

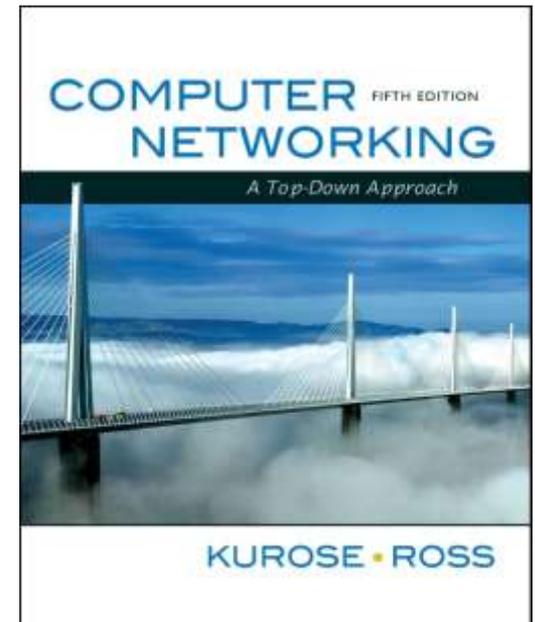
TDTS06: computer Networks

Lecturer: Johannes Schmidt

*The slides are taken from the book's companion
Web site with few modifications:
Computer Networking: A Top Down Approach
5th edition.*

Jim Kurose, Keith Ross
Addison-Wesley, April 2009.

Thank you JFK and KWR!



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

Today's lecture

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

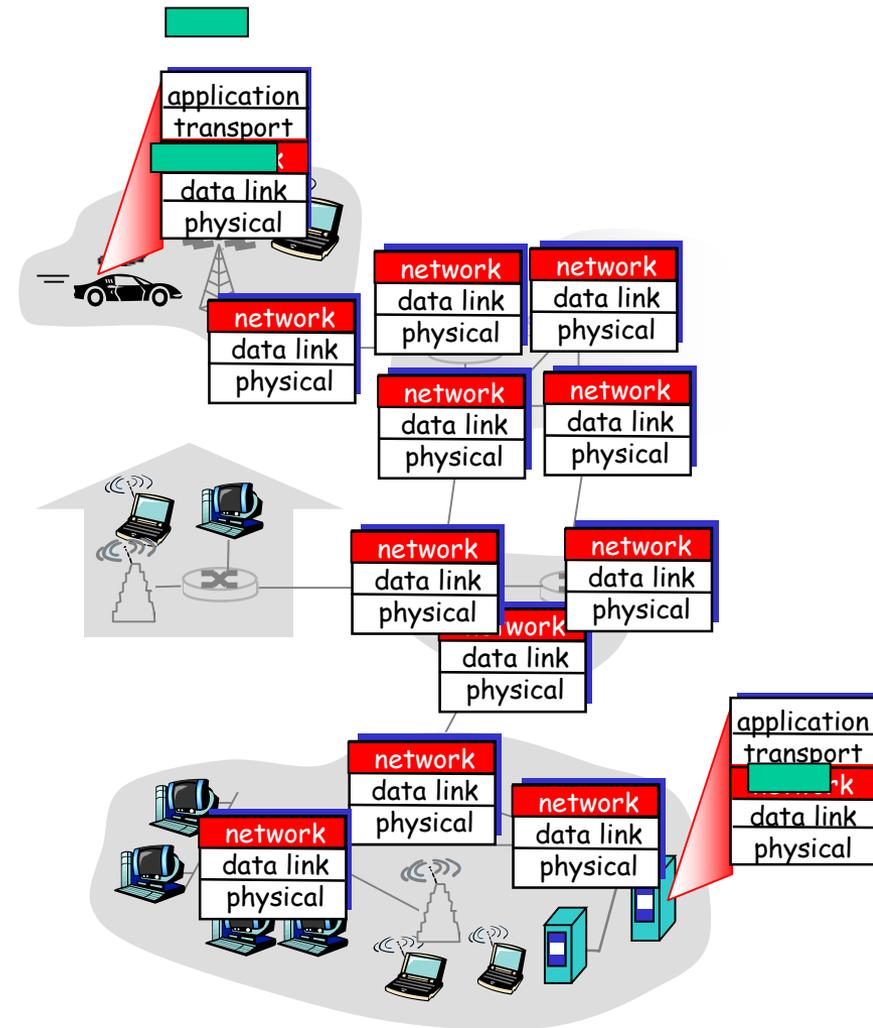
Network layer vs Transport layer

- ❖ *transport layer*: logical communication between **processes**
principal protocols
TCP, UDP
- ❖ *network layer*: logical communication between **hosts**
principal protocols
IP (IPv4, IPv6)

application
transport
network
data link
physical

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



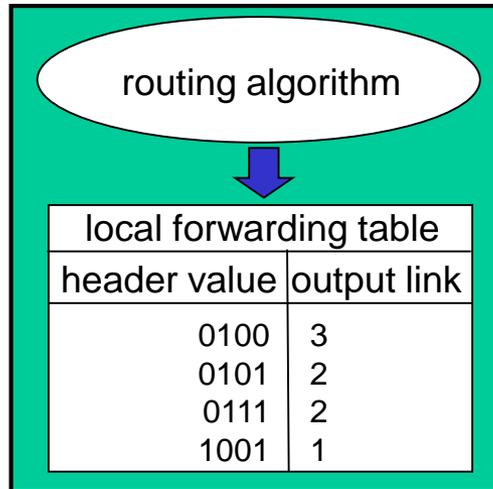
Two Key Network-Layer Functions

- ❖ *forwarding*: move packets from router's input to appropriate router output
- ❖ *routing*: determine route taken by packets from source to dest.
 - *routing algorithms*

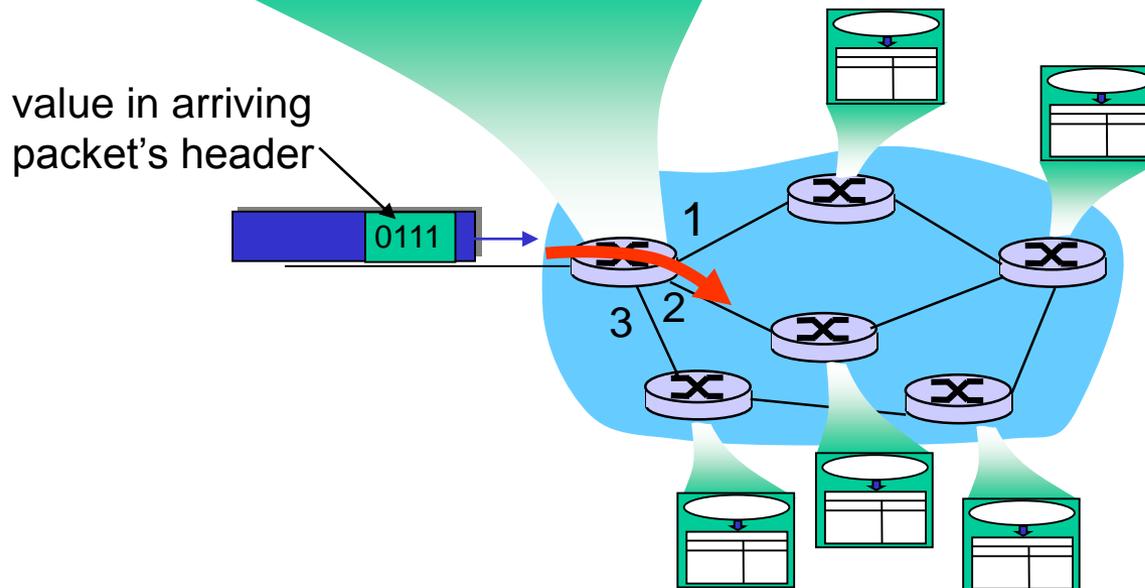
analogy:

- ❖ *routing*: process of planning trip from source to dest
- ❖ *forwarding*: process of getting through single interchange

Interplay between routing and forwarding



- ❖ *forwarding*: look up in routing table
- ❖ *routing*: update/compute routing table via a routing algorithm



Connection setup

- ❖ 3rd 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 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 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 connection and connection-less service

- ❖ **Datagram Network** provides network-layer connectionless service
- ❖ **Virtual Circuit** 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

Virtual circuits

“source-to-dest path behaves much like telephone circuit”

- performance-wise
- network actions along source-to-dest path

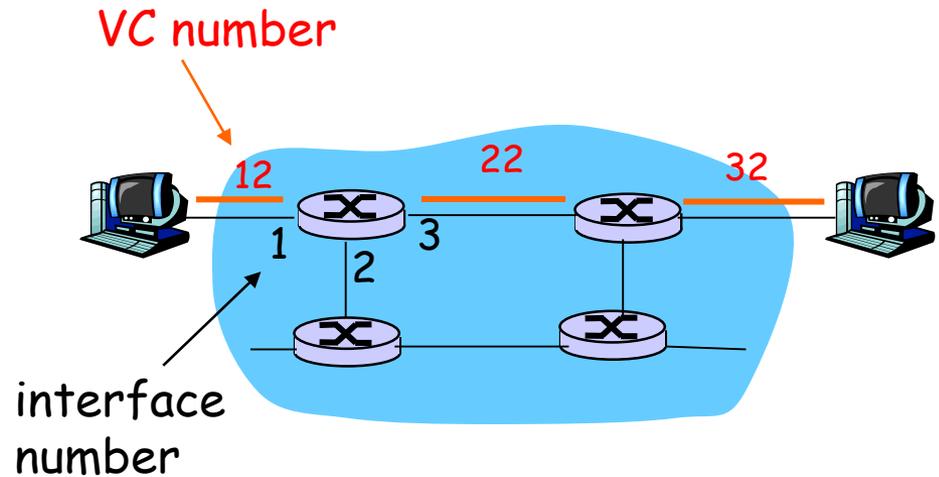
- ❖ call setup, teardown for each call *before* data can flow
- ❖ each packet carries VC identifier (not destination host address)
- ❖ *every* router on source-dest path maintains “state” for each passing connection
- ❖ link, router resources (bandwidth, buffers) may be *allocated* to VC (dedicated resources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

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

VC Forwarding table



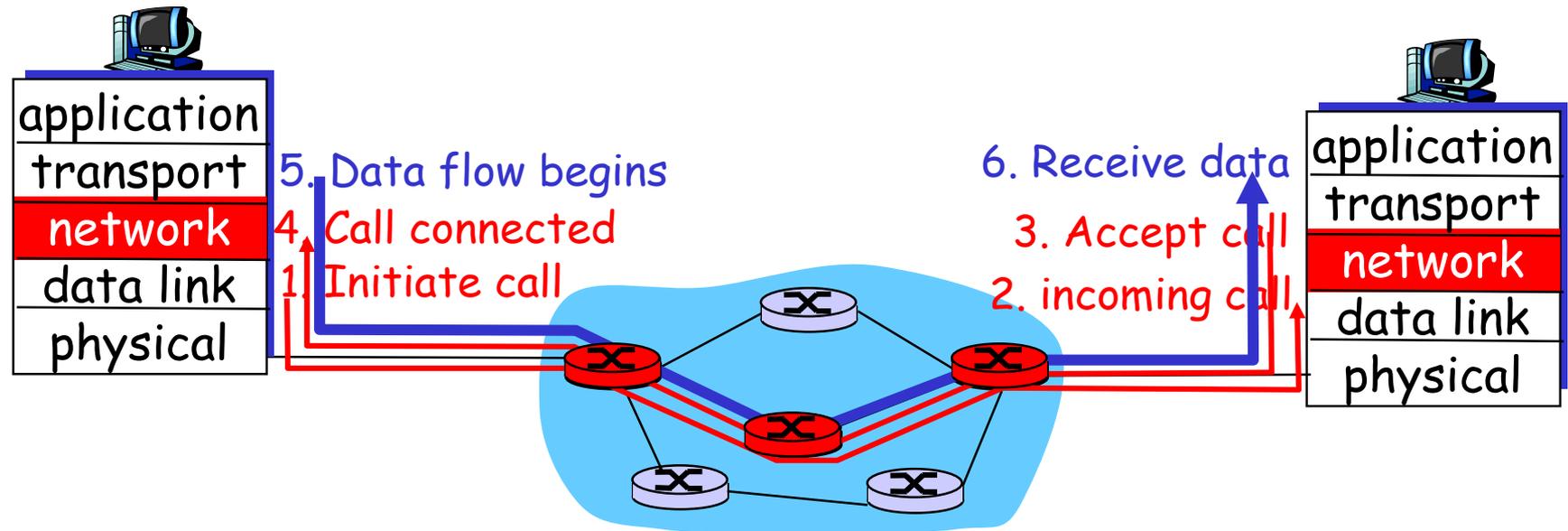
Forwarding table in northwest router:

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!

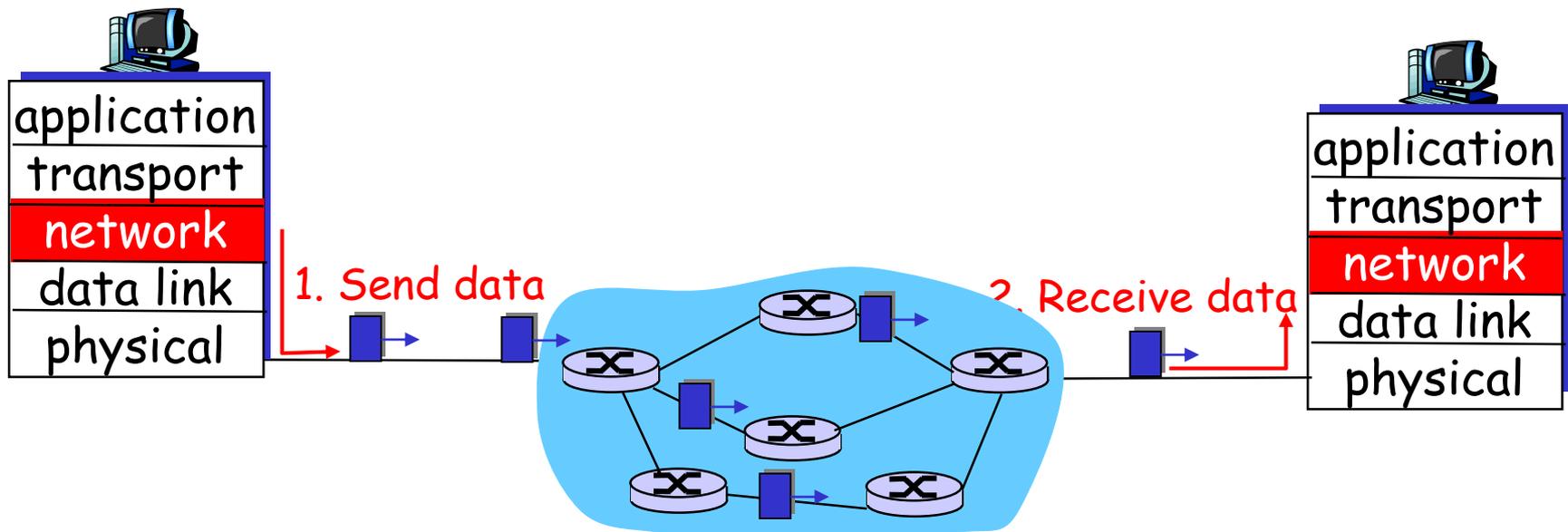
Virtual circuits: signaling protocols

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

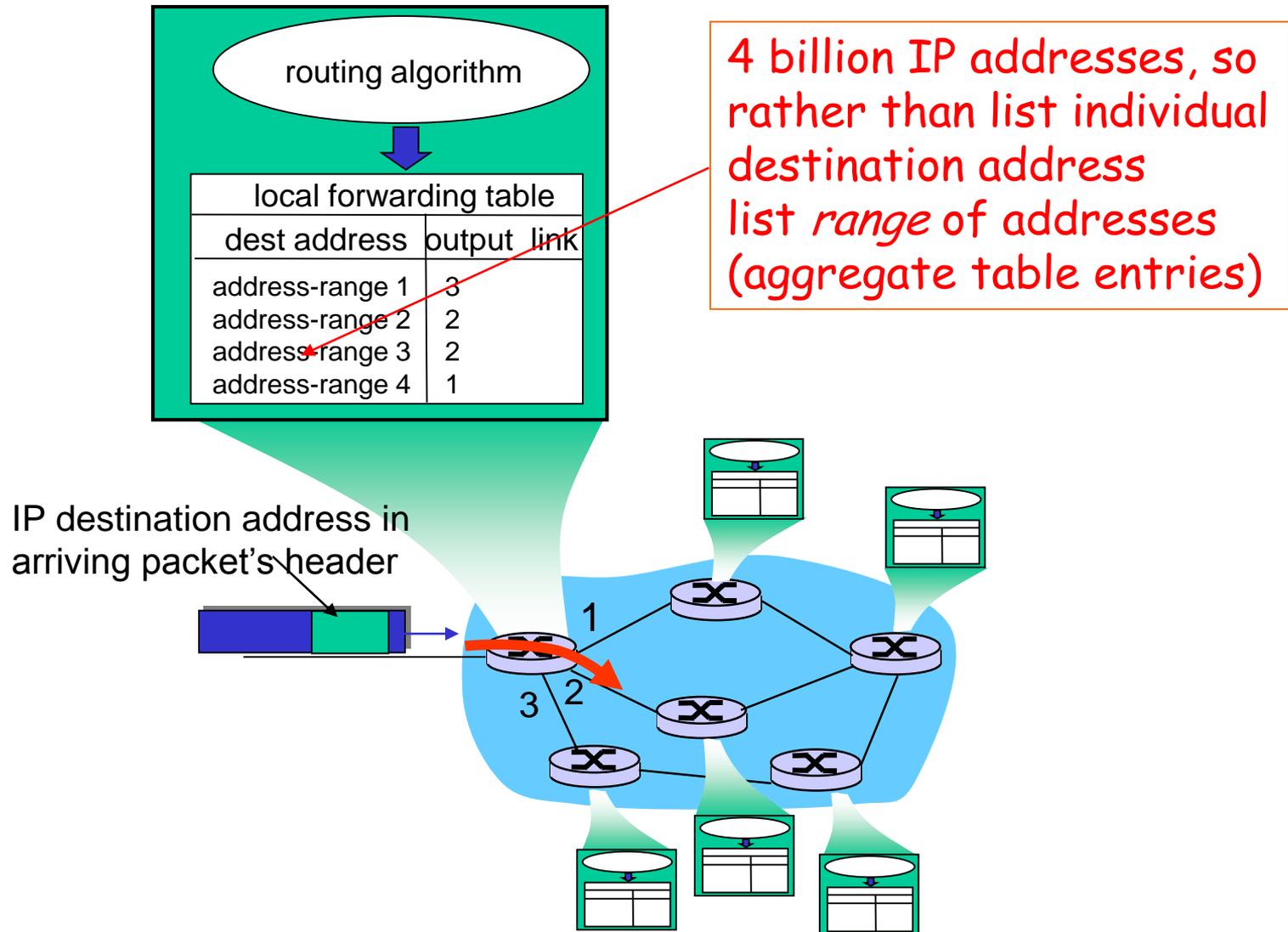


Datagram networks

- ❖ 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



Datagram Forwarding table



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?

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?

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

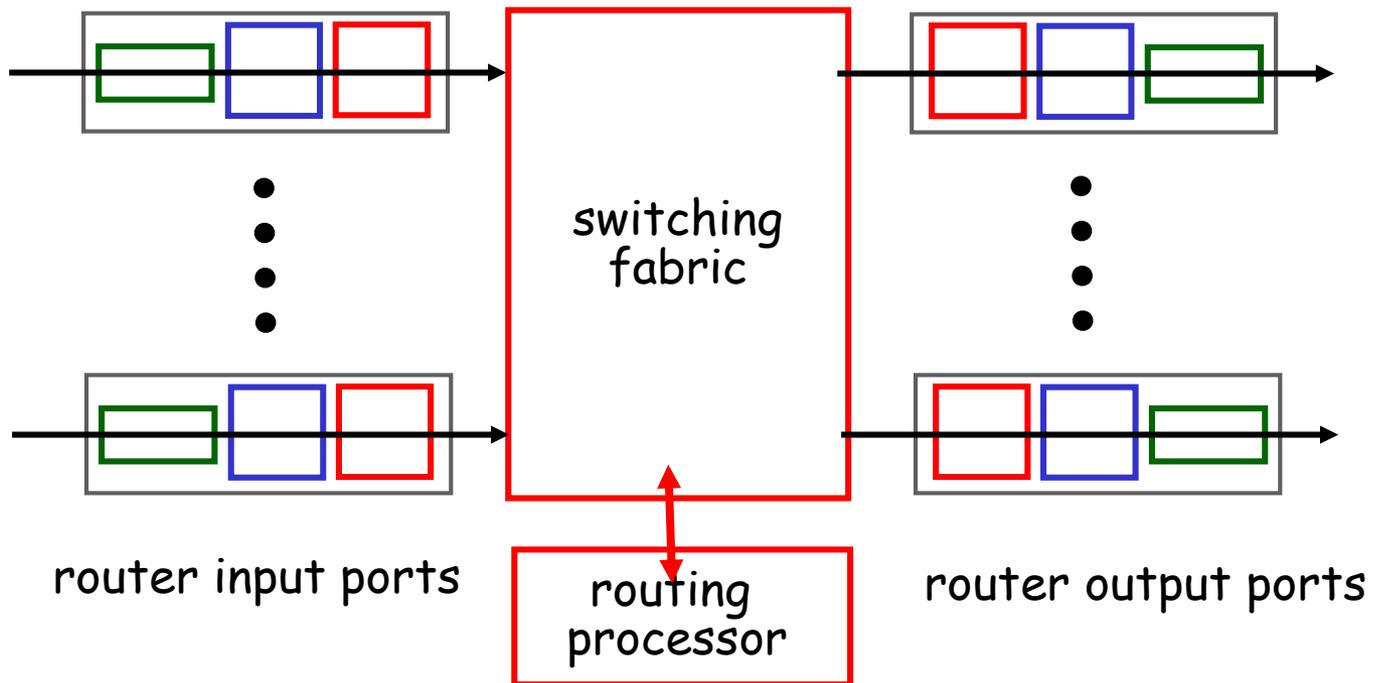
ATM (VC)

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

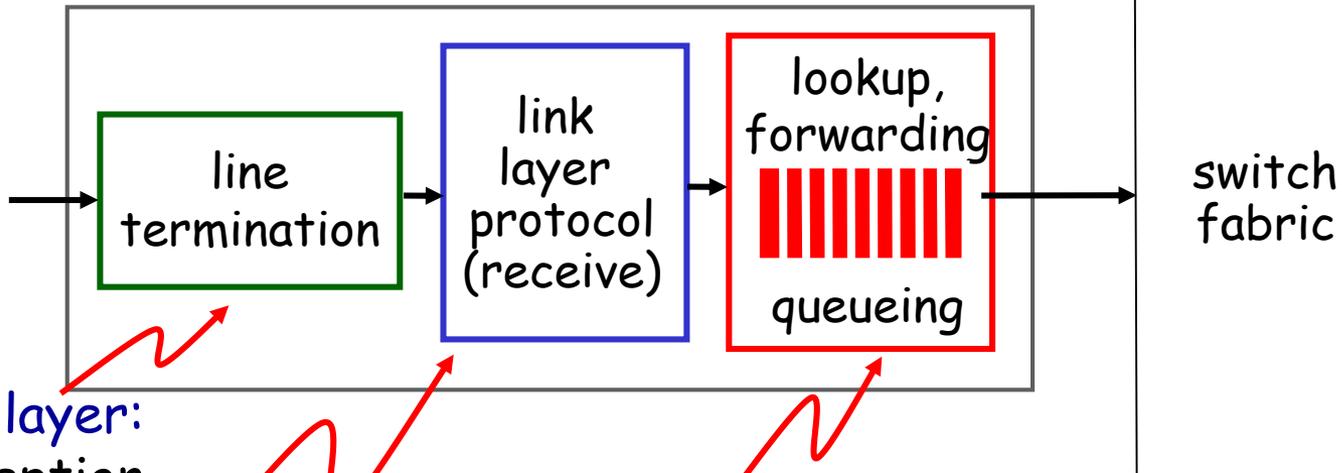
Router Architecture Overview

two key router functions:

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



Input Port Functions



Physical layer:
bit-level reception

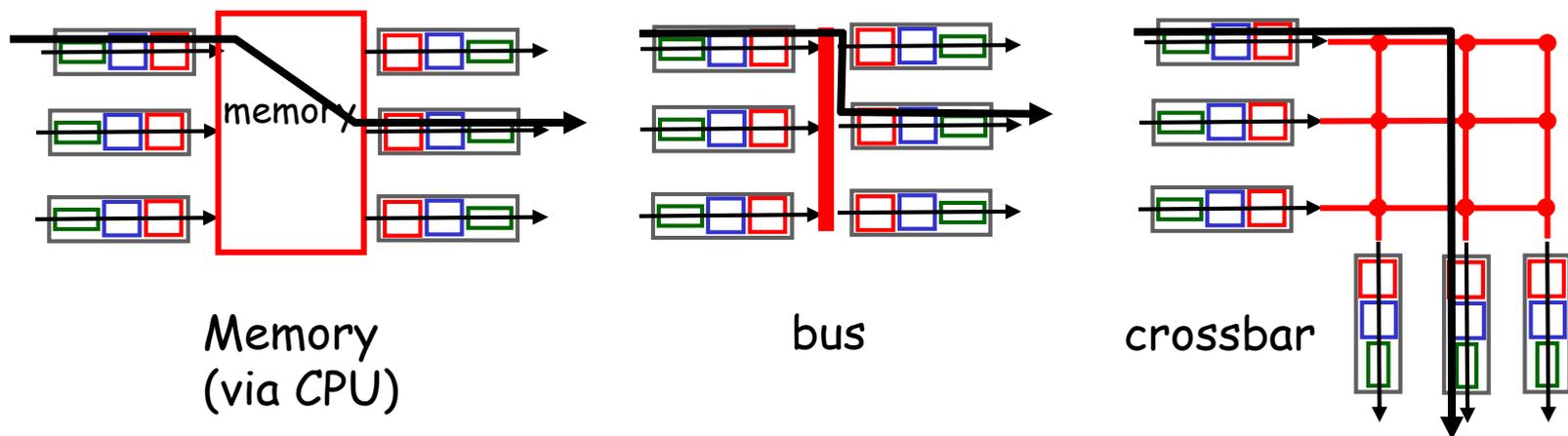
Data link layer:
e.g., Ethernet
see chapter 5

Decentralized switching:

- ❖ given datagram dest., lookup output port using forwarding table in input port memory
- ❖ goal: complete input port processing at "line speed"
- ❖ queuing: if datagrams arrive faster than forwarding rate into switch fabric

Switching fabrics

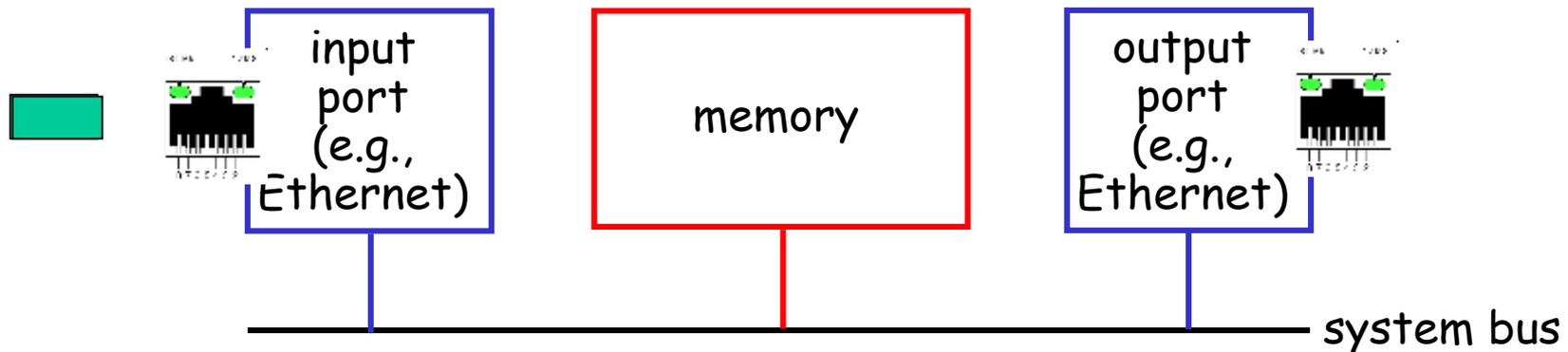
- ❖ transfer packet from input buffer to appropriate output buffer
- ❖ switching rate: rate at which packets can be transferred from inputs to outputs
 - often measured as multiple of input/output line rate
 - N inputs: switching rate N times line rate desirable
- ❖ three types of switching fabrics



Switching Via Memory

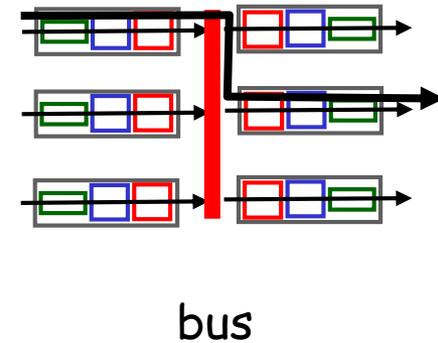
First generation routers:

- ❖ 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)



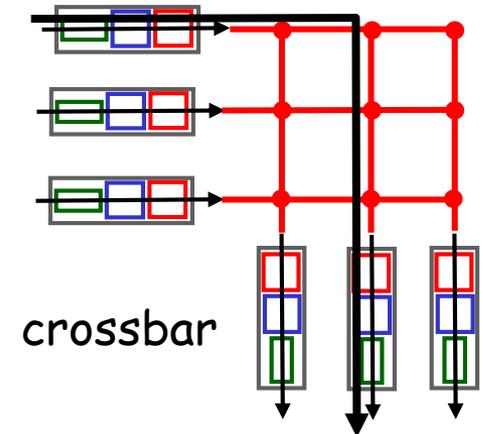
Switching Via a Bus

- ❖ 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

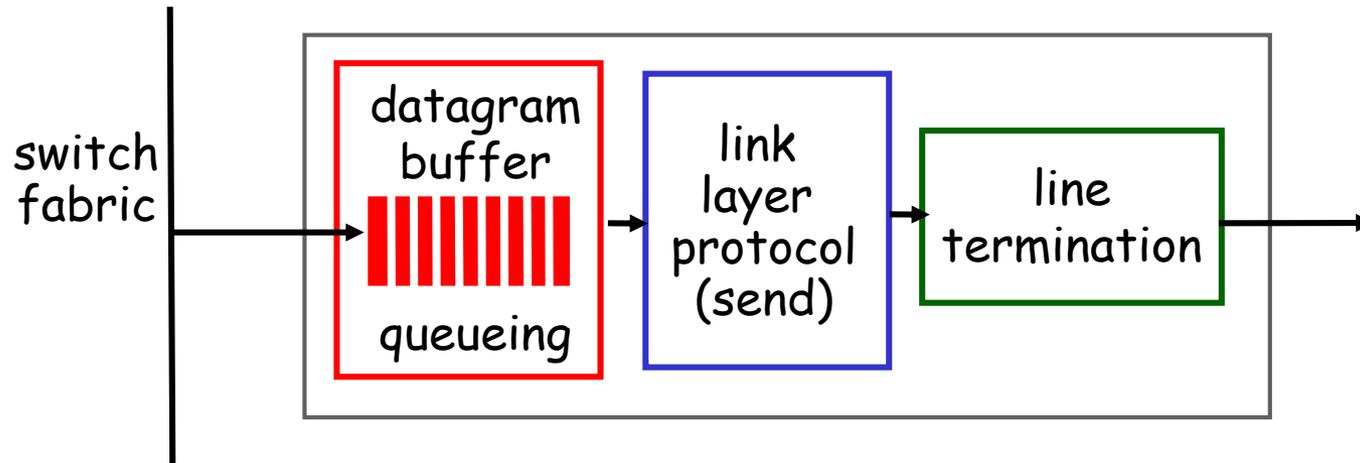


Switching Via An Interconnection Network

- ❖ 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

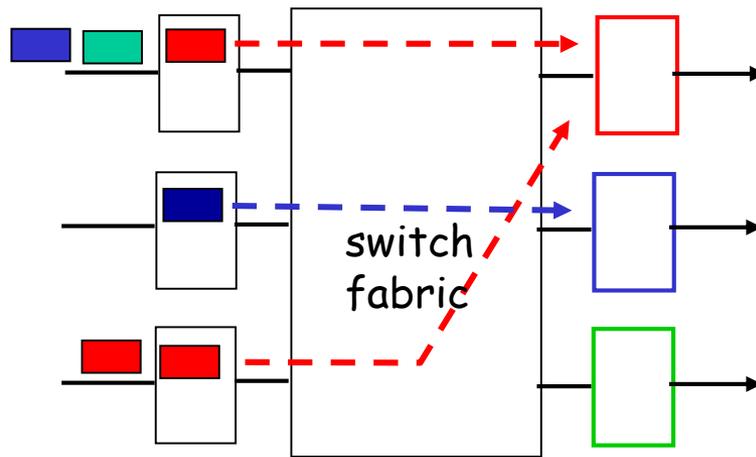


Output Ports

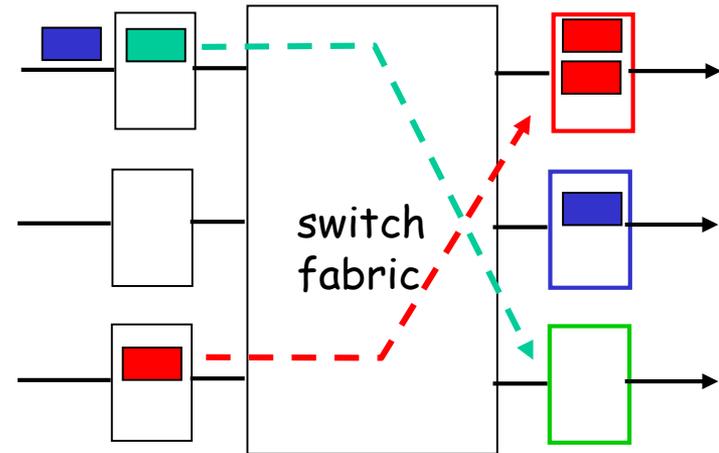


- ❖ *buffering* required when datagrams arrive from fabric faster than the transmission rate
- ❖ *scheduling discipline* chooses among queued datagrams for transmission

Output port queueing



at t , packets move
from input to output



one packet time
later

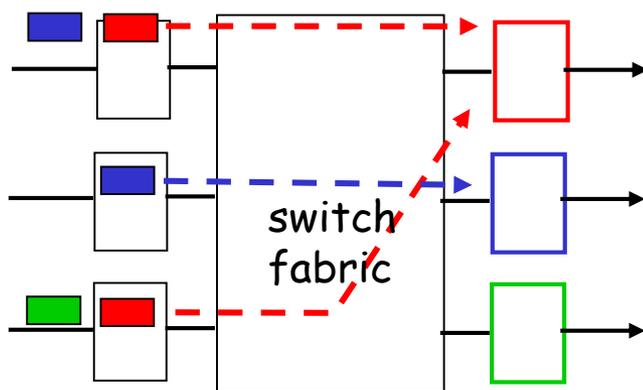
- ❖ buffering when arrival rate via switch exceeds output line speed
- ❖ *queueing (delay) and loss due to output port buffer overflow!*

How much buffering?

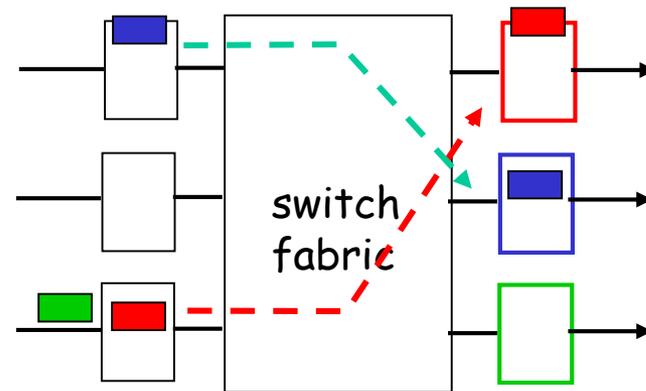
- ❖ RFC 3439 rule of thumb: average buffering equal to "typical" RTT (say 250 msec) times link capacity C
 - e.g., $C = 10$ Gpbs link: 2.5 Gbit buffer
- ❖ recent recommendation: with N flows, buffering equal to $\frac{RTT \cdot C}{\sqrt{N}}$

Input Port Queuing

- ❖ fabric slower than input ports combined -> queuing 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



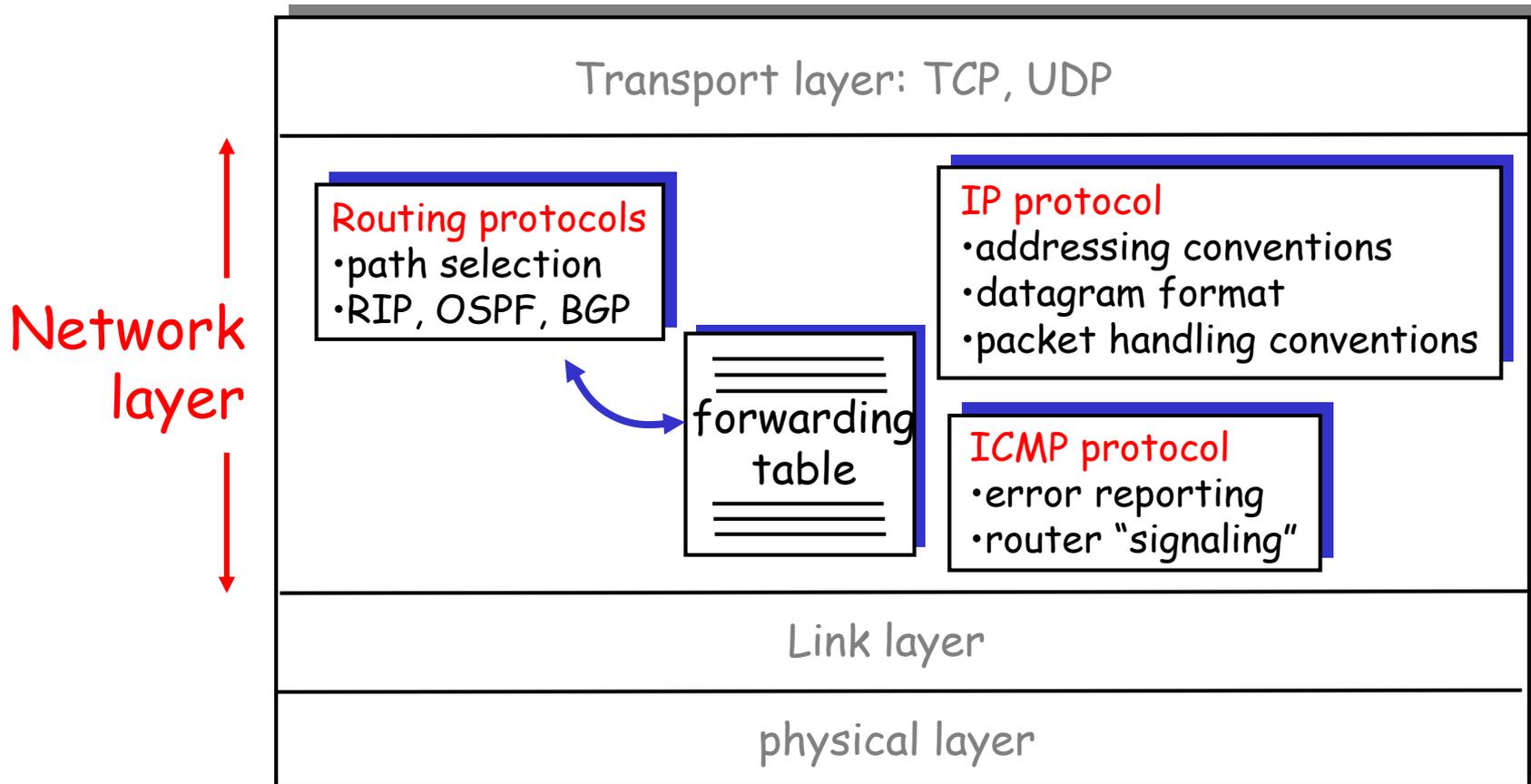
output port contention:
only one red datagram can be
transferred.
lower red packet is blocked



one packet time
later: green packet
experiences HOL
blocking

The Internet Network layer

Host, router network layer functions:



IP datagram format

IP protocol version number

header length (bytes)

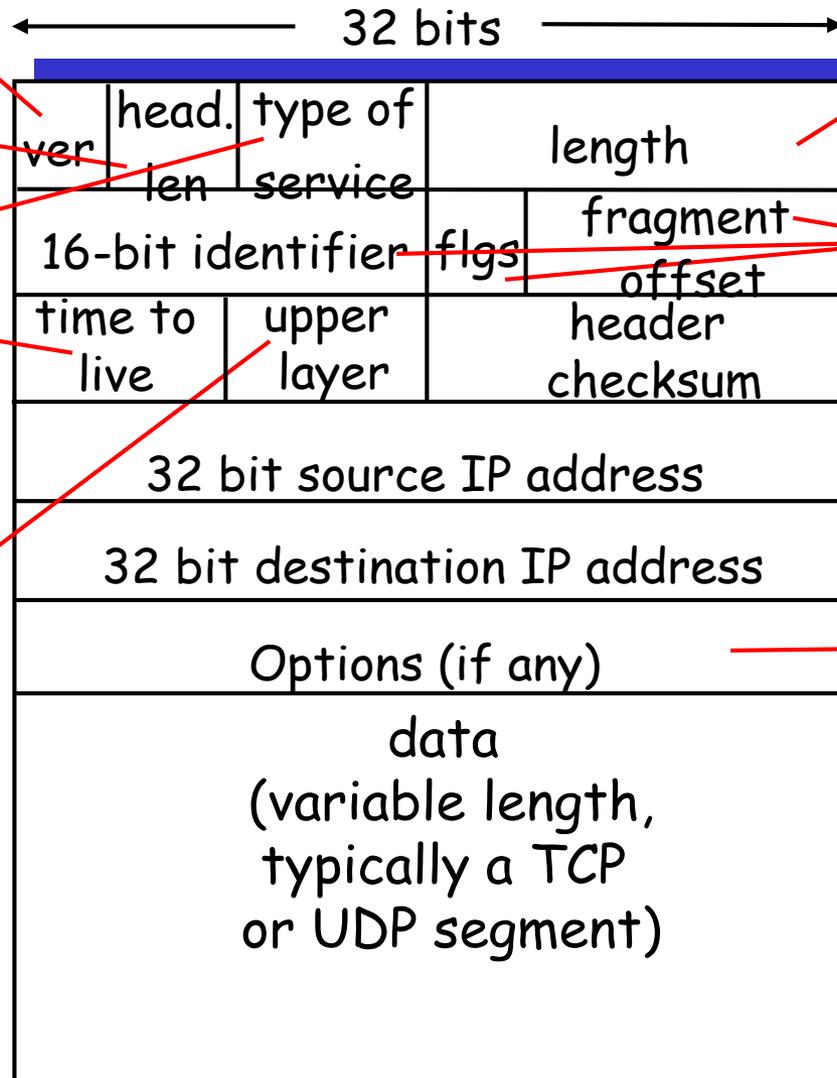
"type" of data

max number remaining hops (decremented at each router)

upper layer protocol to deliver payload to

how much overhead with TCP?

- ❖ 20 bytes of TCP
- ❖ 20 bytes of IP
- ❖ = 40 bytes + app layer overhead



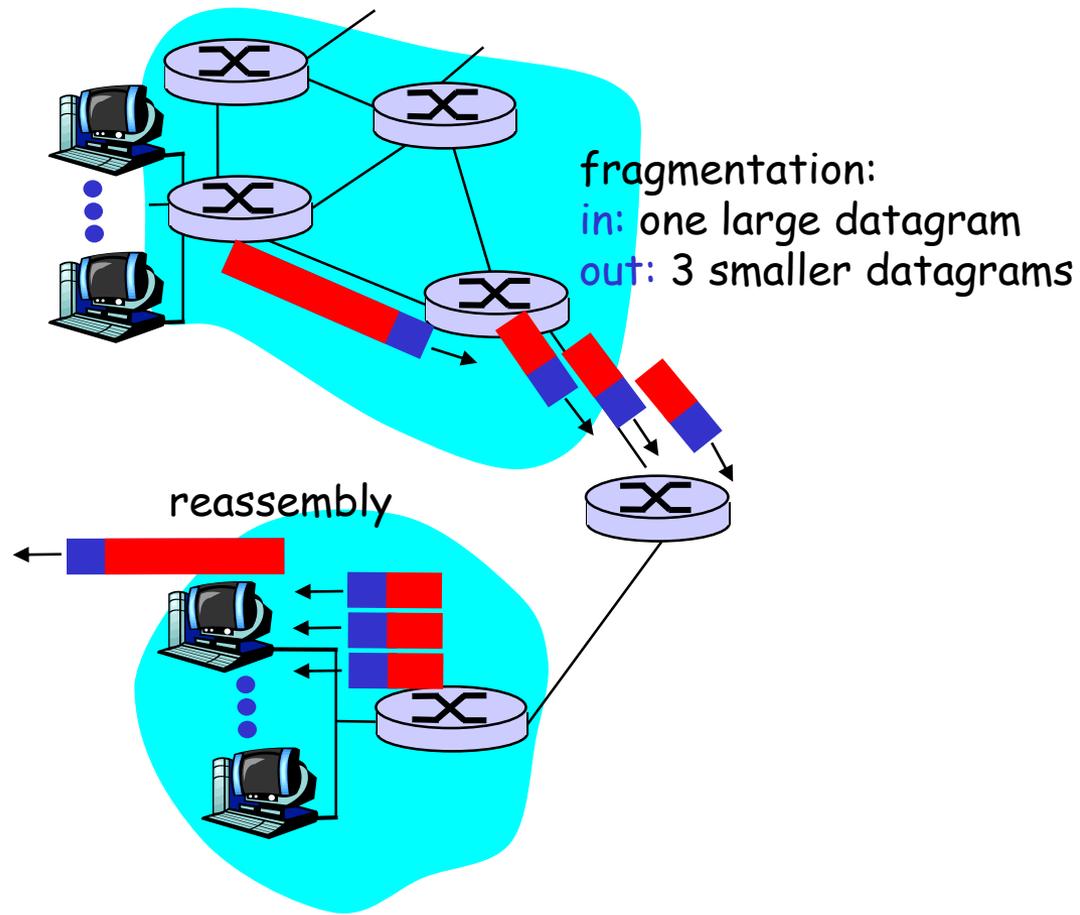
total datagram length (bytes)

for fragmentation/reassembly

E.g. timestamp, record route taken, specify list of routers to visit.

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



IP Fragmentation and Reassembly

Example

- ❖ 4000 byte datagram
- ❖ MTU = 1500 bytes

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

One large datagram becomes several smaller datagrams

1480 bytes in data field

offset =
 $1480/8$

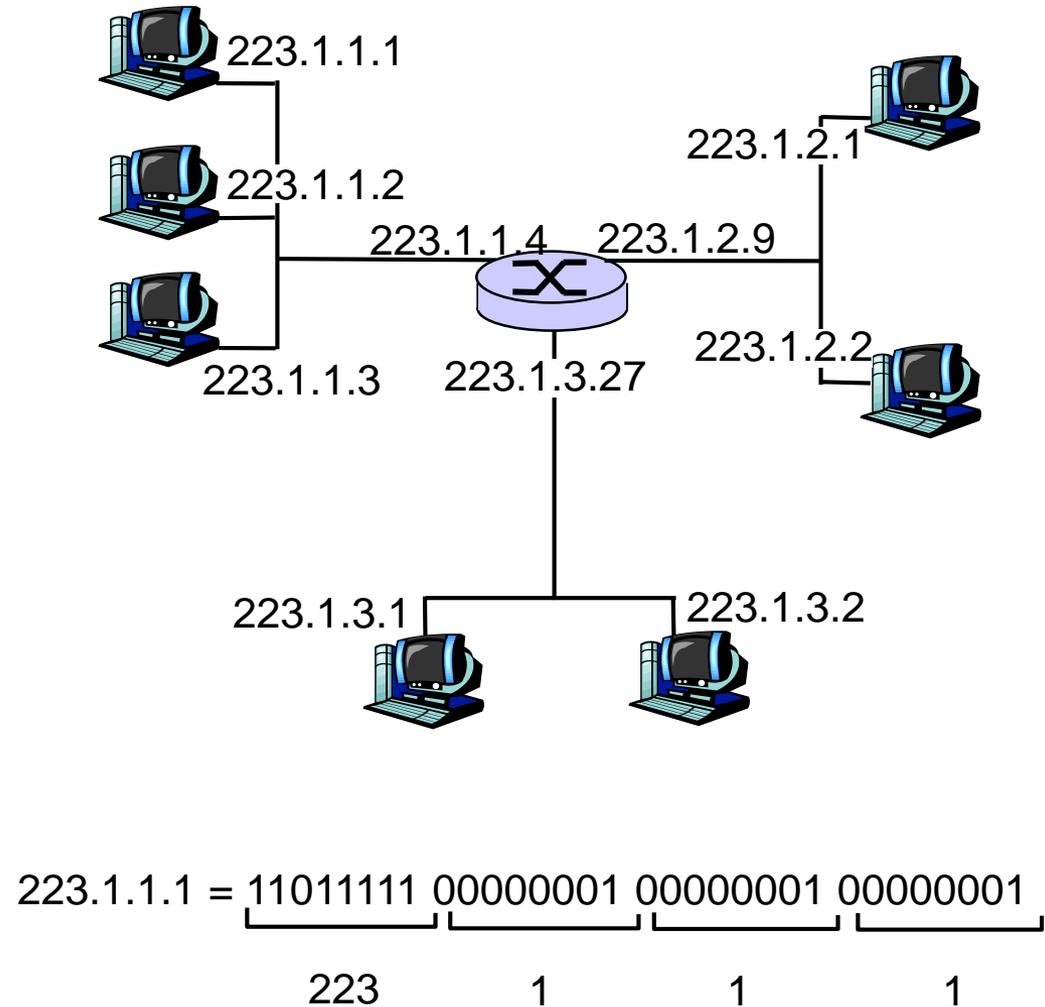
	length	ID	fragflag	offset	
	=1500	=x	=1	=0	

	length	ID	fragflag	offset	
	=1500	=x	=1	=185	

	length	ID	fragflag	offset	
	=1040	=x	=0	=370	

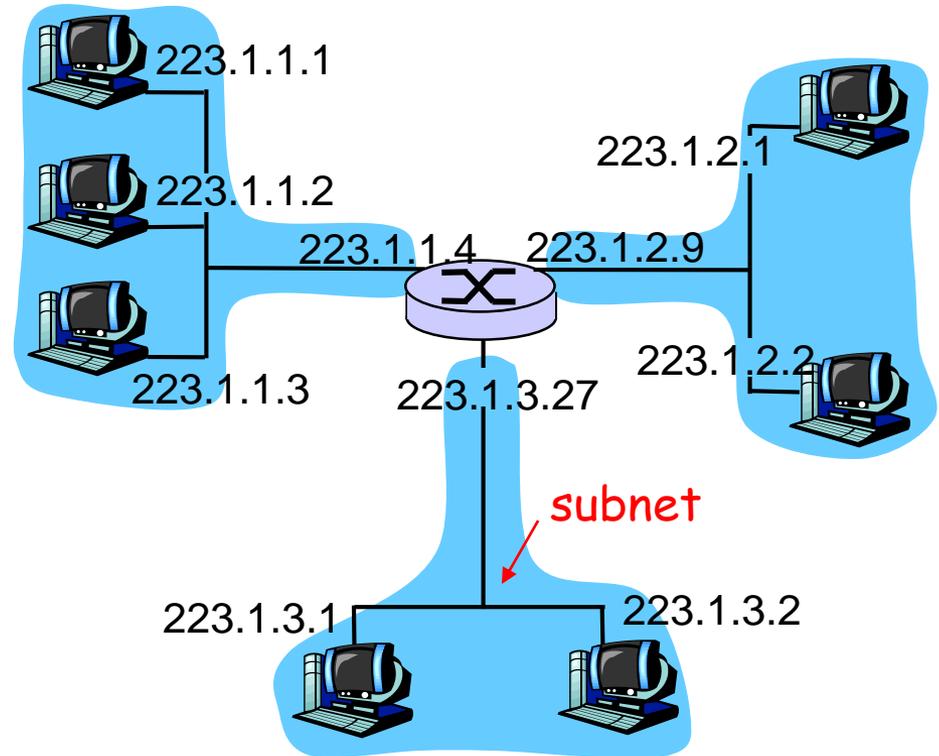
IP Addressing: introduction

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



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

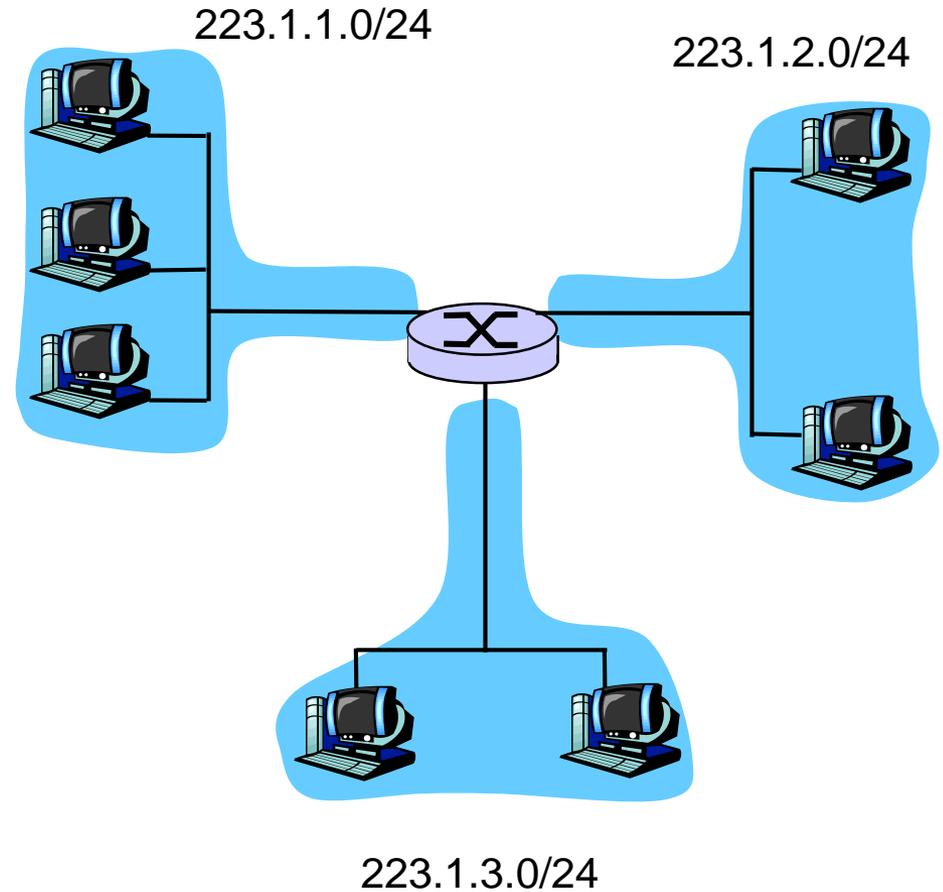


network consisting of 3 subnets

Subnets

Recipe

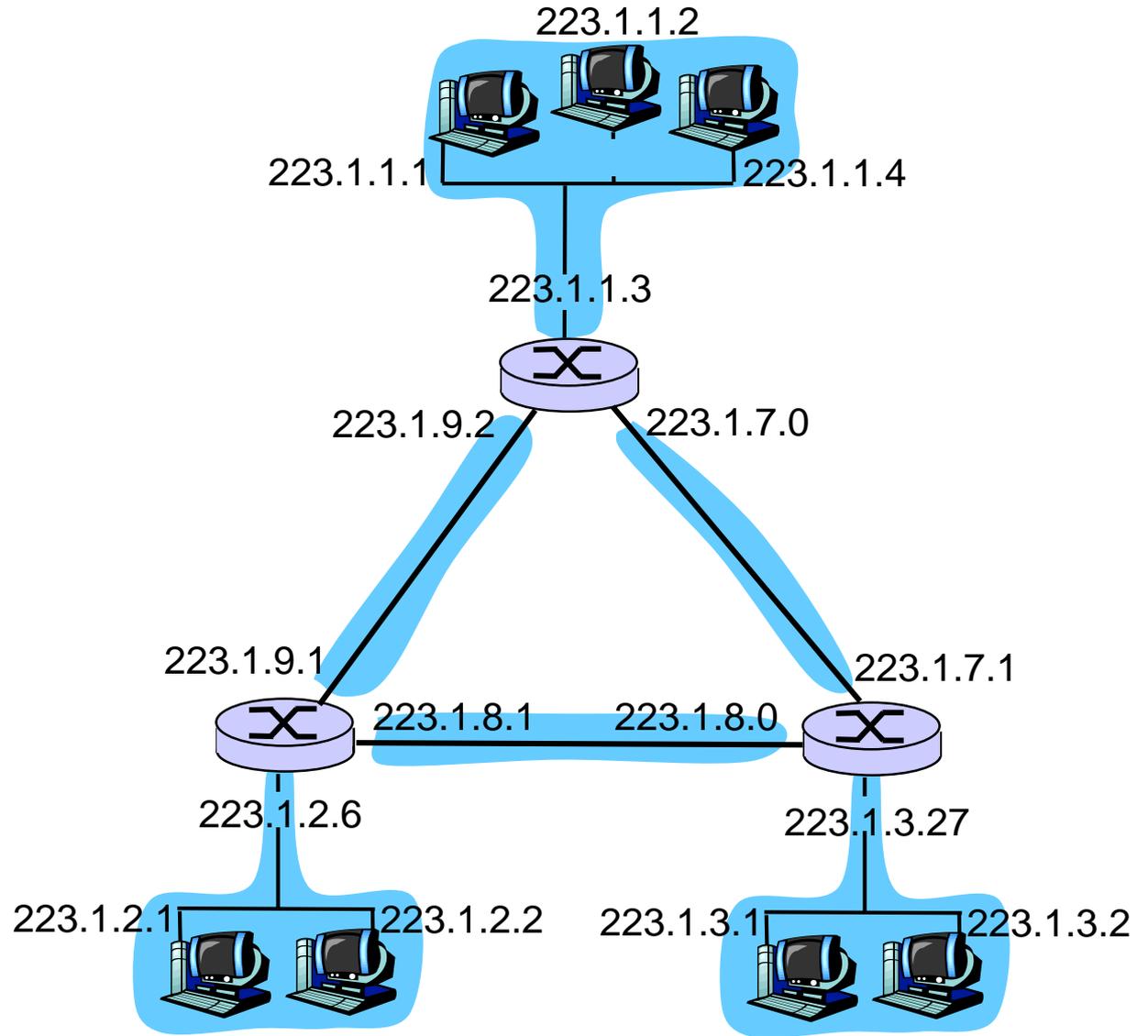
- ❖ to determine the subnets, detach each interface from its host or router, creating islands of isolated networks
- ❖ each isolated network is called a **subnet**.



Subnet mask: /24

Subnets

How many?



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



IP addresses: how to get one?

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

- ❖ 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 a server
 - “plug-and-play”

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

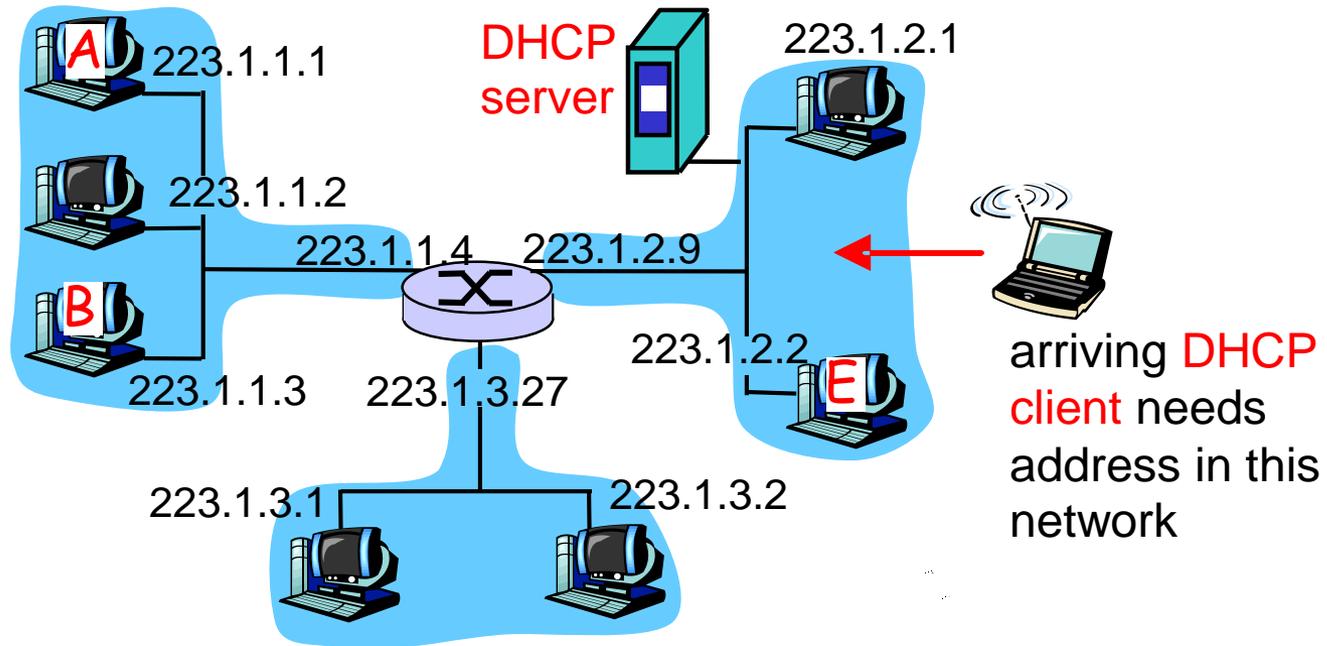
Allows reuse of addresses (only hold address while connected an "on")

Support for mobile users who want to join network (more shortly)

DHCP overview:

- 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

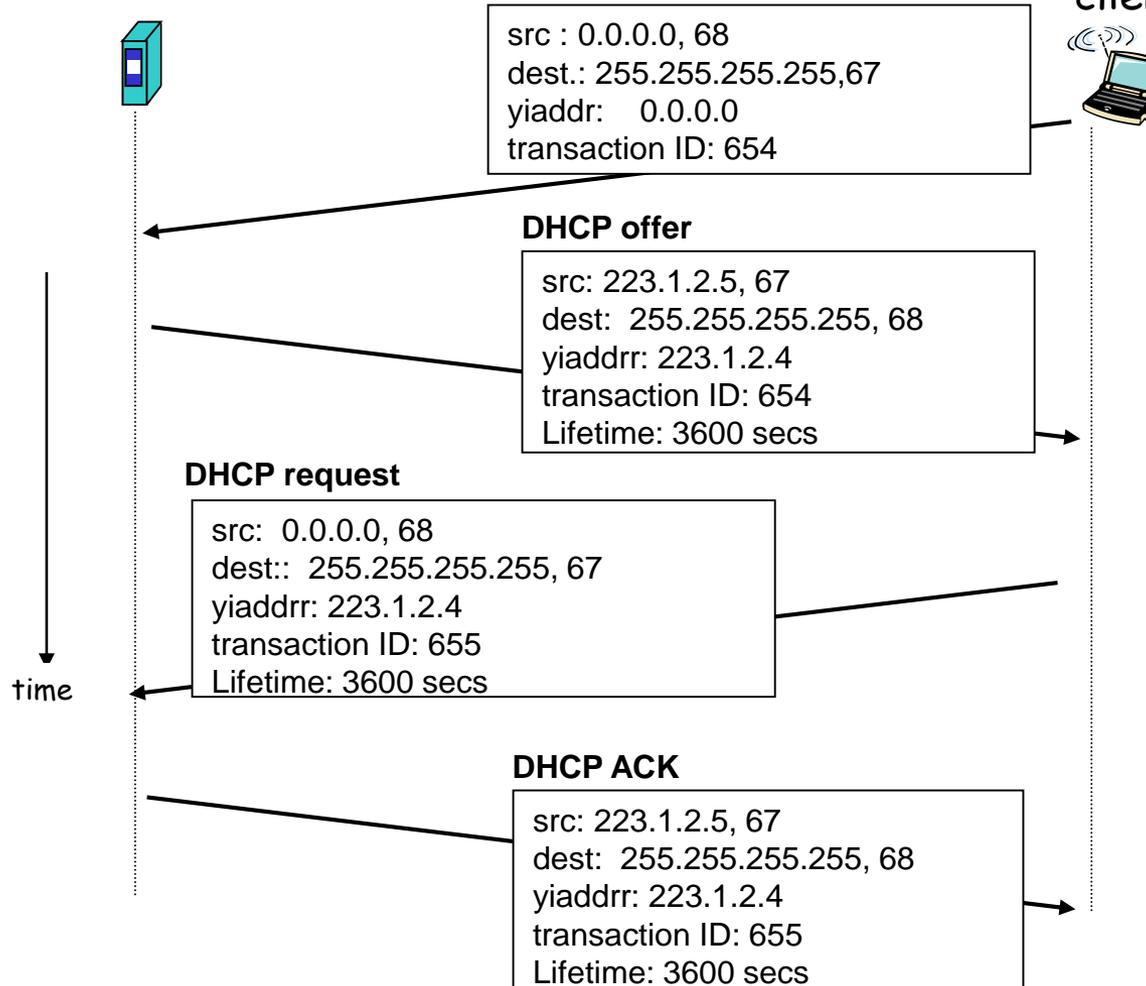
DHCP client-server scenario



DHCP client-server scenario

DHCP server: 223.1.2.5

arriving client

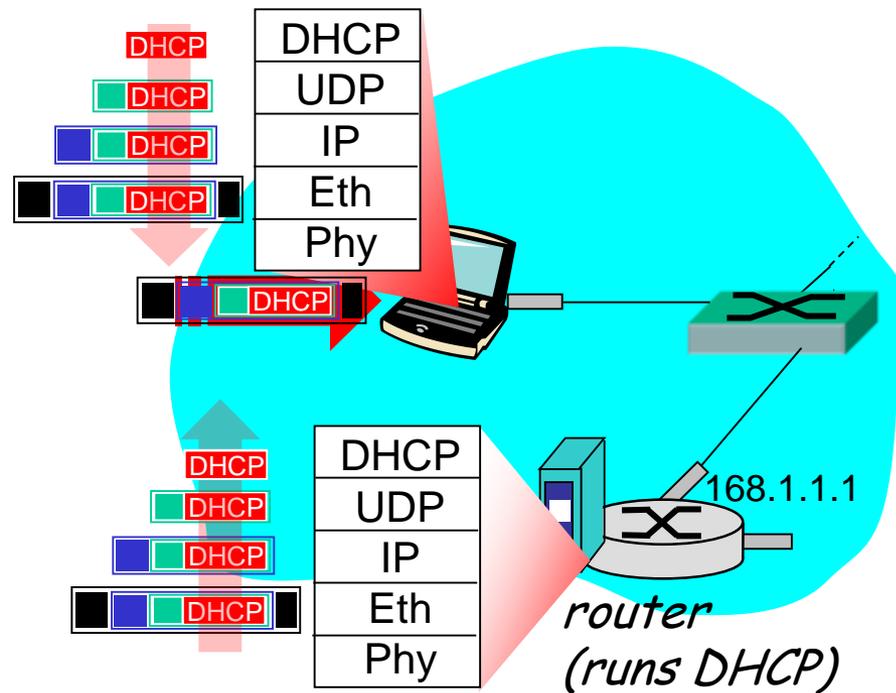


DHCP: more than IP address

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

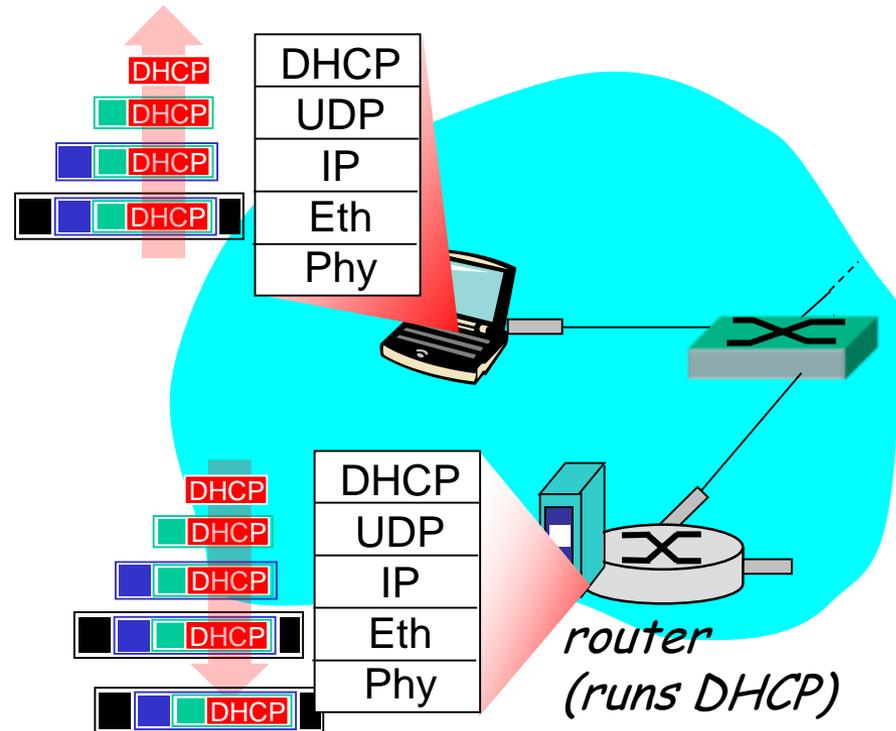
- address of first-hop router for client
- name and IP address of DNS sever
- network mask (indicating network versus host portion of address)

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: FFFFFFFFFFFFFFFF) on LAN, received at router running DHCP server
- ❖ Ethernet demuxed to IP demuxed, UDP demuxed to DHCP

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

IP addresses: how to get one?

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

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

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

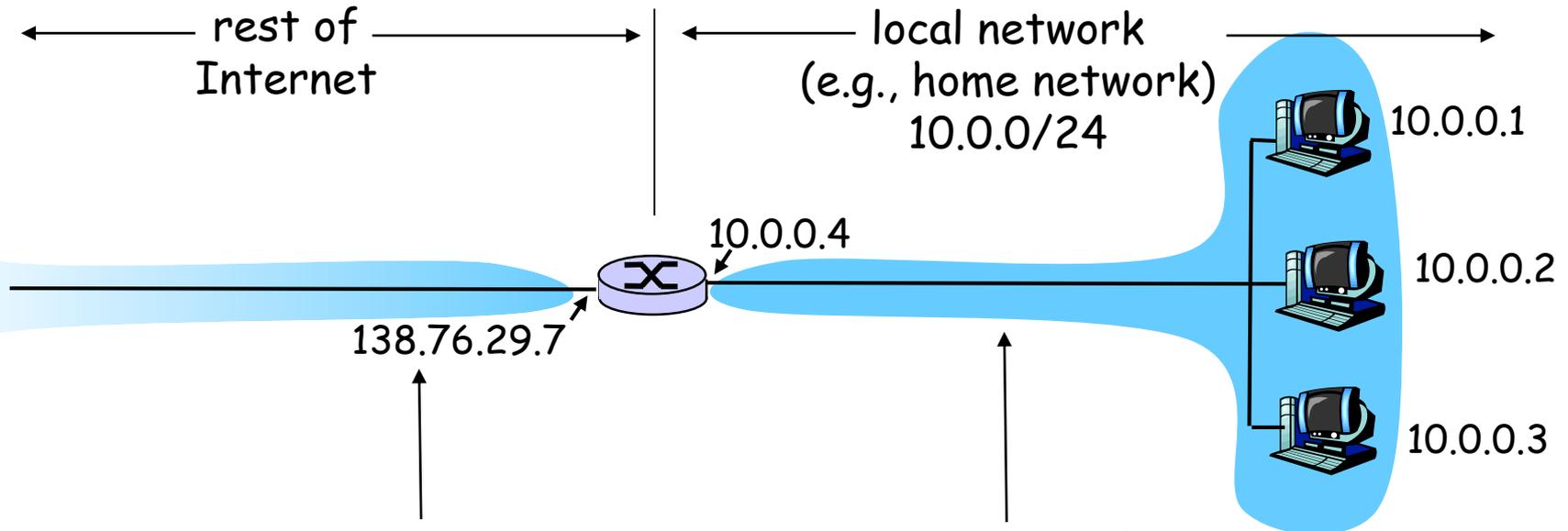
IP addressing: the last word...

Q: How does an ISP get block of addresses?

A: **ICANN:** Internet **C**orporation for **A**ssigned
Names and **N**umbers

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

NAT: Network Address Translation



All datagrams *leaving* local network have *same* single source NAT IP address: 138.76.29.7, different source port numbers

Datagrams with source or destination in this network have 10.0.0/24 address for source, destination (as usual)

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).

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

NAT: Network Address Translation

NAT translation table	
WAN side addr	LAN side addr
138.76.29.7, 5001	10.0.0.1, 3345
.....

1: host 10.0.0.1 sends datagram to 128.119.40.186, 80

S: 10.0.0.1, 3345
D: 128.119.40.186, 80

10.0.0.1
10.0.0.2
10.0.0.3

2: NAT router changes datagram source addr from 10.0.0.1, 3345 to 138.76.29.7, 5001, updates table

S: 138.76.29.7, 5001
D: 128.119.40.186, 80

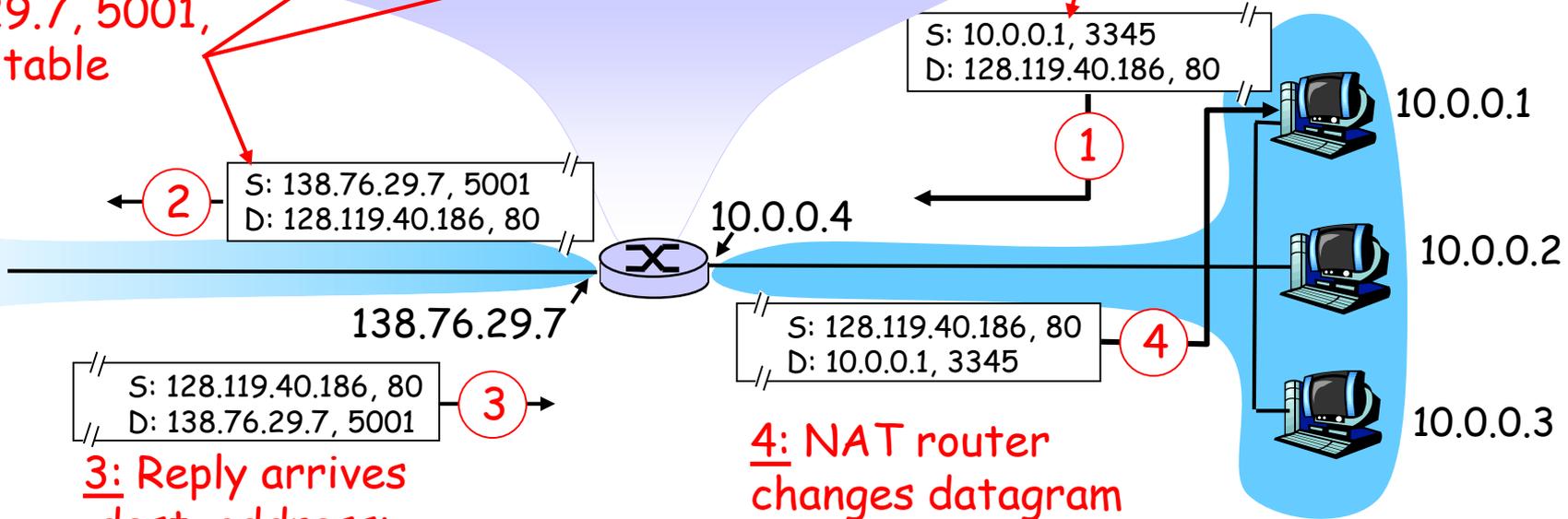
138.76.29.7

S: 128.119.40.186, 80
D: 138.76.29.7, 5001

3: Reply arrives dest. address: 138.76.29.7, 5001

S: 128.119.40.186, 80
D: 10.0.0.1, 3345

4: NAT router changes datagram dest addr from 138.76.29.7, 5001 to 10.0.0.1, 3345

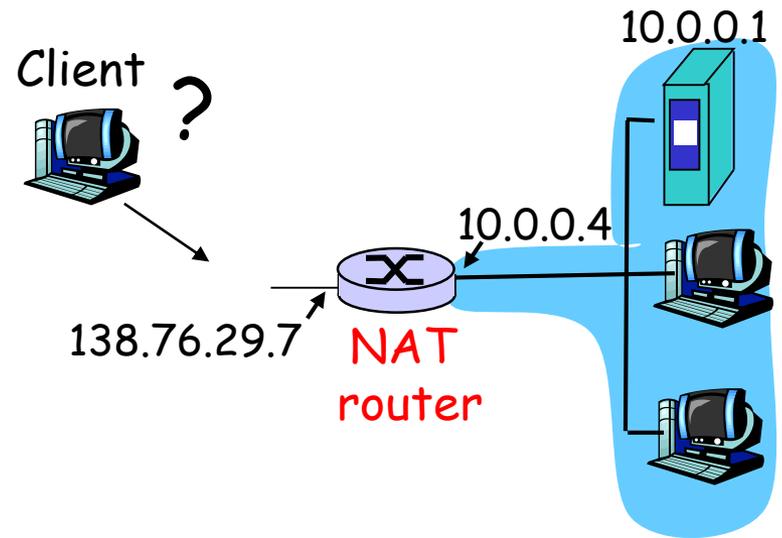


NAT: Network Address Translation

- ❖ 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

NAT traversal problem

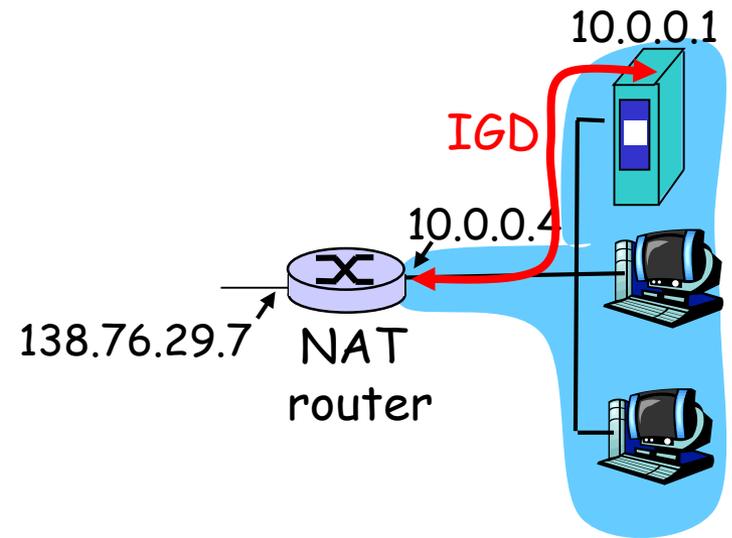
- ❖ client wants to connect to server with address 10.0.0.1
 - server address 10.0.0.1 local to LAN (client can't use it as destination addr)
 - only one externally visible NATed address: 138.76.29.7
- ❖ solution 1: statically configure NAT to forward incoming connection requests at given port to server
 - e.g., (123.76.29.7, port 2500) always forwarded to 10.0.0.1 port 25000



NAT traversal problem

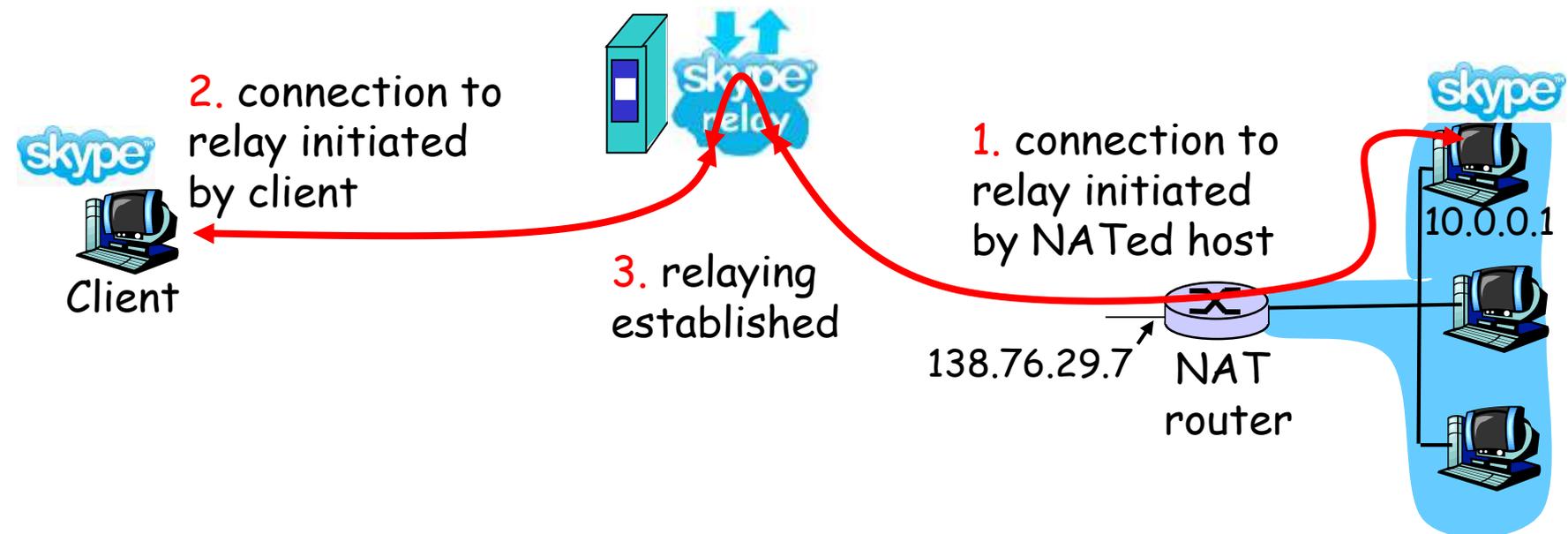
- ❖ solution 2: Universal Plug and Play (UPnP) Internet Gateway Device (IGD) Protocol. Allows NATed host to:
 - ❖ learn public IP address (138.76.29.7)
 - ❖ add/remove port mappings (with lease times)

i.e., automate static NAT port map configuration



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



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

<u>Type</u>	<u>Code</u>	<u>description</u>
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

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.

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

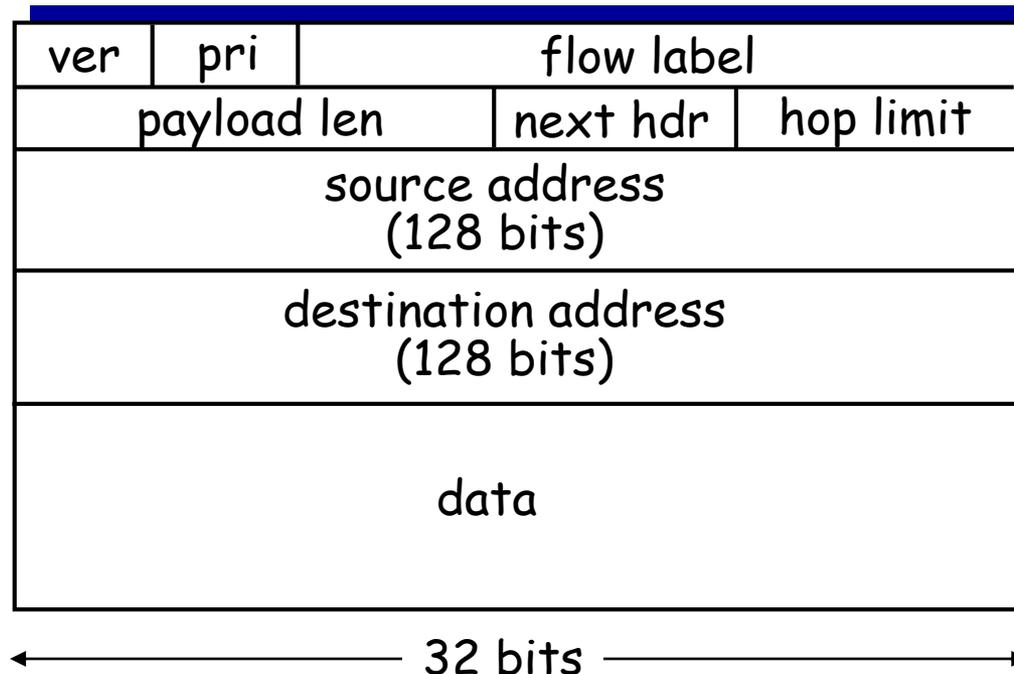
IPv6 Header (Cont)

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



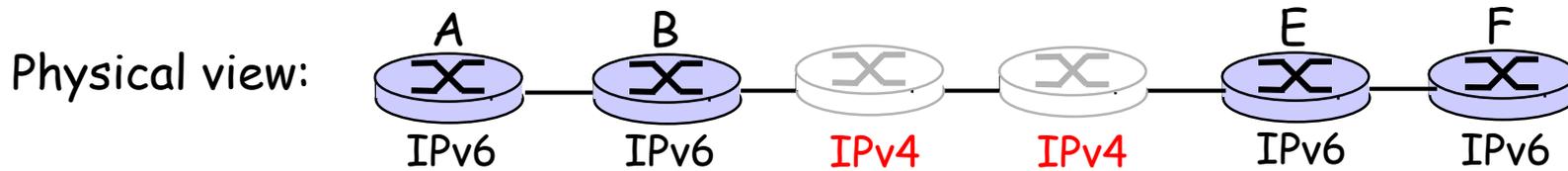
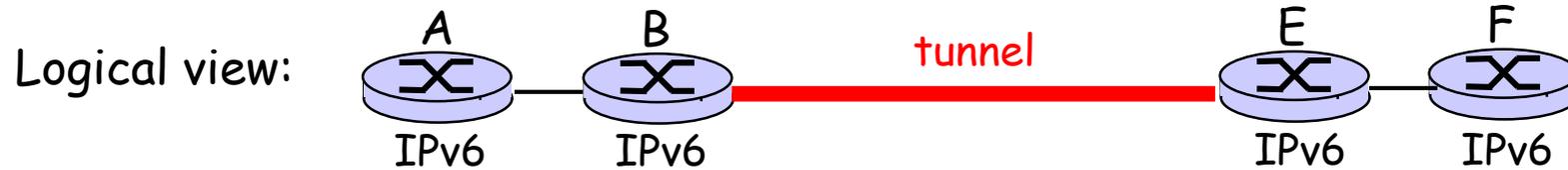
Other Changes from IPv4

- ❖ *No checksum*: removed entirely to reduce processing time at each hop
- ❖ *No fragmentation*: deferred to end systems
- ❖ *No 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

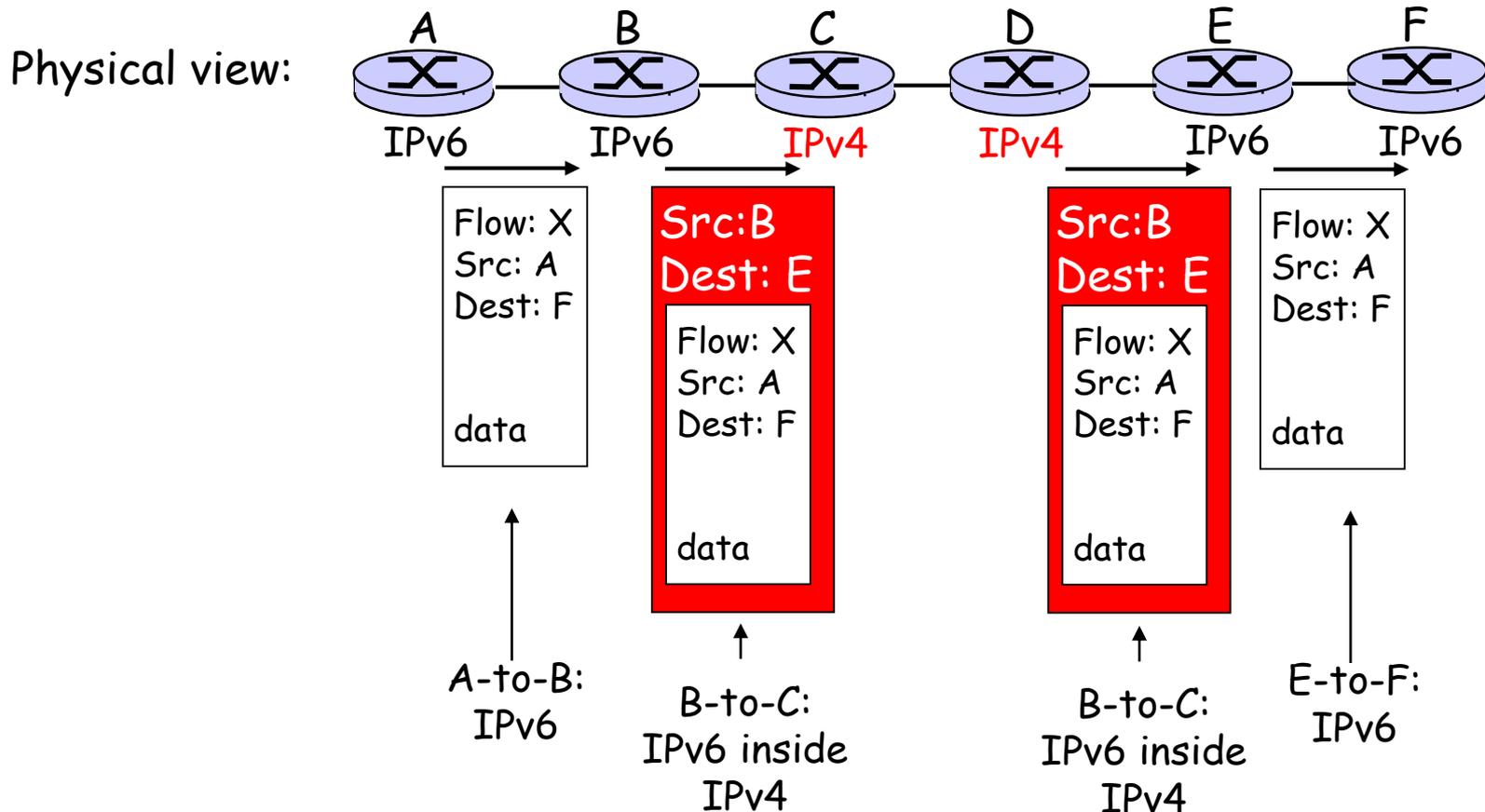
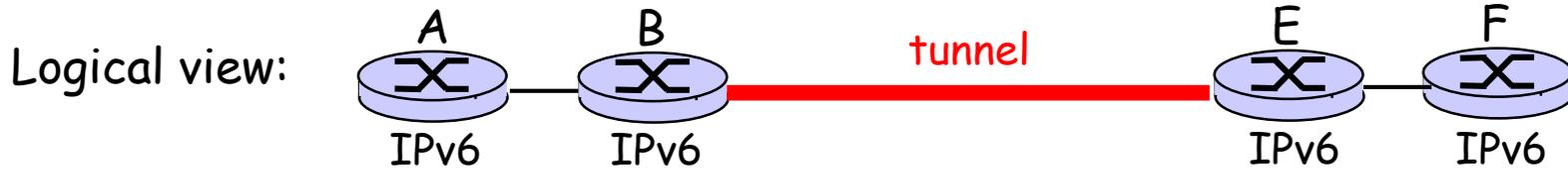
Transition From IPv4 To IPv6

- ❖ Not all routers can be upgraded simultaneous
 - no “flag days”
 - How will the network operate with mixed IPv4 and IPv6 routers?
- ❖ *Dual stack*: datagram starts as IPv6, when encountering IPv4-only router it is transformed to IPv4, later back to IPv6. Problem: information loss (e.g. flow label)
- ❖ *Tunneling*: IPv6 carried as payload in IPv4 datagram among IPv4-only routers

Tunneling



Tunneling



Next Lecture

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