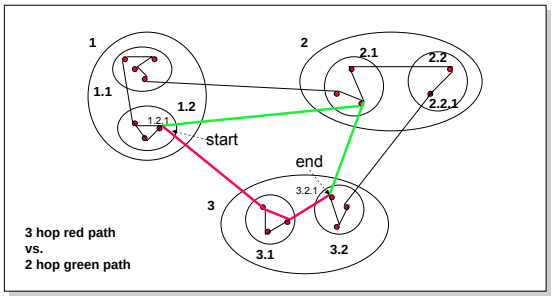


Area Hierarchy Addressing



Can result in "Sub-optimal" Path

- Sub-optimal in the sense of "non-shortest hop"

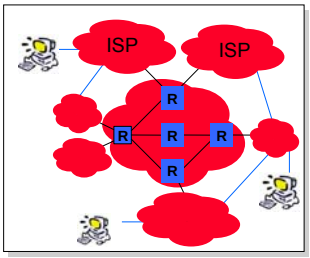


Outline

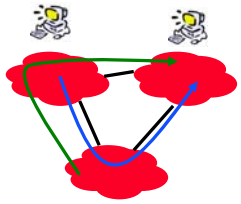
- Routing hierarchy
- Internet structure
- Border Gateway Protocol – BGP
  - External BGP (E-BGP)
  - Internal BGP (I-BGP)

A Logical View of the Internet?

- RIP/OSPF not very scalable → area hierarchies
- But, ISP's aren't equal
  - Size
  - Connectivity
- How else to ISPs differ?

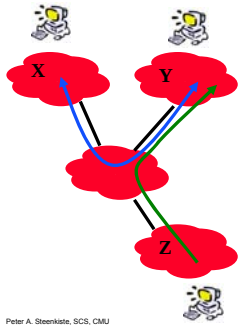


Transit versus Non-transit Networks



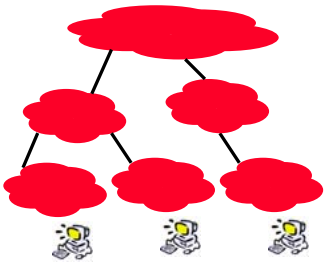
- Transit networks will allow packets to travel through to a destination in a different network.
  - Subject to certain conditions, of course
  - Typically an ISP
- Non-transit networks do not allow transit packets.
  - Typically a corporate or campus network

Selective Transitivity



- Only allow transit for packets between certain subnets.
- Very common in ISPs

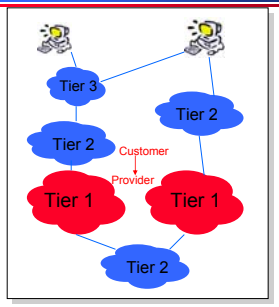
Customer-Provider Relationship



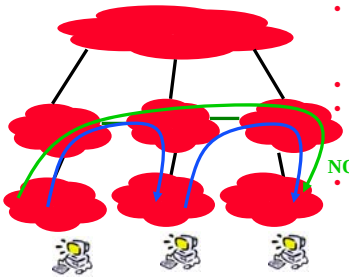
- Customers pay providers for Internet connectivity.
- Smaller providers pay larger providers for connectivity to a larger set of destinations.
- Creates a provider hierarchy

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



Peering

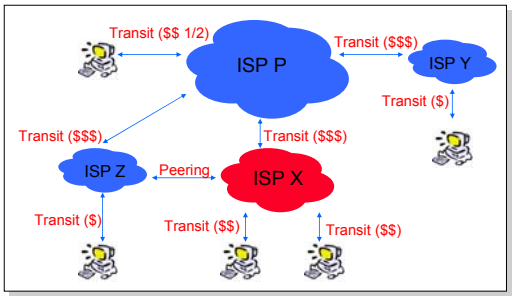


- Two ISPs directly exchange packets between each other's customers.
- No transit traffic
- Can involve exchange of money or not.
  - When traffic flow asymmetric

NO

- Business side can be tricky.
  - Cuts the provider bills
  - Helps other ISP's customers

Transit vs. Peering: Example



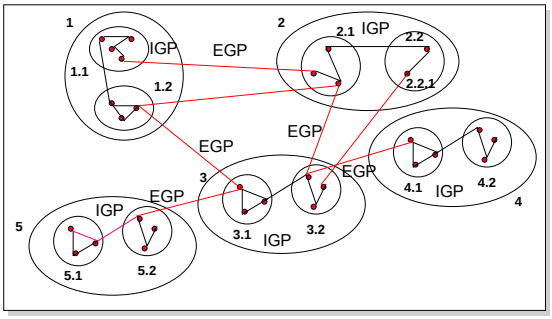
Policy Impact

- "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

Intra- versus Inter-domain Routing

- Intra: inside an ISP
- Inter: between ISPs
- How do they differ?

Example



Outline

- Routing hierarchy
- Internet structure
- Border Gateway Protocol – BGP
  - » External BGP (E-BGP)
  - » Internal BGP (I-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
  - » Path is identified as a sequence of "autonomous systems" (AS)
- 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
- Provides capability for enforcing various policies
  - » Policies are not part of BGP: they are provided to BGP as configuration information
  - » Metrics are local - AS chooses path, protocol ensures no loops

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
  - » Sometimes AS's use multiple IGPs and metrics, but appear as single AS's 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
- JANET: 786
- 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

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**
    - **You would 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**

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**

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

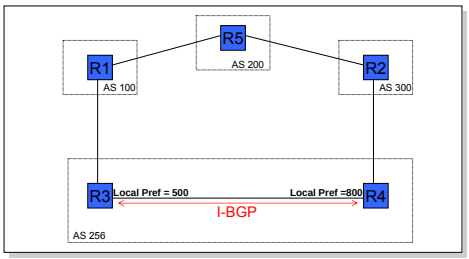
- **List of withdrawn routes**
- **Network layer reachability information**
  - **List of reachable prefixes**
- **Path attributes**
  - **Origin**
  - **Path**
  - **Metrics**
- **All prefixes advertised in message have same path attributes**

Path Selection Criteria

- Path selection is based on path attributes + external (policy) information
- Examples:
  - » Hop count
  - » Policy considerations
    - – Preference for AS
    - – Presence or absence of certain AS
  - » Path origin
  - » Link dynamics
- LOCAL PREF and MED path attributes
  - » LOCAL PREF: outgoing traffic
  - » MED: incoming traffic

LOCAL PREF

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

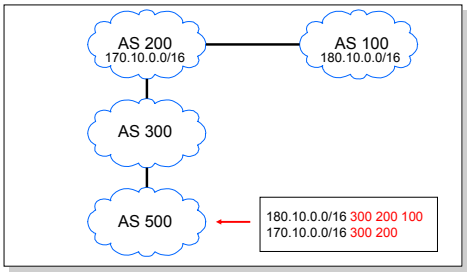


LOCAL PREF – Common Uses

- Handle routes advertised to multi-homed transit customers
  - » Should use direct connection
- 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

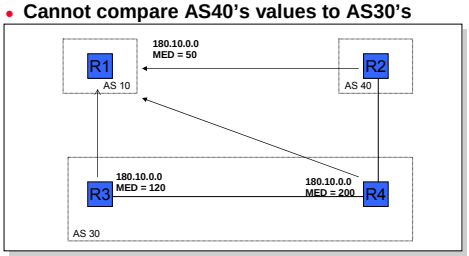


Multi-Exit Discriminator (MED)

- 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

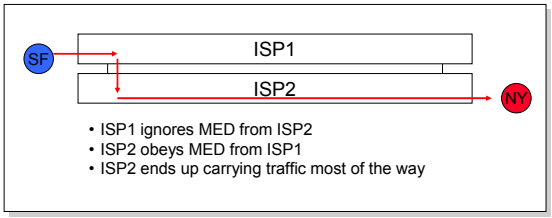
MED

- 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:



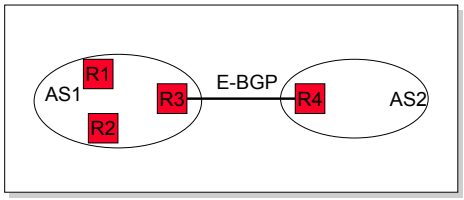
- ISP1 ignores MED from ISP2
- ISP2 obeys MED from ISP1
- ISP2 ends up carrying traffic most of the way

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)**

Internal vs. External BGP

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



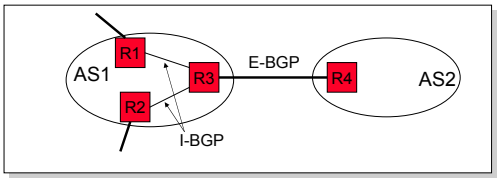
Internal BGP (I-BGP)

- **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.**

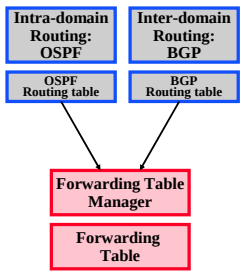
Internal BGP (I-BGP)

- 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



Putting It Together



- **Hierarchy to deal with scalability: inter- versus intra-domain routing**
- **Wide area Internet structure and routing driven by economic considerations**
  - **Customer, providers and peers**
- **BGP designed to allow enforcement of policies in a network hierarchy:**
  - **Path vector – scalable, hides structure from neighbors, detects loops quickly**
  - **IBGP structure/requirements – reuse of BGP, need for a fully connected mesh**