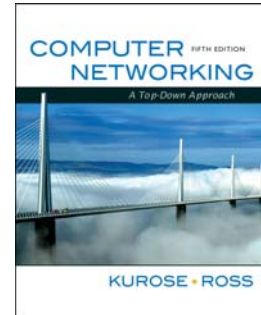


# Chapter 4 Network Layer



## A note on the use of these ppt slides:

We're making these slides freely available to all (faculty, students, readers). They're in PowerPoint form so you can add, modify, and delete slides (including this one) and slide content to suit your needs. They obviously represent a *lot* of work on our part. In return for use, we only ask the following:

- ❖ If you use these slides (e.g., in a class) in substantially unaltered form, that you mention their source (after all, we'd like people to use our book!)
- ❖ If you post any slides in substantially unaltered form on a www site, that you note that they are adapted from (or perhaps identical to) our slides, and note our copyright of this material.

Thanks and enjoy! JFK/KWR

All material copyright 1996-2010  
J.F Kurose and K.W. Ross, All Rights Reserved

*Computer Networking:  
A Top Down Approach  
5<sup>th</sup> edition.  
Jim Kurose, Keith Ross  
Addison-Wesley, April  
2009.*

Network Layer 4-1

## Chapter 4: Network Layer

### Chapter goals:

- ❖ understand principles behind network layer services:
  - network layer service models
  - forwarding versus routing
  - how a router works
  - routing (path selection)
  - broadcast, multicast
- ❖ instantiation, implementation in the Internet

Network Layer 4-2

## Chapter 4: Network Layer

### 4.1 Introduction

#### 4.2 Virtual circuit and datagram networks

#### 4.3 What's inside a router

#### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

#### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

#### 4.6 Routing in the Internet

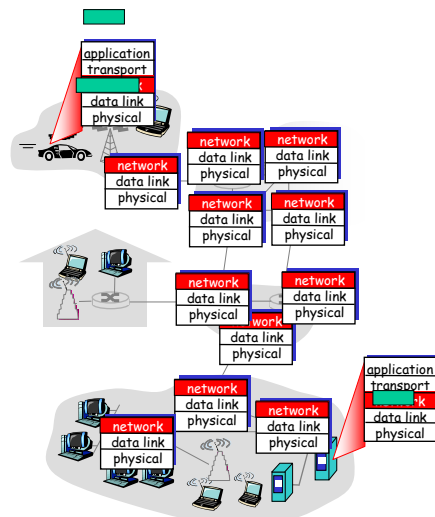
- RIP
- OSPF
- BGP

#### 4.7 Broadcast and multicast routing

Network Layer 4-3

## Network layer

- ❖ transport segment from sending to receiving host
- ❖ on sending side encapsulates segments into datagrams
- ❖ on rcving side, delivers segments to transport layer
- ❖ network layer protocols in *every* host, router
- ❖ router examines header fields in all IP datagrams passing through it



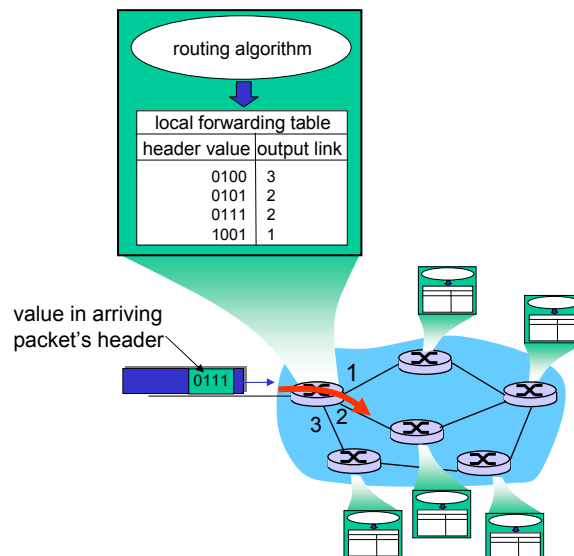
Network Layer 4-4

## Two Key Network-Layer Functions

- ❖ *forwarding*: move packets from router's input to appropriate router output
  - ❖ *routing*: determine route taken by packets from source to dest.
    - *routing algorithms*
- analogy:
- ❖ *routing*: process of planning trip from source to dest
  - ❖ *forwarding*: process of getting through single interchange

Network Layer 4-5

## Interplay between routing and forwarding



Network Layer 4-6

## Connection setup

- ❖ 3<sup>rd</sup> important function in *some* network architectures:
  - ATM, frame relay, X.25
- ❖ before datagrams flow, two end hosts *and* intervening routers establish virtual connection
  - routers get involved
- ❖ network vs transport layer connection service:
  - **network**: between two hosts (may also involve intervening routers in case of VCs)
  - **transport**: between two processes

Network Layer 4-7

## Network service model

**Q:** What *service model* for "channel" transporting datagrams from sender to receiver?

### example services for individual datagrams:

- ❖ guaranteed delivery
- ❖ guaranteed delivery with less than 40 msec delay

### example services for a flow of datagrams:

- ❖ in-order datagram delivery
- ❖ guaranteed minimum bandwidth to flow
- ❖ restrictions on changes in inter-packet spacing

Network Layer 4-8

## Network layer service models:

Network Architecture	Service Model	Guarantees ?				Congestion feedback
		Bandwidth	Loss	Order	Timing	
Internet	best effort	none	no	no	no	no (inferred via loss)
ATM	CBR	constant rate	yes	yes	yes	no congestion
ATM	VBR	guaranteed rate	yes	yes	yes	no congestion
ATM	ABR	guaranteed minimum	no	yes	no	yes
ATM	UBR	none	no	yes	no	no

Network Layer 4-9

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-10

## Network layer connection and connection-less service

- ❖ datagram network provides network-layer connectionless service
- ❖ VC network provides network-layer connection service
- ❖ analogous to the transport-layer services, but:
  - **service:** host-to-host
  - **no choice:** network provides one or the other
  - **implementation:** in network core

Network Layer 4-11

## Virtual circuits

"source-to-dest path behaves much like telephone circuit"

- performance-wise
  - network actions along source-to-dest path
- 
- ❖ call setup, teardown for each call *before* data can flow
  - ❖ each packet carries VC identifier (not destination host address)
  - ❖ *every* router on source-dest path maintains "state" for each passing connection
  - ❖ link, router resources (bandwidth, buffers) may be *allocated* to VC (dedicated resources = predictable service)

Network Layer 4-12

## VC implementation

a VC consists of:

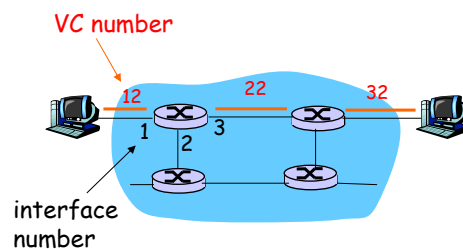
1. path from source to destination
  2. VC numbers, one number for each link along path
  3. entries in forwarding tables in routers along path
- ❖ packet belonging to VC carries VC number (rather than dest address)
  - ❖ VC number can be changed on each link.
    - New VC number comes from forwarding table

Network Layer 4-13

## VC Forwarding table

Forwarding table in northwest router:

Incoming interface	Incoming VC #	Outgoing interface	Outgoing VC #
1	12	3	22
2	63	1	18
3	7	2	17
1	97	3	87
...	...	...	...

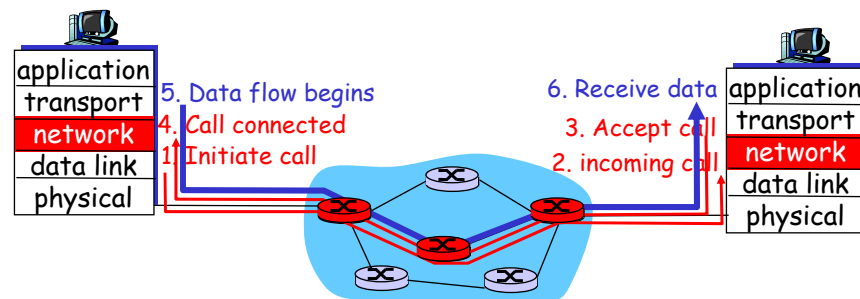


**Routers maintain connection state information!**

Network Layer 4-14

## Virtual circuits: signaling protocols

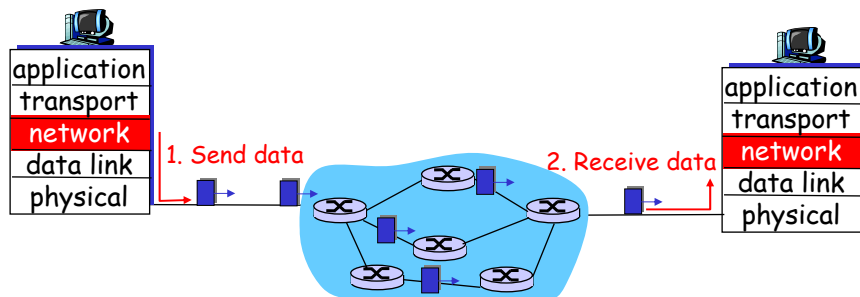
- ❖ used to setup, maintain teardown VC
- ❖ used in ATM, frame-relay, X.25
- ❖ not used in today's Internet



Network Layer 4-15

## Datagram networks

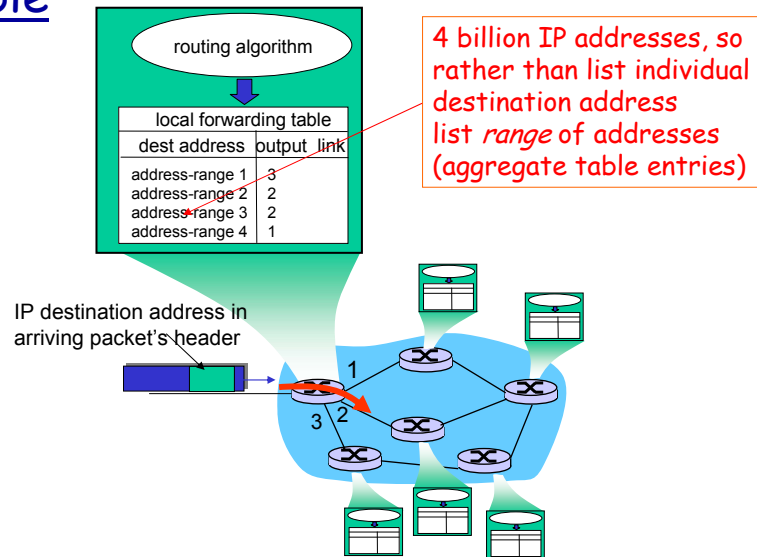
- ❖ no call setup at network layer
- ❖ routers: no state about end-to-end connections
  - no network-level concept of "connection"
- ❖ packets forwarded using destination host address
  - packets between same source-dest pair may take different paths



Network Layer 4-16



## Datagram Forwarding table



Network Layer 4-17

## Datagram Forwarding table

Destination Address Range	Link Interface
11001000 00010111 00010000 00000000 through 11001000 00010111 00010111 11111111	0
11001000 00010111 00011000 00000000 through 11001000 00010111 00011000 11111111	1
11001000 00010111 00011001 00000000 through 11001000 00010111 00011111 11111111	2
otherwise	3

Q: but what happens if ranges don't divide up so nicely?

Network Layer 4-18

## Longest prefix matching

### Longest prefix matching

when looking for forwarding table entry for given destination address, use *longest* address prefix that matches destination address.

Destination Address Range	Link interface
11001000 00010111 00010*** *****	0
11001000 00010111 00011000 *****	1
11001000 00010111 00011*** *****	2
otherwise	3

### Examples:

DA: 11001000 00010111 00010110 10100001

Which interface?

DA: 11001000 00010111 00011000 10101010

Which interface?

Network Layer 4-19

## Datagram or VC network: why?

### Internet (datagram)

- ❖ data exchange among computers
  - "elastic" service, no strict timing req.
- ❖ "smart" end systems (computers)
  - can adapt, perform control, error recovery
  - simple inside network, complexity at "edge"
- ❖ many link types
  - different characteristics
  - uniform service difficult

### ATM (VC)

- ❖ evolved from telephony
- ❖ human conversation:
  - strict timing, reliability requirements
  - need for guaranteed service
- ❖ "dumb" end systems
  - telephones
  - complexity inside network

Network Layer 4-20

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router?

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

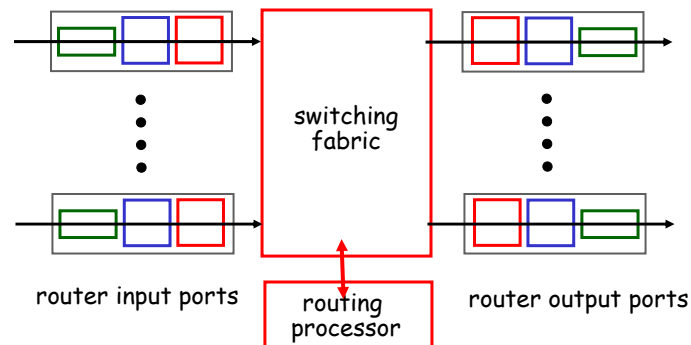
### 4.7 Broadcast and multicast routing

Network Layer 4-21

## Router Architecture Overview

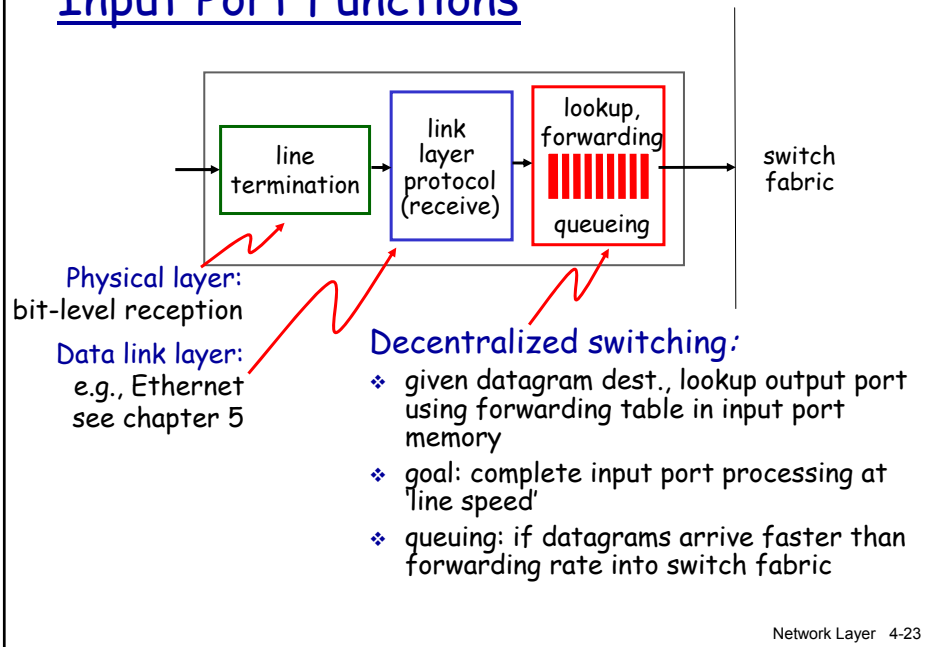
two key router functions:

- ❖ run routing algorithms/protocol (RIP, OSPF, BGP)
- ❖ *forwarding* datagrams from incoming to outgoing link



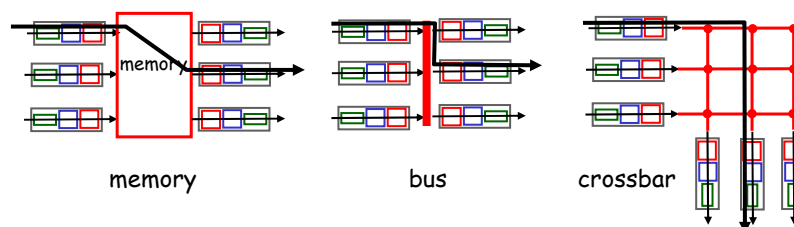
Network Layer 4-22

## Input Port Functions



## Switching fabrics

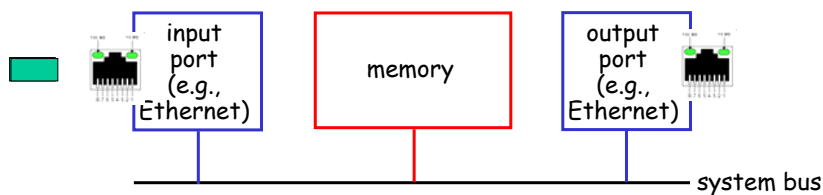
- ❖ transfer packet from input buffer to appropriate output buffer
- ❖ switching rate: rate at which packets can be transfer from inputs to outputs
  - often measured as multiple of input/output line rate
  - N inputs: switching rate N times line rate desirable
- ❖ three types of switching fabrics



## Switching Via Memory

### First generation routers:

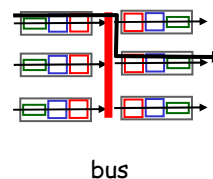
- ❖ traditional computers with switching under direct control of CPU
- ❖ packet copied to system's memory
- ❖ speed limited by memory bandwidth (2 bus crossings per datagram)



Network Layer 4-25

## Switching Via a Bus

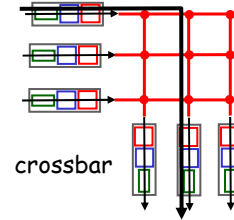
- ❖ datagram from input port memory to output port memory via a shared bus
- ❖ **bus contention:** switching speed limited by bus bandwidth
- ❖ 32 Gbps bus, Cisco 5600: sufficient speed for access and enterprise routers



Network Layer 4-26

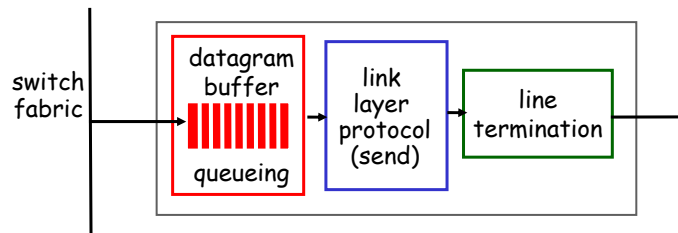
## Switching Via An Interconnection Network

- ❖ overcome bus bandwidth limitations
- ❖ Banyan networks, crossbar, other interconnection nets initially developed to connect processors in multiprocessor
- ❖ advanced design: fragmenting datagram into fixed length cells, switch cells through the fabric.
- ❖ Cisco 12000: switches 60 Gbps through the interconnection network



Network Layer 4-27

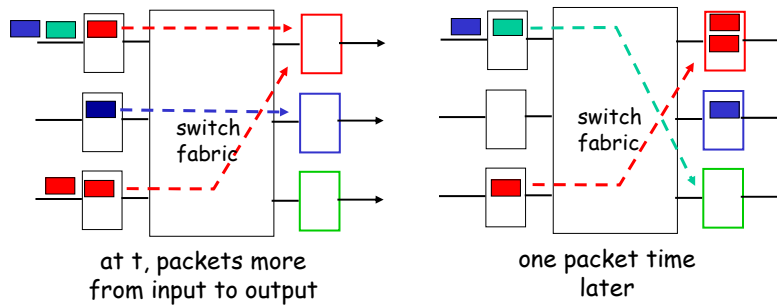
## Output Ports



- ❖ *buffering* required when datagrams arrive from fabric faster than the transmission rate
- ❖ *scheduling discipline* chooses among queued datagrams for transmission

Network Layer 4-28

## Output port queuing



- ❖ buffering when arrival rate via switch exceeds output line speed
- ❖ *queueing (delay) and loss due to output port buffer overflow!*

Network Layer 4-29

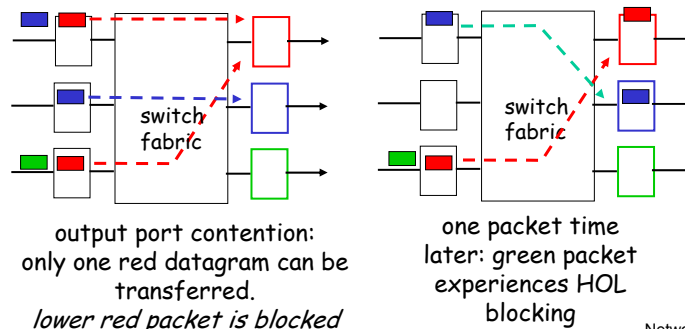
## How much buffering?

- ❖ RFC 3439 rule of thumb: average buffering equal to "typical" RTT (say 250 msec) times link capacity  $C$ 
  - e.g.,  $C = 10$  Gpbs link: 2.5 Gbit buffer
- ❖ recent recommendation: with  $N$  flows, buffering equal to  $\frac{RTT \cdot C}{\sqrt{N}}$

Network Layer 4-30

## Input Port Queuing

- ❖ fabric slower than input ports combined -> queueing may occur at input queues
  - *queueing delay and loss due to input buffer overflow!*
- ❖ **Head-of-the-Line (HOL) blocking:** queued datagram at front of queue prevents others in queue from moving forward



Network Layer 4-31

## Chapter 4: Network Layer

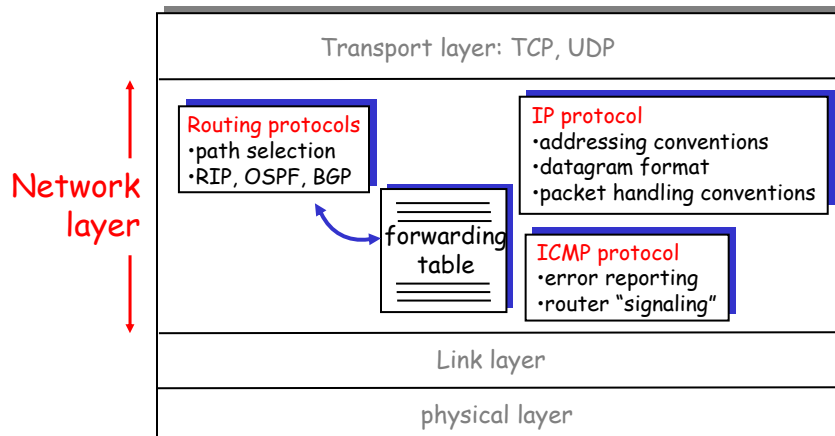
- ❖ 4.1 Introduction
- ❖ 4.2 Virtual circuit and datagram networks
- ❖ 4.3 What's inside a router
- ❖ **4.4 IP: Internet Protocol**
  - Datagram format
  - IPv4 addressing
  - ICMP
  - IPv6
- ❖ 4.5 Routing algorithms
  - Link state
  - Distance Vector
  - Hierarchical routing
- ❖ 4.6 Routing in the Internet
  - RIP
  - OSPF
  - BGP
- ❖ 4.7 Broadcast and multicast routing

Network Layer 4-32



## The Internet Network layer

Host, router network layer functions:



Network Layer 4-33

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

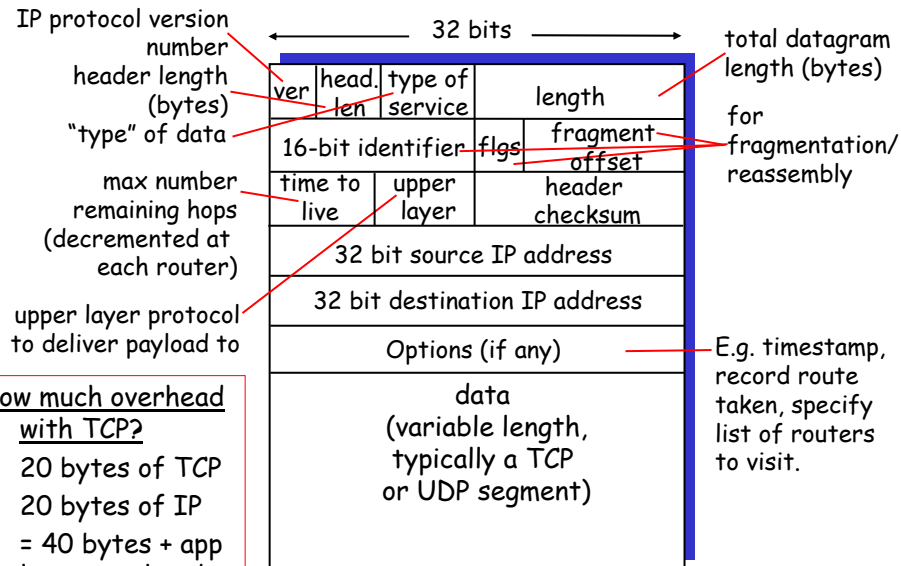
### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-34

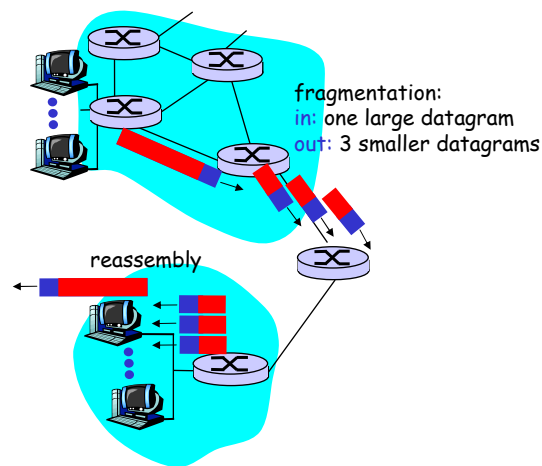
## IP datagram format



Network Layer 4-35

## IP Fragmentation & Reassembly

- ❖ network links have MTU (max.transfer size) - largest possible link-level frame.
  - different link types, different MTUs
- ❖ large IP datagram divided ("fragmented") within net
  - one datagram becomes several datagrams
  - "reassembled" only at final destination
  - IP header bits used to identify, order related fragments



Network Layer 4-36

## IP Fragmentation and Reassembly

### Example

- ❖ 4000 byte datagram
- ❖ MTU = 1500 bytes

1480 bytes in data field

offset =  
 $1480/8$

length	ID	fragflag	offset
=4000	=x	=0	=0

One large datagram becomes several smaller datagrams

length	ID	fragflag	offset
=1500	=x	=1	=0
=1500	=x	=1	=185
=1040	=x	=0	=370

Network Layer 4-37

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

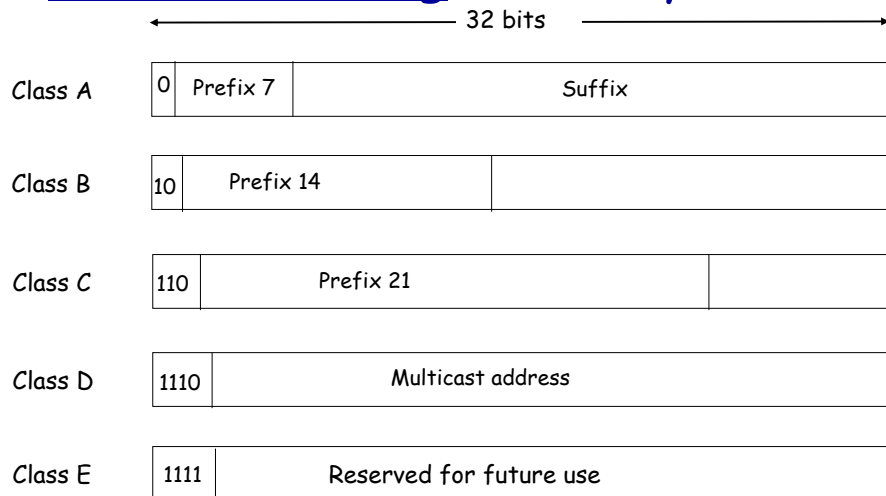
Network Layer 4-38

## IPv4 Addressing old days

- ❖ 32-bit address
- ❖ Theoretically, up to 4G address, practically much less than that because of the way the address is structured
- ❖ 5 different classes

Network Layer 4-39

## IPv4 addressing old days



Network Layer 4-40

## IPv4 Addressing old days

- ❖ The prefix (together with the bits identifying the class) identifies the network
- ❖ The suffix identifies a node in the network
- ❖ Routing is performed on the network part only.
- ❖ Dotted decimal notation is used to represent the IP address
- ❖ For example 130.63.95.218 What class?

Network Layer 4-41

## IPv4 Addressing old days

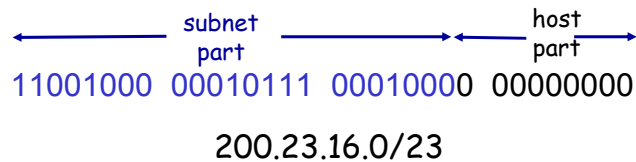
- ❖ A suffix of all zeros means network own address, so 132.187.0.0 means network 132.187, why?
- ❖ Suffix of all 1's means broadcast to this network.
- ❖ Computer own address (all 0's) when the computer does not know its own address (when starting and does not know its own address)
- ❖ Loopback address 127.0.0.1

Network Layer 4-42

## IP addressing: CIDR

### CIDR: Classless InterDomain Routing

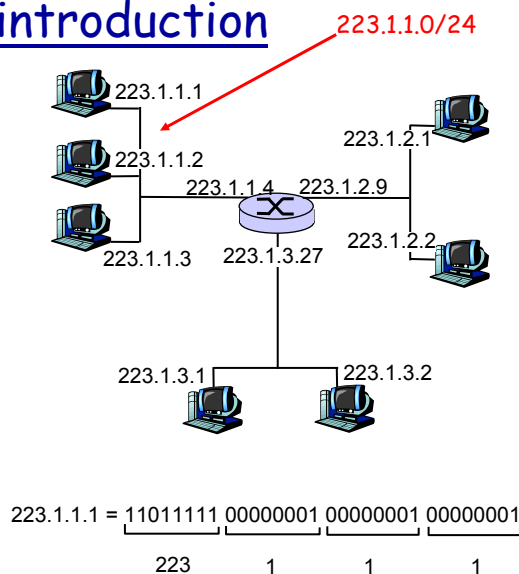
- subnet portion of address of arbitrary length
- address format: **a.b.c.d/x**, where x is # bits in subnet portion of address



Network Layer 4-43

## IP Addressing: introduction

- ❖ **IP address:** 32-bit identifier for host, router *interface*
- ❖ **interface:** connection between host/router and physical link
  - router's typically have multiple interfaces
  - host typically has one interface
  - IP addresses associated with each interface



Network Layer 4-44

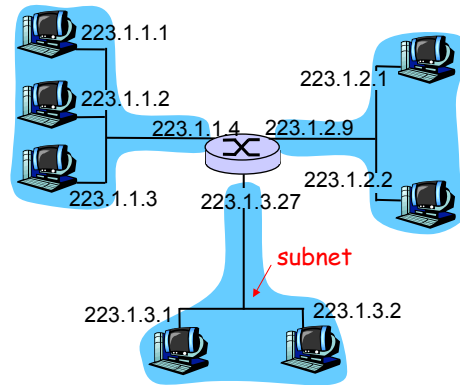
## Subnets

### ❖ IP address:

- subnet part (high order bits)
- host part (low order bits)

### ❖ *What's a subnet ?*

- device interfaces with same subnet part of IP address
- can physically reach each other without intervening router



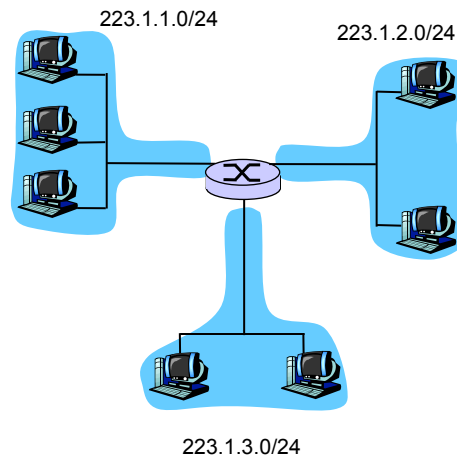
network consisting of 3 subnets

Network Layer 4-45

## Subnets

### Recipe

- ❖ to determine the subnets, detach each interface from its host or router, creating islands of isolated networks
- ❖ each isolated network is called a **subnet**.

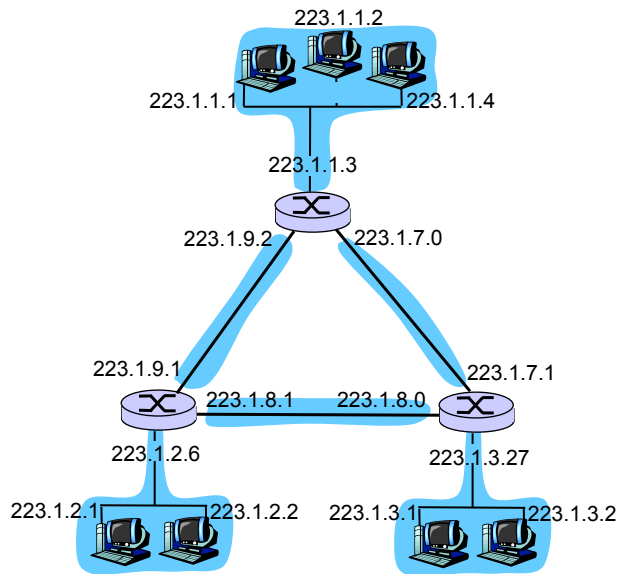


Subnet mask: /24

Network Layer 4-46

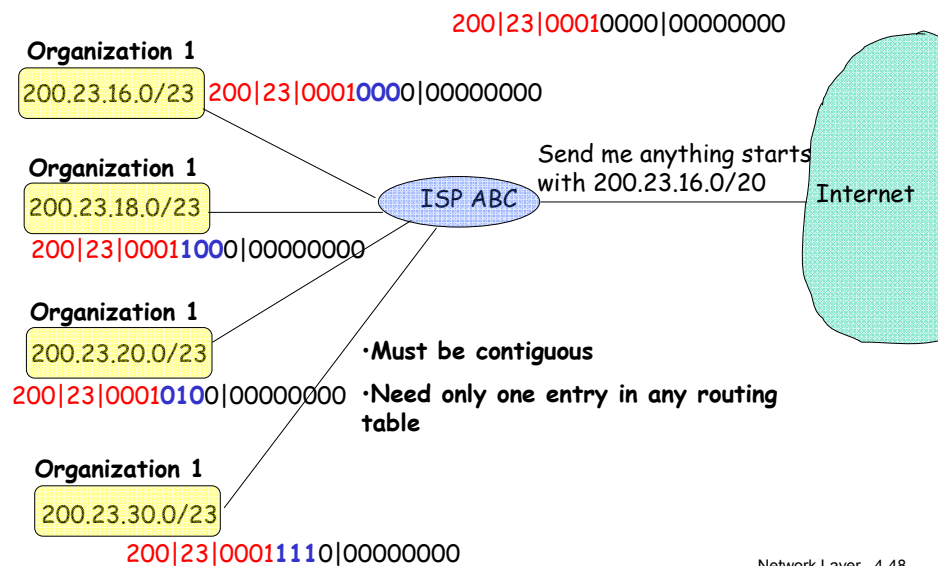
## Subnets

How many?



Network Layer 4-47

## Address Allocation



Network Layer 4-48



## IP addresses: how to get one?

Q: How does a *host* get IP address?

- ❖ Once the organization obtained a chunk of addresses, they can configure it anyway they want
- ❖ Hard-coded by system admin in a file
  - Windows: control-panel->network->configuration->tcp/ip->properties
  - UNIX: /etc/rc.config
- ❖ **DHCP: Dynamic Host Configuration Protocol:** dynamically get address from as server
  - "plug-and-play"

Network Layer 4-49

## DHCP: Dynamic Host Configuration Protocol

Goal: allow host to *dynamically* obtain its IP address from network server when it joins network

- Can renew its lease on address in use
- Especially mobile users come and go, not practical (or even possible) to hardwire (reserve) an IP address for each user.
- Allows reuse of addresses (only hold address while connected an "on")
- Support for mobile users who want to join network (more shortly)

Network Layer 4-50

## DHCP: Dynamic Host Configuration Protocol

DHCP overview [RFC2131]:

- DHCP is a client server protocol
- Ideally, each subnet has a server (or relays messages, by an HDCP agent or router, to the server).
- Host broadcasts "DHCP discover" msg [optional]
- DHCP server responds with "DHCP offer" msg [optional]
- Host requests IP address: "DHCP request" msg
- DHCP server sends address: "DHCP ack" msg

Network Layer 4-51

## DHCP details

- ❖ A host wants to join, sends a DHCP Discover message. It uses UDP, port 67. But to whom?
  - Sends to IP 255.255.255.255 (broadcast) and this host 0.0.0.0. as source address
- ❖ DHCP responds with DHCP offer msg containing the transaction ID, a proposed IP, network mask, and lease time. Then broadcasts it to 255.255.255.255. Client may receive more than one offer

cont.

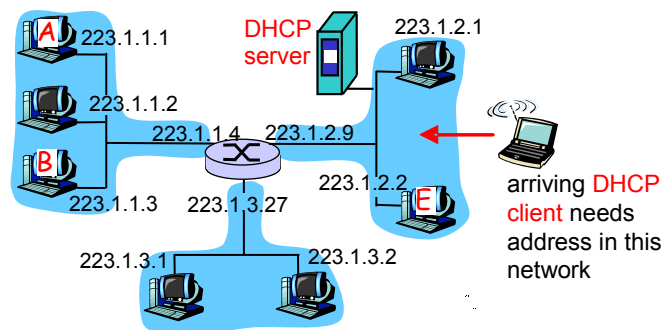
Network Layer 4-52

## DHCP details

- ❖ The client chooses one offer, and sends **DHCP request** msg.
- ❖ The server responds with **DHCP ACK** msg.
- ❖ Now the transaction is complete, and the client knows its IP, and network mask.

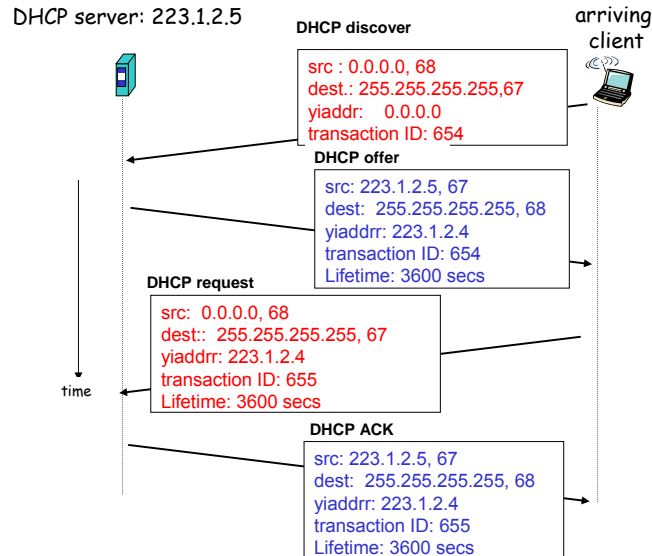
Network Layer 4-53

## DHCP client-server scenario



Network Layer 4-54

## DHCP client-server scenario



irk Layer 4-55

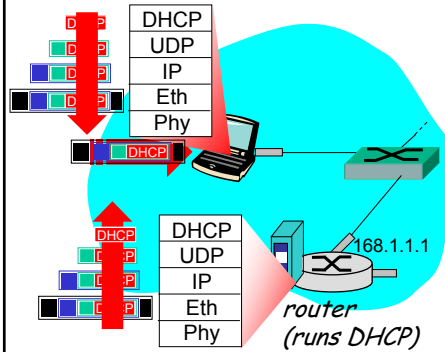
## DHCP: more than IP address

DHCP can return more than just allocated IP address on subnet:

- address of first-hop router for client
- name and IP address of DNS sever
- network mask (indicating network versus host portion of address)

Network Layer 4-56

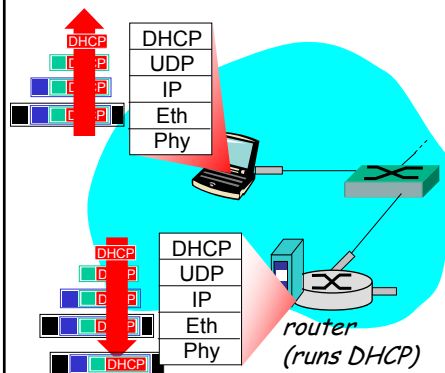
## DHCP: example



- ❖ connecting laptop needs its IP address, addr of first-hop router, addr of DNS server: use DHCP
- ❖ DHCP request encapsulated in UDP, encapsulated in IP, encapsulated in 802.1 Ethernet
- ❖ Ethernet frame broadcast (dest: FFFFFFFF) on LAN, received at router running DHCP server
- ❖ Ethernet demuxed to IP demuxed, UDP demuxed to DHCP

Network Layer 4-57

## DHCP: example



- ❖ DHCP server formulates DHCP ACK containing client's IP address, IP address of first-hop router for client, name & IP address of DNS server
- ❖ encapsulation of DHCP server, frame forwarded to client, demuxing up to DHCP at client
- ❖ client now knows its IP address, name and IP address of DNS server, IP address of its first-hop router

Network Layer 4-58

## DHCP: Wireshark output (home LAN)

Message type: **Boot Request (1)**  
 Hardware type: Ethernet  
 Hardware address length: 6  
 Hops: 0  
**Transaction ID: 0x6b3a11b7**  
 Seconds elapsed: 0  
 Bootp flags: 0x0000 (Unicast)  
 Client IP address: 0.0.0.0 (0.0.0.0)  
 Your (client) IP address: 0.0.0.0 (0.0.0.0)  
 Next server IP address: 0.0.0.0 (0.0.0.0)  
 Relay agent IP address: 0.0.0.0 (0.0.0.0)  
**Client MAC address: Wistron\_23:68:8a (00:16:d3:23:68:8a)**  
 Server host name not given  
 Boot file name not given  
 Magic cookie: (OK)  
**Option: (t=53,l=1) DHCP Message Type = DHCP Request**  
 Option: (61) Client identifier  
 Length: 7; Value: 010016D323688A;  
 Hardware type: Ethernet  
 Client MAC address: Wistron\_23:68:8a (00:16:d3:23:68:8a)  
 Option: (t=50,l=4) Requested IP Address = 192.168.1.101  
 Option: (t=12,l=5) Host Name = "nomad"  
**Option: (55) Parameter Request List**  
 Length: 11; Value: 010F03062C2E2F1F21F92B  
**1 = Subnet Mask; 15 = Domain Name**  
**3 = Router; 6 = Domain Name Server**  
 44 = NetBIOS over TCP/IP Name Server  
 .....

request

Message type: **Boot Reply (2)**  
 Hardware type: Ethernet  
 Hardware address length: 6  
 Hops: 0  
**Transaction ID: 0x6b3a11b7**  
 Seconds elapsed: 0  
 Bootp flags: 0x0000 (Unicast)  
**Client IP address: 192.168.1.101 (192.168.1.101)**  
 Your (client) IP address: 0.0.0.0 (0.0.0.0)  
**Next server IP address: 192.168.1.1 (192.168.1.1)**  
 Relay agent IP address: 0.0.0.0 (0.0.0.0)  
 Client MAC address: Wistron\_23:68:8a (00:16:d3:23:68:8a)  
 Server host name not given  
 Boot file name not given  
 Magic cookie: (OK)  
**Option: (t=53,l=1) DHCP Message Type = DHCP ACK**  
**Option: (t=54,l=4) Server Identifier = 192.168.1.1**  
**Option: (t=1,l=4) Subnet Mask = 255.255.255.0**  
**Option: (t=3,l=4) Router = 192.168.1.1**  
**Option: (6) Domain Name Server**  
 Length: 12; Value: 445747E2445749F244574092;  
 IP Address: 68.87.71.226;  
 IP Address: 68.87.73.242;  
 IP Address: 68.87.64.146  
**Option: (t=15,l=20) Domain Name = "hsd1.ma.comcast.net."**

reply

Network Layer 4-59

## IP addresses: how to get one?

**Q:** How does *network* get subnet part of IP addr?

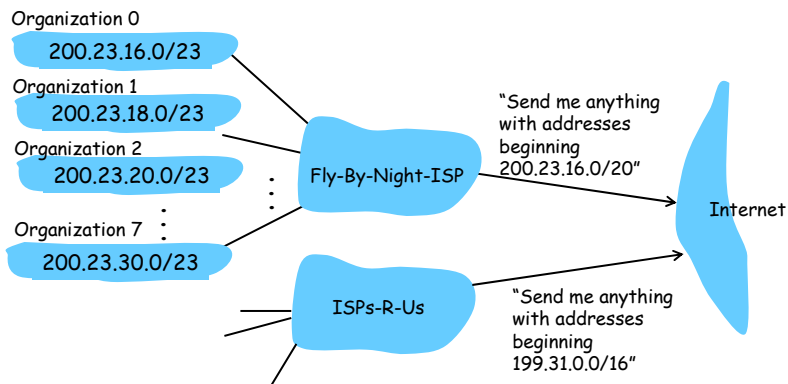
**A:** gets allocated portion of its provider ISP's address space

ISP's block	<u>11001000 00010111 00010000</u> 00000000	200.23.16.0/20
Organization 0	<u>11001000 00010111 00010000</u> 00000000	200.23.16.0/23
Organization 1	<u>11001000 00010111 00010010</u> 00000000	200.23.18.0/23
Organization 2	<u>11001000 00010111 00010100</u> 00000000	200.23.20.0/23
...	.....	.....
Organization 7	<u>11001000 00010111 00011110</u> 00000000	200.23.30.0/23

Network Layer 4-60

## Hierarchical addressing: route aggregation

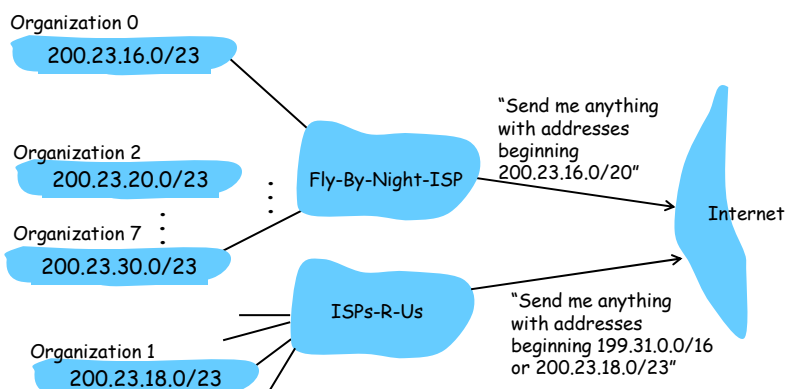
Hierarchical addressing allows efficient advertisement of routing information:



Network Layer 4-61

## Hierarchical addressing: more specific routes

ISPs-R-Us has a more specific route to Organization 1



Network Layer 4-62

## IP addressing: the last word...

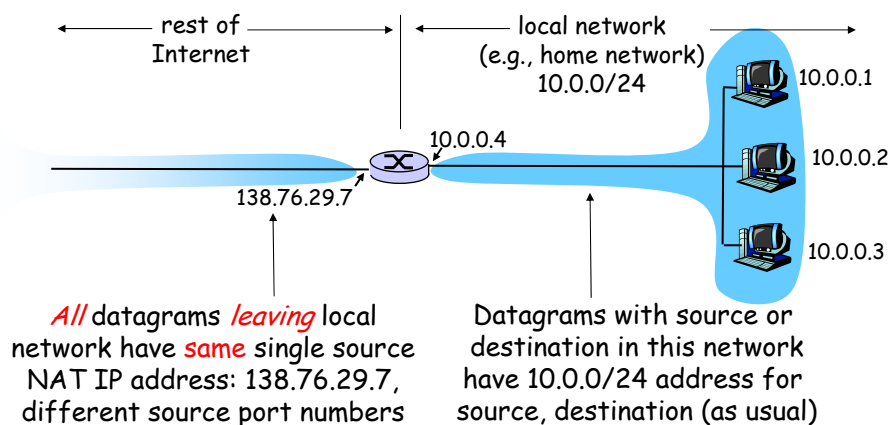
Q: How does an ISP get block of addresses?

A: **ICANN**: Internet Corporation for Assigned Names and Numbers

- allocates addresses
- manages DNS
- assigns domain names, resolves disputes

Network Layer 4-63

## NAT: Network Address Translation



Network Layer 4-64



## NAT: Network Address Translation

- ❖ **Motivation:** local network uses just one IP address as far as outside world is concerned:
  - range of addresses not needed from ISP: just one IP address for all devices
  - can change addresses of devices in local network without notifying outside world
  - can change ISP without changing addresses of devices in local network
  - devices inside local net not explicitly addressable, visible by outside world (a security plus).

Network Layer 4-65

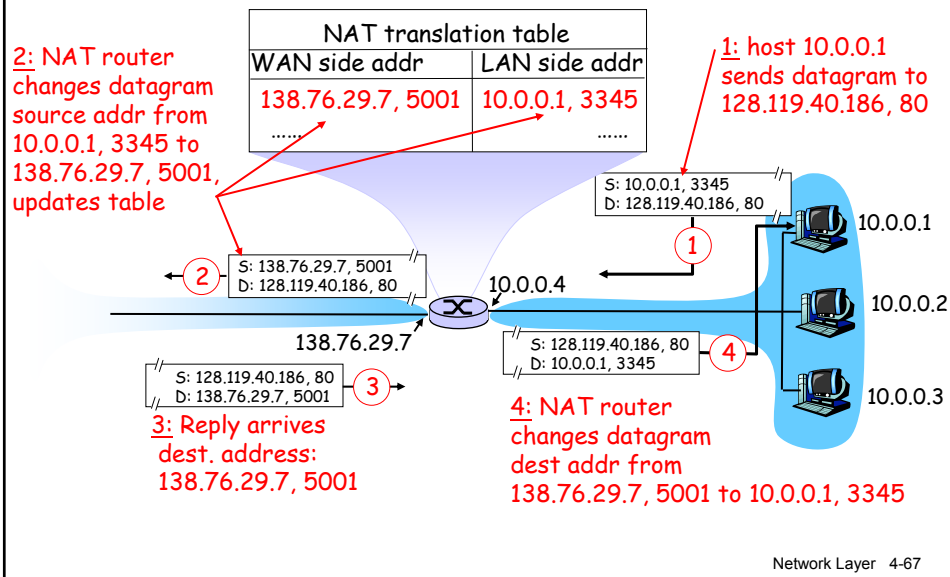
## NAT: Network Address Translation

**Implementation:** NAT router must:

- *outgoing datagrams: replace* (source IP address, port #) of every outgoing datagram to (NAT IP address, new port #)
  - ... remote clients/servers will respond using (NAT IP address, new port #) as destination addr.
- *remember (in NAT translation table)* every (source IP address, port #) to (NAT IP address, new port #) translation pair
- *incoming datagrams: replace* (NAT IP address, new port #) in dest fields of every incoming datagram with corresponding (source IP address, port #) stored in NAT table

Network Layer 4-66

## NAT: Network Address Translation



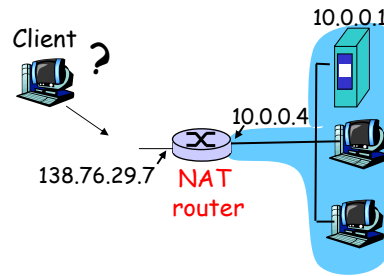
## NAT: Network Address Translation

- ❖ 16-bit port-number field:
  - 60,000 simultaneous connections with a single LAN-side address!
- ❖ NAT is controversial:
  - routers should only process up to layer 3
  - violates end-to-end argument
    - NAT possibility must be taken into account by app designers, e.g., P2P applications
  - address shortage should instead be solved by IPv6

Network Layer 4-68

## NAT traversal problem

- ❖ client wants to connect to server with address 10.0.0.1
  - server address 10.0.0.1 local to LAN (client can't use it as destination addr)
  - only one externally visible NATed address: 138.76.29.7
- ❖ solution 1: statically configure NAT to forward incoming connection requests at given port to server
  - e.g., (138.76.29.7, port 2500) always forwarded to 10.0.0.1 port 25000

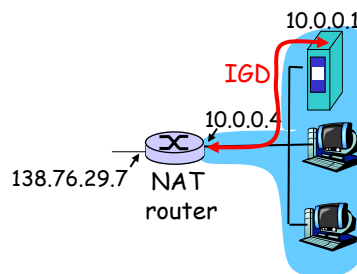


Network Layer 4-69

## NAT traversal problem

- ❖ solution 2: Universal Plug and Play (UPnP) Internet Gateway Device (IGD) Protocol. Allows NATed host to:
  - ❖ learn public IP address (138.76.29.7)
  - ❖ add/remove port mappings (with lease times)

i.e., automate static NAT port map configuration

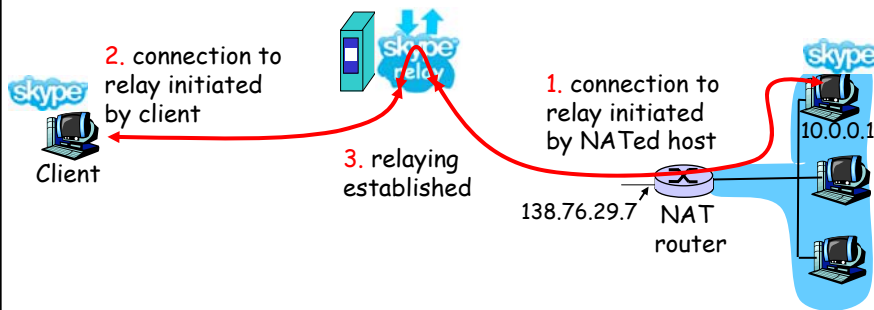


Network Layer 4-70

## NAT traversal problem

### ❖ solution 3: relaying (used in Skype)

- NATed client establishes connection to relay
- External client connects to relay
- relay bridges packets between to connections



Network Layer 4-71

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-72

## ICMP: Internet Control Message Protocol

- ❖ used by hosts & routers to communicate network-level information
  - error reporting: unreachable host, network, port, protocol
  - echo request/reply (used by ping)
- ❖ network-layer "above" IP:
  - ICMP msgs carried in IP datagrams
- ❖ **ICMP message:** type, code plus first 8 bytes of IP datagram causing error

Type	Code	description
0	0	echo reply (ping)
3	0	dest. network unreachable
3	1	dest host unreachable
3	2	dest protocol unreachable
3	3	dest port unreachable
3	6	dest network unknown
3	7	dest host unknown
4	0	source quench (congestion control - not used)
8	0	echo request (ping)
9	0	route advertisement
10	0	router discovery
11	0	TTL expired
12	0	bad IP header

Network Layer 4-73

## Traceroute and ICMP

- ❖ Source sends series of UDP segments to dest
  - first has TTL =1
  - second has TTL=2, etc.
  - unlikely port number
- ❖ When nth datagram arrives to nth router:
  - router discards datagram
  - and sends to source an ICMP message (type 11, code 0)
  - ICMP message includes name of router & IP address
- ❖ when ICMP message arrives, source calculates RTT
- ❖ traceroute does this 3 times
- Stopping criterion**
  - ❖ UDP segment eventually arrives at destination host
  - ❖ destination returns ICMP "port unreachable" packet (type 3, code 3)
  - ❖ when source gets this ICMP, stops.

Network Layer 4-74

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-75

## IPv6

- ❖ **Initial motivation:** 32-bit address space soon to be completely allocated.
  - ❖ **Additional motivation:**
    - header format helps speed processing/forwarding
    - header changes to facilitate QoS
- IPv6 datagram format:**
- fixed-length 40 byte header
  - no fragmentation allowed

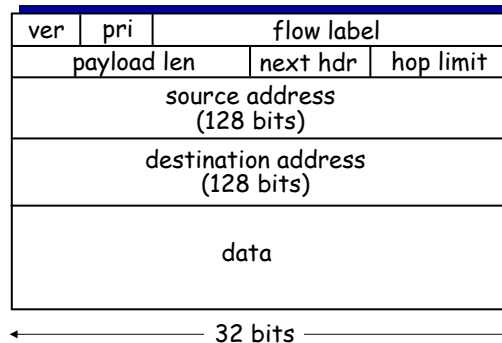
Network Layer 4-76

## IPv6 Header (Cont)

*Priority:* identify priority among datagrams in flow

*Flow Label:* identify datagrams in same "flow."  
(concept of "flow" not well defined).

*Next header:* identify upper layer protocol for data



Network Layer 4-77

## Other Changes from IPv4

- ❖ *Checksum:* removed entirely to reduce processing time at each hop
- ❖ *Options:* allowed, but outside of header, indicated by "Next Header" field
- ❖ *ICMPv6:* new version of ICMP
  - additional message types, e.g. "Packet Too Big"
  - multicast group management functions

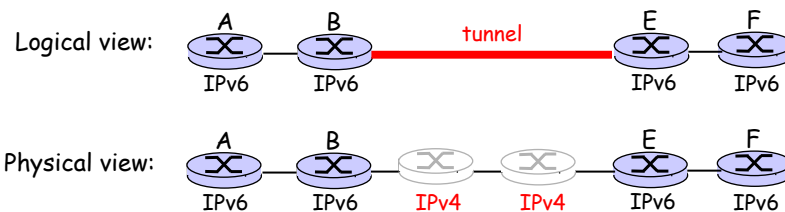
Network Layer 4-78

## Transition From IPv4 To IPv6

- ❖ Not all routers can be upgraded simultaneous
  - no “flag days”
  - How will the network operate with mixed IPv4 and IPv6 routers?
- ❖ *Tunneling*: IPv6 carried as payload in IPv4 datagram among IPv4 routers

Network Layer 4-79

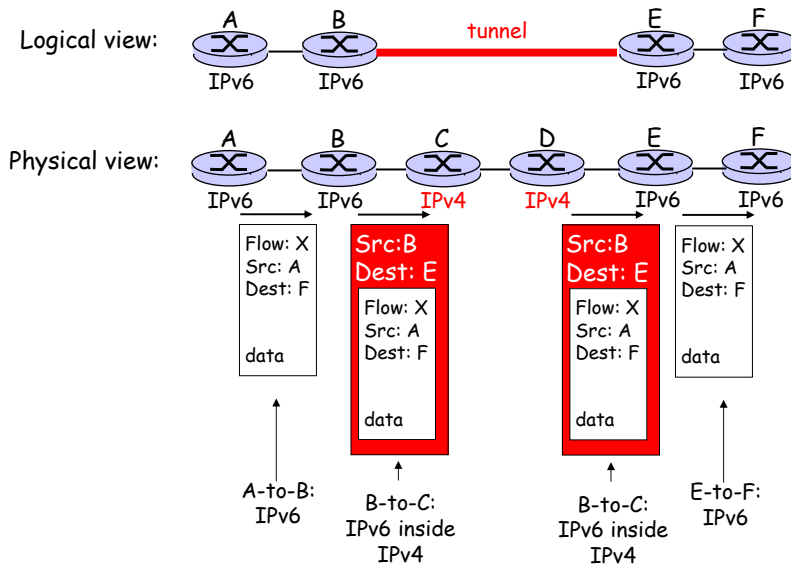
## Tunneling



Network Layer 4-80



## Tunneling



Network Layer 4-81

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-82

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

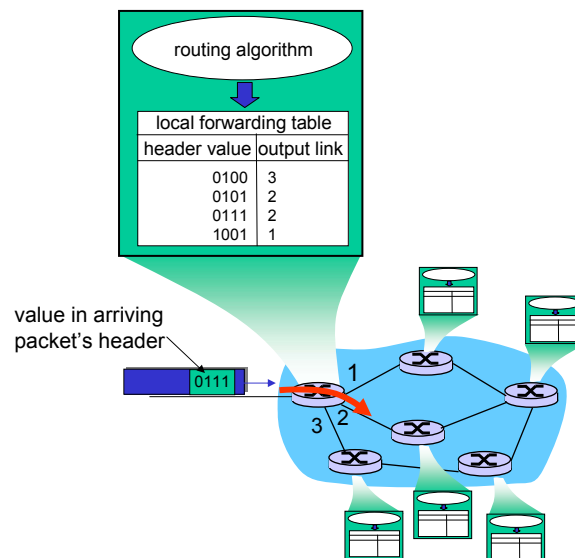
### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

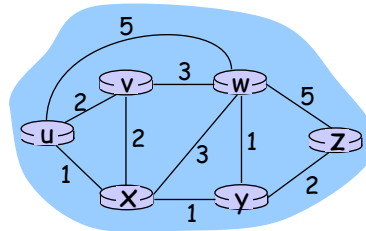
Network Layer 4-83

## Interplay between routing, forwarding



Network Layer 4-84

## Graph abstraction



Graph:  $G = (N, E)$

$N$  = set of routers =  $\{ u, v, w, x, y, z \}$

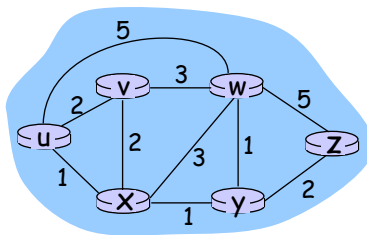
$E$  = set of links =  $\{ (u,v), (u,x), (v,x), (v,w), (x,w), (x,y), (w,y), (w,z), (y,z) \}$

Remark: Graph abstraction is useful in other network contexts

Example: P2P, where  $N$  is set of peers and  $E$  is set of TCP connections

Network Layer 4-85

## Graph abstraction: costs



•  $c(x, x')$  = cost of link  $(x, x')$

- e.g.,  $c(w, z) = 5$

• cost could always be 1, or  
inversely related to bandwidth,  
or inversely related to  
congestion

Cost of path  $(x_1, x_2, x_3, \dots, x_p) = c(x_1, x_2) + c(x_2, x_3) + \dots + c(x_{p-1}, x_p)$

Question: What's the least-cost path between  $u$  and  $z$  ?

Routing algorithm: algorithm that finds least-cost path

Network Layer 4-86

## Routing Algorithm classification

### Global or decentralized information?

#### Global:

- ❖ all routers have complete topology, link cost info

- ❖ "link state" algorithms

#### Decentralized:

- ❖ router knows physically-connected neighbors, link costs to neighbors
- ❖ iterative process of computation, exchange of info with neighbors
- ❖ "distance vector" algorithms

### Static or dynamic?

#### Static:

- ❖ routes change slowly over time

#### Dynamic:

- ❖ routes change more quickly
  - periodic update
  - in response to link cost changes

Network Layer 4-87

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-88

## A Link-State Routing Algorithm

### Dijkstra's algorithm

- ❖ net topology, link costs known to all nodes
  - accomplished via "link state broadcast"
  - all nodes have same info
- ❖ computes least cost paths from one node ("source") to all other nodes
  - gives *forwarding table* for that node
- ❖ iterative: after k iterations, know least cost path to k dest.'s

### Notation:

- ❖  $c(x,y)$ : link cost from node x to y;  $= \infty$  if not direct neighbors
- ❖  $D(v)$ : current value of cost of path from source to dest. v
- ❖  $p(v)$ : predecessor node along path from source to v
- ❖  $N'$ : set of nodes whose least cost path definitively known

Network Layer 4-89

## Dijkstra's Algorithm

### 1 Initialization:

```
2  N' = {u}
3  for all nodes v
4    if v adjacent to u
5      then D(v) = c(u,v)
6    else D(v) = ∞
7
```

### 8 Loop

```
9  find w not in N' such that D(w) is a minimum
10 add w to N'
11 update D(v) for all v adjacent to w and not in N' :
12    $D(v) = \min( D(v), D(w) + c(w,v) )$ 
13   /* new cost to v is either old cost to v or known
14      shortest path cost to w plus cost from w to v */
15 until all nodes in N'
```

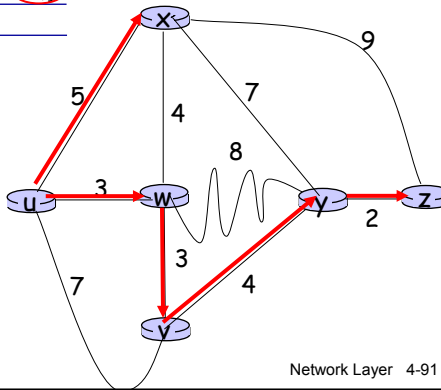
Network Layer 4-90

## Dijkstra's algorithm: example

Step	N'	D(v)	D(w)	D(x)	D(y)	D(z)
		p(v)	p(w)	p(x)	p(y)	p(z)
0	u	7,u	3,u	5,u	$\infty$	$\infty$
1	uw	6,w		5,u	11,w	$\infty$
2	uwx	6,w			11,w	14,x
3	uwxv				10,v	14,x
4	uwxvy					12,y
5	uwxvyz					

### Notes:

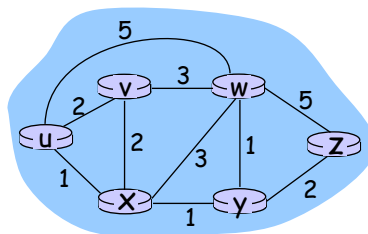
- ❖ construct shortest path tree by tracing predecessor nodes
- ❖ ties can exist (can be broken arbitrarily)



Network Layer 4-91

## Dijkstra's algorithm: another example

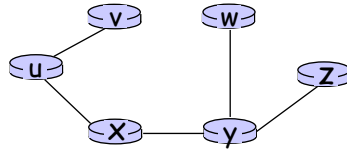
Step	N'	D(v),p(v)	D(w),p(w)	D(x),p(x)	D(y),p(y)	D(z),p(z)
0	u	2,u	5,u	1,u	$\infty$	$\infty$
1	ux	2,u	4,x		2,x	$\infty$
2	uxy	2,u	3,y			4,y
3	uxyv		3,y			4,y
4	uxyvw					4,y
5	uxyvwz					



Network Layer 4-92

## Dijkstra's algorithm: example (2)

Resulting shortest-path tree from u:



Resulting forwarding table in u:

destination	link
v	(u,v)
x	(u,x)
y	(u,x)
w	(u,x)
z	(u,x)

Network Layer 4-93

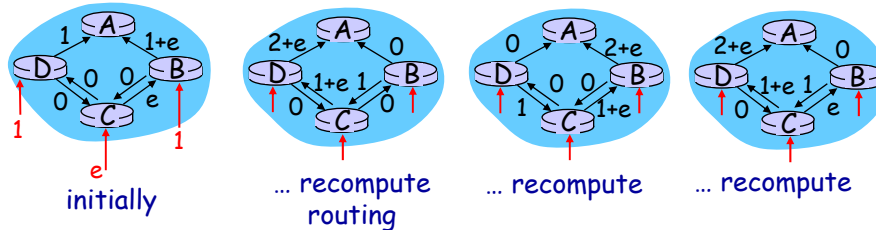
## Dijkstra's algorithm, discussion

**Algorithm complexity:** n nodes

- ❖ each iteration: need to check all nodes, w, not in N
- ❖  $n(n+1)/2$  comparisons:  $O(n^2)$
- ❖ more efficient implementations possible:  $O(n \log n)$

**Oscillations possible:**

- ❖ e.g., link cost = amount of carried traffic



Network Layer 4-94

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- **Distance Vector**
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-95

## Distance Vector Algorithm

### Bellman-Ford Equation (dynamic programming)

Define

$d_x(y) :=$  cost of least-cost path from  $x$  to  $y$

Then

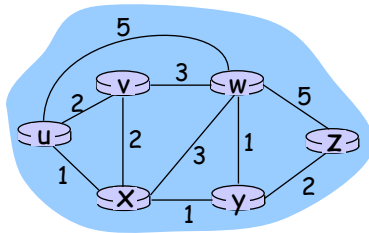
$$d_x(y) = \min_v \{c(x,v) + d_v(y)\}$$

where min is taken over all neighbors  $v$  of  $x$

Network Layer 4-96



## Bellman-Ford example



Clearly,  $d_v(z) = 5$ ,  $d_x(z) = 3$ ,  $d_w(z) = 3$

B-F equation says:

$$\begin{aligned}
 d_u(z) &= \min \{ c(u,v) + d_v(z), \\
 &\quad c(u,x) + d_x(z), \\
 &\quad c(u,w) + d_w(z) \} \\
 &= \min \{ 2 + 5, \\
 &\quad 1 + 3, \\
 &\quad 5 + 3 \} = 4
 \end{aligned}$$

Node that achieves minimum is next hop in shortest path forwarding table

Network Layer 4-97

## Distance Vector Algorithm

- ❖  $D_x(y)$  = estimate of least cost from x to y
  - x maintains distance vector  $D_x = [D_x(y): y \in N]$
- ❖ node x:
  - knows cost to each neighbor v:  $c(x,v)$
  - maintains its neighbors' distance vectors. For each neighbor v, x maintains  $D_v = [D_v(y): y \in N]$

Network Layer 4-98

## Distance vector algorithm (4)

### Basic idea:

- ❖ from time-to-time, each node sends its own distance vector estimate to neighbors
- ❖ when  $x$  receives new DV estimate from neighbor, it updates its own DV using B-F equation:

$$D_x(y) \leftarrow \min_v \{c(x,v) + D_v(y)\} \quad \text{for each node } y \in N$$

- ❖ under minor, natural conditions, the estimate  $D_x(y)$  converge to the actual least cost  $d_x(y)$

Network Layer 4-99

## Distance Vector Algorithm (5)

### Iterative, asynchronous:

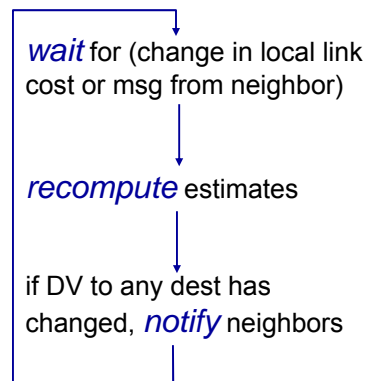
each local iteration caused by:

- ❖ local link cost change
- ❖ DV update message from neighbor

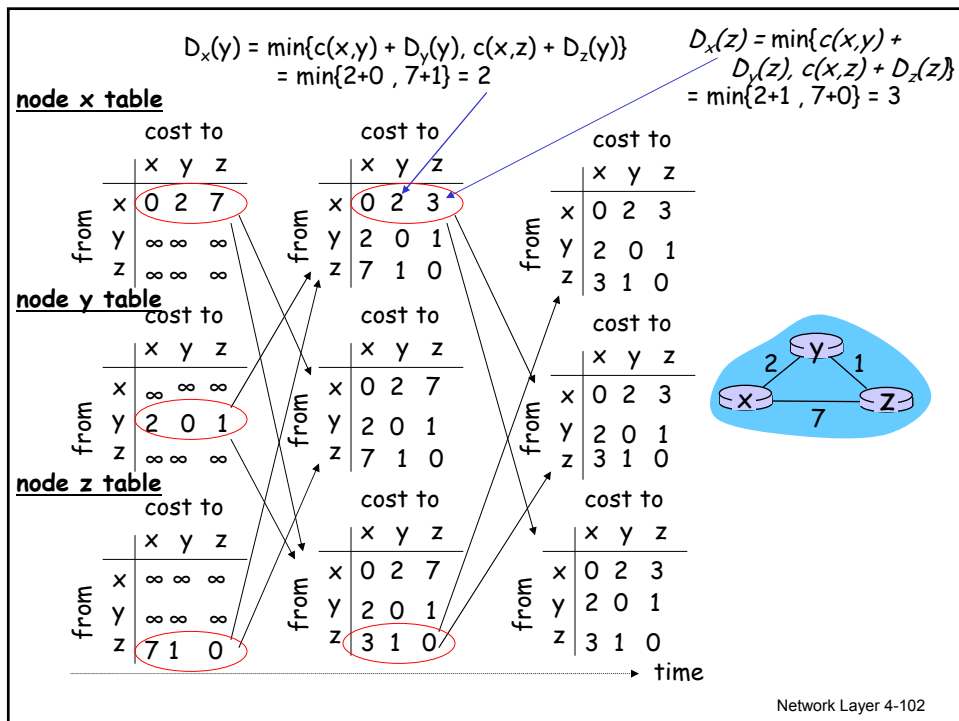
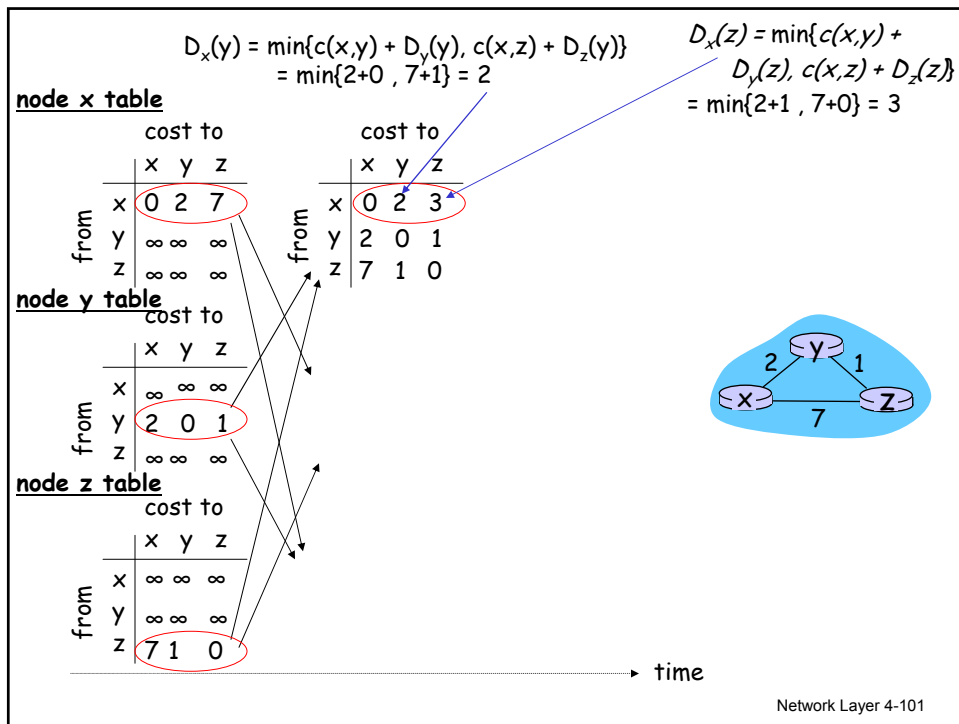
### Distributed:

- ❖ each node notifies neighbors *only* when its DV changes
  - neighbors then notify their neighbors if necessary

### Each node:



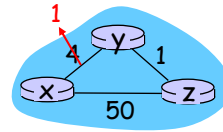
Network Layer 4-100



## Distance Vector: link cost changes

### Link cost changes:

- ❖ node detects local link cost change
- ❖ updates routing info, recalculates distance vector
- ❖ if DV changes, notify neighbors



"good  
news  
travels  
fast"

$t_0$ :  $y$  detects link-cost change, updates its DV, informs its neighbors.

$t_1$ :  $z$  receives update from  $y$ , updates its table, computes new least cost to  $x$ , sends its neighbors its DV.

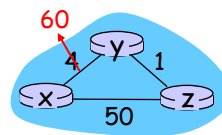
$t_2$ :  $y$  receives  $z$ 's update, updates its distance table.  $y$ 's least costs do *not* change, so  $y$  does *not* send a message to  $z$ .

Network Layer 4-103

## Distance Vector: link cost changes

### Link cost changes:

- ❖ good news travels fast
- ❖ bad news travels slow - "count to infinity" problem!
- ❖ 44 iterations before algorithm stabilizes: see text



### Poisoned reverse:

- ❖ If  $Z$  routes through  $Y$  to get to  $X$ :
  - $Z$  tells  $Y$  its ( $Z$ 's) distance to  $X$  is infinite (so  $Y$  won't route to  $X$  via  $Z$ )
- ❖ will this completely solve count to infinity problem?

Network Layer 4-104

## Comparison of LS and DV algorithms

### Message complexity

- ❖ **LS:** with  $n$  nodes,  $E$  links,  $O(nE)$  msgs sent
- ❖ **DV:** exchange between neighbors only
  - convergence time varies

### Speed of Convergence

- ❖ **LS:**  $O(n^2)$  algorithm requires  $O(nE)$  msgs
  - may have oscillations
- ❖ **DV:** convergence time varies
  - may be routing loops
  - count-to-infinity problem

### Robustness: what happens if router malfunctions?

#### **LS:**

- node can advertise incorrect *link* cost
- each node computes only its *own* table

#### **DV:**

- DV node can advertise incorrect *path* cost
- each node's table used by others
  - error propagate thru network

Network Layer 4-105

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-106

## Hierarchical Routing

Our routing study thus far - idealization

- ❖ all routers identical
- ❖ network "flat"
- ... *not* true in practice

**scale:** with 200 million destinations:

- ❖ can't store all dest's in routing tables!
- ❖ routing table exchange would swamp links!

**administrative autonomy**

- ❖ internet = network of networks
- ❖ each network admin may want to control routing in its own network

Network Layer 4-107

## Hierarchical Routing

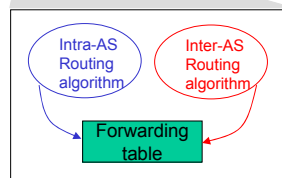
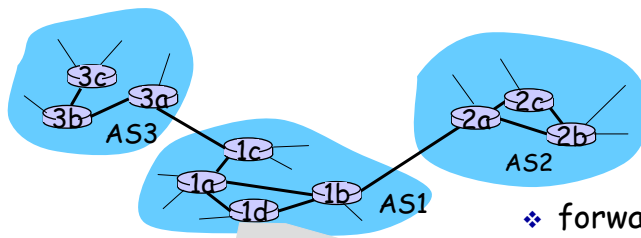
- ❖ aggregate routers into regions, "**autonomous systems**" (AS)
- ❖ routers in same AS run same routing protocol
  - "**intra-AS**" routing protocol
  - routers in different AS can run different intra-AS routing protocol

**gateway router**

- ❖ at "edge" of its own AS
- ❖ has link to router in another AS

Network Layer 4-108

## Interconnected ASes



- ❖ forwarding table configured by both intra- and inter-AS routing algorithm
  - intra-AS sets entries for internal dests
  - inter-AS & intra-AS sets entries for external dests

Network Layer 4-109

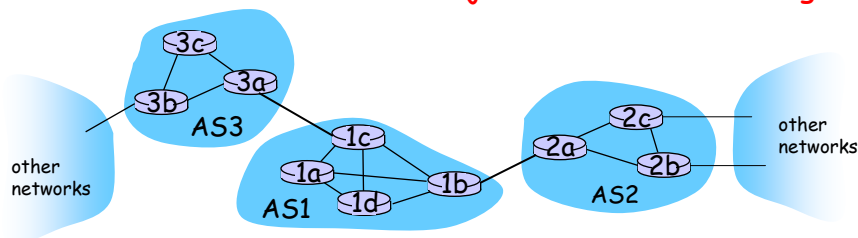
## Inter-AS tasks

- ❖ suppose router in AS1 receives datagram destined outside of AS1:
  - router should forward packet to gateway router, but which one?

### AS1 must:

1. learn which dests are reachable through AS2, which through AS3
2. propagate this reachability info to all routers in AS1

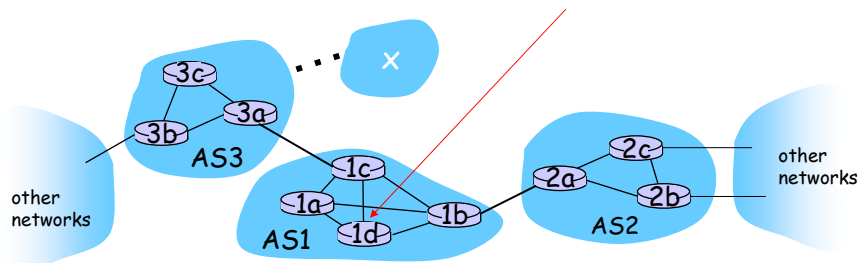
**job of inter-AS routing!**



Network Layer 4-110

### Example: Setting forwarding table in router 1d

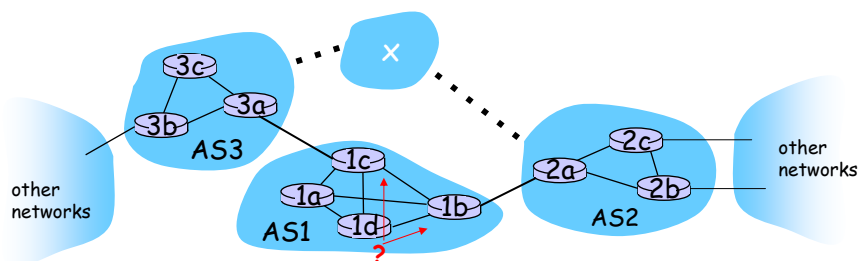
- ❖ suppose AS1 learns (via inter-AS protocol) that subnet  $x$  reachable via AS3 (gateway 1c) but not via AS2.
  - inter-AS protocol propagates reachability info to all internal routers
- ❖ router 1d determines from intra-AS routing info that its interface  $I$  is on the least cost path to 1c.
  - installs forwarding table entry  $(x, I)$



Network Layer 4-111

### Example: Choosing among multiple ASes

- ❖ now suppose AS1 learns from inter-AS protocol that subnet  $x$  is reachable from AS3 *and* from AS2.
- ❖ to configure forwarding table, router 1d must determine which gateway it should forward packets towards for dest  $x$ 
  - this is also job of inter-AS routing protocol!

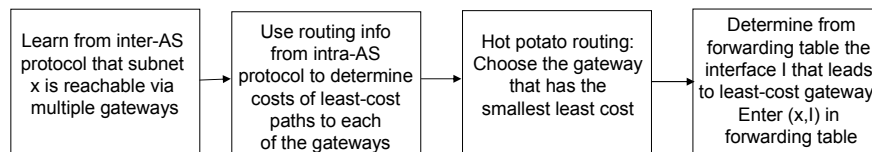


Network Layer 4-112



## Example: Choosing among multiple ASes

- ❖ now suppose AS1 learns from inter-AS protocol that subnet **x** is reachable from AS3 *and* from AS2.
- ❖ to configure forwarding table, router 1d must determine towards which gateway it should forward packets for dest **x**.
  - this is also job of inter-AS routing protocol!
- ❖ **hot potato routing**: send packet towards closest of two routers.



Network Layer 4-113

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-114

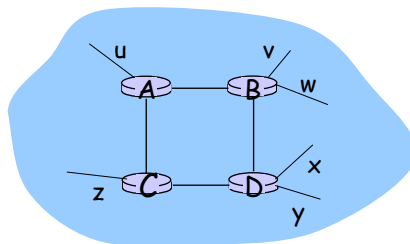
## Intra-AS Routing

- ❖ also known as **Interior Gateway Protocols (IGP)**
- ❖ most common Intra-AS routing protocols:
  - RIP: Routing Information Protocol
  - OSPF: Open Shortest Path First
  - IGRP: Interior Gateway Routing Protocol (Cisco proprietary)

Network Layer 4-115

## RIP (Routing Information Protocol)

- ❖ included in BSD-UNIX distribution in 1982
- ❖ distance vector algorithm
  - distance metric: # hops (max = 15 hops), each link has cost 1
  - DVs exchanged with neighbors every 30 sec in response message (aka **advertisement**)
  - each advertisement: list of up to 25 destination **subnets** (in IP addressing sense)

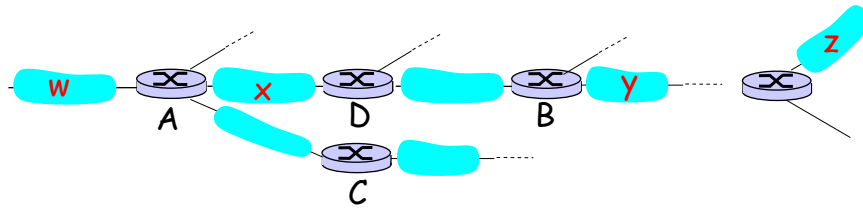


from router A to destination **subnets**:

subnet	hops
u	1
v	2
w	2
x	3
y	3
z	2

Network Layer 4-116

## RIP: Example



routing table in router D

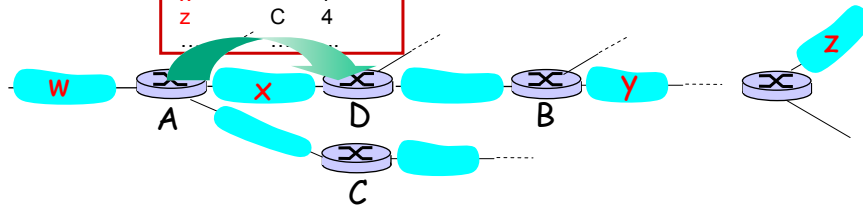
destination subnet	next router	# hops to dest
W	A	2
Y	B	2
X	--	1
....	....	....

Network Layer 4-117

## RIP: Example

A-to-D advertisement

dest	next	hops
W	-	1
X	-	1
Z	C	4



routing table in router D

destination subnet	next router	# hops to dest
W	A	2
Y	B	2
Z	<del>B</del> A	<del>7</del> 5
X	--	1
....	....	....

Network Layer 4-118

## RIP: Link Failure and Recovery

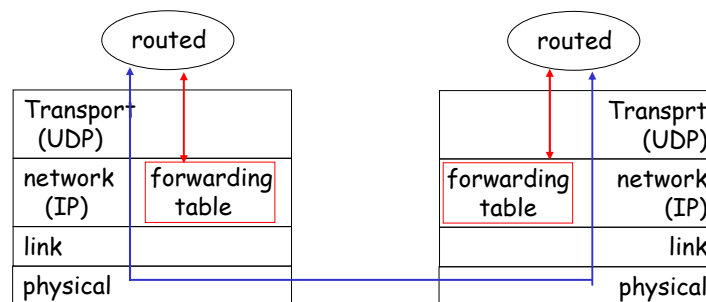
If no advertisement heard after 180 sec -->  
neighbor/link declared dead

- routes via neighbor invalidated
- new advertisements sent to neighbors
- neighbors in turn send out new advertisements (if tables changed)
- link failure info quickly (?) propagates to entire net
- *poison reverse* used to prevent ping-pong loops (infinite distance = 16 hops)

Network Layer 4-119

## RIP Table processing

- ❖ RIP routing tables managed by **application-level** process called route-d (daemon)
- ❖ advertisements sent in UDP packets, periodically repeated



Network Layer 4-120

## OSPF (Open Shortest Path First)

- ❖ "open": publicly available
- ❖ uses Link State algorithm
  - LS packet dissemination
  - topology map at each node
  - route computation using Dijkstra's algorithm
- ❖ OSPF advertisement carries one entry per neighbor router
- ❖ advertisements disseminated to **entire** AS (via flooding)
  - carried in OSPF messages directly over IP (rather than TCP or UDP)

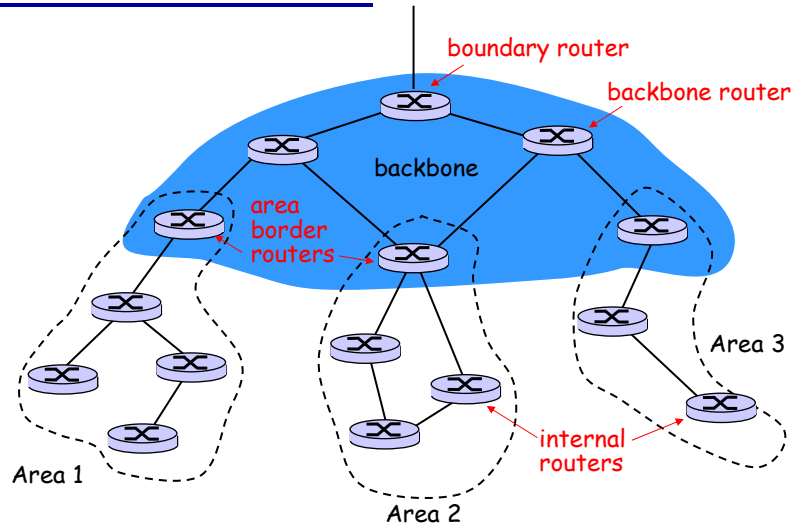
Network Layer 4-121

## OSPF "advanced" features (not in RIP)

- ❖ **security**: all OSPF messages authenticated (to prevent malicious intrusion)
- ❖ **multiple** same-cost **paths** allowed (only one path in RIP)
- ❖ for each link, multiple cost metrics for different **TOS** (e.g., satellite link cost set "low" for best effort ToS; high for real time ToS)
- ❖ integrated uni- and **multicast** support:
  - Multicast OSPF (MOSPF) uses same topology data base as OSPF
- ❖ **hierarchical** OSPF in large domains.

Network Layer 4-122

## Hierarchical OSPF



Network Layer 4-123

## Hierarchical OSPF

- ❖ **two-level hierarchy:** local area, backbone.
  - link-state advertisements only in area
  - each nodes has detailed area topology; only know direction (shortest path) to nets in other areas.
- ❖ **area border routers:** "summarize" distances to nets in own area, advertise to other Area Border routers.
- ❖ **backbone routers:** run OSPF routing limited to backbone.
- ❖ **boundary routers:** connect to other AS's.

Network Layer 4-124

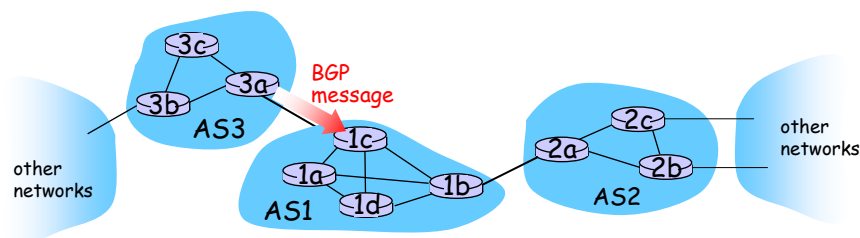
## Internet inter-AS routing: BGP

- ❖ **BGP (Border Gateway Protocol):** *the de facto* inter-domain routing protocol
  - "glue that holds the Internet together"
- ❖ BGP provides each AS a means to:
  - **eBGP:** obtain subnet reachability information from neighboring ASs.
  - **iBGP:** propagate reachability information to all AS-internal routers.
  - determine "good" routes to other networks based on reachability information and policy.
- ❖ allows subnet to advertise its existence to rest of Internet: *"I am here"*

Network Layer 4-125

## BGP basics

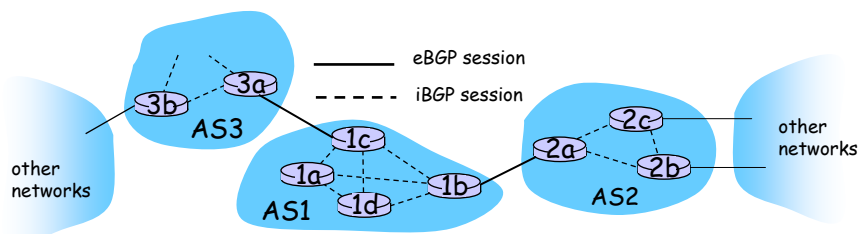
- ❖ **BGP session:** two BGP routers ("peers") exchange BGP messages:
  - advertising *paths* to different destination network prefixes ("path vector" protocol)
  - exchanged over semi-permanent TCP connections
- ❖ when AS3 advertises a prefix to AS1:
  - AS3 *promises* it will forward datagrams towards that prefix
  - AS3 can aggregate prefixes in its advertisement



Network Layer 4-126

## BGP basics: distributing path information

- ❖ using eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.
  - 1c can then use iBGP to distribute new prefix info to all routers in AS1
  - 1b can then re-advertise new reachability info to AS2 over 1b-to-2a eBGP session
- ❖ when router learns of new prefix, it creates entry for prefix in its forwarding table.



Network Layer 4-127

## Path attributes & BGP routes

- ❖ advertised prefix includes BGP attributes
  - prefix + attributes = "route"
- ❖ two important attributes:
  - **AS-PATH**: contains ASs through which prefix advertisement has passed: e.g., AS 67, AS 17
  - **NEXT-HOP**: indicates specific internal-AS router to next-hop AS. (may be multiple links from current AS to next-hop AS)
- ❖ gateway router receiving route advertisement uses **import policy** to accept/decline
  - e.g., never route through AS x
  - **policy-based** routing

Network Layer 4-128



## BGP route selection

- ❖ router may learn about more than 1 route to destination AS, selects route based on:
  1. local preference value attribute: policy decision
  2. shortest AS-PATH
  3. closest NEXT-HOP router: hot potato routing
  4. additional criteria

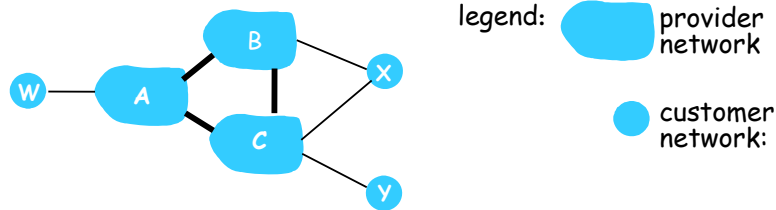
Network Layer 4-129

## BGP messages

- ❖ BGP messages exchanged between peers over TCP connection
- ❖ BGP messages:
  - **OPEN**: opens TCP connection to peer and authenticates sender
  - **UPDATE**: advertises new path (or withdraws old)
  - **KEEPALIVE**: keeps connection alive in absence of UPDATES; also ACKs OPEN request
  - **NOTIFICATION**: reports errors in previous msg; also used to close connection

Network Layer 4-130

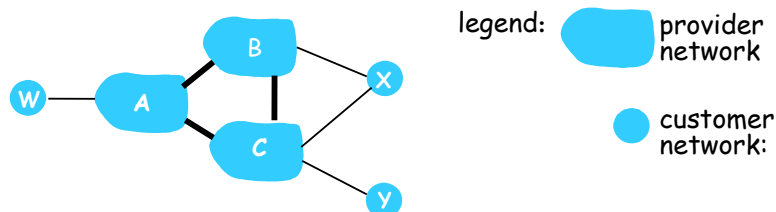
## BGP routing policy



- ❖ A,B,C are **provider networks**
- ❖ X,W,Y are customer (of provider networks)
- ❖ X is **dual-homed**: attached to two networks
  - X does not want to route from B via X to C
  - .. so X will not advertise to B a route to C

Network Layer 4-131

## BGP routing policy (2)



- ❖ A advertises path AW to B
- ❖ B advertises path BAW to X
- ❖ Should B advertise path BAW to C?
  - No way! B gets no "revenue" for routing CBAW since neither W nor C are B's customers
  - B wants to force C to route to w via A
  - B wants to route **only** to/from its customers!

Network Layer 4-132

## Why different Intra- and Inter-AS routing ?

### Policy:

- ❖ Inter-AS: admin wants control over how its traffic routed, who routes through its net.
- ❖ Intra-AS: single admin, so no policy decisions needed

### Scale:

- ❖ hierarchical routing saves table size, reduced update traffic

### Performance:

- ❖ Intra-AS: can focus on performance
- ❖ Inter-AS: policy may dominate over performance

Network Layer 4-133

## Chapter 4: Network Layer

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

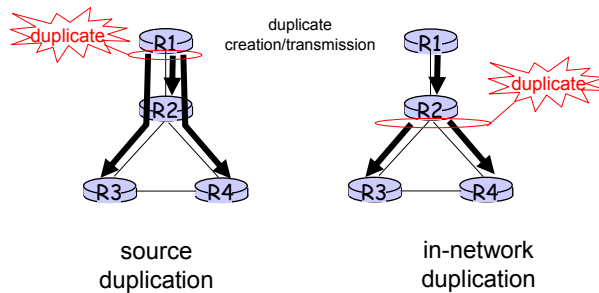
- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing

Network Layer 4-134

## Broadcast Routing

- ❖ deliver packets from source to all other nodes
- ❖ source duplication is inefficient:



- ❖ source duplication: how does source determine recipient addresses?

Network Layer 4-135

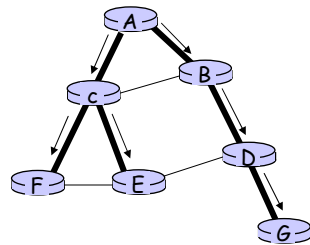
## In-network duplication

- ❖ flooding: when node receives broadcast packet, sends copy to all neighbors
  - problems: cycles & broadcast storm
- ❖ controlled flooding: node only broadcasts pkt if it hasn't broadcast same packet before
  - node keeps track of packet ids already broadcasted
  - or reverse path forwarding (RPF): only forward packet if it arrived on shortest path between node and source
- ❖ spanning tree
  - No redundant packets received by any node

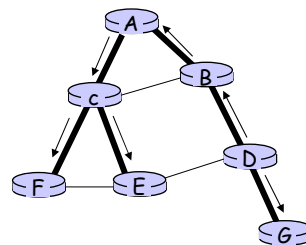
Network Layer 4-136

## Spanning Tree

- ❖ First construct a spanning tree
- ❖ Nodes forward copies only along spanning tree



(a) Broadcast initiated at A

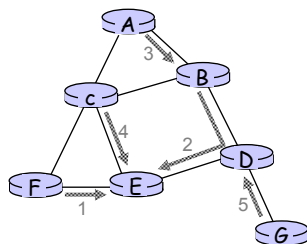


(b) Broadcast initiated at D

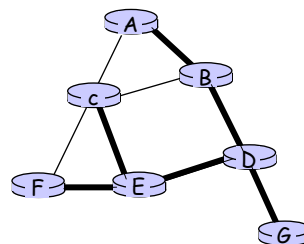
Network Layer 4-137

## Spanning Tree: Creation

- ❖ center node
- ❖ each node sends unicast join message to center node
  - message forwarded until it arrives at a node already belonging to spanning tree



(a) Stepwise construction of spanning tree

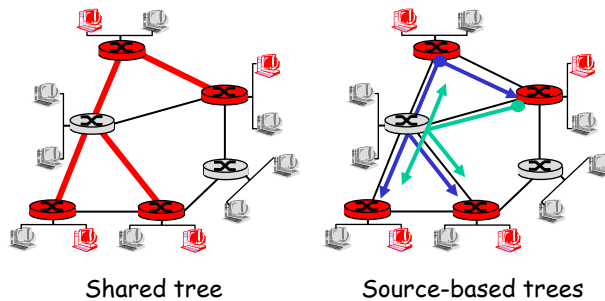


(b) Constructed spanning tree

Network Layer 4-138

## Multicast Routing: Problem Statement

- ❖ **Goal:** find a tree (or trees) connecting routers having local mcast group members
  - tree: not all paths between routers used
  - source-based: different tree from each sender to rcvrs
  - shared-tree: same tree used by all group members



## Approaches for building mcast trees

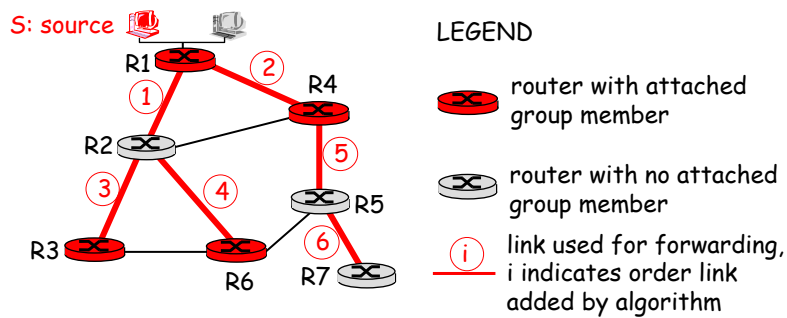
Approaches:

- ❖ **source-based tree:** one tree per source
  - shortest path trees
  - reverse path forwarding
- ❖ **group-shared tree:** group uses one tree
  - minimal spanning (Steiner)
  - center-based trees

...we first look at basic approaches, then specific protocols adopting these approaches

## Shortest Path Tree

- ❖ mcast forwarding tree: tree of shortest path routes from source to all receivers
  - Dijkstra's algorithm

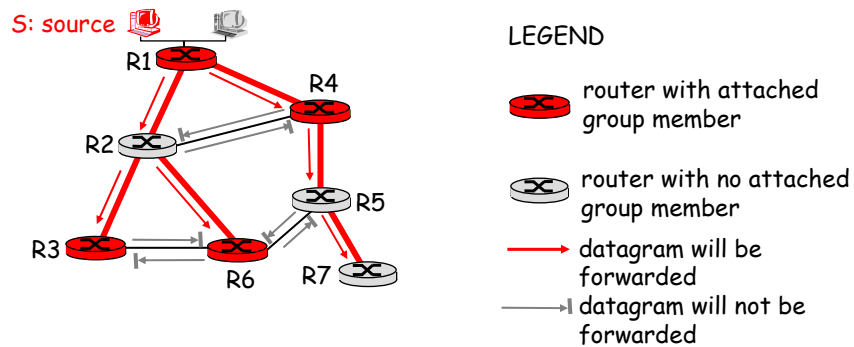


## Reverse Path Forwarding

- ❖ rely on router's knowledge of unicast shortest path from it to sender
- ❖ each router has simple forwarding behavior:

*if* (mcast datagram received on incoming link on shortest path back to center)  
*then* flood datagram onto all outgoing links  
*else* ignore datagram

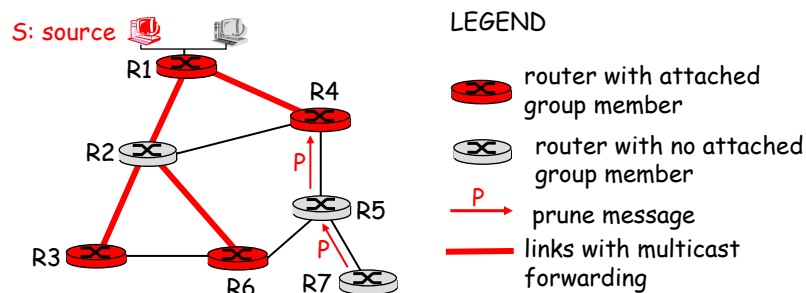
## Reverse Path Forwarding: example



- ❖ result is a source-specific *reverse* SPT
  - may be a bad choice with asymmetric links

## Reverse Path Forwarding: pruning

- ❖ forwarding tree contains subtrees with no mcast group members
  - no need to forward datagrams down subtree
  - "prune" msgs sent upstream by router with no downstream group members





## Shared-Tree: Steiner Tree

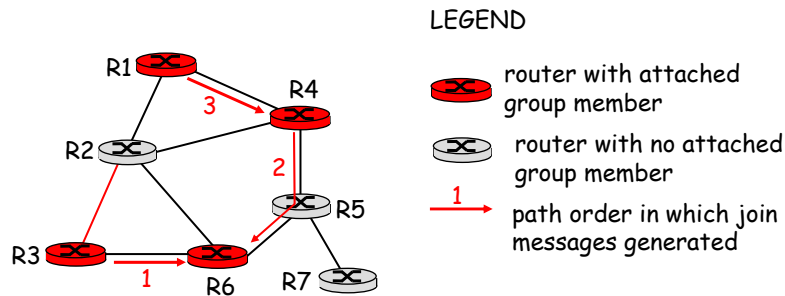
- ❖ **Steiner Tree:** minimum cost tree connecting all routers with attached group members
- ❖ problem is NP-complete
- ❖ excellent heuristics exists
- ❖ not used in practice:
  - computational complexity
  - information about entire network needed
  - monolithic: rerun whenever a router needs to join/leave

## Center-based trees

- ❖ single delivery tree shared by all
- ❖ one router identified as "**center**" of tree
- ❖ to join:
  - edge router sends unicast *join-msg* addressed to center router
  - *join-msg* "processed" by intermediate routers and forwarded towards center
  - *join-msg* either hits existing tree branch for this center, or arrives at center
  - path taken by *join-msg* becomes new branch of tree for this router

## Center-based trees: an example

Suppose R6 chosen as center:



## Internet Multicasting Routing: DVMRP

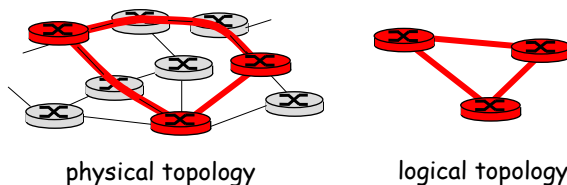
- ❖ **DVMRP**: distance vector multicast routing protocol, RFC1075
- ❖ **flood and prune**: reverse path forwarding, source-based tree
  - RPF tree based on DVMRP's own routing tables constructed by communicating DVMRP routers
  - no assumptions about underlying unicast
  - initial datagram to mcast group flooded everywhere via RPF
  - routers not wanting group: send upstream prune msgs

## DVMRP: continued...

- ❖ soft state: DVMRP router periodically (1 min.) "forgets" branches are pruned:
  - mcast data again flows down unpruned branch
  - downstream router: reprune or else continue to receive data
- ❖ routers can quickly regraft to tree
  - following IGMP join at leaf
- ❖ odds and ends
  - commonly implemented in commercial routers
  - Mbone routing done using DVMRP

## Tunneling

Q: How to connect "islands" of multicast routers in a "sea" of unicast routers?



- ❖ mcast datagram encapsulated inside "normal" (non-multicast-addressed) datagram
- ❖ normal IP datagram sent thru "tunnel" via regular IP unicast to receiving mcast router
- ❖ receiving mcast router unencapsulates to get mcast datagram

## PIM: Protocol Independent Multicast

- ❖ not dependent on any specific underlying unicast routing algorithm (works with all)
- ❖ two different multicast distribution scenarios :

### Dense:

- ❖ group members densely packed, in "close" proximity.
- ❖ bandwidth more plentiful

### Sparse:

- ❖ # networks with group members small wrt # interconnected networks
- ❖ group members "widely dispersed"
- ❖ bandwidth not plentiful

## Consequences of Sparse-Dense Dichotomy:

### Dense

- ❖ group membership by routers *assumed* until routers explicitly prune
- ❖ *data-driven* construction on mcast tree (e.g., RPF)
- ❖ bandwidth and non-group-router processing *profligate*

### Sparse:

- ❖ no membership until routers explicitly join
- ❖ *receiver-driven* construction of mcast tree (e.g., center-based)
- ❖ bandwidth and non-group-router processing *conservative*

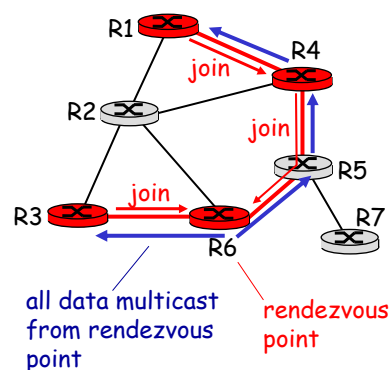
## PIM- Dense Mode

**flood-and-prune RPF**, similar to DVMRP but

- ❖ underlying unicast protocol provides RPF info for incoming datagram
- ❖ less complicated (less efficient) downstream flood than DVMRP reduces reliance on underlying routing algorithm
- ❖ has protocol mechanism for router to detect it is a leaf-node router

## PIM - Sparse Mode

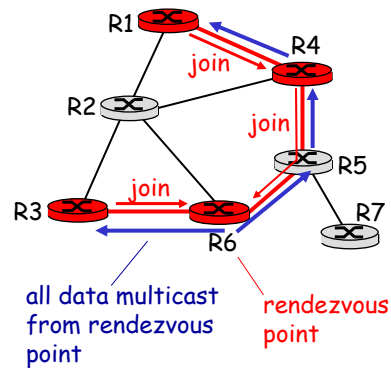
- ❖ center-based approach
- ❖ router sends *join* msg to rendezvous point (RP)
  - intermediate routers update state and forward *join*
- ❖ after joining via RP, router can switch to source-specific tree
  - increased performance: less concentration, shorter paths



## PIM - Sparse Mode

sender(s):

- ❖ unicast data to RP, which distributes down RP-rooted tree
- ❖ RP can extend mcast tree upstream to source
- ❖ RP can send *stop* msg if no attached receivers
  - "no one is listening!"



## Chapter 4: summary

### 4.1 Introduction

### 4.2 Virtual circuit and datagram networks

### 4.3 What's inside a router

### 4.4 IP: Internet Protocol

- Datagram format
- IPv4 addressing
- ICMP
- IPv6

### 4.5 Routing algorithms

- Link state
- Distance Vector
- Hierarchical routing

### 4.6 Routing in the Internet

- RIP
- OSPF
- BGP

### 4.7 Broadcast and multicast routing