Chapter 26

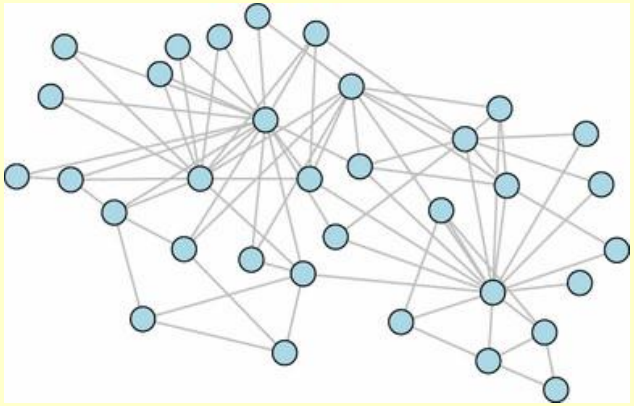
Internetwork Routing

(Static and automatic routing;

route propagation; BGP, RIP, OSPF;

multicast routing)

Terminology



- Forwarding (data plane)
  - Refers to datagram transfer
  - Performed by host or router
  - Uses routing table
- Routing (control plane)
  - Refers to propagation of routing information
  - Performed by routers
  - Inserts / changes values in routing table

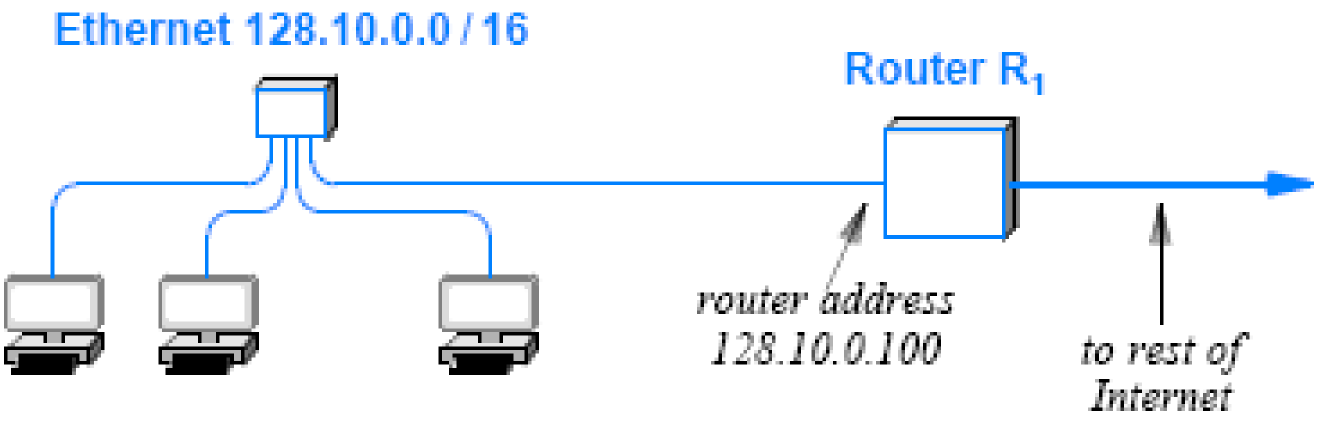
Two Forms of Internet Routing

- Static routing
  - Table initialized when system boots
  - No further changes
- Dynamic (Automatic) routing
  - Table initialized when system boots
  - Routing software learns routes and updates table
  - Continuous changes possible

Static Routing

- Used by most Internet hosts
- Typical routing table has two entries:
  - Local network → direct delivery
  - *Default* → nearest router

Example of Static Routing



(a)

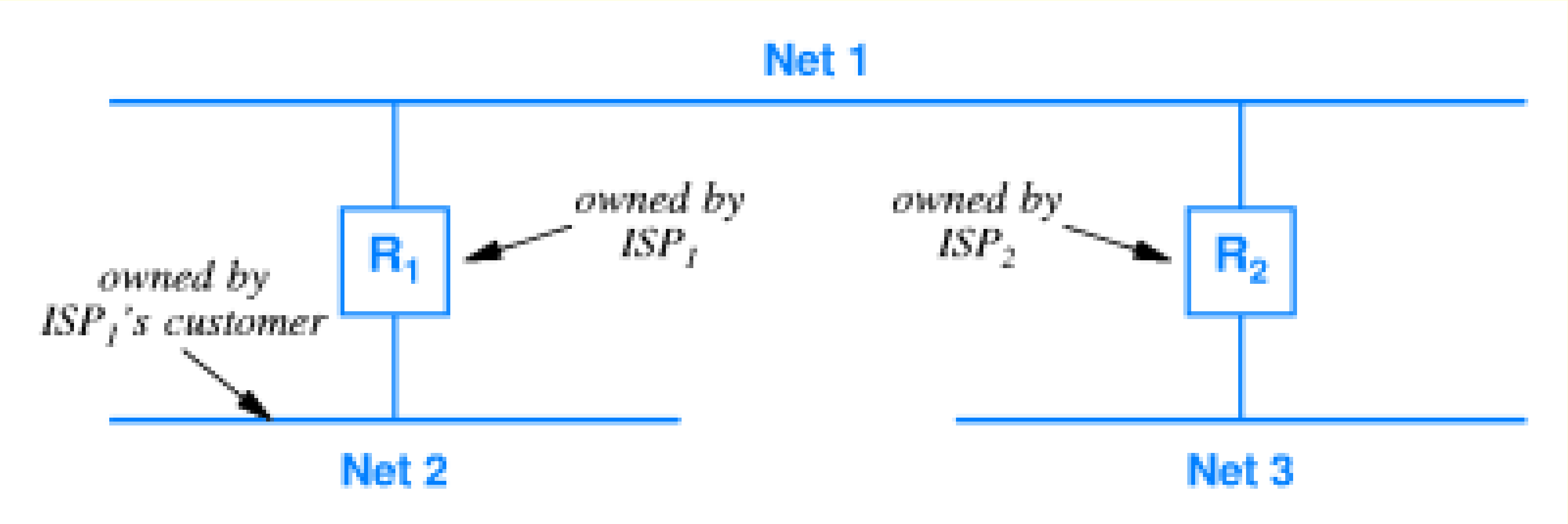
Net	Mask	Next hop
128.10.0.0	255.255.0.0	direct
default	0.0.0.0	128.10.0.100

(b)

Dynamic Routing

- Used by IP routers
- Requires special software
- Each router communicates with neighbors
- Pass routing information
- Typically use the Link-state or Distance-vector algorithm
- Use route propagation protocol

Example of Route Propagation



- Each router advertises destinations that lie beyond it

The Point of Routing Exchange

Each router runs routing software that learns about destinations other routers can reach and informs other routers about destinations that it can reach.

The routing software uses incoming information to update the local routing table continuously.

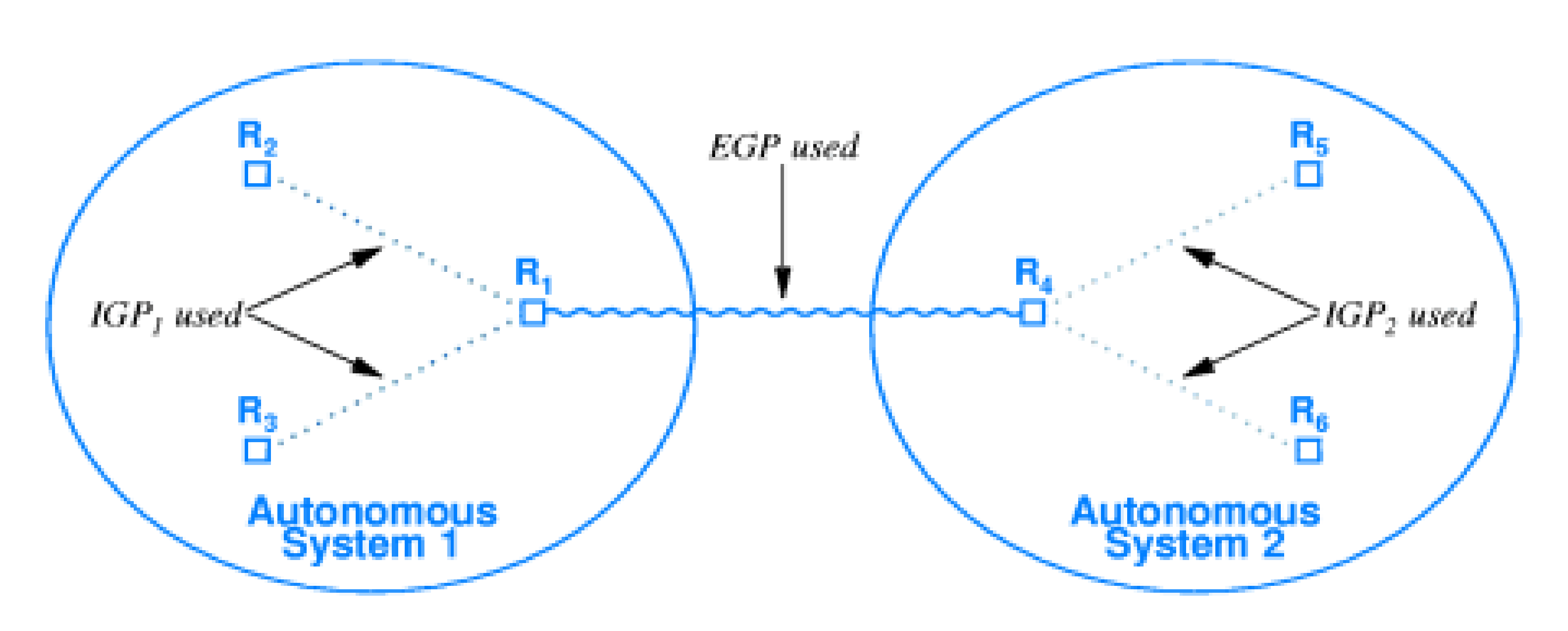
Autonomous System Concept

- An Autonomous System (AS) is a set of networks and routers under one administrative authority
- Flexible, soft definition
- Intuition: a single corporation
- Needed because no routing protocol can scale to the entire Internet
- Each AS chooses an IGP routing protocol
- Today (Feb 2024) there are about 75 000 ASs on the internet

Classifications of Internet Routing Protocols

- Two broad classes
- Interior Gateway Protocols (IGPs)
  - Used among routers within autonomous system
  - Destinations lie within the autonomous system
- Exterior Gateway Protocols (EGPs)
  - Used among autonomous systems
  - Destinations lie throughout Internet

Illustration of IGP / EGP Use



Optimal Routes, Routing Metrics, and IGPs

- **Routing software should find all possible paths and then choose one that is optimal**
- **Although the Internet usually has multiple paths between any source and destination**
  - **there is no universal agreement about which path is optimal**
- **Consider the requirements of various applications**
  - **For a remote desktop application**
    - a path with least delay is optimal
  - **For a browser downloading a large graphics file**
    - a path with maximum throughput is optimal
  - **For an audio webcast application that receives real-time audio**
    - a path with least jitter is optimal

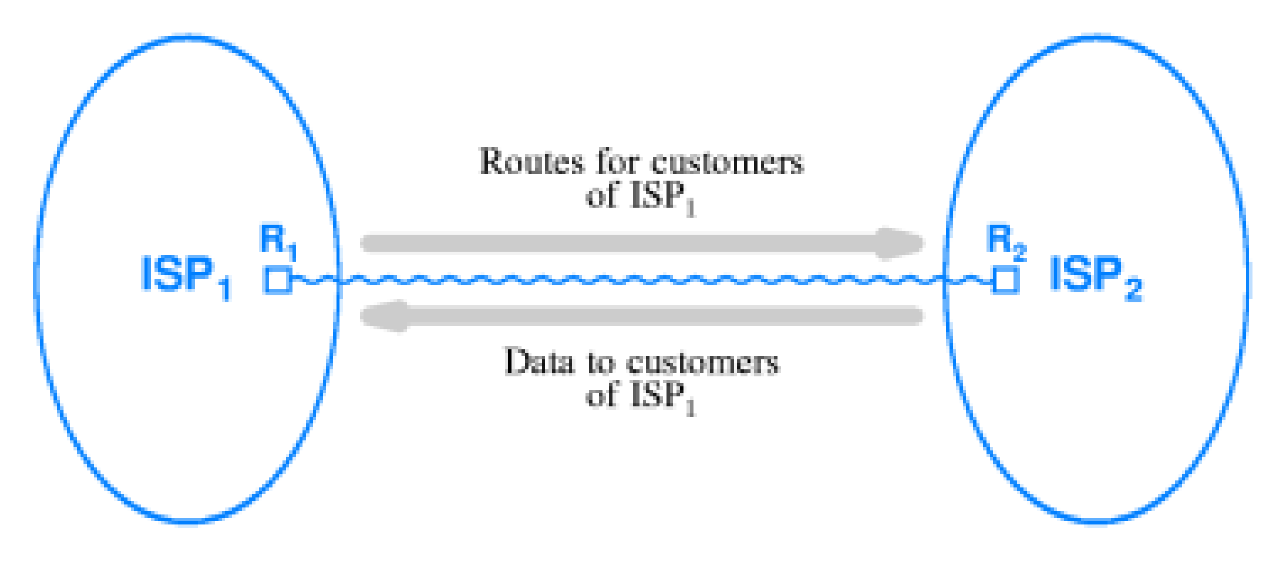
Optimal Routes, Routing Metrics, and IGPs

- The routing metric refers to a measure of the path
  - that routing software uses when choosing a route
- It is possible to use throughput, delay, jitter or any other measurable property as a routing metric
  - but most Internet routing software does not use them as it is
- Typical Internet routing uses a combination of two metrics:
  - administrative cost and hop count
- A hop corresponds to an intermediate network (or router)
  - the hop count for a destination gives the number of intermediate networks on the path to the destination
- Administrative costs are assigned manually
  - Often to control which paths traffic can use
  - Routing software will choose the path with the lower cost
    - Thus, traffic will follow the corporate policy

Optimal Routes, Routing Metrics, and IGPs

- **IGPs and EGPs differ in an important way with respect to routing metrics:**
    - **IGPs use routing metrics, but EGPs do not**
        - each AS chooses a routing metric and arranges internal routing software to send the metric with each route so receiver can use the metric to choose optimal paths
- **Outside an AS, an EGP does not attempt to choose an optimal path**
    - **Instead, the EGP merely finds a path**
- **Each AS is free to choose a routing metric internally**
- **An EGP cannot make meaningful comparisons**
    - **Suppose one AS reports the number of hops along a path to destination D and another AS reports the throughput along a different path to D**
    - **An EGP that receives the two reports cannot choose which of the two paths has least cost because there is no way to convert from hops to throughput**
- **Thus, an EGP can only report the existence of a path and not its cost**

The Concept of Route and Data Flow



Each ISP is an autonomous system that uses an Exterior Gateway Protocol to advertise its customers' networks to other ISPs. After an ISP advertises destination D, datagrams destined for D can begin to arrive

Specific Routing Protocols

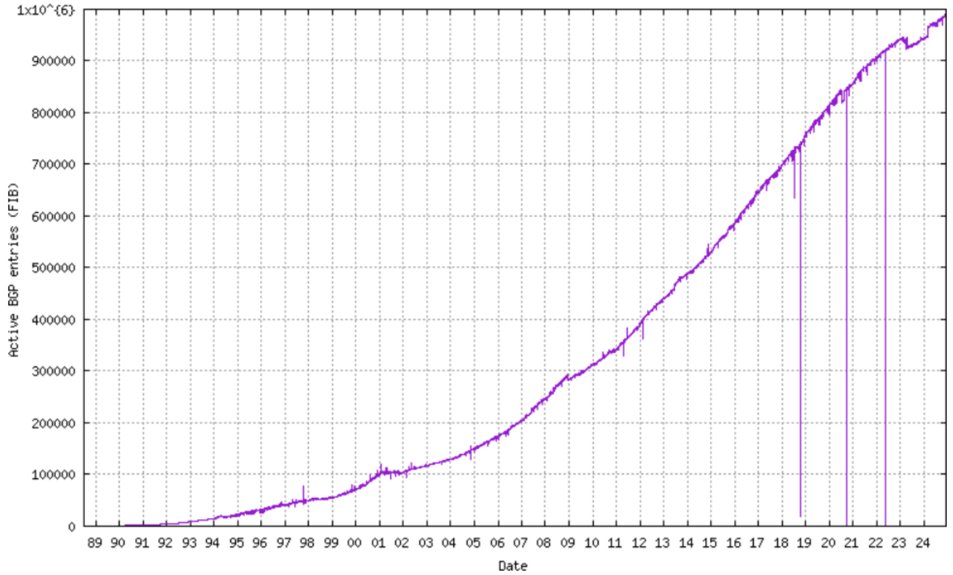
- Border Gateway Protocol (BGP)
- Routing Information Protocol (RIP)
- Open Shortest Path First Protocol (OSPF)
- Intermediate System – Intermediate System (IS-IS)
- And more we won't look at, e.g., IGRP and EIGRP

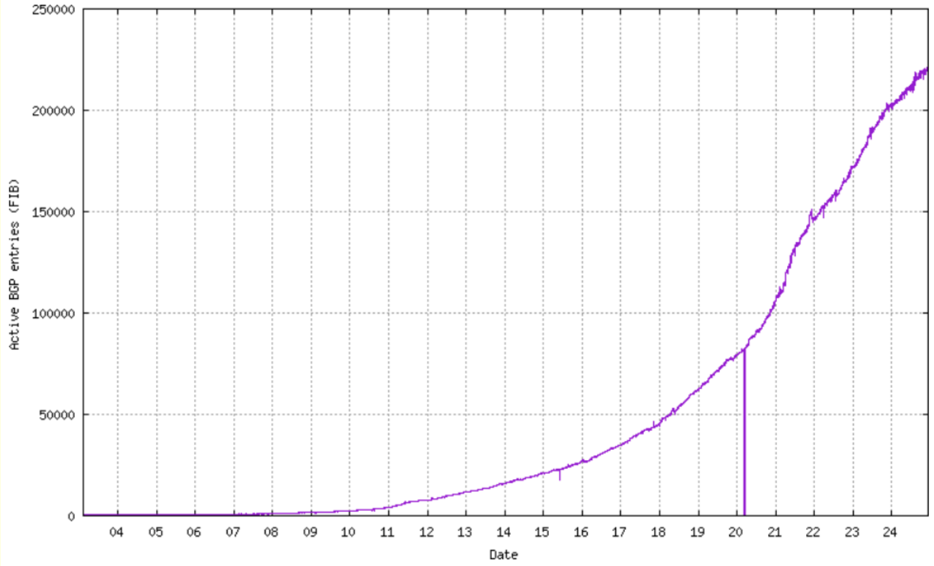
Border Gateway Protocol (BGP)

- Provides routing among autonomous systems (EGP)
- Policies to control routes advertised
- Uses reliable transport (TCP)
- Gives path of autonomous systems for each destination
- Currently the EGP of choice in the Internet
- Current version is BGP version 4 (since 1994), but with later additions to handle e.g. IPv6
- Supports Classless Inter-Domain Routing and uses route aggregation to decrease the size of routing tables

Border Gateway Protocol (BGP)

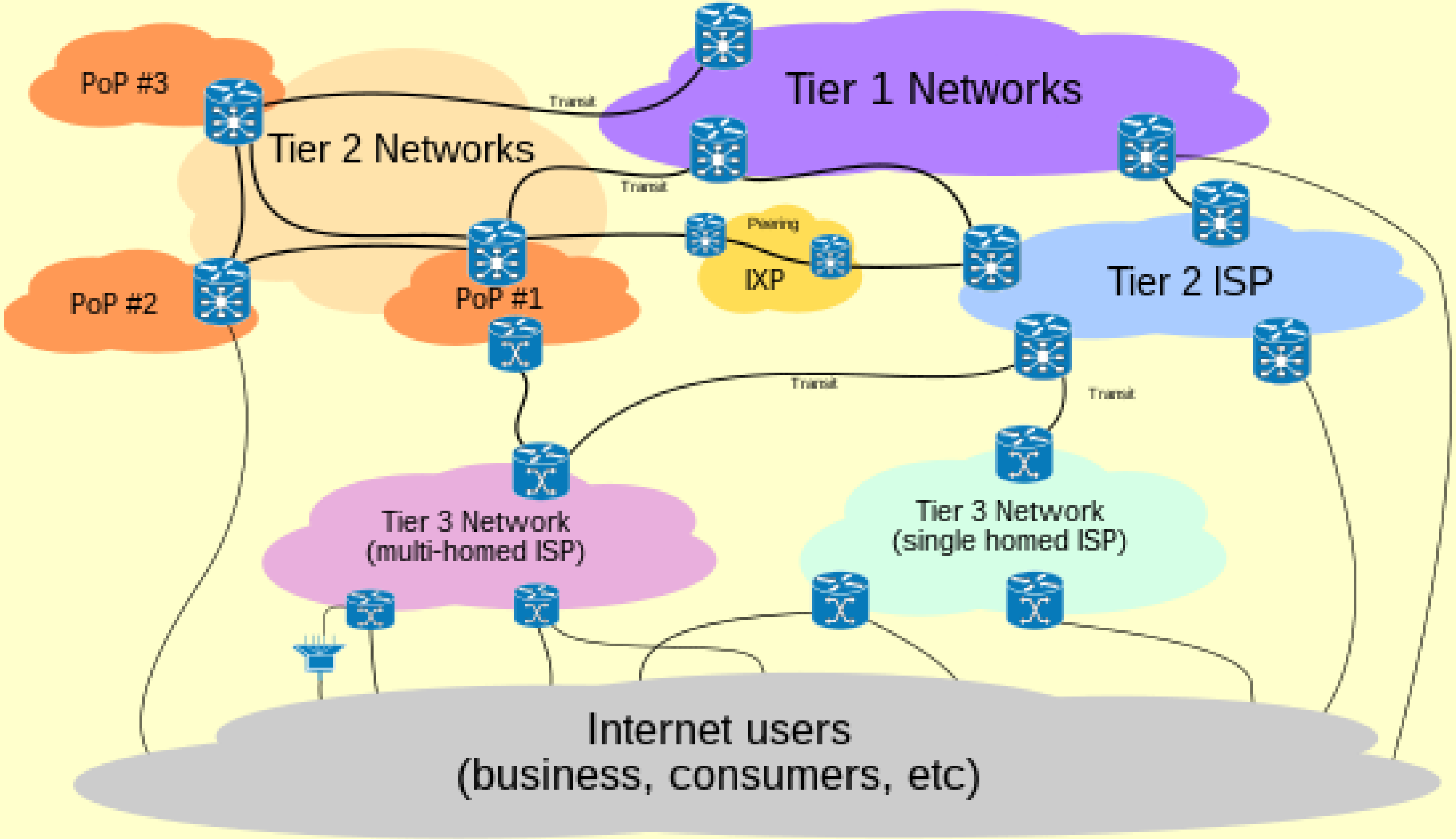
- As of December 2024, the backbone Internet active routing table has approximately 989 698 entries for IPv4 and 220 340 for IPv6
- For IPv4 that is a 5 % growth and for IPv6 it is a 9 % growth compared with one year ago.





IPv4 http://bgp.potaroo.net/index-bgp.html IPv6

Internet infrastructure



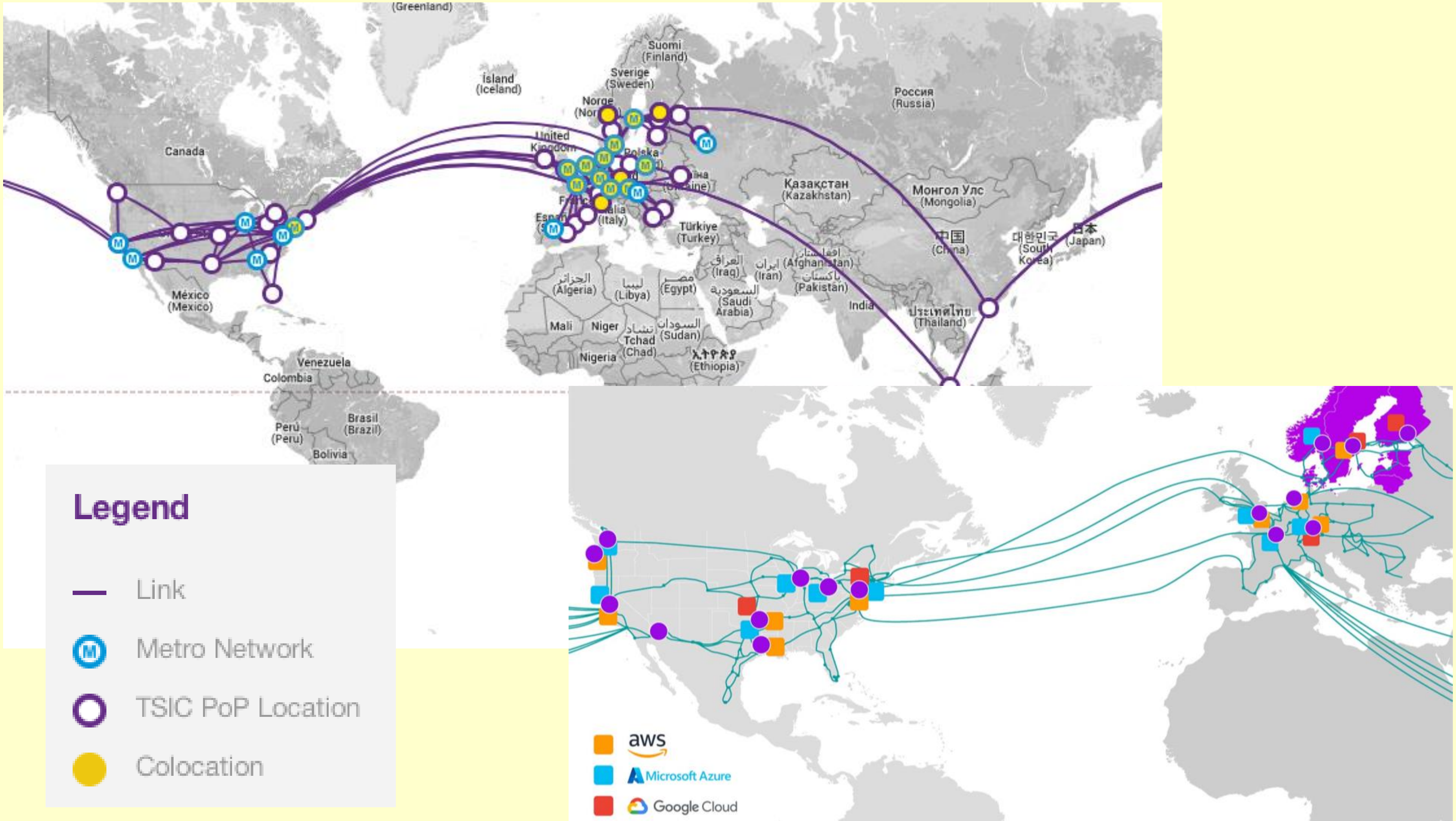
- Tier 1 networks can exchange traffic with other Tier 1 networks without paying any fees
- Tier 2 network: A network that peers for free with some networks, but still purchases IP transit or pays for peering to reach at least some portion of the Internet.
- Tier 3 network: A network that solely purchases transit/peering from other networks to participate in the Internet.
- PoP: Point of Presence, physical location of ISP's equipment
- IXP: Internet eXchange Point

Backbone of the Internet

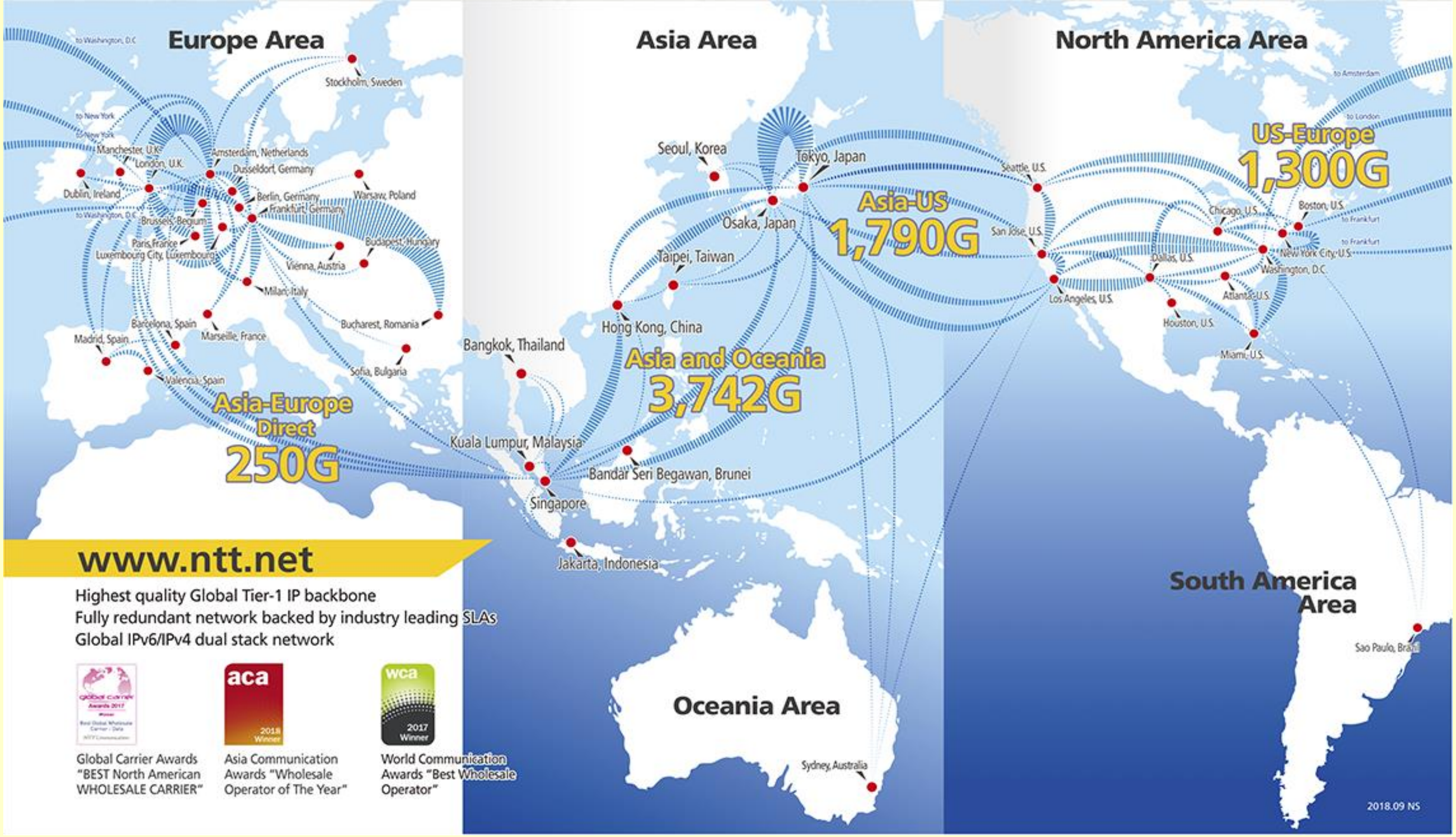
- The backbone consist of networks from a number of interconnected big Network operators
- They are called Tier 1 operators and they are:

Name	Headquarters	Fiber Route km
AT&T	United States	660 000
Deutsche Telekom	Germany	250 000
GTT Communications	United States	232 934
Liberty Globa	United Kingdom	800 000
Lumen Technologies	United States	885 139
NTT	Japan	?
Orange	France	495 000
PCCW Global	Hong Kong	?
T-Mobile US (formerly Sprint)	United States	30 000
Tata Communications	India	700 000
Telecom Italia Sparkle	Italy	560 000
Arelion (formerly Telia Carrier)	Sweden	70 000
Telxius	Spain	65 000
Verizon Enterprise Solutions	United States	805 000
Zayo Group	United States	196 339

Arelion Tier 1 network map



NTT Tier 1 network map



www.ntt.net

Highest quality Global Tier-1 IP backbone
Fully redundant network backed by industry leading SLAs
Global IPv6/IPv4 dual stack network

Global Carrier Awards "BEST North American WHOLESALE CARRIER"

Asia Communication Awards "Wholesale Operator of The Year"

World Communication Awards "Best Wholesale Operator"

BGP Hijacking

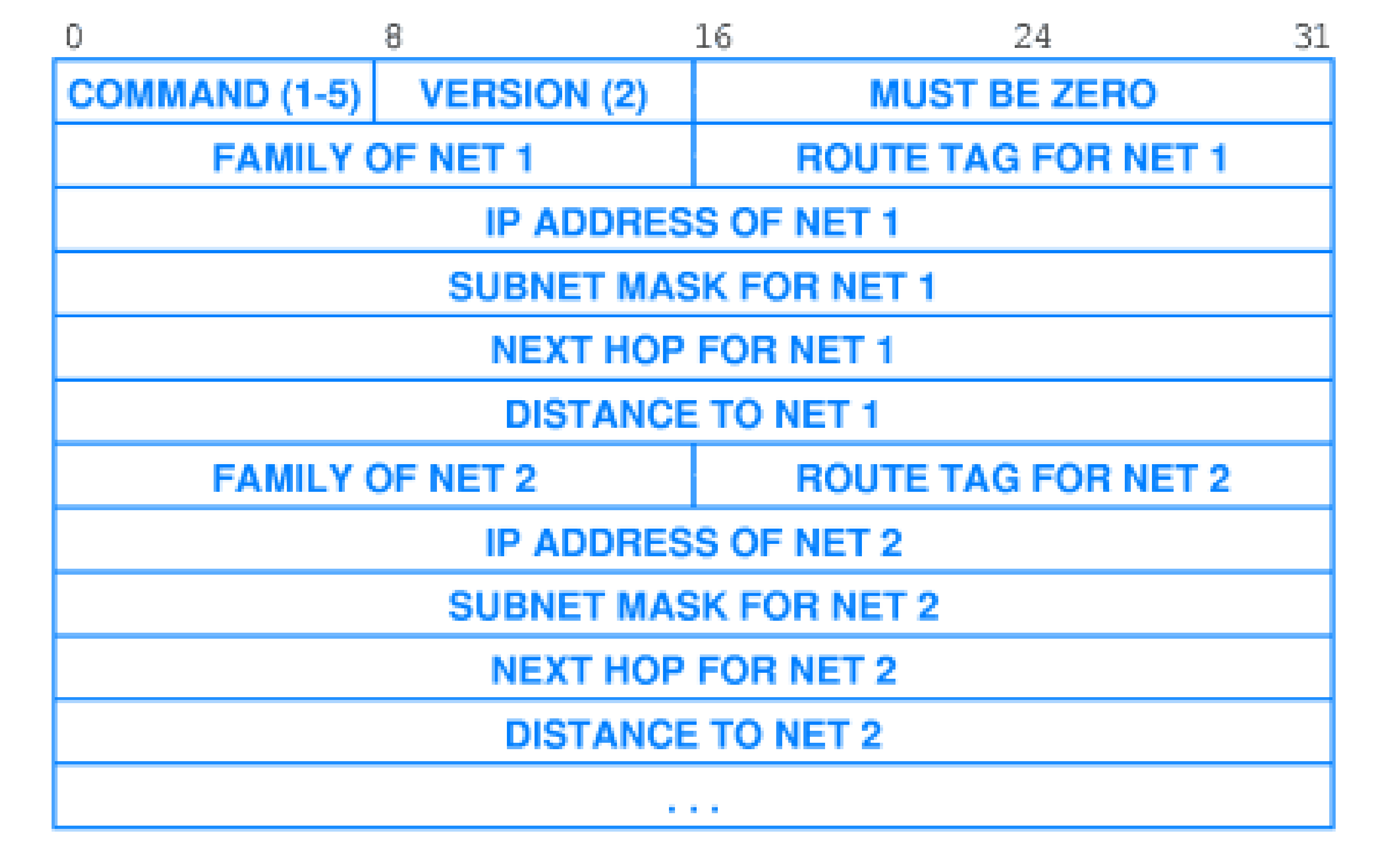
- If you advertise access to networks through BGP, you can get large parts of traffic to these networks routed through your network. This makes it possible to monitor and eavesdrop on the traffic.
- This happens now and then, often claimed to be by mistake.
- There are some efforts to mitigate these problems, based on the idea that new routes and network operators must be authenticated to be able to advertise new routes.

https://www.zdnet.com/article/russian-telco-hijacks-internet-traffic-for-google-aws-cloudflare-and-others/
https://blog.apnic.net/2021/02/05/bgp-rpki-and-manrs-2020-in-review/

The Routing Information Protocol (RIP)

- Routing within an autonomous system (IGP)
- Hop count metric
- Unreliable transport (uses UDP)
- Broadcast (ver. 1) or multicast (ver. 2) delivery
- Distance vector algorithm
- Can propagate a default route
- Implemented by Unix program routed
- Passive version for hosts

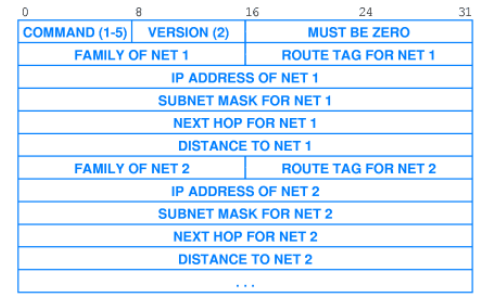
RIP Packet Format



<table>
<tr><th>0</th><th>8</th><th>16</th><th>24</th><th>31</th></tr>
<tr><td>COMMAND (1-5)</td><td>VERSION (2)</td><td colspan="3">MUST BE ZERO</td></tr>
<tr><td colspan="2">FAMILY OF NET 1</td><td colspan="3">ROUTE TAG FOR NET 1</td></tr>
<tr><td colspan="5">IP ADDRESS OF NET 1</td></tr>
<tr><td colspan="5">SUBNET MASK FOR NET 1</td></tr>
<tr><td colspan="5">NEXT HOP FOR NET 1</td></tr>
<tr><td colspan="5">DISTANCE TO NET 1</td></tr>
<tr><td colspan="2">FAMILY OF NET 2</td><td colspan="3">ROUTE TAG FOR NET 2</td></tr>
<tr><td colspan="5">IP ADDRESS OF NET 2</td></tr>
<tr><td colspan="5">SUBNET MASK FOR NET 2</td></tr>
<tr><td colspan="5">NEXT HOP FOR NET 2</td></tr>
<tr><td colspan="5">DISTANCE TO NET 2</td></tr>
<tr><td colspan="5">. . .</td></tr>
</table>

Family of net = 2 mean IP, Family of net = FFFF mean authentication, Route tag is for advertise routes learn from an external AS

RIP Packet Format



0	8	16	24	31
COMMAND (1-5)	VERSION (2)	MUST BE ZERO		
FAMILY OF NET 1		ROUTE TAG FOR NET 1		
IP ADDRESS OF NET 1				
SUBNET MASK FOR NET 1				
NEXT HOP FOR NET 1				
DISTANCE TO NET 1				
FAMILY OF NET 2		ROUTE TAG FOR NET 2		
IP ADDRESS OF NET 2				
SUBNET MASK FOR NET 2				
NEXT HOP FOR NET 2				
DISTANCE TO NET 2				
. . .				

Command: request or response message

Version: This field indicates the RIP version used.

Must be zero: Not used in RIPv1. This field provides backward compatibility with pre-standard varieties of RIP. The default value is zero.

Address family identifier (AFI): This field indicates the type of address. RIP can carry routing information for several different protocols. For IP, it is 2.

Route tag: allows a distinction between routes learned from the RIP protocol and routes learned from other protocols

Address: This field indicates the IP address for the packet.

Metric: This field specifies the number of hops to the destination.

Mask This field specifies the IP address mask.

Next hop: This field specifies the IP address of the next router along the path to the destination.

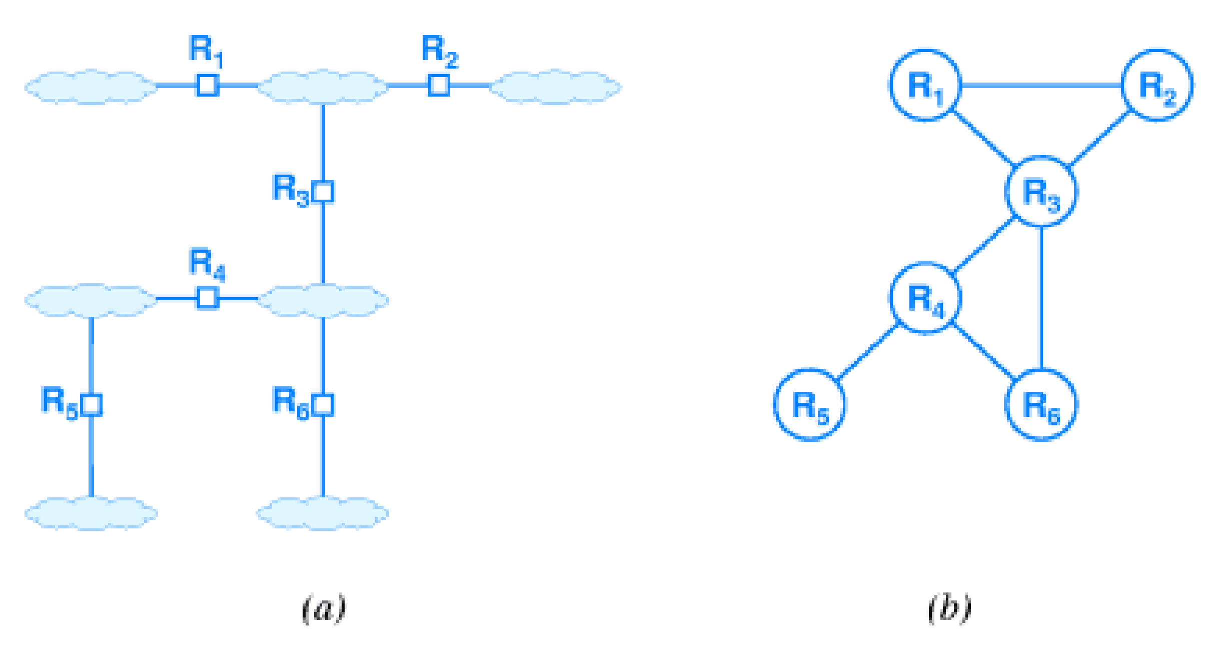
The Open Shortest Path First Protocol (OSPF)

- Routing within an autonomous system (IGP)
- Full CIDR and subnet support
- Authenticated message exchange
- Allows routes to be imported from outside the autonomous system
- Uses link-status (SPF) algorithm
- Support for multi-access networks (e.g., Ethernet)

Link-Status on the internet

- The router corresponds to a node in the graph
- Network corresponds to the edge
- Adjacent pair of routers periodically
  - Test connectivity
  - Broadcast link-status information to area
- Each router uses link-status messages to compute the shortest paths

Illustration of OSPF Graph



- *(a)* an interconnect of routers and networks, and
- *(b)* an equivalent OSPF graph
- Router corresponds to a node in the graph

OSPF Areas and Efficiency

- Allows subdivision of AS into areas
- Link-status information propagated within area
- Routes summarized before being propagated to another area
- Reduces overhead (less broadcast traffic)

OSPF and Scale

Because it allows a manager to partition the routers and networks in an autonomous system into multiple areas, OSPF can scale to handle a much larger number of routers than other IGPs

Some networks are too big even for OSPF. For these networks, BGP (then called IBGP) can be used as IGP.

Intermediate System – Intermediate System (IS-IS)

- Originally designed by DEC to be part of DECNET V
- The IS-IS is an IGP
  - naming follows Digital's terminology in which a router was called an IS and a host was called an End System
- IS-IS was created around the same time as OSPF
- The two protocols are similar in many ways
  - Both use the link-state approach
    - employ Dijkstra's algorithm to compute the shortest paths
  - Both protocols require two adjacent routers to periodically test the link between them and broadcast a status message

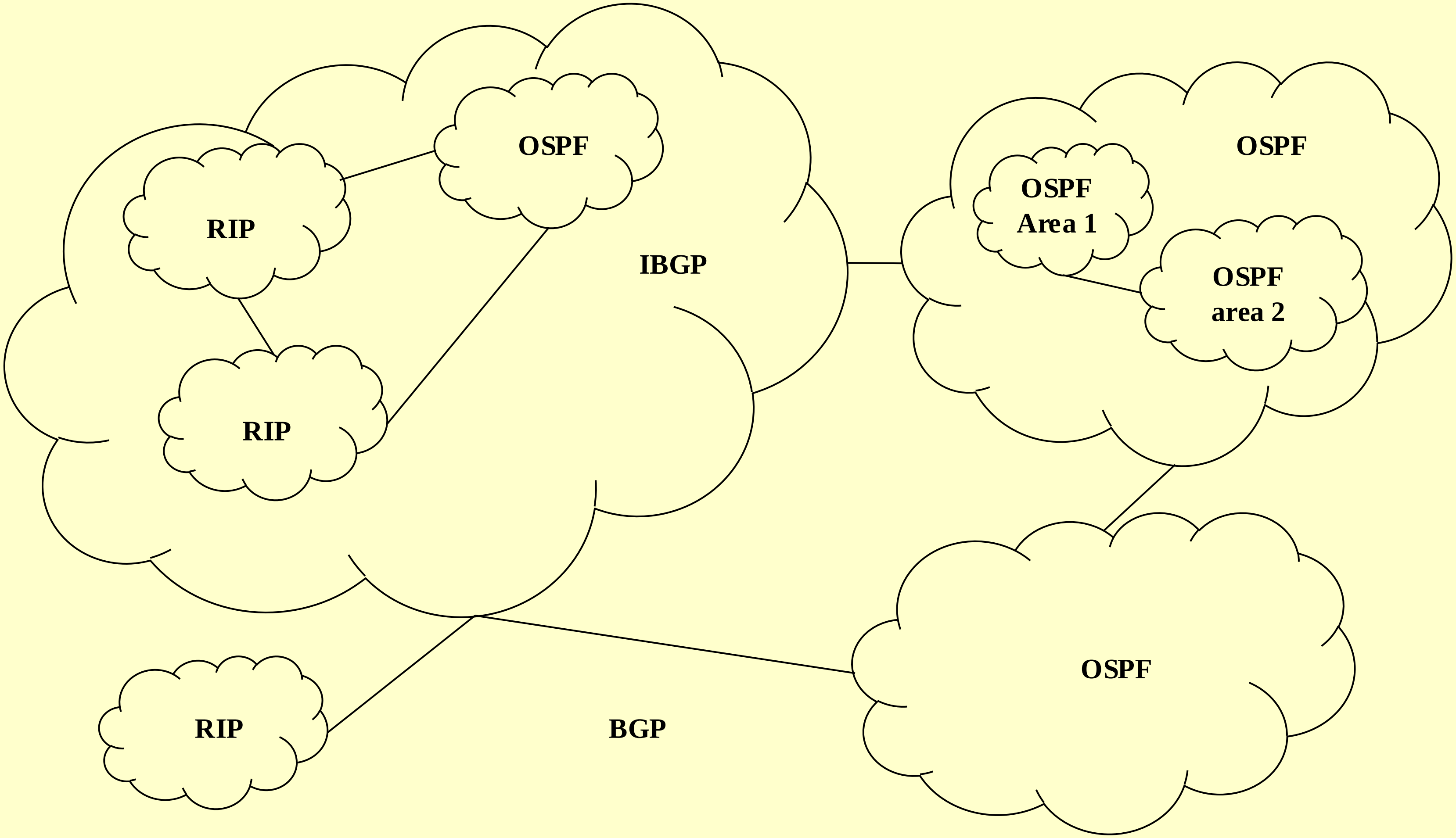
Intermediate System – Intermediate System (IS-IS)

- The main differences between OSPF and the original IS-IS can be summarized as:
- IS-IS was proprietary (owned by DEC)
  - and OSPF was created as an open standard available to all vendors
- OSPF was designed to run over IP
  - IS-IS was designed to run over CLNS (part of the OSI protocol stack)
- OSPF was designed to propagate IPv4 routes
  - IS-IS was designed to propagate routes for OSI protocols
- Over time, OSPF gained many features
  - As a result, IS-IS now has less overhead
- When the protocols were initially invented
  - OSPF's openness made it much more popular than IS-IS
  - In fact, IS-IS was almost completely forgotten

Intermediate System – Intermediate System (IS-IS)

- As the years progressed
    - OSPF's popularity encouraged the IETF to add additional features
- Ironically, in the early 2000s, ten years after the protocols were designed
    - several things changed to give IS-IS a second chance
- DEC had dissolved, and IS-IS was no longer considered valuable proprietary property
- A newer version of IS-IS was defined to integrate it with IP and the Internet
- Because OSPF was built for IPv4
    - a completely new version had to be developed to handle larger IPv6 addresses
- The largest ISPs have grown to a size where the extra overhead in OSPF makes IS-IS more attractive
- As a result, IS-IS has started to make a comeback

Relations between different routing protocols



Routing Problems
Where Intuition Fails

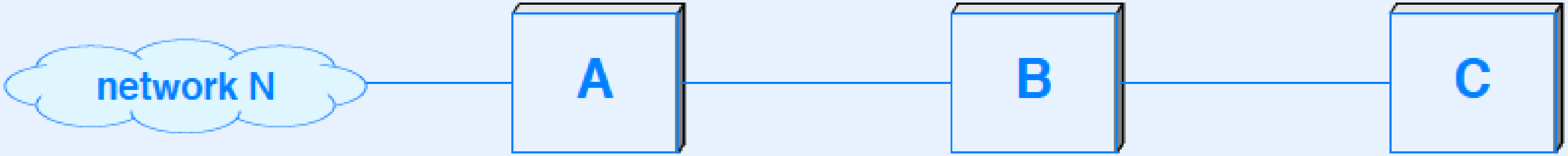
- Routing is *not* like water flowing through pipes or traffic on highways
  - Multi-path routing is difficult
  - Capacity can go unused if not along the “shortest” path
- The fewest hops may not always be the best
  - Compare two Ethernet hops and one satellite hop
- Routing around congestion is *not* straightforward and does *not* always yield a big improvement
  - Can cause out-of-order packets (TCP reacts)
  - Can result in route flapping

Routing Problems
Loops And Convergence

- Routing loop
  - Circular routes
  - Can be caused if “good news” flows backwards
- Slow convergence (count to infinity) problem arises
  - Routes fail to converge after a change
  - Can cause a routing loop to persist

Routing Problems
How Good News Can Backwash

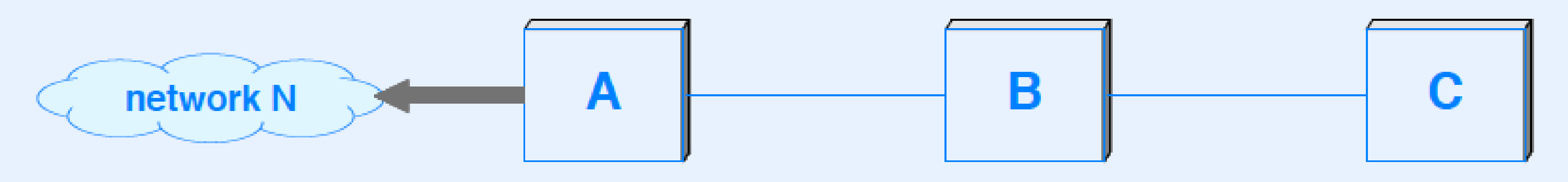
- A story with three routers and a network



Three routers A,B and C are connected each other

Routing Problems
How Good News Can Backwash

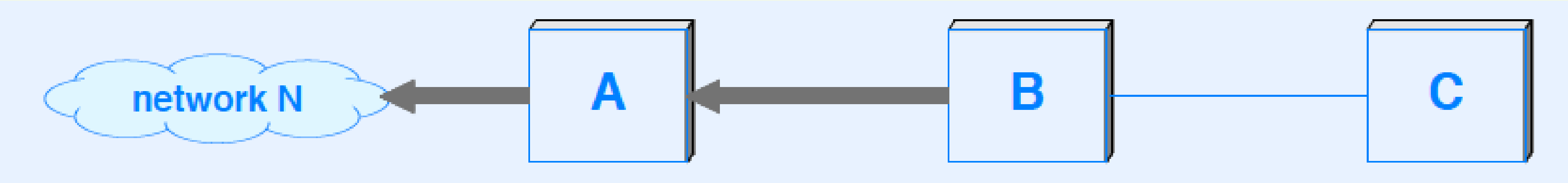
- A story with three routers and a network



Router A can reach network N

Routing Problems
How Good News Can Backwash

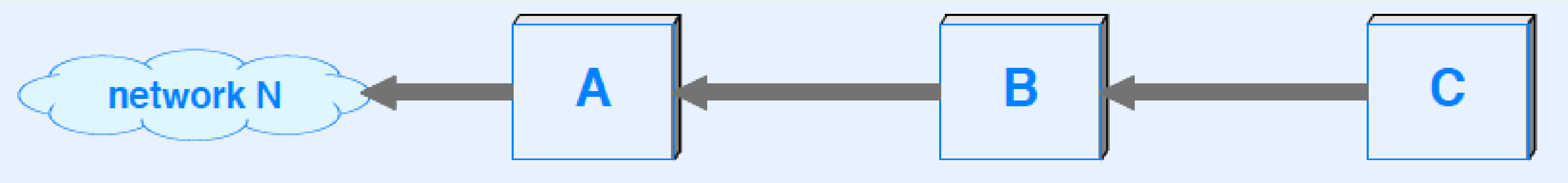
- A story with three routers and a network



Router B can reach network N trough router A

Routing Problems
How Good News Can Backwash

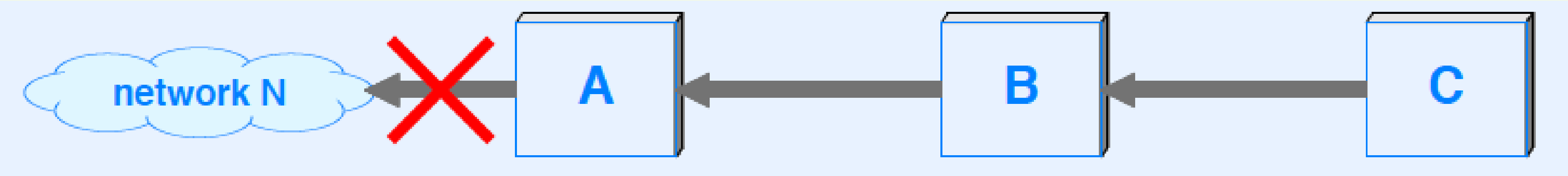
- A story with three routers and a network



Router C can reach network N trough routers B and A, now they all know they can reach network N

Routing Problems
How Good News Can Backwash

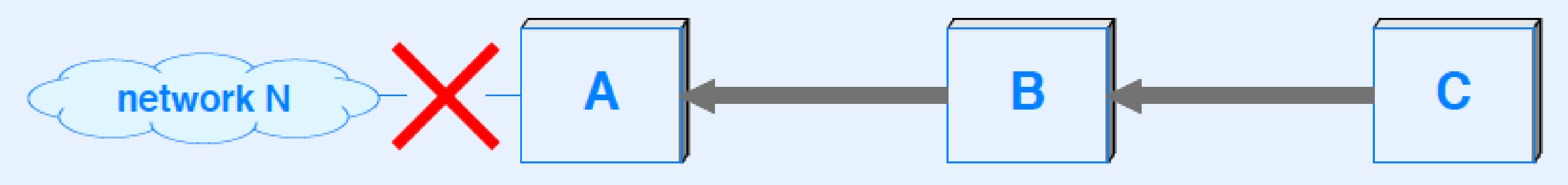
- A story with three routers and a network



Router A loooses it connection to network N

Routing Problems
How Good News Can Backwash

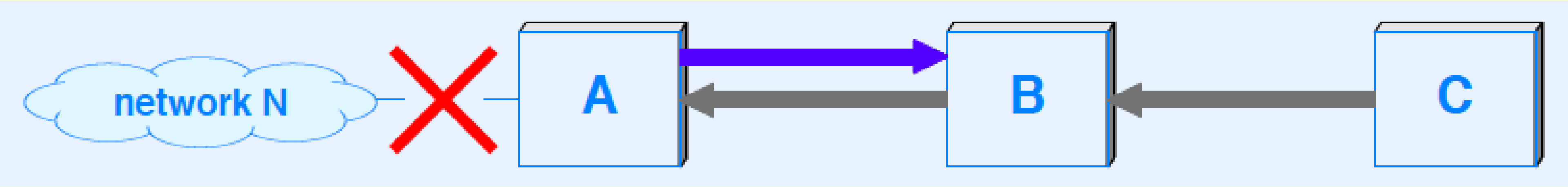
- A story with three routers and a network



Router A removes the info to reach network N directly from it’s routing table

Routing Problems
How Good News Can Backwash

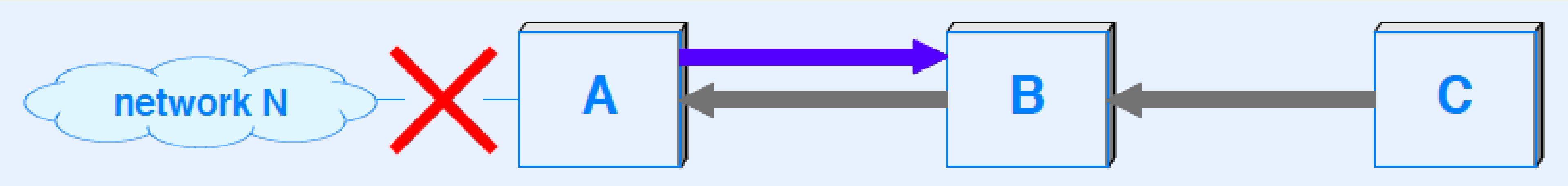
- A story with three routers and a network



“Luckily” Router A gets info from router B that it can still reach (incorrectly) network N by sending it to router B

Routing Problems
How Good News Can Backwash

- A story with three routers and a network



- In practice, modern DV protocols employ heuristics that
  - Eliminate backflow
  - Lock down changes after a failure

Other Routing Problems

- Black hole
  - Routing system sends packets for a set of destinations to a location where they are silently discarded
  - This can be caused if routing update packets are lost
- Route flapping (lack of convergence)
  - Routes continue to oscillate
  - This can be caused by equal-length paths or by routes continuously appearing and disappearing

Routing Overhead

- Traffic from routing protocols is “overhead”
- Specific cases
  - DV advertisements tend to be large
  - LSR uses broadcast
- Fundamental trade-off
  - Decreasing the frequency of routing exchanges lowers overhead
  - Increasing the frequency of routing exchanges reduces the time between a failure and rerouting around the failure

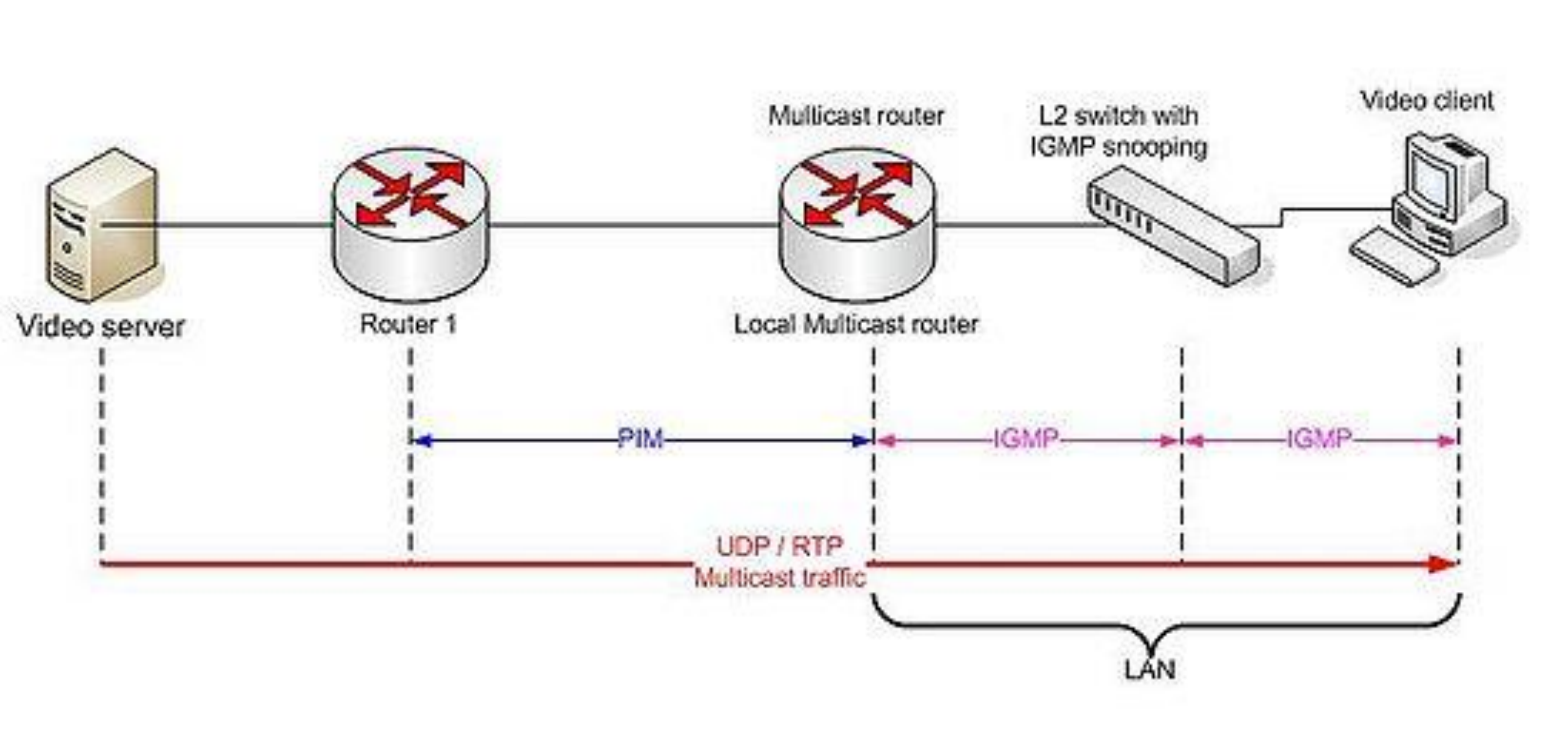
Internet Multicast Routing

- IPv4 Multicast
  - Defined early; informally called “Deering multicast” (after Steve Deering inventor of IP Multicast)
  - Provides Internet-wide multicast dissemination
  - Uses IPv4 addresses 224.0.0.0 through 239.255.255.255 (the original Class D address space)
- IPv6 Multicast
  - Fundamental part of IPv6
  - IPv6 prohibits broadcast, but defines multicast groups that are equivalent

Internet Multicast Routing

- Difficult because Internet multicast allows
  - Arbitrary computer to join a multicast group at any time
  - Arbitrary member to leave the multicast group at any time
  - Arbitrary computer to send message to a group (even if not a member)
- Internet Group Multicast Protocol (IGMP)
  - Used between computer and local router
  - Specifies multicast group membership

Multicast architecture



- IGMP is used both by the client computer and the adjacent network switches to connect the client to a local multicast router. Different protocols are then used between the local and remote multicast routers, to direct multicast traffic from servers to many multicast clients.

IP Multicast And Ethernet Delivery

- When sending IP multicast across Ethernet
  - Can use Ethernet multicast capability
  - IP multicast address is mapped to an Ethernet multicast address
- Problem
  - Most interface hardware limits the number of Ethernet multicast addresses that can be used simultaneously
  - Trick: use a few multicast addresses and allow the software to decide how a given packet should be processed (multiplexing)

Multicast Routing
Forwarding and Discovery Techniques

- When a router learns that a host on one of its networks has joined a multicast group
  - the router must establish a path to the group and propagate datagrams it receives for the group to the host
  - thus, routers, not hosts, have responsibility for the propagation of multicast routing information
- Dynamic group membership and support for anonymous senders make general-purpose multicast routing extremely difficult
- Moreover, the size and topology of groups vary considerably among applications
  - For example, teleconferencing often creates small groups
    - who may be geographically dispersed or in the same organization
  - A webcast application can potentially create a group with millions of members that span the globe

Multicast Routing
Forwarding and Discovery Techniques

- To accommodate dynamic membership
  - multicast routing protocols must be able to change routes quickly and continually
- For example, if a user in France joins a multicast group that has members in the U.S. and Japan
  - multicast routing software must first find other members of the group, and then create an optimal forwarding structure
- An arbitrary user can send a datagram to the group
  - information about routes must extend beyond group members
- In practice, multicast protocols have followed three different approaches for datagram forwarding:
  - Flood-and-Prune
  - Configuration-and-Tunnelling
  - Core-Based Discovery

Multicast Routing
Forwarding and Discovery Techniques

Flood-and-Prune

- Flood-and-prune is ideal in a situation where the group is small, and all members are attached to contiguous LANs
- Initially, routers forward each datagram to all networks
    - That is, when a multicast datagram arrives
        - a router transmits the datagram on all directly attached LANs via hardware multicast
- To avoid routing loops
    - flood-and-prune protocols use a technique known as Reverse Path Broadcasting (RPB) that breaks cycles
- While the flooding stage proceeds
    - routers exchange information about group membership
- If a router learns that no hosts on a given network are members of the group
    - the router stops forwarding multicast to the network (i.e., “prunes” the network from the set)

Multicast Routing
Forwarding and Discovery Techniques

Configuration-and-Tunnelling

- Configuration-and-tunnelling is ideal in a situation where the group is geographically dispersed

  (i.e., has a few members at each site, with sites far apart)
- A router at each site is configured to know about other sites
- When a multicast datagram arrives
  - the router at a site transmits the datagram on all directly attached LANs via hardware multicast
- The router then consults its configuration table to
  - determine which remote sites should receive a copy
  - and uses IP-in-IP tunnelling to transfer a copy of the multicast datagram to each of the remote sites

Multicast Routing
Forwarding and Discovery Techniques

Core-Based Discovery

- Flood-and-prune and configuration-and-tunnelling handle extreme cases well
- But a technique is needed that allows multicast to scale gracefully from a small group in one area to a large group with members at arbitrary locations
- To provide smooth growth
  - some multicast routing protocols designate a core unicast address for each multicast group.
- Whenever a router R_1 receives a multicast datagram that must be transmitted to a group
  - R_1 encapsulates the multicast datagram in a unicast datagram
    - and forwards the unicast datagram to the group's core unicast address

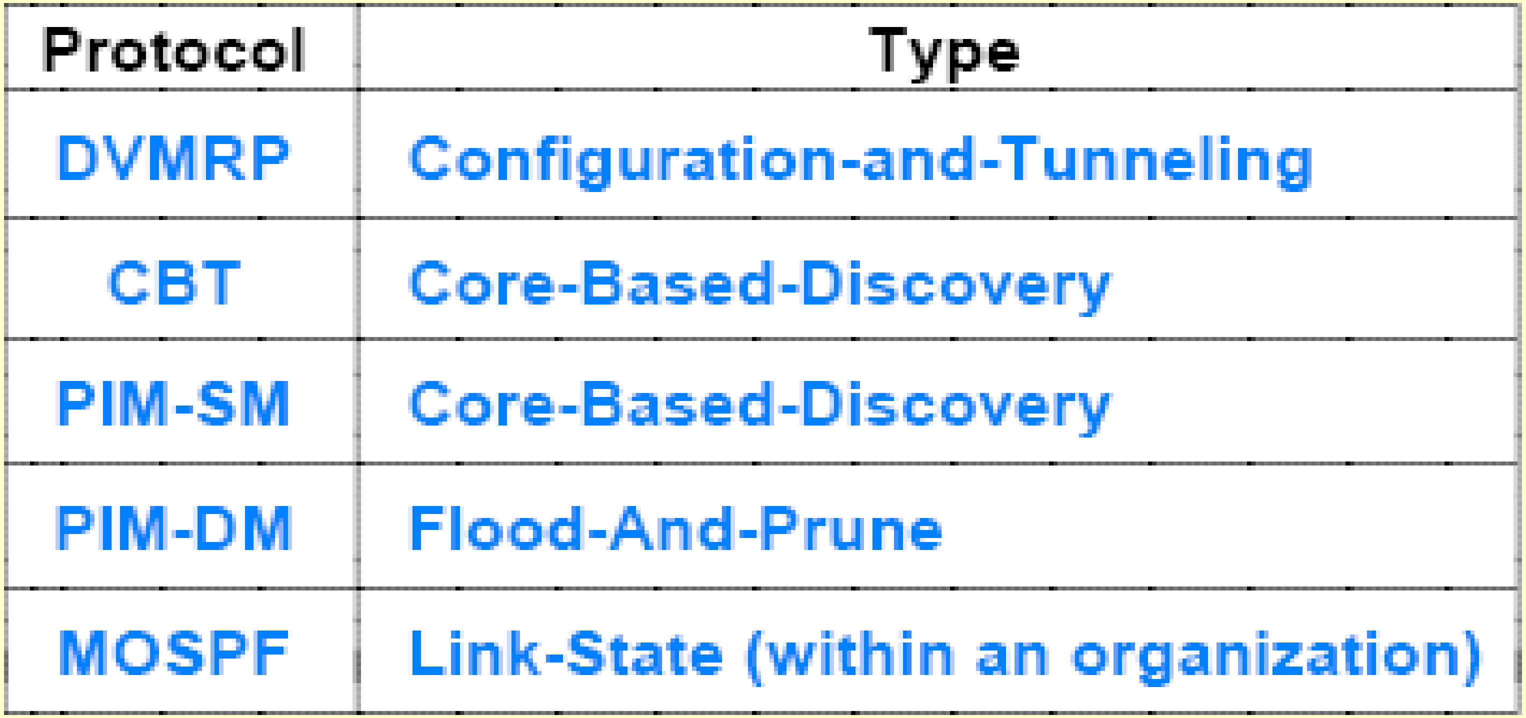
Multicast Routing
Forwarding and Discovery Techniques

- As the unicast datagram travels through the Internet
  - each router examines the contents
- When the datagram reaches a router R_2 that participates in the group
  - R_2 removes and processes the multicast message
  - R_2 uses multicast routing to forward the datagram to members of the group
- Requests to join the group follow the same pattern
  - if it receives a request to join a group
    - R_2 adds a new route to its multicast forwarding table and begins to forward a copy of each multicast datagram to R_1
- The set of routers receiving a particular multicast group grows from the core outward
  - In graph-theoretic terms, the routers form a tree

Multicast Routing Protocols

- Several protocols exist
  - Distance Vector Multicast Routing Protocol (DVMRP)
  - Core Based Trees (CBT)
  - Protocol Independent Multicast – Sparse Mode (PIM-SM)
  - Protocol Independent Multicast – Dense Mode (PIM-DM)
  - Multicast extensions to the Open Shortest Path First (MOSPF)
- None best in all circumstances

Multicast Routing Protocols



Protocol	Type
DVMRP	Configuration-and-Tunneling
CBT	Core-Based-Discovery
PIM-SM	Core-Based-Discovery
PIM-DM	Flood-And-Prune
MOSPF	Link-State (within an organization)

Multicast routing today

- IP multicast is widely deployed in enterprises, commercial stock exchanges, and multimedia content delivery networks.
- No mechanism has yet been demonstrated allowing the IP multicast model to scale to the full Internet. For this reason, and also reasons of economics, IP multicast is not, in general, used in commercial Internet backbones.

Summary

- Static routing used by hosts
- Routers require automatic routing
- Internet divided into autonomous systems
- Two broad classes of routing protocols
  - Interior Gateway Protocols (IGPs) provide routing within an autonomous system
  - Exterior Gateway Protocols (EGPs) provide routing among autonomous systems

Summary (continued)

- Border Gateway Protocol (BGP version 4) is the current EGP used in Internet
- Interior Gateway Protocols include:
  - Routing Information Protocol (RIP)
  - Open Shortest Path First protocol (OSPF)
- Internet multicast routing difficult
  - Protocols proposed include: DVMRP, PIM-SM, PIM-DM, MOSPF