PART XIII

ROUTING: CORES, PEERS, AND ALGORITHMS

Internet Routing
(review)

- IP implements datagram forwarding
- Both hosts and routers
  - Have an IP module
  - Forward datagrams
- IP forwarding is table-driven
- Table known as *routing table*

How / When Are IP Routing Tables Built?

- Depends on size / complexity of internet
- Static routing
  - Fixes routes at boot time
  - Useful only for simplest cases
- Dynamic routing
  - Table initialized at boot time
  - Values inserted / updated by protocols that propagate route information
  - Necessary in large internets

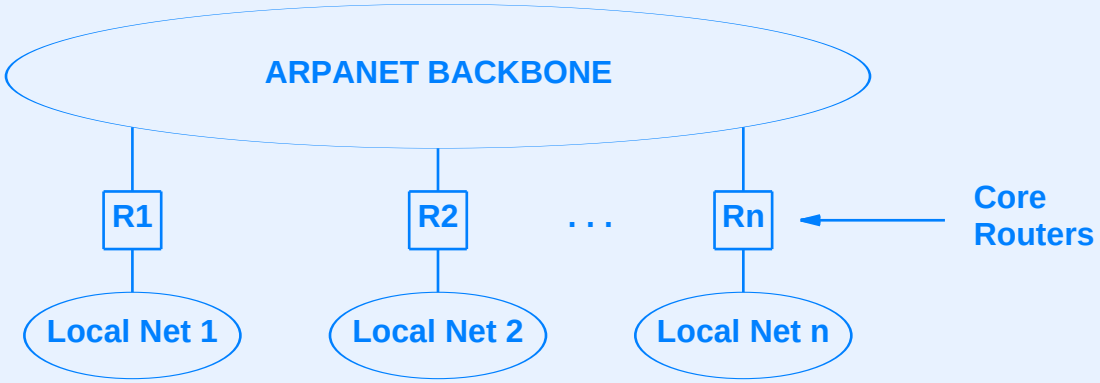
Routing Tables

- Two sources of information
  - Initialization (e.g., from disk)
  - Update (e.g., from protocols)
- Hosts tend to freeze the routing table after initialization
- Routers use protocols to learn new information and update their routing table dynamically

Routing With Partial Information

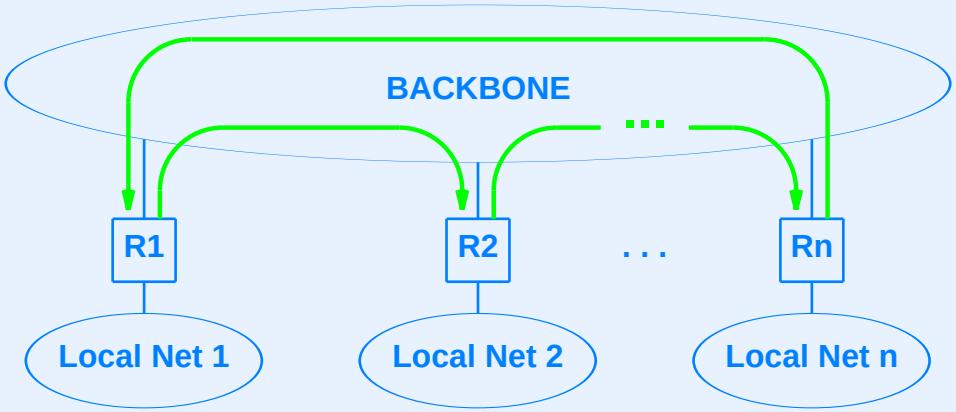
A host can forward datagrams successfully even if it only has partial routing information because it can rely on a router.

Original Internet



- Backbone network plus routers each connecting a local network

Worst Case If All Routers Contain A Default Route



- Datagram sent to nonexistent destination loops until TTL expires

Original Routing Architecture

- Small set of ‘‘core’’ routers with complete information about all destinations
- Other routers know local destinations and use the core as central router

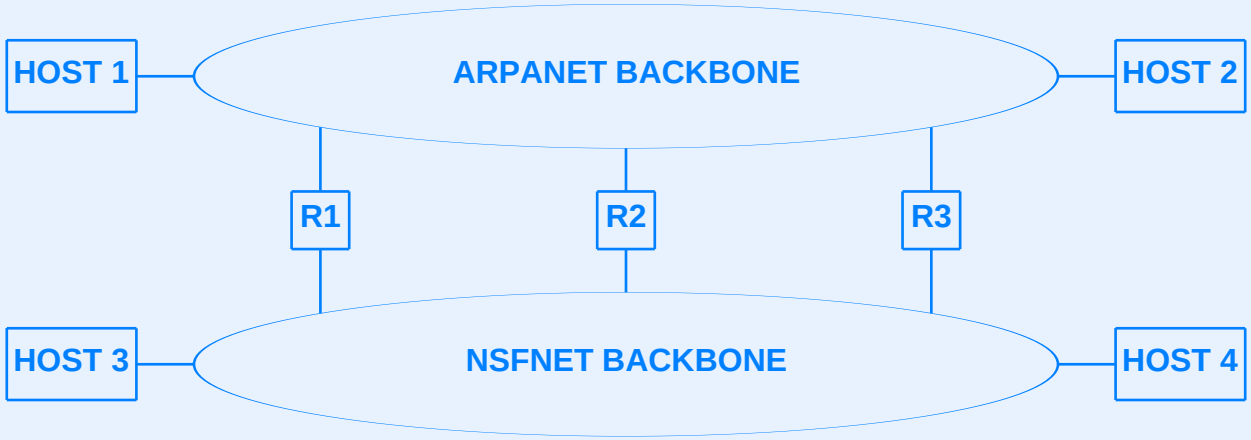
Disadvantage Of Original Core

- Central bottleneck for all traffic
- No shortcut routes possible
- Does not scale

Beyond A Core Architecture

- Single core insufficient in world where multiple ISPs each have a wide-area backbone
- Two backbones first appeared when NSF and ARPA funded separate backbone networks
- Known as *peer backbones*

Illustration Of Peer Backbones



When A Core Routing Architecture Works

A core routing architecture assumes a centralized set of routers serves as the repository of information about all possible destinations in an internet. Core systems work best for internets that have a single, centrally managed backbone. Expanding the topology to multiple backbones makes routing complex; attempting to partition the core architecture so that all routers use default routes introduces potential routing loops.

General Idea

- Have a set of core routers know routes to all locations
- Devise a mechanism that allows other routers to contact the core to learn routes (spread necessary routing information automatically)
- Continually update routing information

Automatic Route Propagation

- Two basic algorithms used by routing update protocols
  - Distance-vector
  - Link-state
- Many variations in implementation details

Distance-Vector Algorithm

- Initialize routing table with one entry for each directly-connected network
- Periodically run a distance-vector update to exchange information with routers that are reachable over directly-connected networks

Dynamic Update With Distance-Vector

- One router sends list of its routes to another
- List contains pairs of destination network and distance
- Receiver replaces entries in its table by routes to the sender if routing through the sender is less expensive than the current route
- Receiver propagates new routes next time it sends out an update
- Algorithm has well-known shortcomings (we will see an example later)

Example Of Distance-Vector Update

Destination	Distance	Route
Net 1	0	direct
Net 2	0	direct
Net 4	8	Router L
Net 17	5	Router M
Net 24	6	Router J
Net 30	2	Router Q
Net 42	2	Router J

(a)

	Destination	Distance
	Net 1	2
→	Net 4	3
	Net 17	6
→	Net 21	4
	Net 24	5
	Net 30	10
→	Net 42	3

(b)

- (a) is existing routing table
- (b) incoming update (marked items cause change)

Link-State Algorithm

- Alternative to distance-vector
- Distributed computation
  - Broadcast information
  - Allow each router to compute shortest paths
- Avoids problem where one router can damage the entire internet by passing incorrect information
- Also called *Shortest Path First* (SPF)

Link-State Update

- Participating routers learn internet topology
- Think of routers as nodes in a graph, and networks connecting them as edges or links
- Pairs of directly-connected routers periodically
  - Test link between them
  - Propagate (broadcast) status of link
- All routers
  - Receive link status messages
  - Recompute routes from their local copy of information

Summary

- Routing tables can be
  - Initialized at startup (host or router)
  - Updated dynamically (router)
- Original Internet used core routing architecture
- Current Internet accommodates peer backbones
- Two important routing algorithms
  - Distance-vector
  - Link state


Questions?

PART XIV

ROUTING: EXTERIOR GATEWAY PROTOCOLS AND AUTONOMOUS SYSTEMS (BGP)

General Principle

Although it is desirable for routers to exchange routing information, it is impractical for all routers in an arbitrarily large internet to participate in a single routing update protocol.

- Consequence: routers must be divided into groups

Autonomous System Concept (AS)

- Group of networks under one administrative authority
- Free to choose internal routing update mechanism
- Connects to one or more other autonomous systems

Modern Internet Architecture

A large TCP/IP internet has additional structure to accommodate administrative boundaries: each collection of networks and routers managed by one administrative authority is considered to be a single autonomous system that is free to choose an internal routing architecture and protocols.

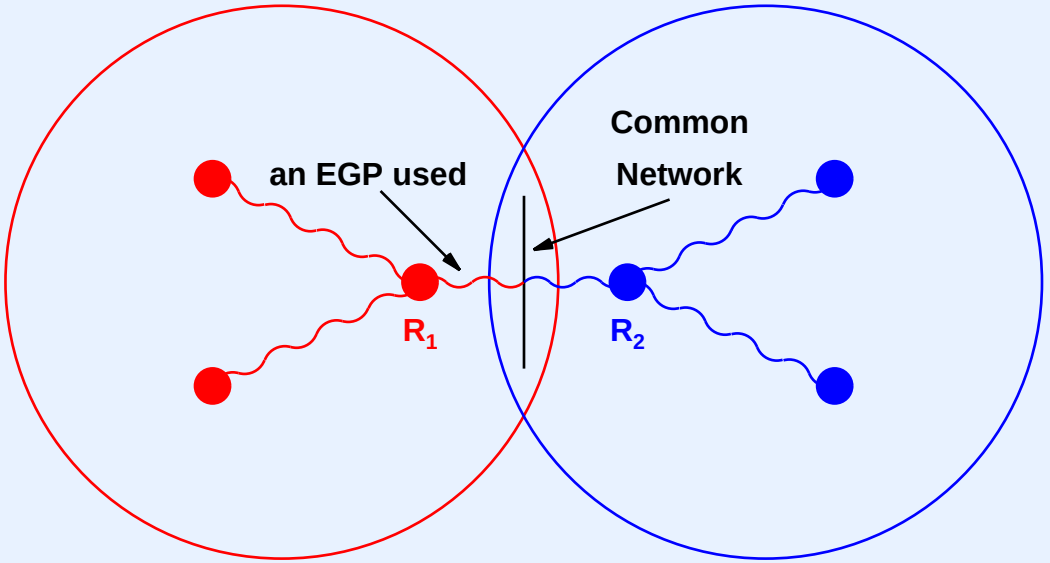
EGPs: Exterior Gateway Protocols

- Originally a single protocol for communicating routes between two autonomous systems
- Now refers to any exterior routing protocol
- Solves two problems
  - Allows router outside a group to advertise networks hidden in another autonomous system
  - Allows router outside a group to learn destinations in the group

Border Gateway Protocol

- The most popular (virtually the only) EGP in use in the Internet
- Current version is BGP-4
- Allows two autonomous systems to communicate routing information
- Supports CIDR (mask accompanies each route)
- Each AS designates a *border router* to speak on its behalf
- Two border routers become *BGP peers*

Illustration Of An EGP (Typically BGP)



Key Characteristics Of BGP

- Provides inter-autonomous system communication
- Propagates reachability information
- Follows next-hop paradigm
- Provides support for policies
- Sends path information
- Permits incremental updates
- Allows route aggregation
- Allows authentication

Additional BGP Facts

- Uses reliable transport (i.e., TCP)
  - Unusual: most routing update protocols use connectionless transport (e.g., UDP)
- Sends *keepalive* messages so other end knows connection is valid (even if no new routing information is needed)

Four BGP Message Types

Type Code	Message Type	Description
1	OPEN	Initialize communication
2	UPDATE	Advertise or withdraw routes
3	NOTIFICATION	Response to an incorrect message
4	KEEPALIVE	Actively test peer connectivity

Metric Interpretation

- Each AS can use its own routing protocol
- Metrics differ
  - Hop count
  - Delay
  - Policy-based values
- EGP communicates between two separate autonomous systems

Key Restriction On An EGP

An exterior gateway protocol does not communicate or interpret distance metrics, even if metrics are available.

- Interpretation: ‘‘my autonomous system provides a path to this network’’

The Point About EGPs

Because an Exterior Gateway Protocol like BGP only propagates reachability information, a receiver can implement policy constraints, but cannot choose a least cost route. A sender must only advertise paths that traffic should follow.

Summary

- Internet is too large for all routers to participate in one routing update protocol
- Group of networks and routers under one administrative authority is called *Autonomous System* (*AS*)
- Each AS chooses its own interior routing update protocol
- Exterior Gateway Protocol (EGP) is used to communicate routing information between two autonomous systems
- Current exterior protocol is Border Gateway Protocol version 4, BGP-4
- An EGP provides reachability information, but does not associate metrics with each route


Questions?

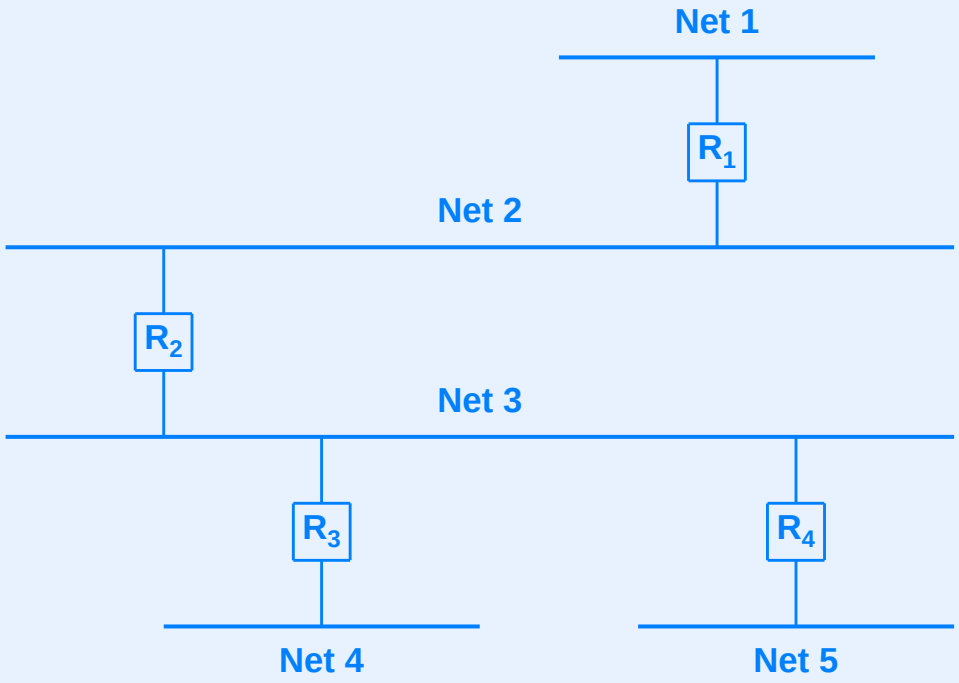
PART XV

ROUTING: INSIDE AN AUTONOMOUS SYSTEM (RIP, OSPF, HELLO)

Static Vs. Dynamic Interior Routes

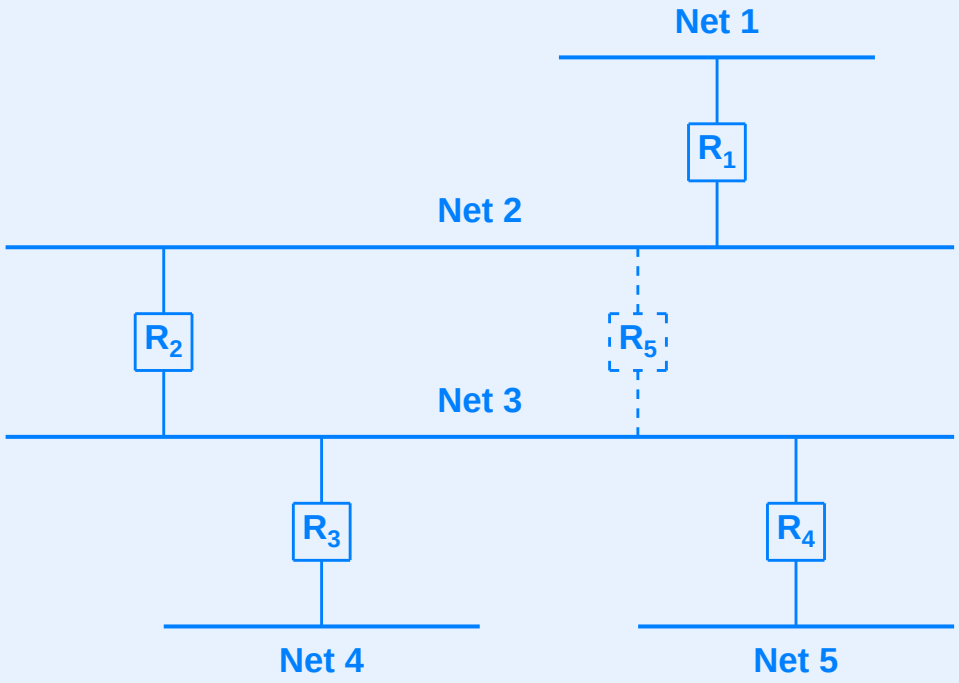
- Static routes
  - Initialized at startup
  - Never change
  - Typical for host
  - Sometimes used for router
- Dynamic router
  - Initialized at startup
  - Updated by route propagation protocols
  - Typical for router
  - Sometimes used in host

Illustration Of Topology In Which Static Routing Is Optimal



- Only one route exists for each destination

Illustration Of Topology In Which Dynamic Routing Is Needed

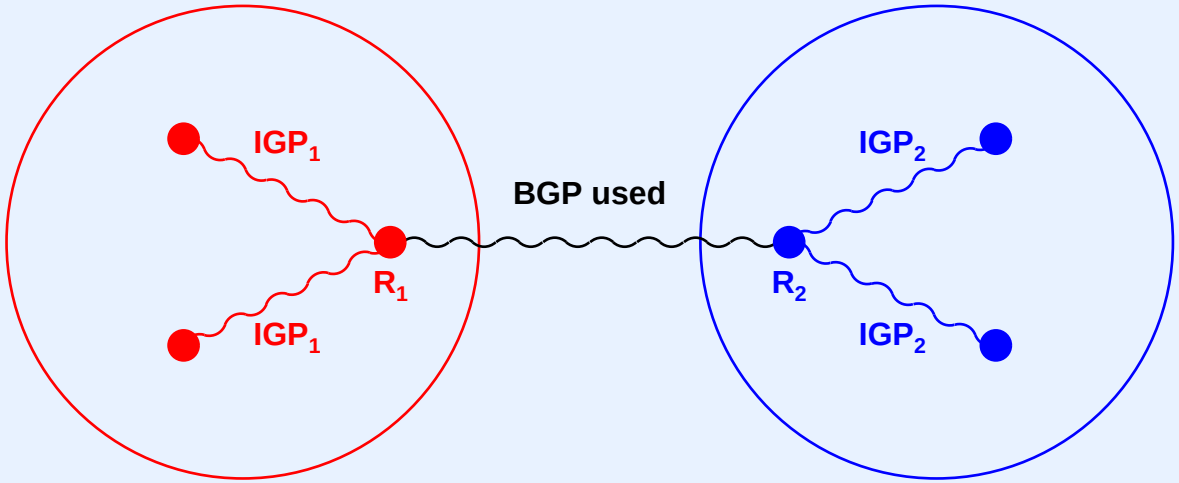


- Additional router introduces multiple paths

Exchanging Routing Information Within An Autonomous System

- Mechanisms called interior gateway protocols, IGPs
- Choice of IGP is made by autonomous system
- Note: if AS connects to rest of the world, a router in the AS must use an EGP to advertise network reachability to other autonomous systems.

Example Of Two Autonomous Systems And the Routing Protocols Used



Example IGPs

- RIP
- HELLO
- OSPF

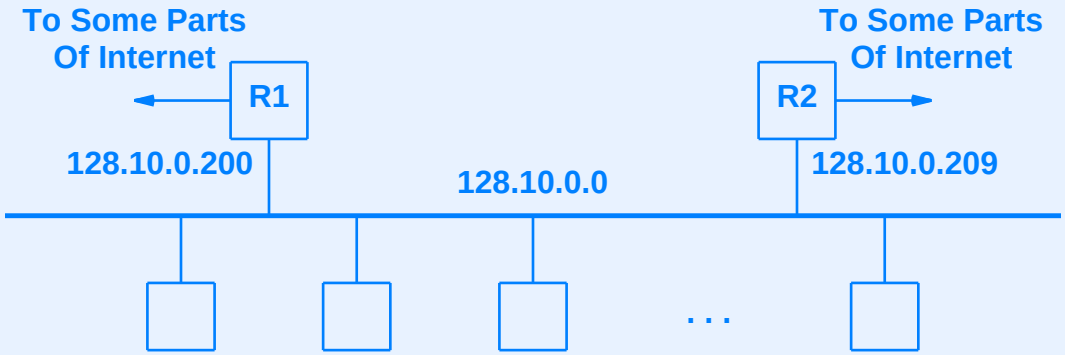
Routing Information Protocol (RIP)

- Implemented by UNIX program *routed*
- Uses hop count metric
- Distance-vector protocol
- Relies on broadcast
- Assumes low-delay local area network
- Uses split horizon and poison reverse techniques to solve inconsistencies
- Current standard is RIP2

Two Forms Of RIP

- Active
  - Form used by routers
  - Broadcasts routing updates periodically
  - Uses incoming messages to update routes
- Passive
  - Form used by hosts
  - Uses incoming messages to update routes
  - Does not send updates

Illustration Of Hosts Using Passive RIP



- Host routing table initialized to:

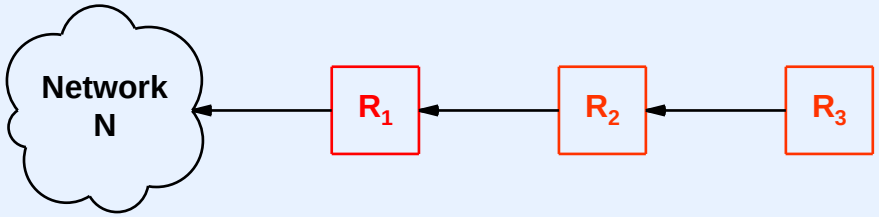
Destination	Route
128.10.0.0	direct
default	128.10.0.200

- Host listens for RIP broadcast and uses data to update table
- Eliminates ICMP redirects

RIP Operation

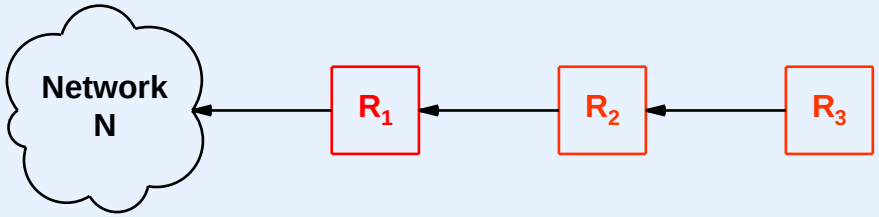
- Each router sends update every 30 seconds
- Update contains pairs of
  (destination address, distance)
- Distance of 16 is *infinity* (i.e., no route)

Slow Convergence Problem (Count To Infinity)

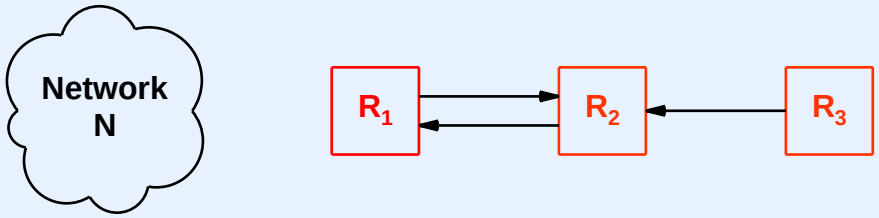


Routers with routes to network N

Slow Convergence Problem (Count To Infinity)



Routers with routes to network N



R_1 erroneously routes to R_2 after failure

Changes To RIP
In Version 2

- Update includes subnet mask
- Authentication supported
- Explicit next-hop information
- Messages can be multicast (optional)
  - IP multicast address is 224.0.0.9

Measures Of Distance That Have Been Used

- Hops
  - Zero-origin
  - One-origin (e.g., RIP)
- Delay
- Throughput
- Jitter

HELLO: A Protocol That Used Delay

- Developed by Dave Mills
- Measured delay in milliseconds
- Used by NSFNET fuzzballs
- Now historic

How HELLO Worked

- Participants kept track of delay between pairs of routers
- HELLO propagated delay information across net
- Route chosen to minimize total delay

Route Oscillation

- Effective delay depends on traffic (delay increases as traffic increases)
- Using delay as metric means routing traffic where delay is low
- Increased traffic raises delay, which means route changes
- Routes tend to oscillate

Why HELLO Worked

- HELLO used only on NSFNET backbone
- All paths had equal throughput
- Route changes damped to avoid oscillation

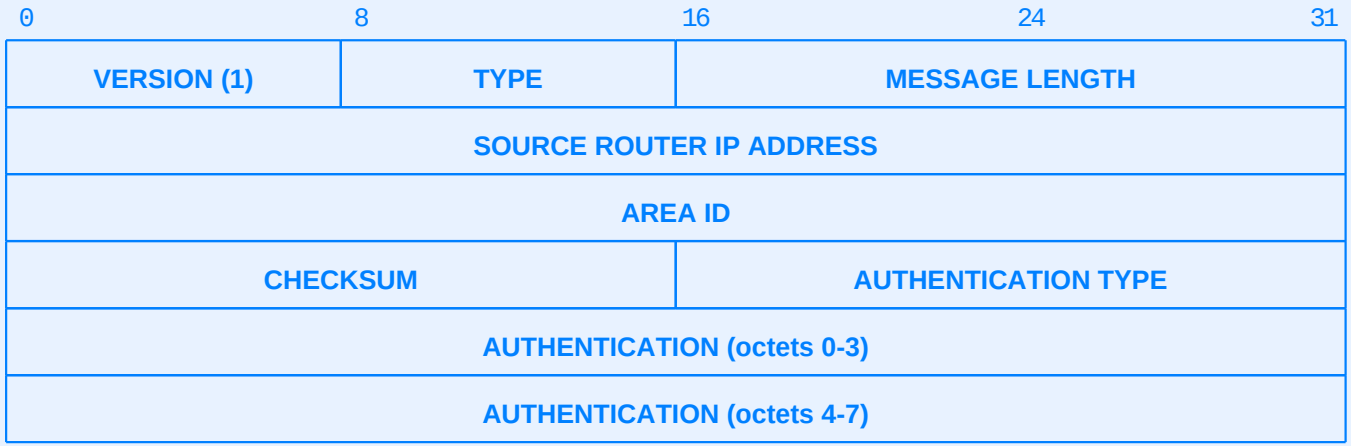
Open Shortest Path First (OSPF)

- Developed by IETF in response to vendors' proprietary protocols
- Uses SPF (link-state) algorithm
- More powerful than most predecessors
- Permits hierarchical topology
- More complex to install and manage

OSPF Features

- Type of service routing
- Load balancing across multiple paths
- Networks partitioned into subsets called *areas*
- Message authentication
- Network-specific, subnet-specific, host-specific, and CIDR routes
- Designated router optimization for shared networks
- Virtual network topology abstracts away details
- Can import external routing information

OSPF Message Header



- Each message starts with same header

OSPF Message Types

Type	Meaning
1	Hello (used to test reachability)
2	Database description (topology)
3	Link status request
4	Link status update
5	Link status acknowledgement

Summary

- Interior Gateway Protocols (IGPs) used within an AS
- Popular IGPs include
  - RIP (distance vector algorithm)
  - OSPF (link-state algorithm)


Questions?

PART XVI

INTERNET MULTICASTING

IP Multicast

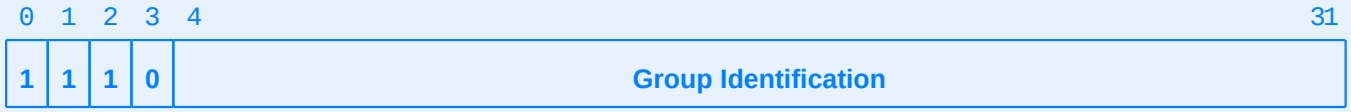
- Group address: each multicast group assigned a unique class D address
- Up to 2^{28} simultaneous multicast groups
- Dynamic group membership: host can join or leave at any time
- Uses hardware multicast where available
- Best-effort delivery semantics (same as IP)
- Arbitrary sender (does not need to be a group member)

Facilities Needed For Internet Multicast

- Multicast addressing scheme
- Effective notification and delivery mechanism
- Efficient Internet forwarding facility

IP Multicast Addressing

- Class D addresses reserved for multicast
- General form:



- Two types
  - Well-known (address reserved for specific protocol)
  - Transient (allocated as needed)

Multicast Addresses

- Address range

  224.0.0.0 through 239.255.255.255

- Notes
  - 224.0.0.0 is reserved (never used)
  - 224.0.0.1 is ‘‘all systems’’
  - 224.0.0.3 is ‘‘all routers’’
  - Address up through 224.0.0.255 used for multicast routing protocols

Example Multicast Address Assignments

Address	Meaning
224.0.0.0	Base Address (Reserved)
224.0.0.1	All Systems on this Subnet
224.0.0.2	All Routers on this Subnet
224.0.0.3	Unassigned
224.0.0.4	DVMRP Routers
224.0.0.5	OSPFIGP All Routers
224.0.0.6	OSPFIGP Designated Routers
224.0.0.7	ST Routers
224.0.0.8	ST Hosts
224.0.0.9	RIP2 Routers
224.0.0.10	IGRP Routers
224.0.0.11	Mobile-Agents
224.0.0.12	DHCP Server / Relay Agent
224.0.0.13	All PIM Routers
224.0.0.14	RSVP-Encapsulation
224.0.0.15	All-CBT-Routers
224.0.0.16	Designated-Sbm
224.0.0.17	All-Sbms
224.0.0.18	VRRP

Example Multicast Address Assignments (continued)

Address	Meaning
224.0.0.19 through 224.0.0.255	Other Link Local Addresses
224.0.1.0 through 238.255.255.255	Globally Scoped Addresses
239.0.0.0 through 239.255.255.255	Scope restricted to one organization

Mapping An IP Multicast Address To An Ethernet Multicast Address

- Place low-order 23 bits of IP multicast address in low-order 23 bits of the special Ethernet address:

$$01.00.5E.00.00.00_{16}$$

- Example IP multicast address 224.0.0.2 becomes Ethernet multicast address

$$01.00.5E.00.00.02_{16}$$

Transmission Of Multicast Datagrams

- Host does *not* install route to multicast router
- Host uses hardware multicast to transmit multicast datagrams
- If multicast router is present on net
  - Multicast router receives datagram
  - Multicast router uses destination address to determine routing

Multicast Scope

- Refers to range of members in a group
- Defined by set of networks over which multicast datagrams travel to reach group
- Two techniques control scope
  - IP's TTL field (TTL of 1 means local net only)
  - Administrative scoping

Host Participation In IP Multicast

- Host can participate in one of three ways:

Level	Meaning
0	Host can neither send nor receive IP multicast
1	Host can send but not receive IP multicast
2	Host can both send and receive IP multicast

- Note: even level 2 requires additions to host software

Host Details For Level 2 Participation

- Host uses *Internet Group Management Protocol* (*IGMP*) to announce participation in multicast
- If multiple applications on a host join the same multicast group, each receives a copy of messages sent to the group
- Group membership is associated with a specific network:

A host joins a specific IP multicast group on a specific network.

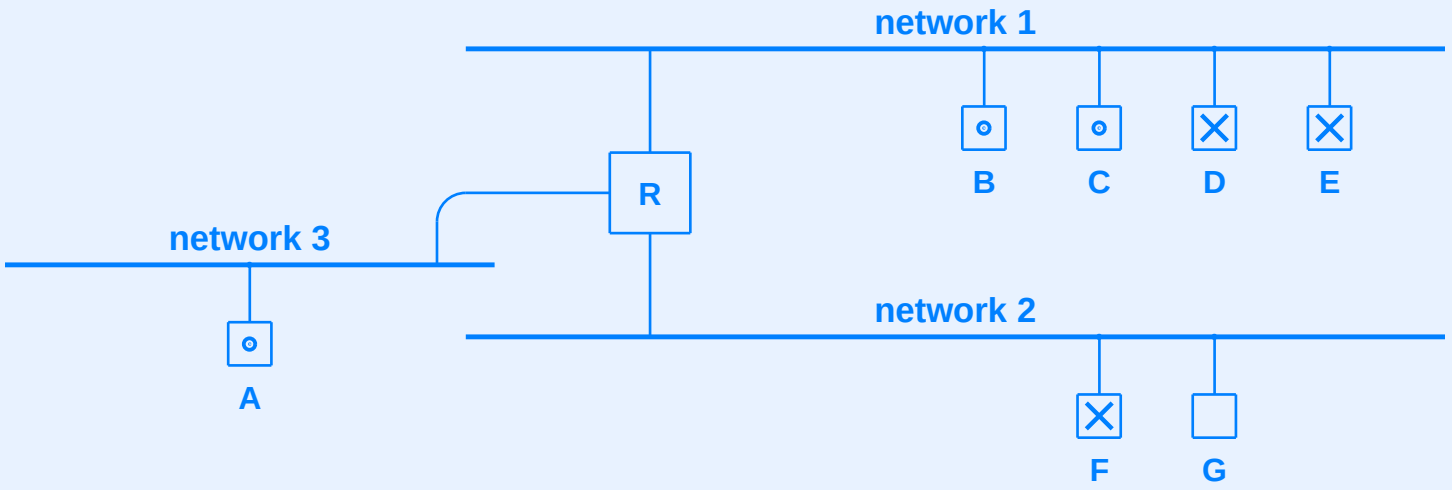
IGMP

- Allows host to register participation in a group
- Two conceptual phases
  - When it joins a group, host sends message declaring membership
  - Multicast router periodically polls a host to determine if any host on the network is still a member of a group

IGMP Implementation

- All communication between host and multicast router uses hardware multicast
- Single query message probes for membership in all active groups
- Default polling rate is every 125 seconds
- If multiple multicast routers attach to a shared network, one is elected to poll
- Host waits random time before responding to poll (to avoid simultaneous responses)
- Host listens to other responses, and suppresses unnecessary duplicate responses

Multicast Forwarding Example



- Hosts marked with dot participate in one group
- Hosts marked with X participate in another group
- Forwarding depends on group membership

The Complexity Of Multicast Routing

Unlike unicast routing in which routes change only when the topology changes or equipment fails, multicast routes can change simply because an application program joins or leaves a multicast group.

Multicast Forwarding Complication

Multicast forwarding requires a router to examine more than the destination address.

- In most cases, forwarding depends on the source address as well as the destination address

Final Item That Complicates IP Multicast

A multicast datagram may originate on a computer that is not part of the multicast group, and may be forwarded across networks that do not have any group members attached.

Multicast Routing Paradigms

- Two basic approaches
- Flood-and-prune
  - Send a copy to all networks
  - Only stop forwarding when it is known that no participant lies beyond a given point
- Multicast trees
  - Routers interact to form a ‘‘tree’’ that reaches all networks of a given group
  - Copy traverses branches of the tree

Reverse Path Forwarding

- Early flood-and-prune approach
- Actual algorithm is *Truncated Reverse Path Forwarding* (*TRPF*)

Multicast Trees

A multicast forwarding tree is defined as a set of paths through multicast routers from a source to all members of a multicast group. For a given multicast group, each possible source of datagrams can determine a different forwarding tree.

Examples Of Multicast Routing Protocols

- Reverse Path Multicasting (RPM)
- Distance-Vector Multicast Routing Protocol (DVMRP)
- Core-Based Trees (CBT)
- Protocol Independent Multicast - Dense Mode (PIM-DM)
- Protocol Independent Multicast - Sparse Mode (PIM-SM)

Reverse Path Multicasting (RPM)

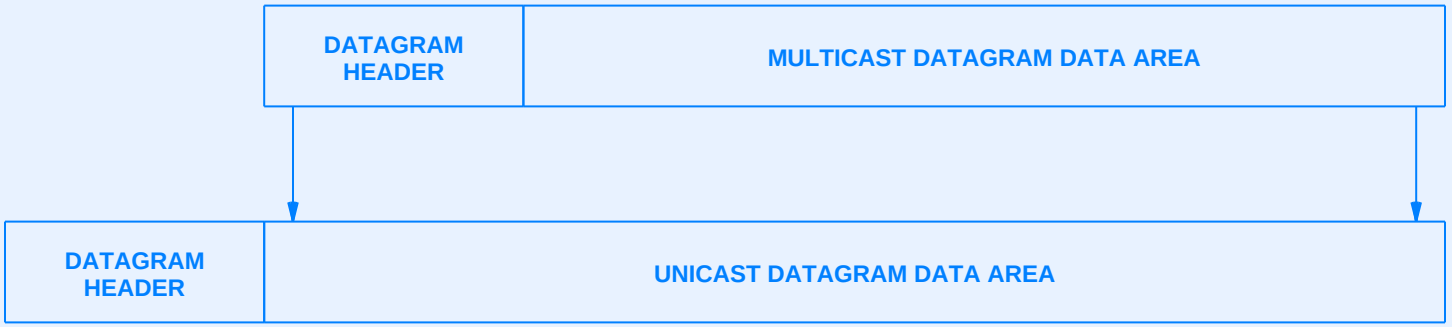
- Early form
- Routers flood datagrams initially
- Flooding pruned as group membership information learned

Distance-Vector Multicast Routing Protocol (DVMRP)

- Early protocol
- Defines extension of IGMP that routers use to exchange multicast routing information
- Implemented by Unix *mrouted* program
  - Configures tables in kernel
  - Supports tunneling
  - Used in Internet's *Multicast backBONE* (*MBONE*)

Topology In Which Tunneling Needed

Encapsulation Used With Tunneling



- IP travels in IP

Core-Based Trees (CBT)

- Proposed protocol
- Better for sparse network
- Does not forward to a net until host on the net joins a group
- Request to join a group sent to ‘‘core’’ of network
- Multiple cores used for large Internet

Division Of Internet

Because CBT uses a demand-driven paradigm, it divides the internet into regions and designates a core router *for each region; other routers in the region dynamically build a forwarding tree by sending join requests to the core.*

Protocol Independent Multicast - Dense Mode (PIM-DM)

- Allows router to build multicast forwarding table from information in conventional routing table
- Term ‘‘dense’’ refers to density of group members
- Best for high density areas
- Uses flood-and-prune approach

Protocol Independent Multicast - Sparse Mode (PIM-SM)

- Allows router to build multicast forwarding table from information in conventional routing table
- Term ‘‘sparse’’ refers to relative density of group members
- Best for situations with ‘‘islands’’ of participating hosts separated by networks with no participants
- Uses tree-based approach

Question For Discussion

- How can we provide reliable multicast?

Summary

- IP multicasting uses hardware multicast for delivery
- Host uses Internet Group Management Protocol (IGMP) to communicate group membership to local multicast router
- Two forms of multicast routing used
  - Flood-and-prune
  - Tree-based

Summary
(continued)

- Many multicast routing protocols have been proposed
  - TRPF
  - DVMRP
  - CBT
  - PIM-DM
  - PIM-SM


Questions?