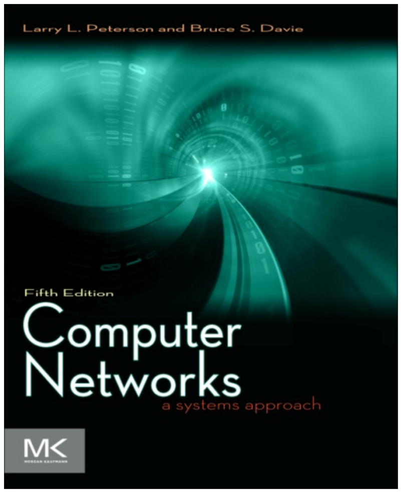


Routing: Multiple Chapters

Routing

- Forwarding table VS Routing table
  - Forwarding table
    - Used when a packet is being forwarded and so must contain enough information to accomplish the forwarding function
    - A row in the forwarding table contains the mapping from a network number to an outgoing interface and some MAC information, such as Ethernet Address of the next hop
  - Routing table
    - Built by the routing algorithm as a precursor to build the forwarding table
    - Generally contains mapping from network numbers to next hops

(a)

Prefix/Length	Next Hop
18/8	171.69.245.10

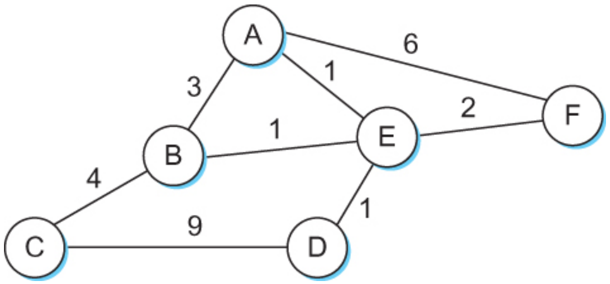
(b)

Prefix/Length	Interface	MAC Address
18/8	if0	8:0:2b:e4:b:1:2

Example rows from
(a) routing and
(b) forwarding tables

Routing

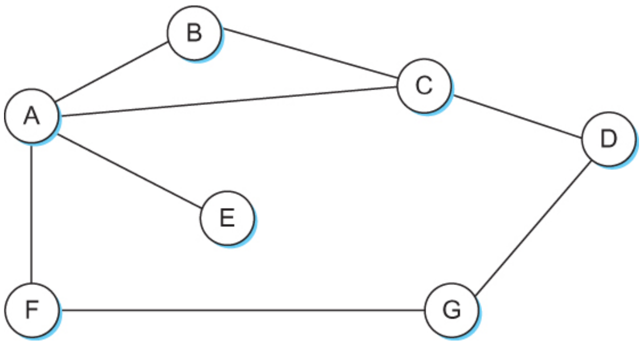
- Network as a Graph



- The basic problem of routing is to find the lowest-cost path between any two nodes
  - Where the cost of a path equals the sum of the costs of all the edges that make up the path
  - Topology is dynamic: Need a distributed and dynamic protocol
  - Two main classes of protocols
    - Distance Vector
    - Link State

Distance Vector (Bellman-Ford algorithm)

- Each node constructs a one dimensional array (a vector) containing the “distances” (costs) to all other nodes and distributes that vector to its immediate neighbors
- Starting assumption is that each node knows the cost of the link to each of its directly connected neighbors
- Every T seconds or when large change detected (1) each router sends its table to its neighbor; (2) each router then updates table based on the new information
- Problems include fast response to good new and slow response to bad news (count to infinite) Also too many messages to update
- Solutions: *split horizon with poison reverse*
  - When current route goes bad, advertise negative information in the route to ensure upstream switches quickly



Link State Routing

Strategy: Send to all nodes (not just neighbors) information about directly connected links (not entire routing table).

- Link State Packet (LSP)
  - id of the node that created the LSP
  - cost of link to each directly connected neighbor
  - sequence number (SEQNO)
  - time-to-live (TTL) for this packet
- Reliable Flooding
  - store most recent LSP from each node
  - forward LSP to all nodes but one that sent it
  - generate new LSP periodically; increment SEQNO
  - start SEQNO at 0 when reboot
  - decrement TTL of each stored LSP; discard when TTL=0

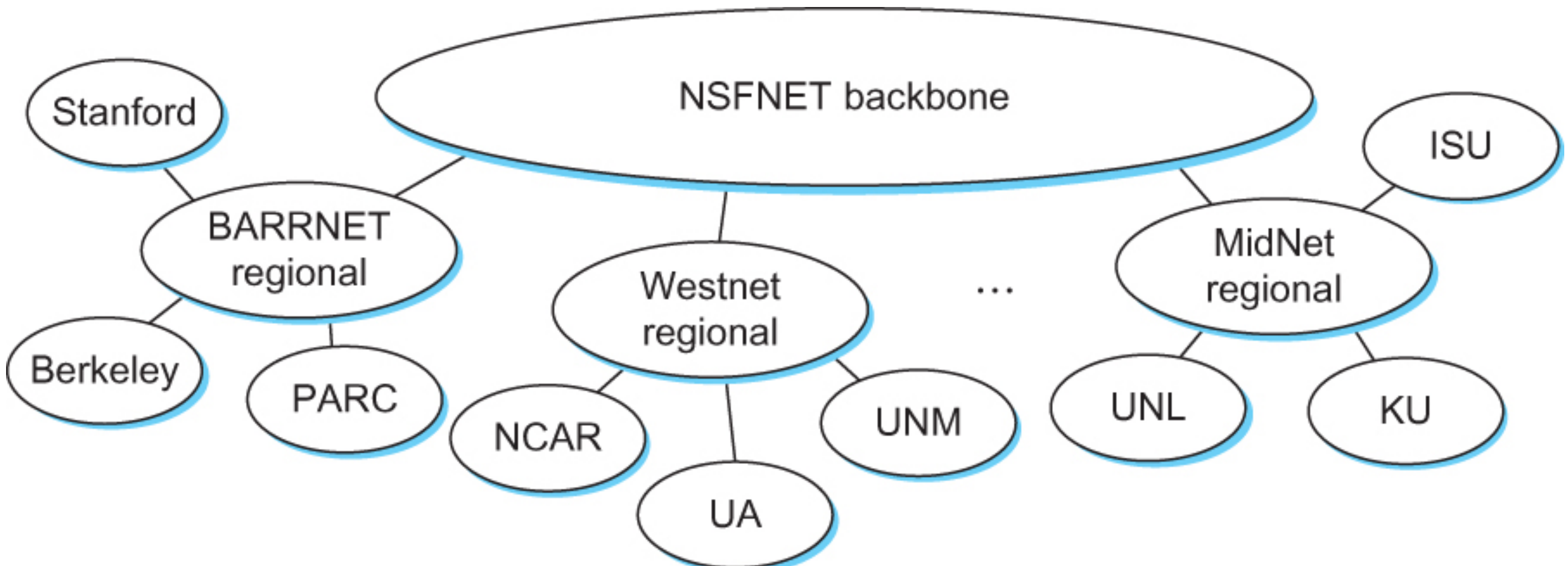
Shortest Path Routing

- Dijkstra's Algorithm - Assume non-negative link weights
  - N: set of nodes in the graph
  - l((i, j): the non-negative cost associated with the edge between nodes i, j ∈N and l(i, j) = ∝ if no edge connects i and j
  - Let s ∈N be the starting node which executes the algorithm to find shortest paths to all other nodes in N
  - Two variables used by the algorithm
    - M: set of nodes incorporated so far by the algorithm
    - C(n) : the cost of the path from s to each node n
    - The algorithm

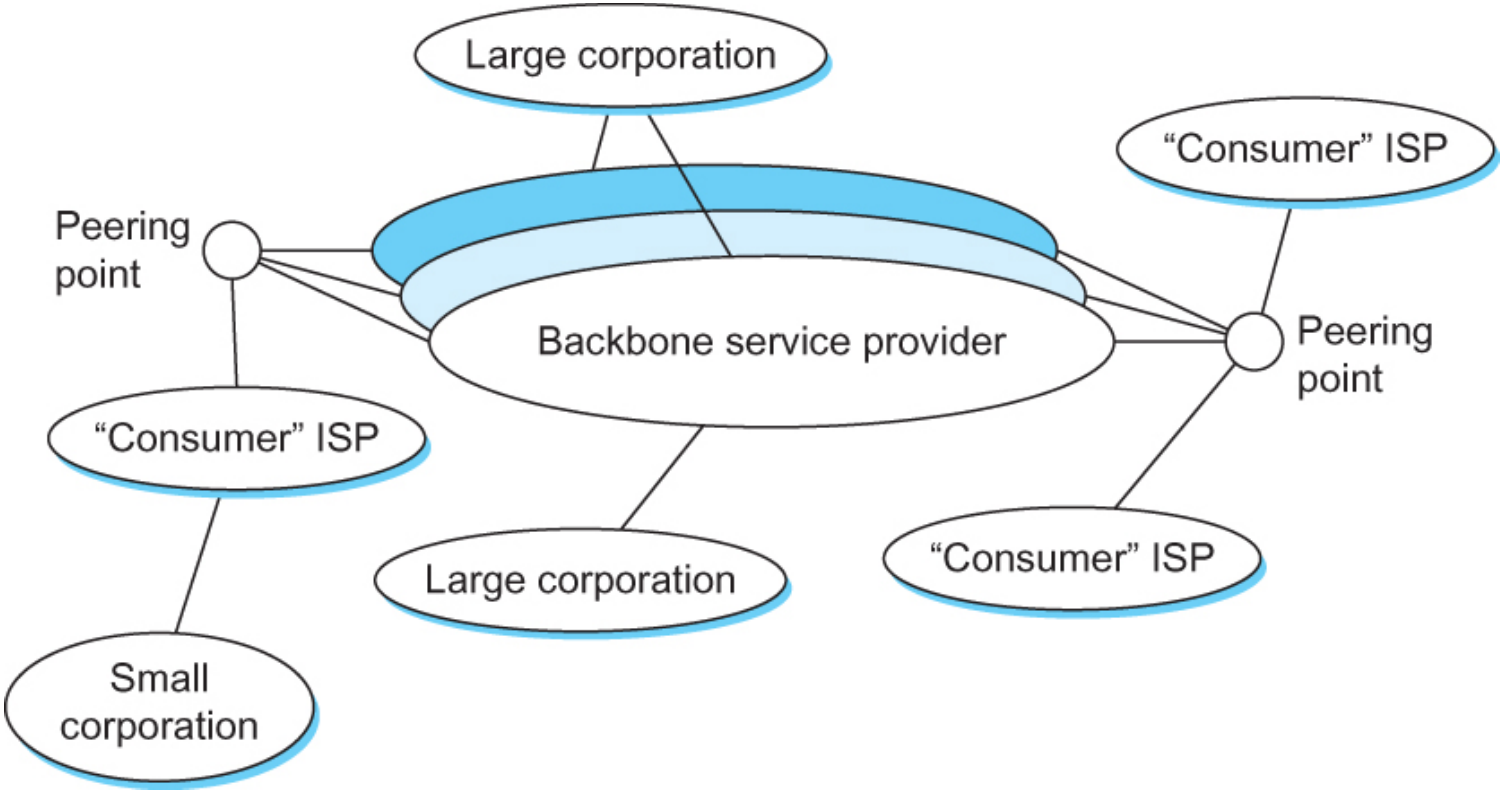
```
M = {s}
For each n in N - {s}
    C(n) = l(s, n)
while ( N ≠ M)
  M = M ∪ {w} such that C(w) is the
    minimum
                        for all w in (N-M)
  For each n in (N-M)
          C(n) = MIN (C(n), C(w) + l(w, n))
```

Inter-AS Routing

- Neither DV or LS routing scaled to a global scale Internet even when it looked like this in the 90s
- How do we build a routing system that can handle hundreds of thousands of networks and billions of end nodes?



The Global Internet



A simple multi-provider Internet

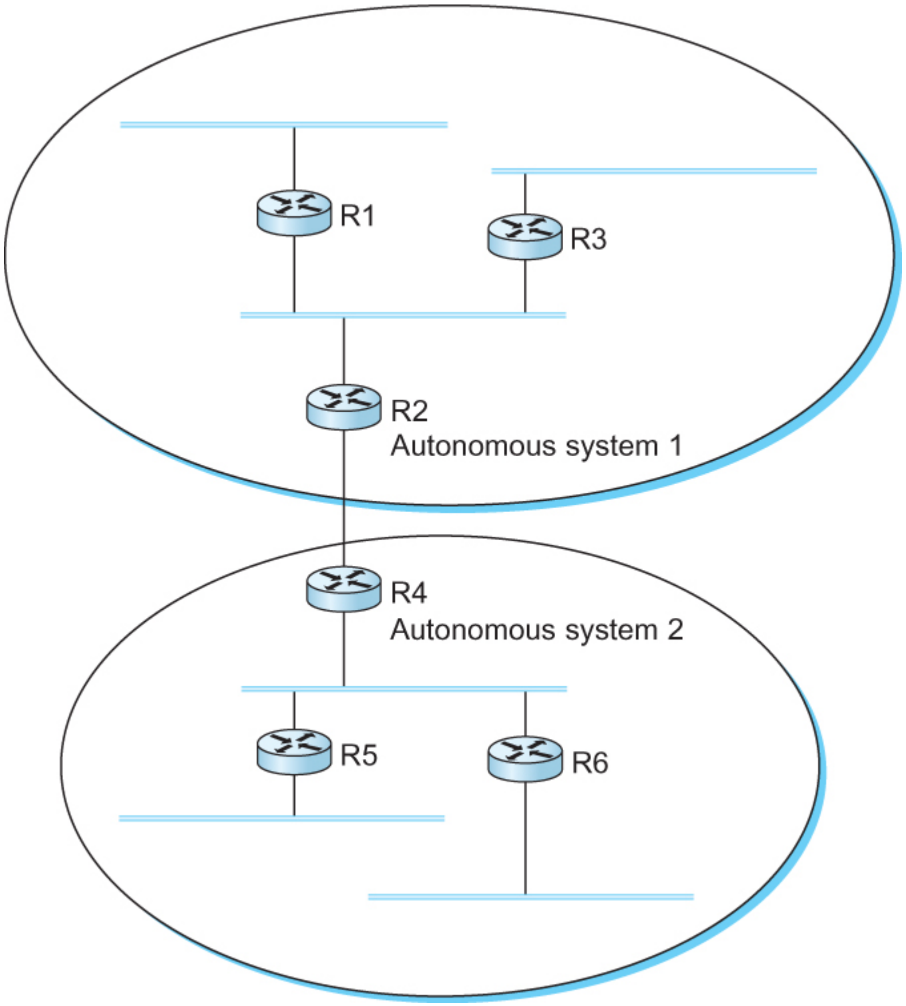
Interdomain Routing (BGP)

- Internet is organized as autonomous systems (AS) each of which is under the control of a single administrative entity

- Autonomous System (AS)
  - corresponds to an administrative domain
  - examples: University, company, backbone network

- A corporation's internal network might be a single AS, as may the network of a single Internet service provider

Interdomain Routing



A network with two autonomous system

Route Propagation

- Idea: Provide an additional way to hierarchically aggregate routing information is a large internet.
  - Improves scalability
- Divide the routing problem in two parts:
  - Routing within a single autonomous system
  - Routing between autonomous systems
- Another name for autonomous systems in the Internet is routing domains
  - Two-level route propagation hierarchy
    - Inter-domain routing protocol (Internet-wide standard)
    - Intra-domain routing protocol (each AS selects its own)

EGP and BGP

- Inter-domain Routing Protocols
  - OLD Exterior Gateway Protocol (EGP)
    - Forced a tree-like topology onto the Internet
    - Did not allow for the topology to become general
      - Tree like structure: there is a single backbone and autonomous systems are connected only as parents and children and not as peers
  - Border Gateway Protocol (BGP)
    - Assumes that the Internet is an arbitrarily interconnected set of ASs.
    - Today's Internet consists of an interconnection of multiple backbone networks (they are usually called service provider networks, and they are operated by private companies rather than the government)
    - Sites are connected to each other in arbitrary ways

BGP

- Some large corporations connect directly to one or more of the backbone, while others connect to smaller, non-backbone service providers.
- Many service providers exist mainly to provide service to “consumers” (individuals with PCs in their homes), and these providers must connect to the backbone providers
- Often many providers arrange to interconnect with each other at a single “peering point”

BGP-4: Border Gateway Protocol

- Assumes the Internet is an arbitrarily interconnected set of AS's.
- Define *local traffic* as traffic that originates at or terminates on nodes within an AS, and *transit traffic* as traffic that passes through an AS.
- We can classify AS's into three types:
  - *Stub AS*: an AS that has only a single connection to one other AS; such an AS will only carry local traffic (*small corporation in the figure of the previous page*).
  - *Multihomed AS*: an AS that has connections to more than one other AS, but refuses to carry transit traffic (*large corporation at the top in the figure of the previous page*).
  - *Transit AS*: an AS that has connections to more than one other AS, and is designed to carry both transit and local traffic (*backbone providers in the figure of the previous page*).

BGP

- The goal of Inter-domain routing is to find any path to the intended destination that is loop free
  - We are concerned with reachability than optimality
  - Finding path anywhere close to optimal is considered to be a great achievement
- Why?

BGP

- Scalability: An Internet backbone router must be able to forward any packet destined anywhere in the Internet
  - Having a routing table that will provide a match for any valid IP address
- Autonomous nature of the domains
  - It is impossible to calculate meaningful path costs for a path that crosses multiple ASs
  - A cost of 1000 across one provider might imply a great path but it might mean an unacceptable bad one from another provid
- Issues of trust
  - Provider A might be unwilling to believe certain advertisements from provider B

BGP

Each AS has:

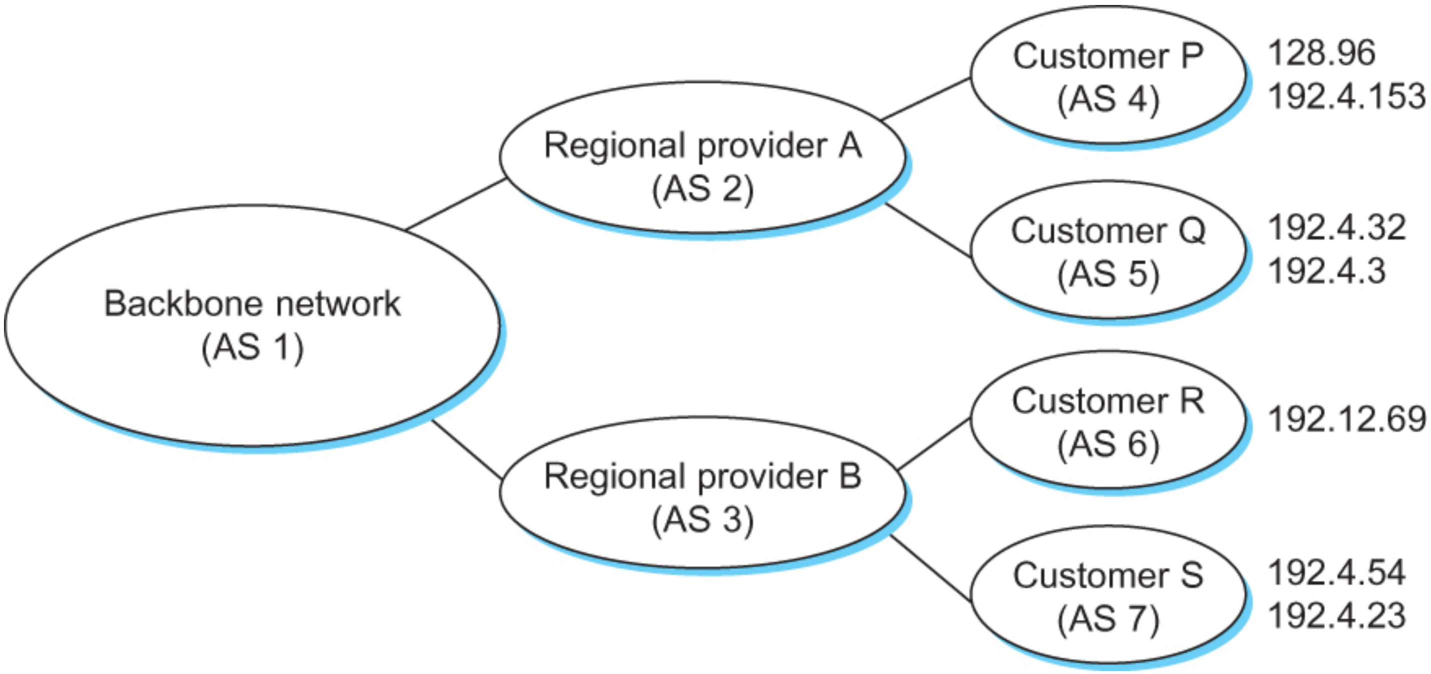
- One BGP *speaker* that advertises:
  - local networks
  - other reachable networks (transit AS only)
  - gives *path* information
- In addition to the BGP speakers, the AS has one or more border “gateways” which need not be the same as the speakers
- The border gateways are the routers through which packets enter and leave the AS

BGP

- BGP does not belong to either of the two main classes of routing protocols (distance vectors and link-state protocols)

- BGP advertises *complete paths* as an enumerated lists of ASs to reach a particular network

BGP Example



Example of a network running BGP

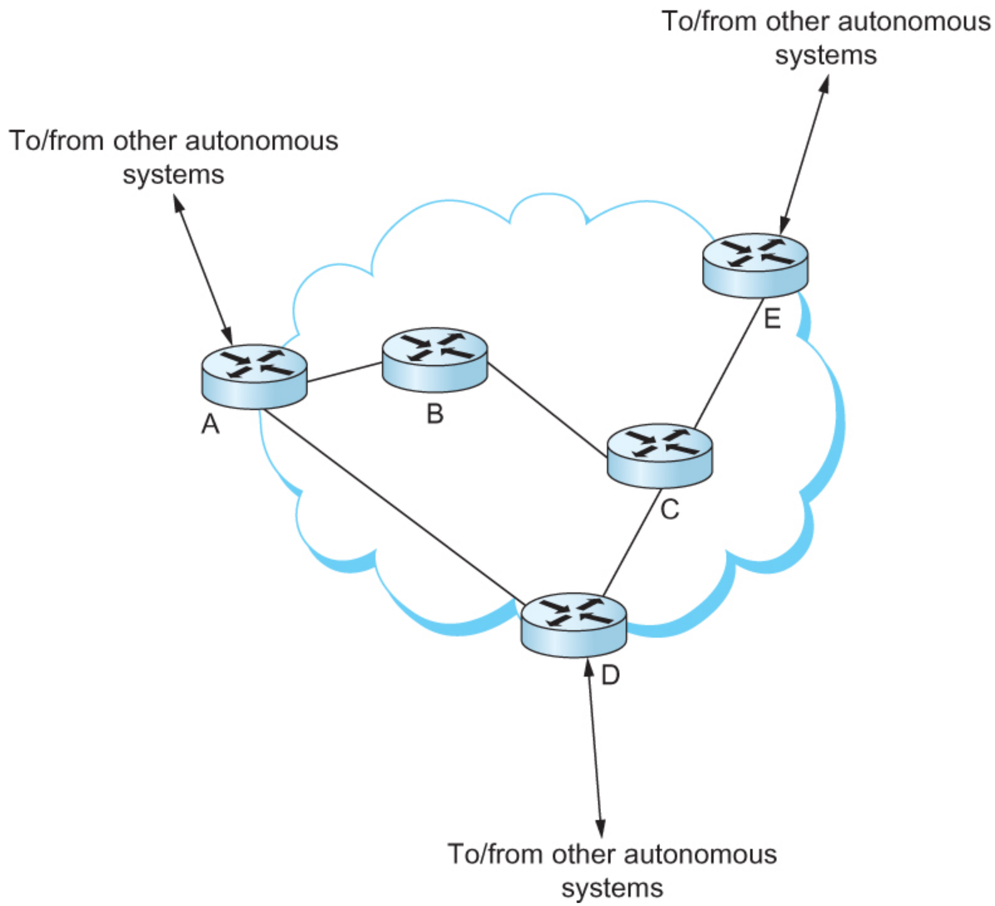
BGP Example

- Speaker for AS 2 advertises reachability to P and Q
  - Network 128.96, 192.4.153, 192.4.32, and 192.4.3, can be reached directly from AS 2.
- Speaker for backbone network then advertises
  - Networks 128.96, 192.4.153, 192.4.32, and 192.4.3 can be reached along the path <AS 1, AS 2>.
- Speaker can also cancel previously advertised paths

BGP Issues

- It should be apparent that the AS numbers carried in BGP need to be unique
- For example, AS 2 can only recognize itself in the AS path in the example if no other AS identifies itself in the same way
- AS numbers are 16-bit numbers assigned by a central authority

Integrating Interdomain and Intradomain Routing



All routers run iBGP and an intradomain routing protocol. Border routers (A, D, E) also run eBGP to other ASs

Integrating Interdomain and Intradomain Routing

Prefix	BGP Next Hop
18.0/16	E
12.5.5/24	A
128.34/16	D
128.69./16	A

BGP table for the AS

Router	IGP Path
A	A
C	C
D	C
E	C

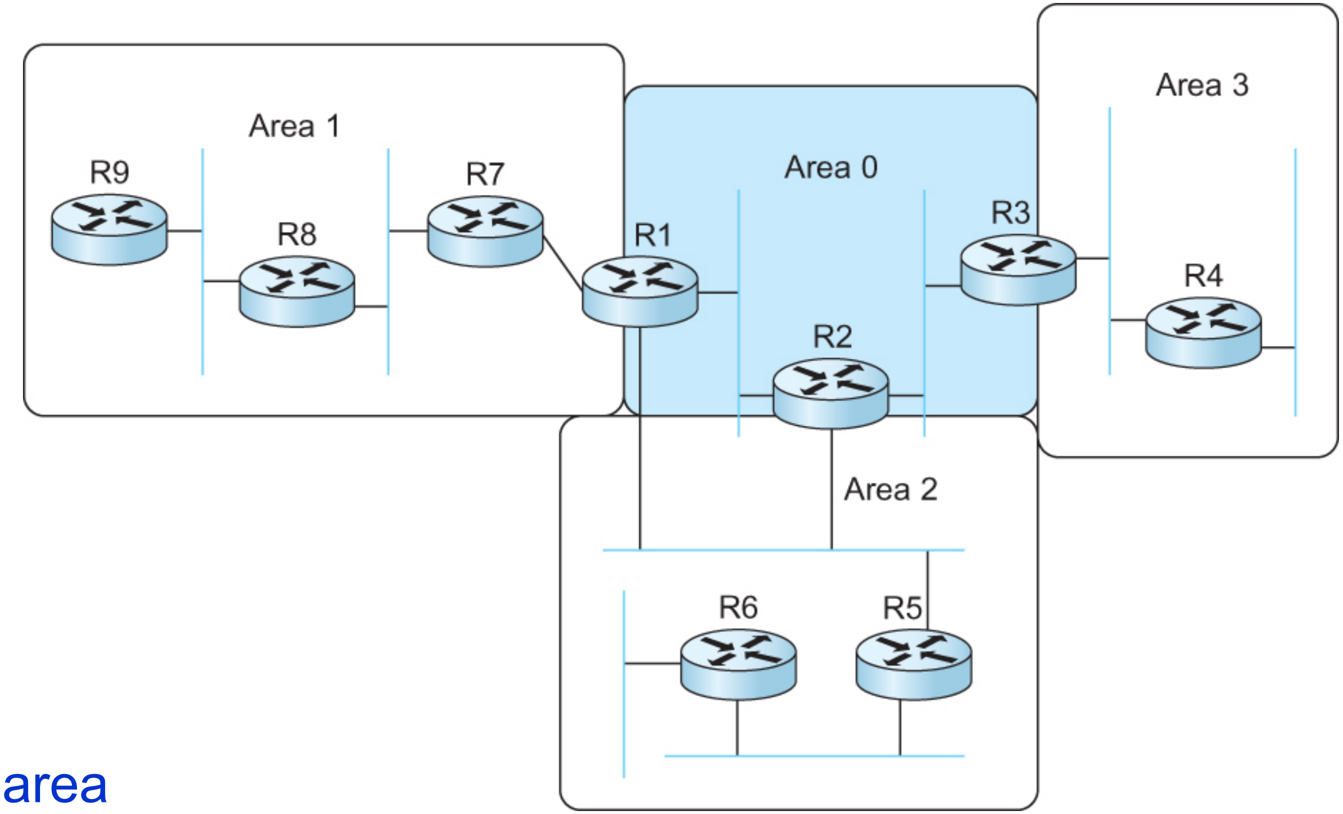
IGP table for router B

Prefix	IGP Path
18.0/16	C
12.5.5/24	A
128.34/16	C
128.69./16	A

Combined table for router B

BGP routing table, IGP routing table, and combined table at router B

Routing Areas



Backbone area

Area border router (ABR)

A domain divided into area

Internet Multicast

Overview

- Without support for multicast
  - A source needs to send a separate packet with the identical data to each member of the group
    - This redundancy consumes more bandwidth
    - Redundant traffic is not evenly distributed, concentrated near the sending host
  - Source needs to keep track of the IP address of each member in the group
    - Group may be dynamic
- To support many-to-many and one-to-many IP provides an IP-level multicast

Overview

- Using IP multicast to send the identical packet to each member of the group
  - A host sends a single copy of the packet addressed to the group's multicast address
  - The sending host does not need to know the individual unicast IP address of each member
  - Sending host does not send multiple copies of the packet

Overview

- A host signals its desire to join or leave a multicast group by communicating with its local router using a special protocol
  - In IPv4, the protocol is Internet Group Management Protocol (IGMP)
  - In IPv6, the protocol is Multicast Listener Discovery (MLD)

- The router has the responsibility for making multicast behave correctly with regard to the host

Multicast Routing

- A router's unicast forwarding tables indicate for any IP address, which link to use to forward the unicast packet
- To support multicast, a router must additionally have multicast forwarding tables that indicate, based on multicast address, which links to use to forward the multicast packet
- Unicast forwarding tables collectively specify a set of paths
- Multicast forwarding tables collectively specify a set of trees
  - Multicast distribution trees

Multicast Routing

- To support source specific multicast, the multicast forwarding tables must indicate which links to use based on the combination of multicast address and the unicast IP address of the source

- Multicast routing is the process by which multicast distribution trees are determined

Distance-Vector Multicast

- Each router already knows that shortest path to source S goes through router N.
- When receive multicast packet from S, forward on all outgoing links (except the one on which the packet arrived), iff packet arrived from N.
- Eliminate duplicate broadcast packets by only letting
  - "parent" for LAN (relative to S) forward
    - shortest path to S (learn via distance vector)
    - smallest address to break ties

Distance-Vector Multicast

Reverse Path Broadcast (RPB) – flood and prune

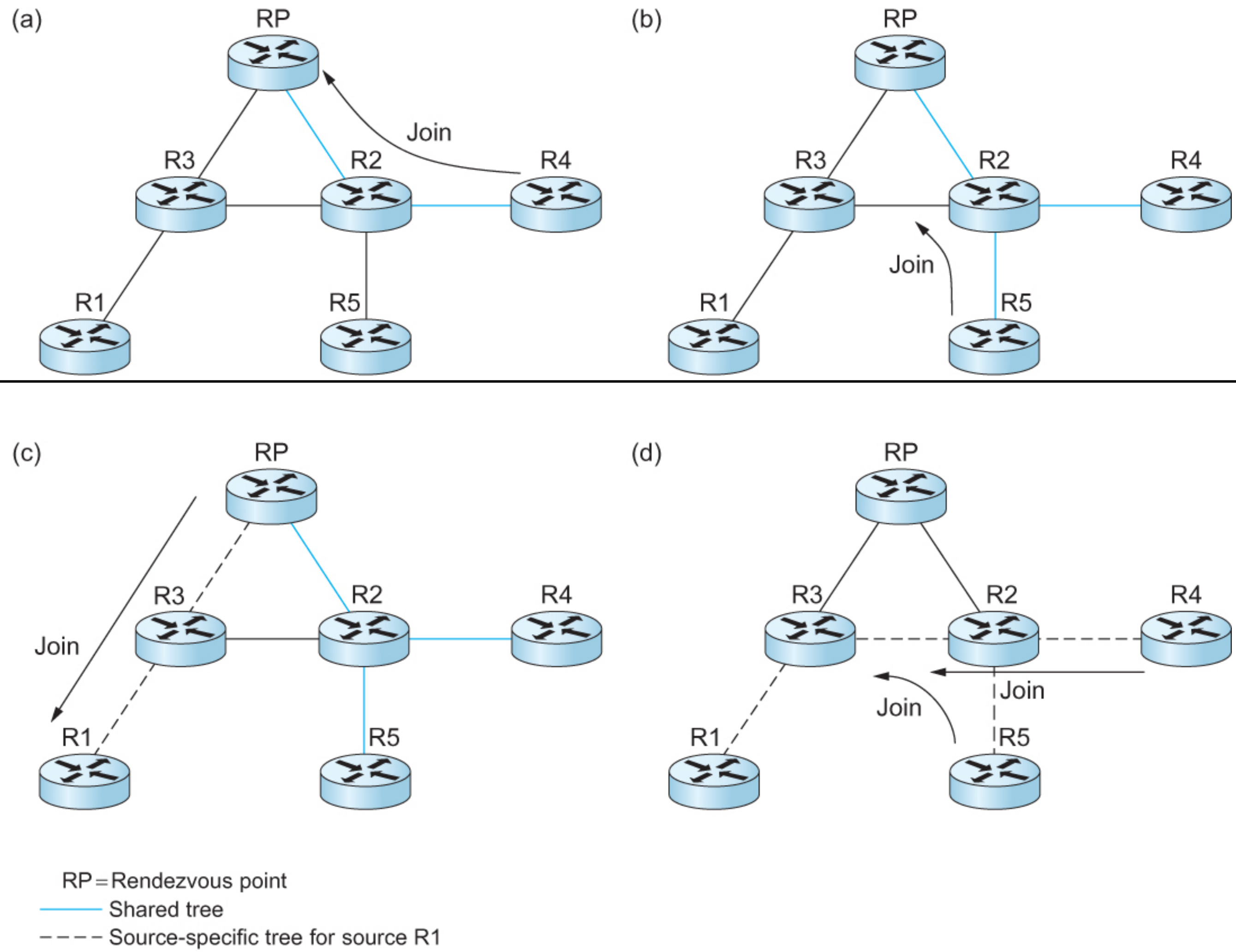
- Goal: Prune networks that have no hosts in group G
- Step 1: Determine of LAN is a *leaf* with no members in G
  - leaf if parent is only router on the LAN
  - determine if any hosts are members of G using IGMP
- Step 2: Propagate “no members of G here” information
  - augment `<Destination, Cost>` update sent to neighbors with set of groups for which this network is interested in receiving multicast packets.
  - only happens when multicast address becomes active.

Protocol Independent Multicast (PIM)

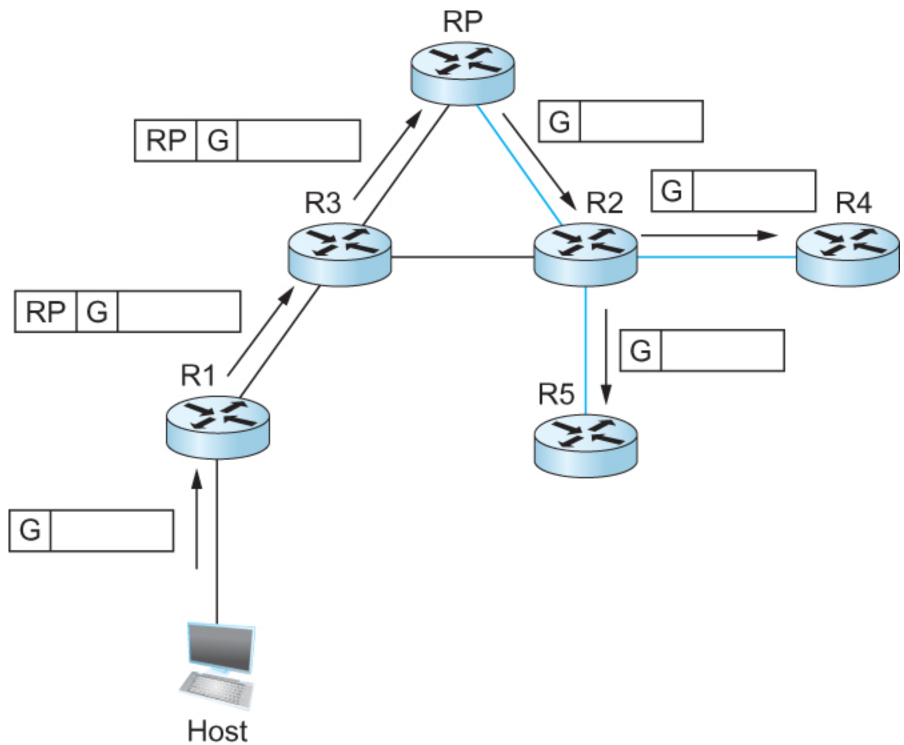
Flooding to find members and create tree does not scale

Shared Tree

Source specific tree



Protocol Independent Multicast (PIM)



Delivery of a packet along a shared tree. R1 tunnels the packet to the RP, which forwards it along the shared tree to R4 and R5.

Inter-domain Multicast

Multicast Source Discovery Protocol (MSDP)