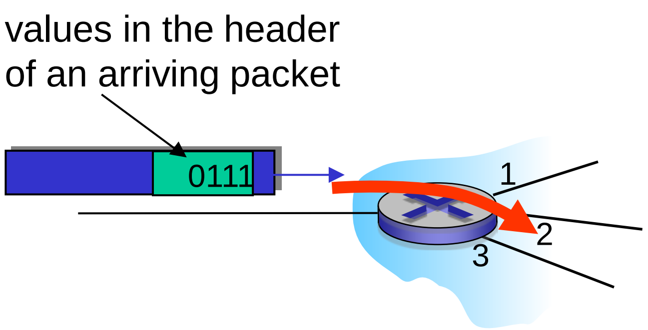
Chapter 14, 19

Network Layer
(Routing and Forwarding)

Network layer: Data plane, Control plane

Data plane

- Forwarding
  - forwards a datagram arriving on input port to an appropriate output port of the router, according to routing decision

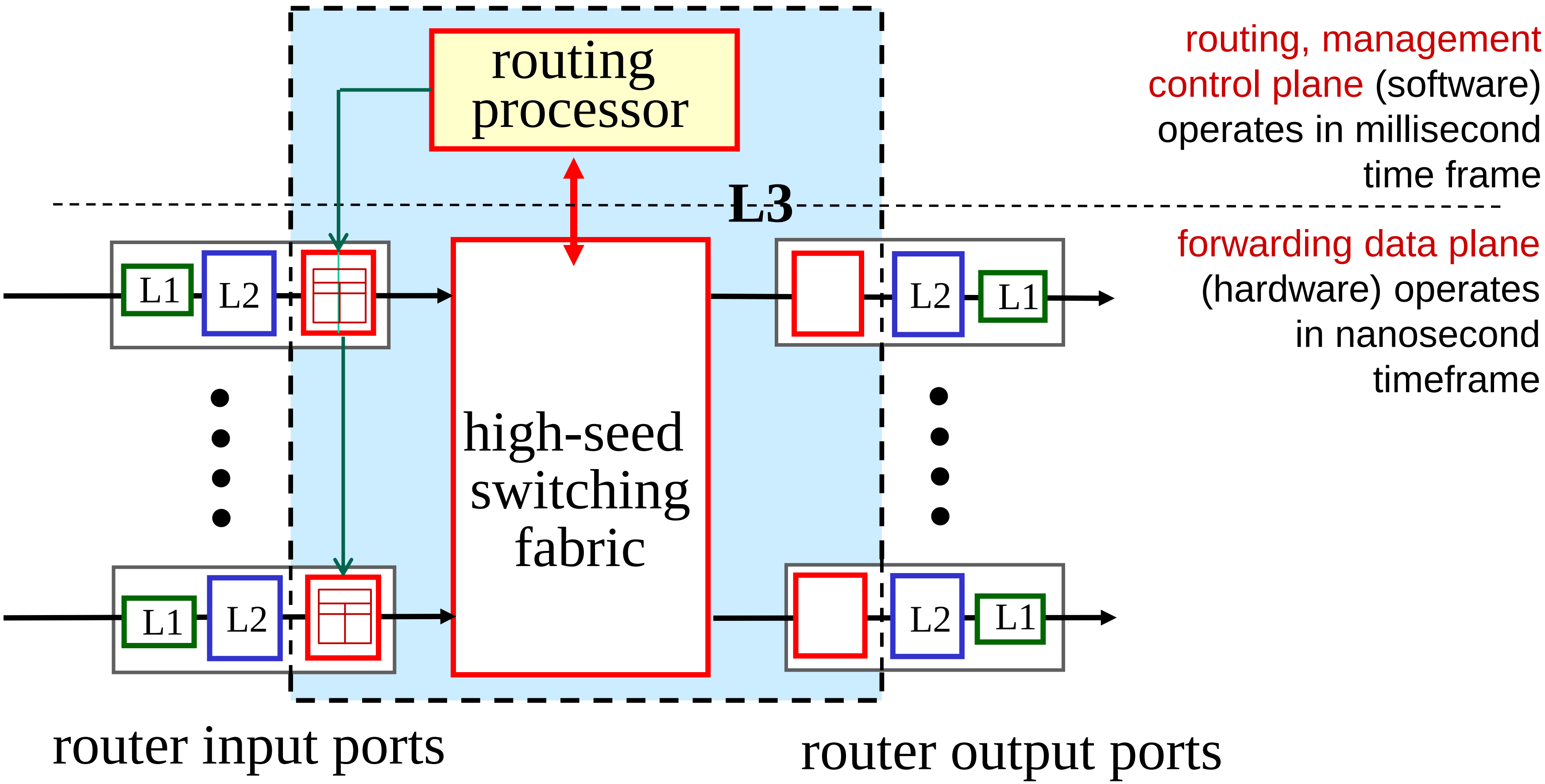


Control plane

- Routing:
  - determines how datagram is routed from source host to destination host
- two approaches:
  - Distributed Routing
    - traditional per-router control plane
    - implemented in routers
  - Centralized Routing
    - software-defined networking (SDN)
    - implemented in (remote) servers

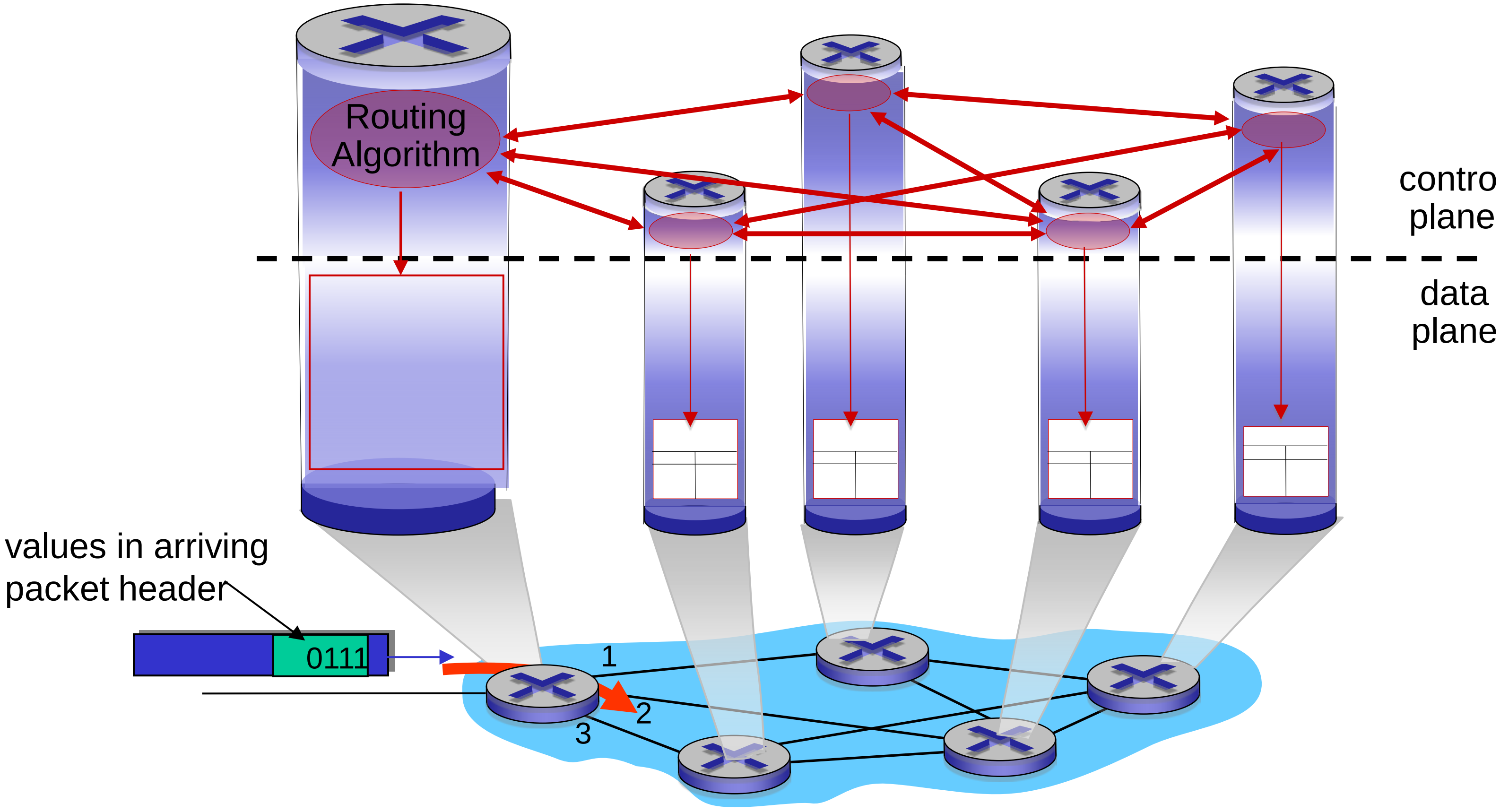
Router architecture overview

- High-level view of generic router architecture:



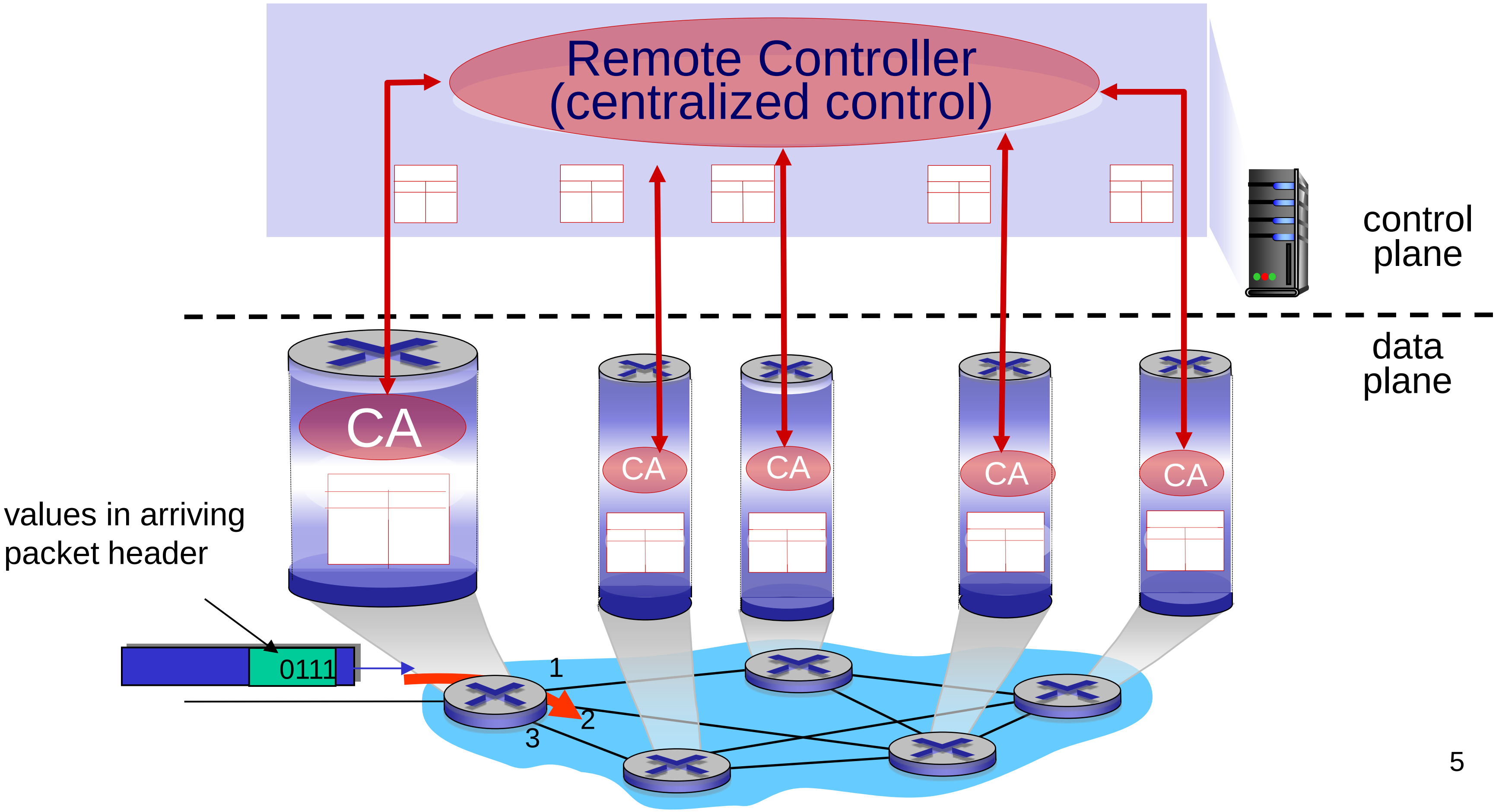
Per-router control plane

Individual routing algorithm components in each and every router interact in the control plane



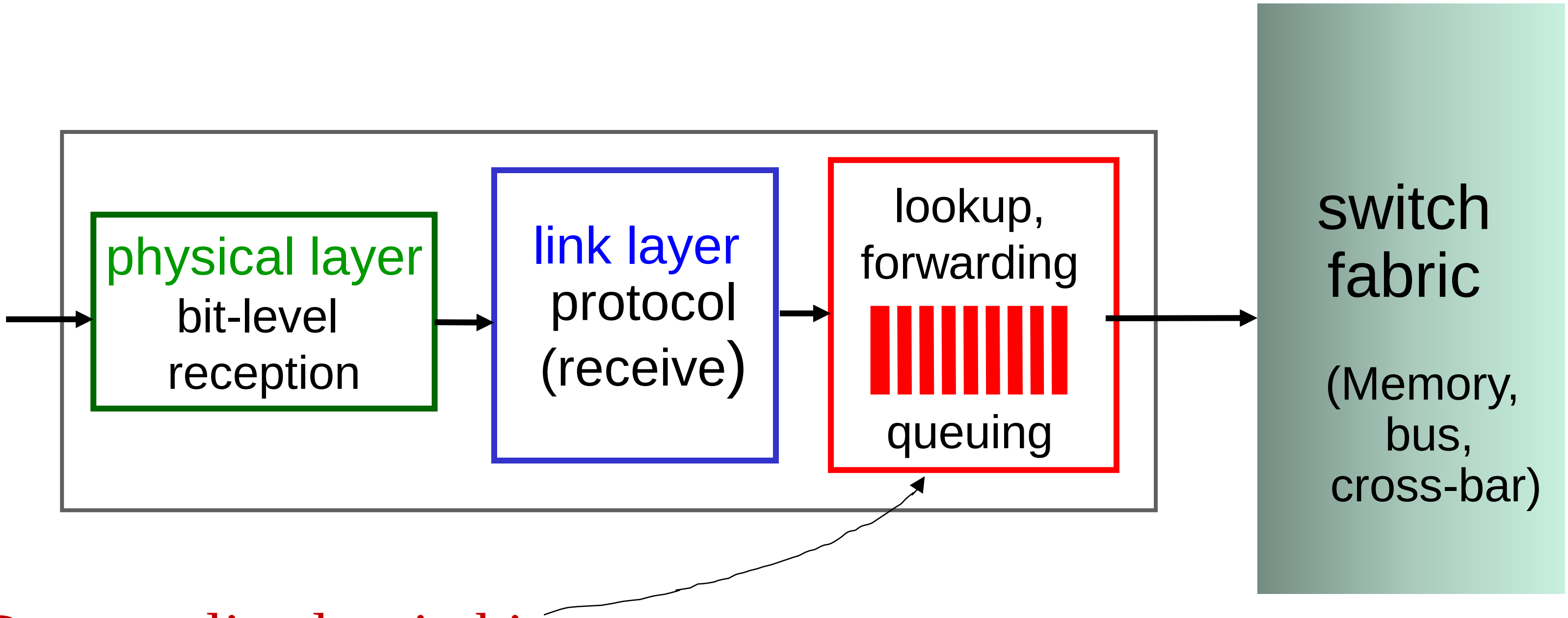
Logically centralized control plane

A distinct (typically remote) controller interacts with local control agents (CAs)



Forwarding

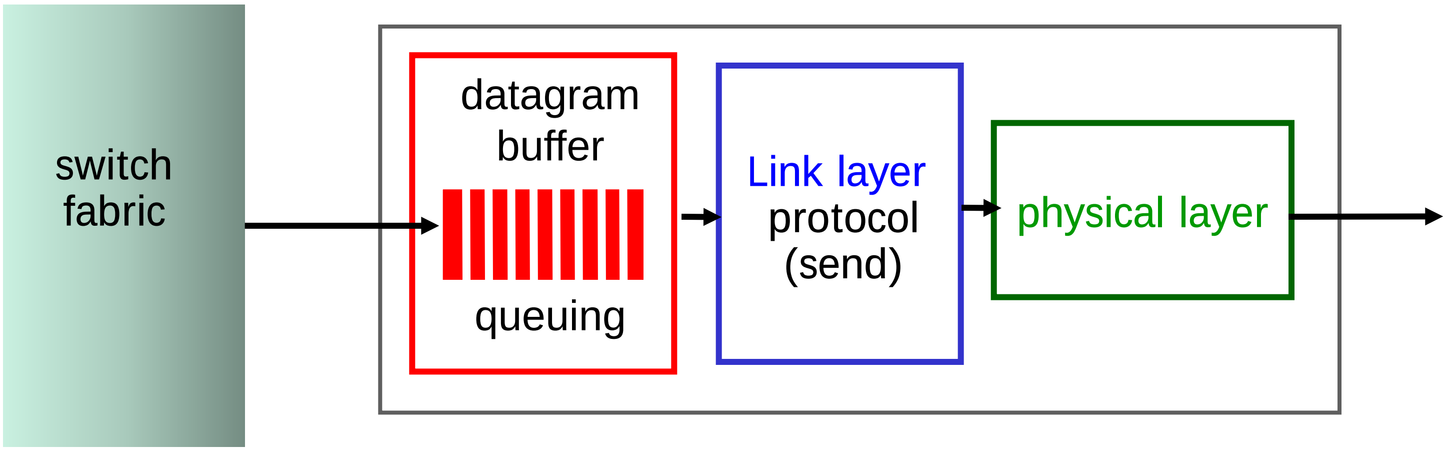
Input port functions



Decentralized switching:

- Based on header field values, lookup output port using forwarding table in input port memory ("match plus action")
  - destination-based forwarding: forward based only on destination IP address (traditional)
  - generalized forwarding: forward based on any set of header field values

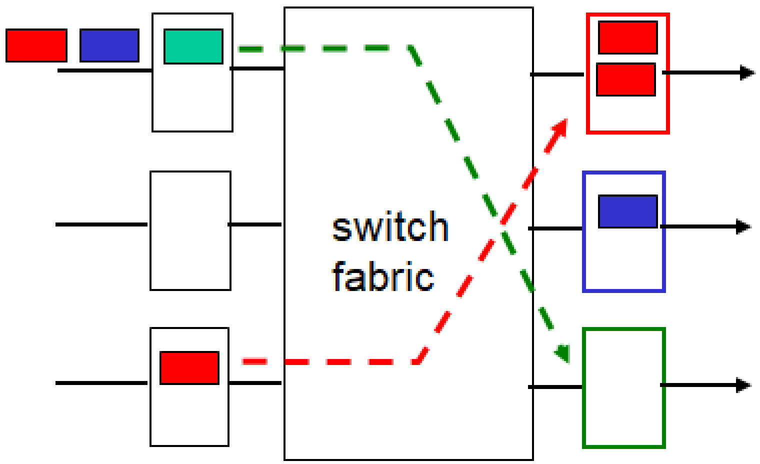
Output ports



- Buffering: when datagrams arrive from fabric faster than the transmission rate
  – Datagram (packets) can be lost due to lack of buffers by congestion
- Scheduling discipline: which one among queued datagrams is chosen for transmission (priority scheduling or FIFO)

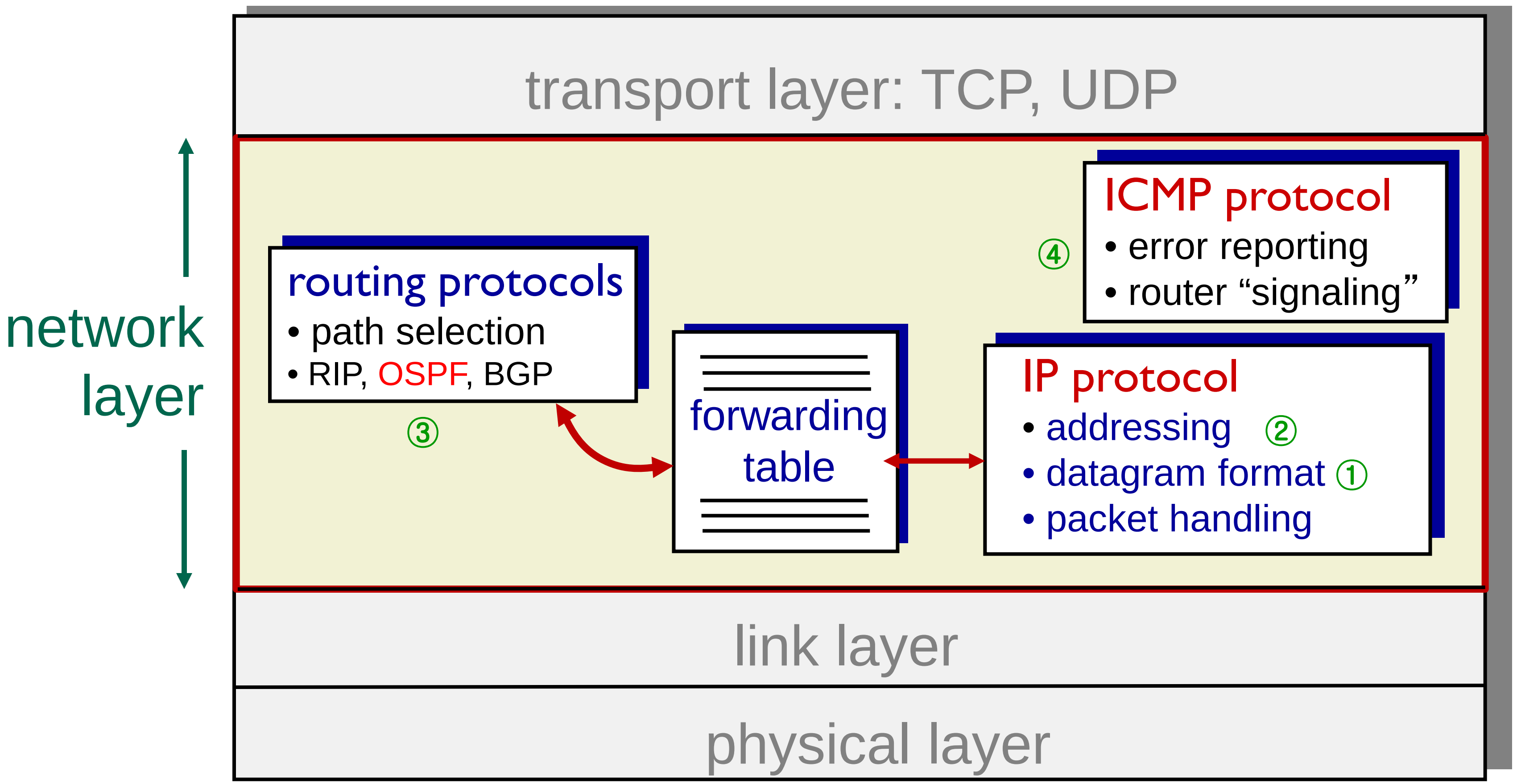
Queuing

- Input queuing
  - Switch fabric is slower than the arrival rates from input ports -> queuing may occur at input queues
- Output port queuing
  - buffering when arrival rate to output line via switch exceeds the output line speed
- queuing (delay) and loss due to output port buffer overflow!

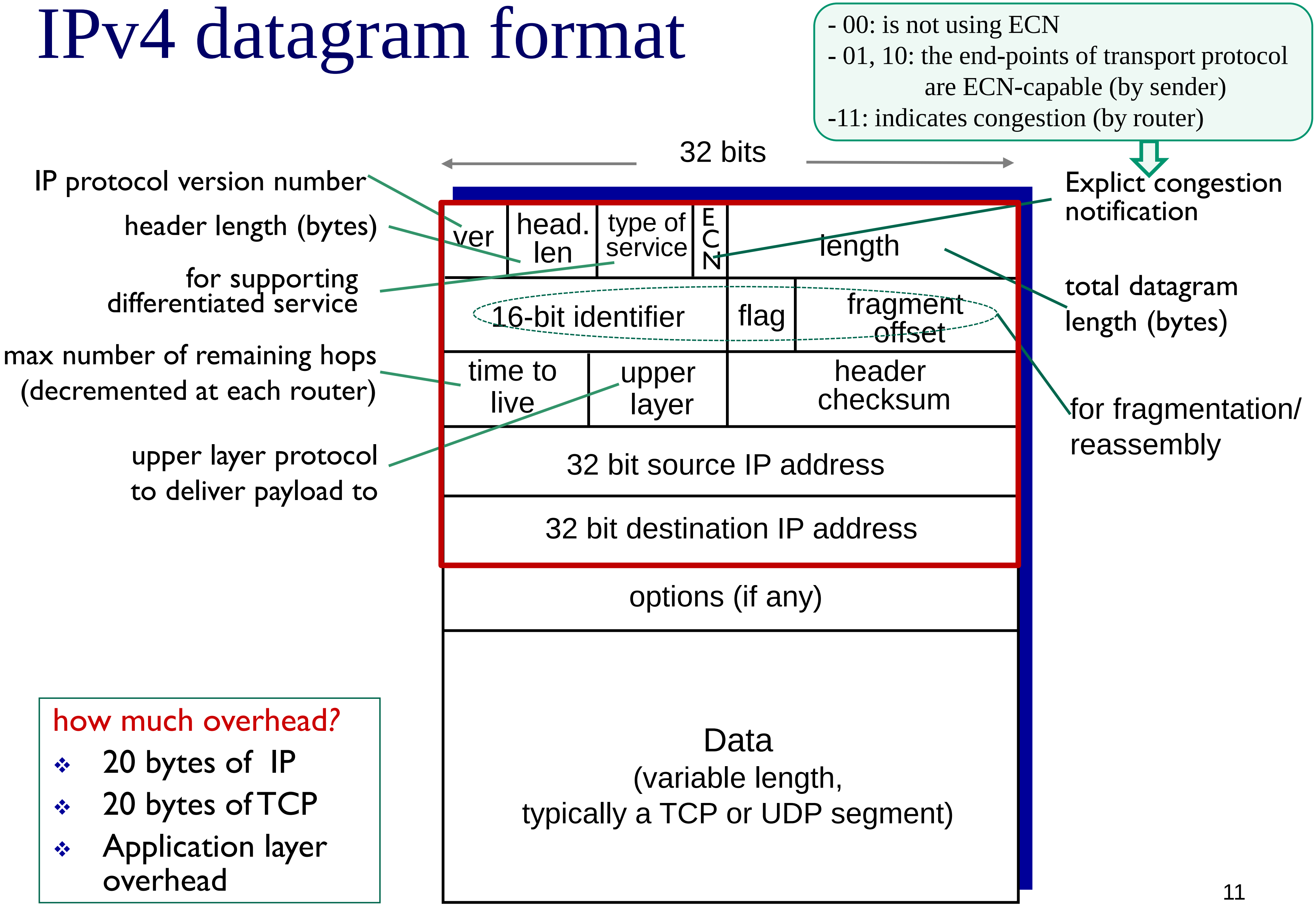


Internet network layer

Network layer functions of hosts and routers:



IPv4 datagram format

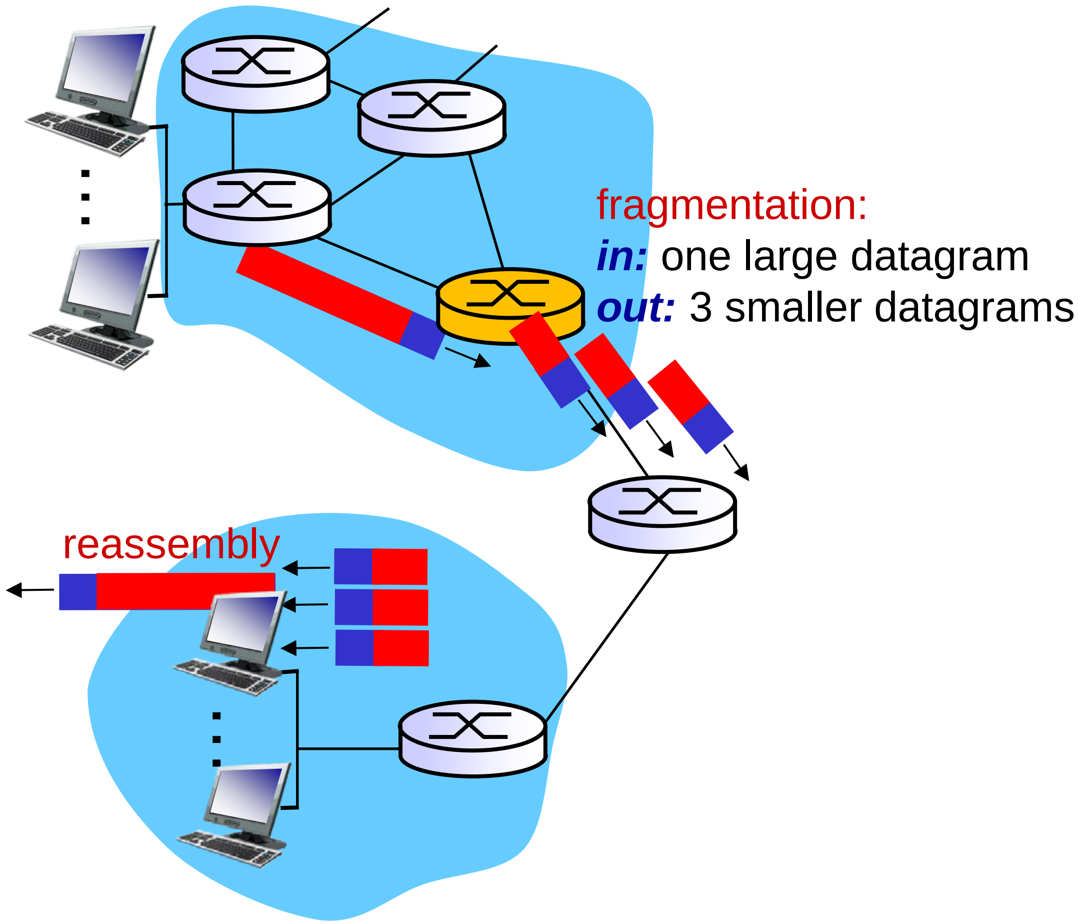


how much overhead?

- 20 bytes of IP
- 20 bytes of TCP
- Application layer overhead

IP fragmentation/reassembly

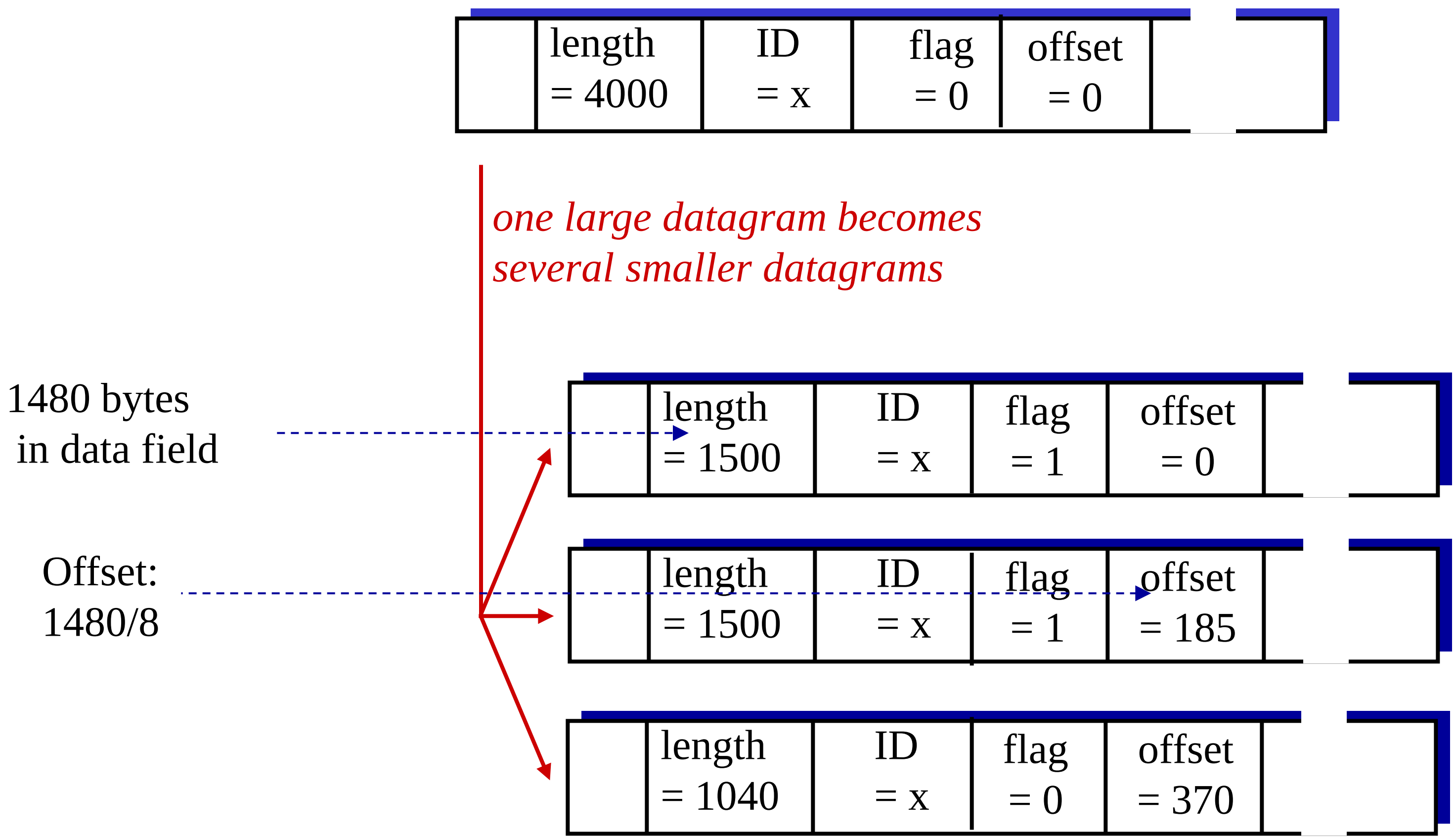
- Network links have MTU (max. transfer size) - largest possible link-level frame
- If the size of an IP datagram exceed the MTU of output network link, the datagram is divided ("fragmented")
  - one datagram becomes several small datagrams (fragments)
  - "reassembled" only at final destination
  - IP header bits are used to identify and order related fragments



IP fragmentation/reassembly

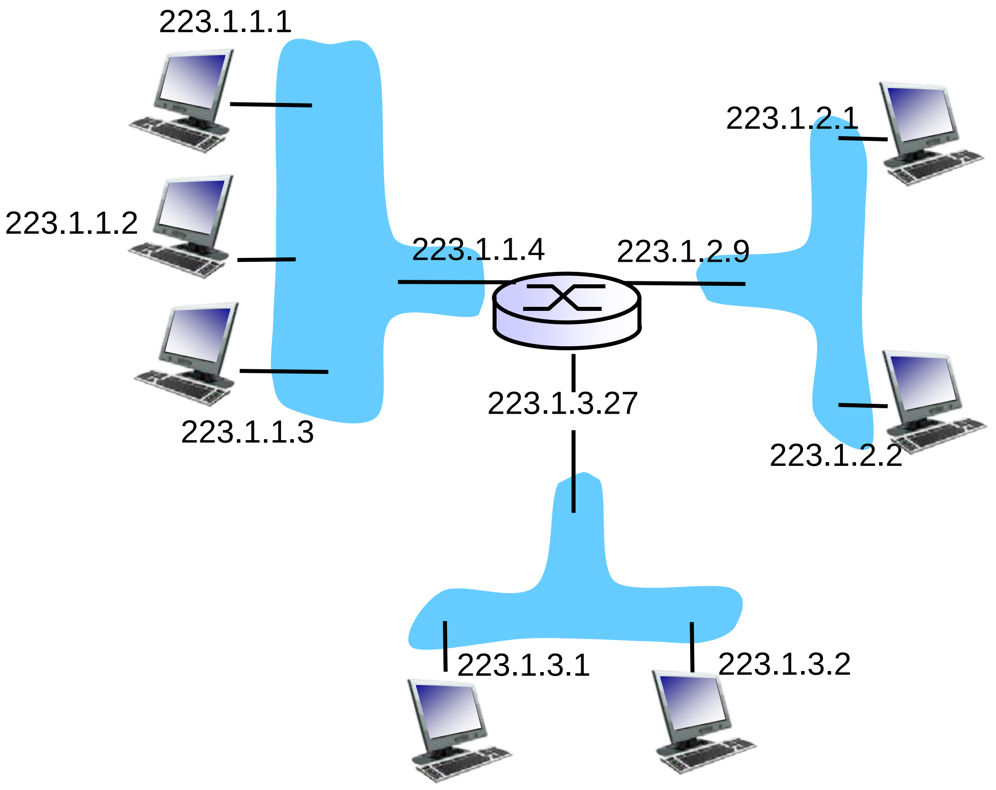
Example:

- 4000 byte datagram; MTU = 1500 bytes



IP addressing: introduction

- IP address: 32-bit identifier for host, router interface
- interface: connection between host/router and physical link
  - router typically have multiple interfaces
  - host typically has one or two interfaces (e.g., wired Ethernet, wireless 802.11)
- IP addresses associated with each interface

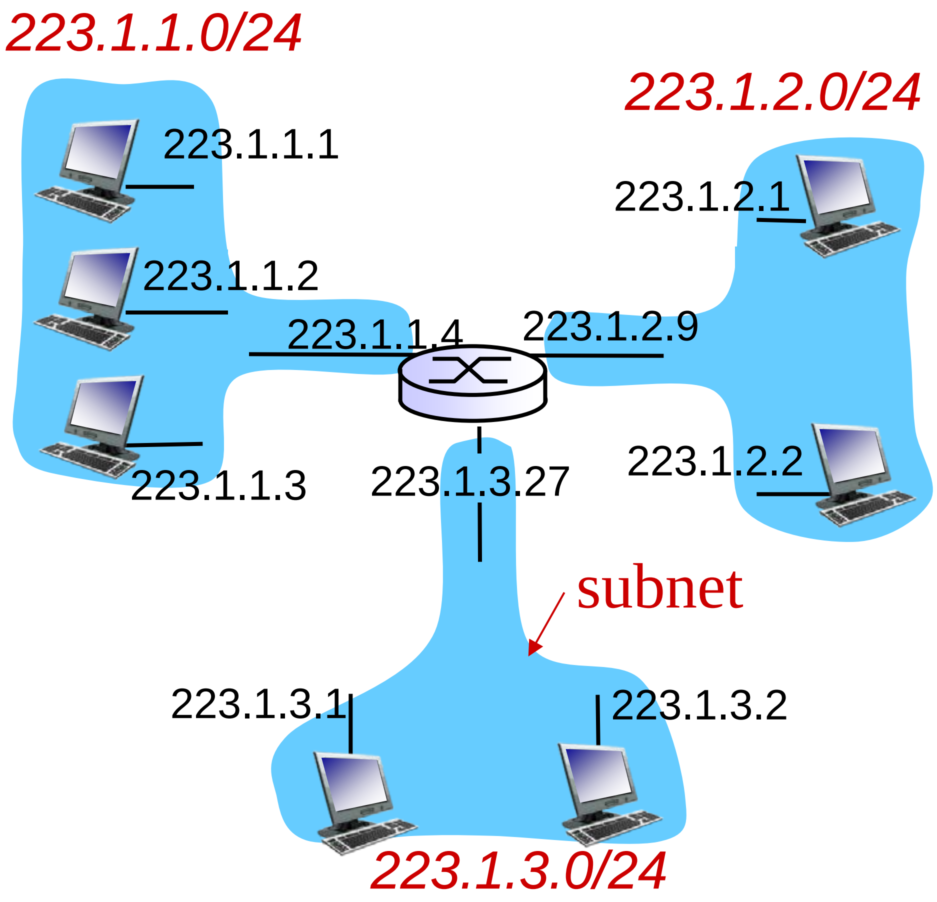


223.1.1.1 = 11011111 00000001 00000001 00000001

223 1 1 1

IP addressing: 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
  - (example) network consisting of 3 subnets
- subnet mask: /24



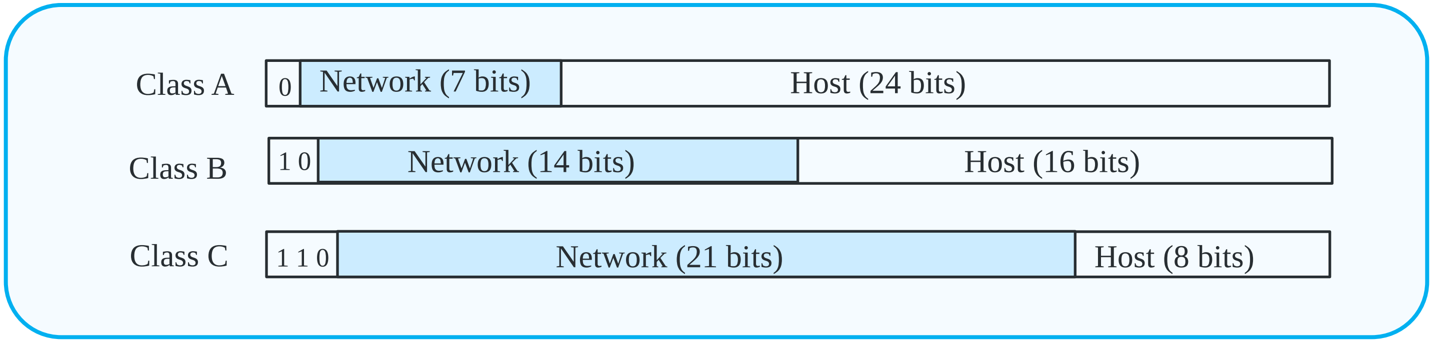
IP addressing: CIDR

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

← subnet part → ← host part →

11001000 00010111 00010000 00000000 : 200.23.16.0/23



IP addressing: 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:

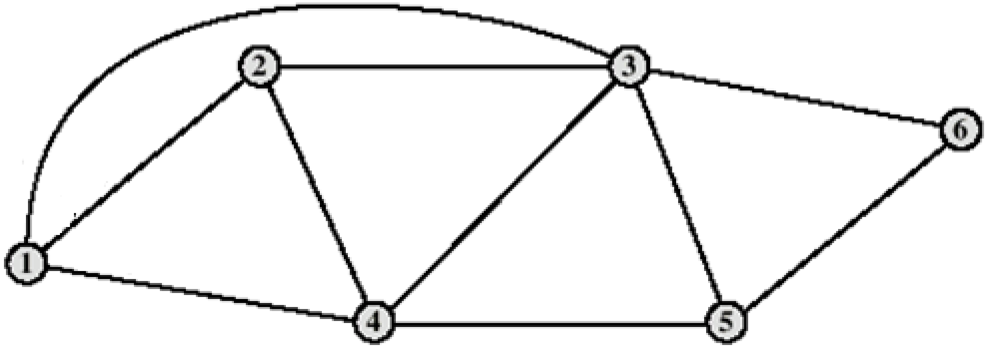
DA: 11001000 00010111 00010110 10100001 which interface?

DA: 11001000 00010111 00011000 10101010 which interface?

Routing

Routing Basics

- Routing: complex, crucial aspect of packet switched networks
- Routing criteria: for selection of route
  - Minimum hop
  - Least cost: shortest path algorithm
- Graph Modeling

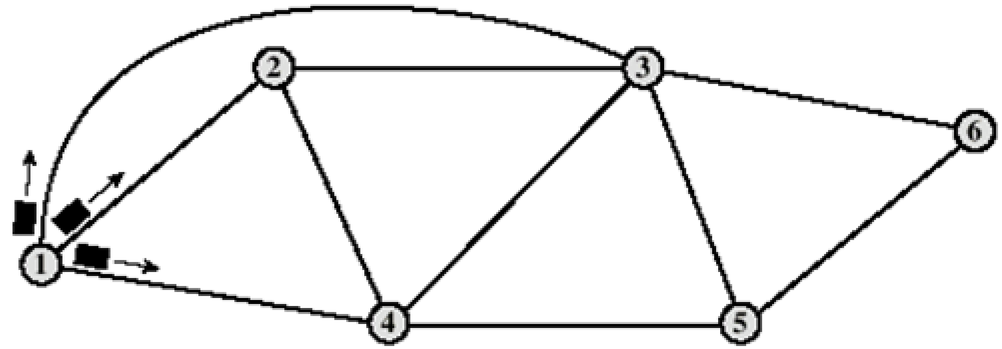


Basics: Routing Strategy

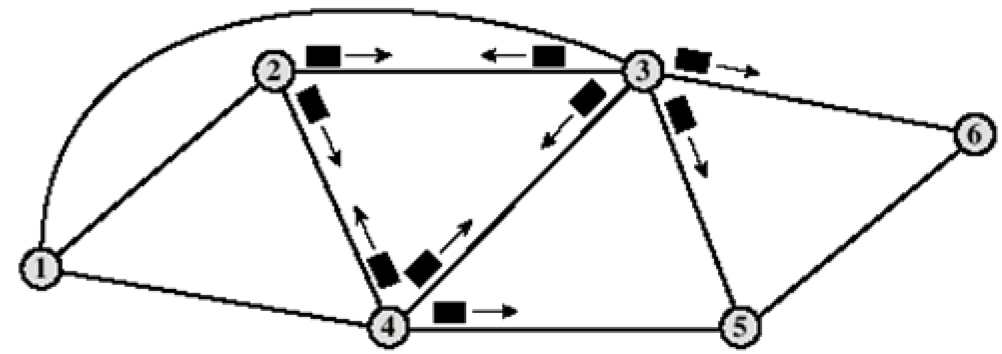
- Without Routing Table
  - Flooding
  - Random routing
- With Routing Table
  - Who is responsible for making a routing table
    - Centralized routing: a specialized central node
    - Distributed routing: each node makes its routing table
  - When the routing table is updated
    - Fixed: little updated
    - Adaptive: regular updates

Flooding

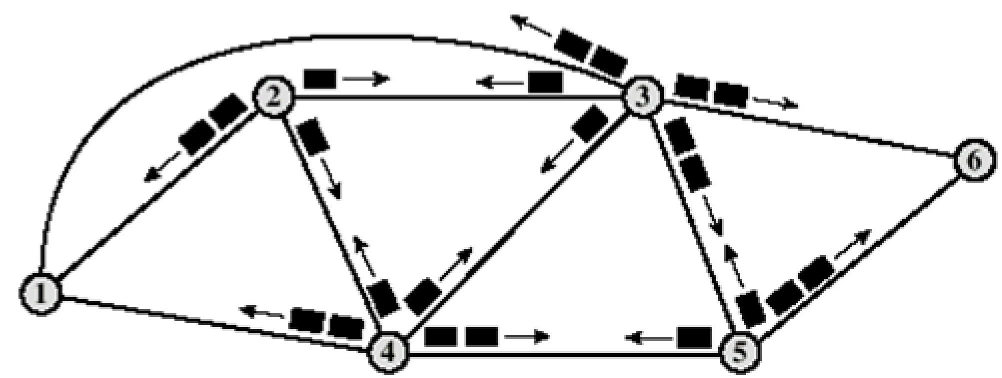
- Incoming packets retransmitted on every link except incoming link
- No network info required
- Eventually a number of copies will arrive at destination
- Each packet is uniquely numbered so duplicates can be discarded
- Nodes can remember packets already forwarded to keep network load in bounds
- Can include a hop count in packets



(a) First hop



(b) Second hop



(c) Third hop

Properties of Flooding

- All possible routes are tried
  - Very robust
- At least one packet will have taken minimum hop count route
  - Can be used to set up virtual circuit
- All nodes are visited
  - Useful to distribute information (e.g., routing)
- When the network topology dynamically changes

Strengths of Flooding-based routing in Ad-Hoc network:

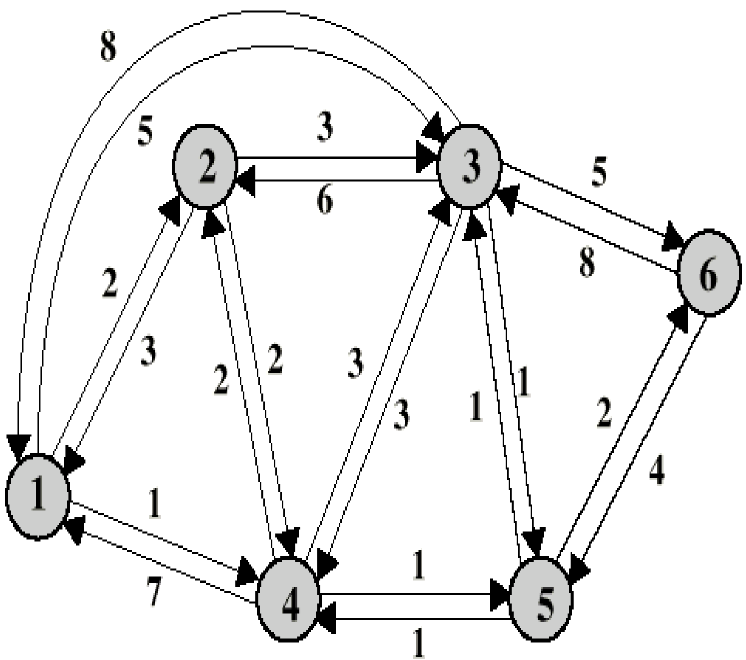
- Small-scale network
- Dynamic topology change (intermittent burst traffic)
- Broadcast property of wireless transmission (low overhead)

With Routing Tables

Making Routing Tables

CENTRAL ROUTING DIRECTORY

To Node \ From Node	1	2	3	4	5	6
1	—	1	5	2	4	5
2	2	—	5	2	4	5
3	4	3	—	5	3	5
4	4	4	5	—	4	5
5	4	4	5	5	—	5
6	4	4	5	5	6	—



Node 1 Directory

Destination	Next Node
2	2
3	4
4	4
5	4
6	4

Node 2 Directory

Destination	Next Node
1	1
3	3
4	4
5	4
6	4

Node 3 Directory

Destination	Next Node
1	5
2	5
4	5
5	5
6	5

Node 4 Directory

Destination	Next Node
1	2
2	2
3	5
5	5
6	5

Node 5 Directory

Destination	Next Node
1	4
2	4
3	3
4	4
6	6

Node 6 Directory

Destination	Next Node
1	5
2	5
3	5
4	5
5	5

Adaptive Routing

- Routing decisions change as the conditions on network change
  - Failure or Congestion
- Requires info about network
  - Information source
    - Adjacent nodes
    - All nodes
- Tradeoff between quality of network info and overhead
  - Reacting too quickly can cause oscillation
  - Reacting too slowly can be irrelevant

Internet Routing

- Per-router (distributed routing)
- With Routing Table
- Adaptive Routing

Internet Routing Architecture

- Internet architecture from routing's views
  - It is unrealistic to apply a single routing protocol to the worldwide Internet because of its size.
  - So, the worldwide Internet is divided into many groups, which are administered independently.
  - These independent groups of networks are called the autonomous systems (ASs) which are assigned 16 bits long AS number.
  - AS
    - a group of sub-networks and routers controlled by a single administrative authority
  - Each AS needs to inform its routing information of other ASs. For this purpose each AS has more than one border routers.

Internet Autonomous System (AS)

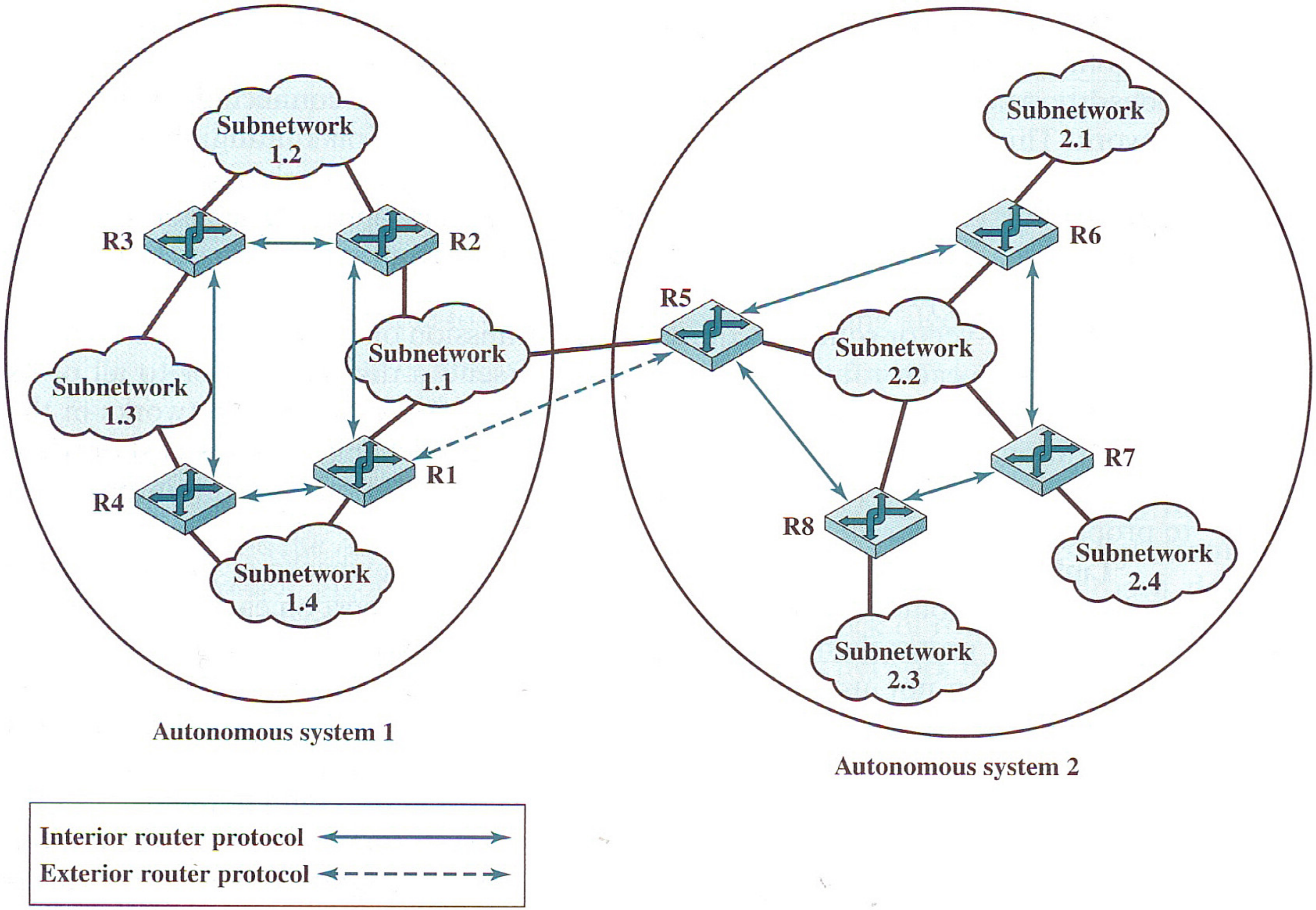


Figure 19.5 Application of Exterior and Interior Routing Protocols

Internet Routing Protocols

- Interior Gateway Protocol (IGP)
  - IGP is operated within each AS.
  - Each AS can operate its own IGP.
  - Most well-known IGPs
    - OSPF(Open Shortest Path First)
      - Link State Routing (information from all nodes)
      - Dijekstra's Algorithm
    - RIP(Routing Information Protocol)
      - Distance Vector Routing (information exchange with neighbors)
      - Bellman-Ford Algorithm
- Exterior Gateway Protocol (EGP)
  - To exchange packets between ASs, the border routers should exchange the routing information.
  - EGP is the routing protocol between ASs.
    - BGP(Border Gateway Protocol)

OSPF: Basic Principles

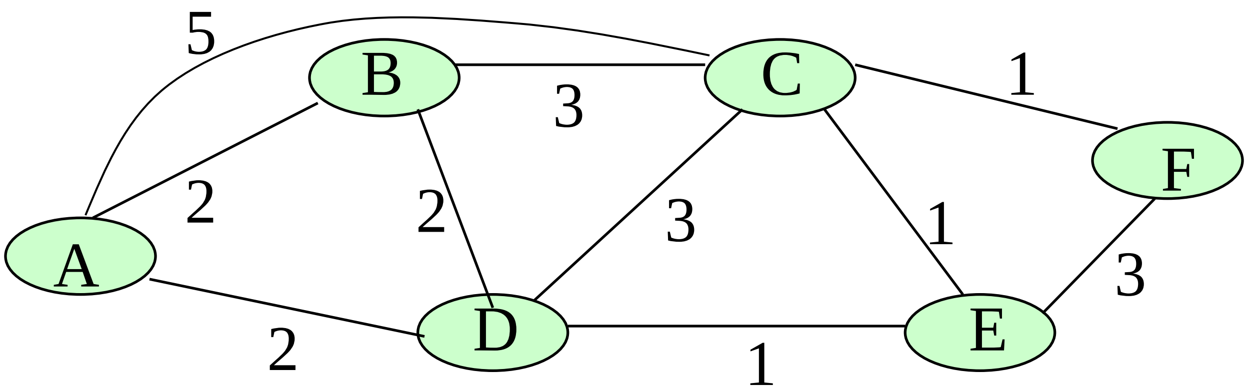
1. Each router establishes a relationship ("adjacency") with its neighbors (Hello packet exchange)
2. Each router generates link state advertisement (LSA) and distributes to all routers (Flooding)

   Router LSA: its presence, the links and metrics to neighbor routers, …
3. Each router maintains a database of all received LSAs (topological database or link state database), which describes the network as a graph with weighted edges
4. Each router uses its link state database to run a shortest path algorithm (Dijekstra's algorithm) to produce the shortest path to each router

Example: Link State Routing

A router should collect the link state information from the other routers.

- Make the link state advertisement packet and flooding



Link state advertisements

A	
seq#	
age	
B	2
C	5
D	2

B	
seq#	
age	
A	2
C	3
D	2

C	
seq#	
age	
A	5
B	3
D	3
E	1
F	1

D	
seq#	
age	
A	2
B	2
C	3
E	1

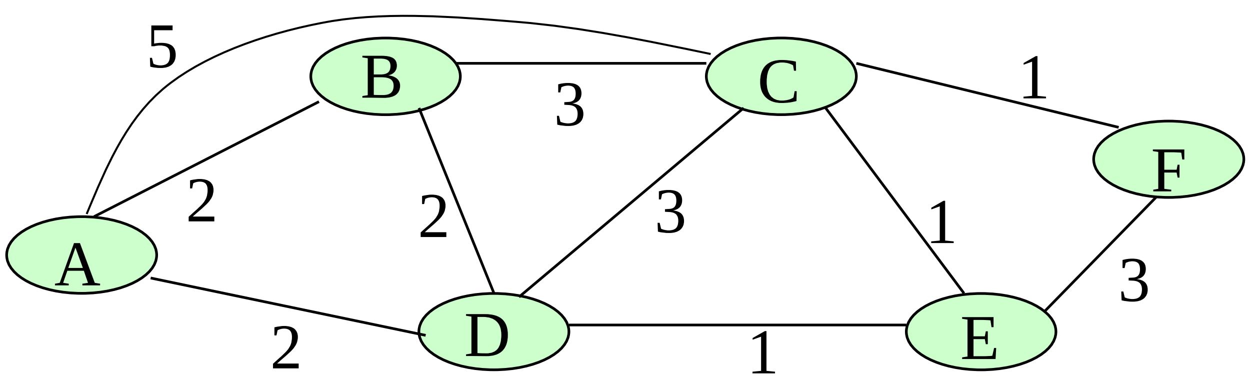
E	
seq#	
age	
C	1
D	1
F	3

F	
seq#	
age	
C	1
E	3

Example: Link State Database

- Based on the collected link state information, the router (node) makes the link state database, which represents the whole network topology.

Link #	Cost	Link #	Cost	Link #	Cost
A-B	2	C-B	3	D-E	1
A-C	5	C-D	3	E-C	1
A-D	2	C-E	1	E-D	1
B-A	2	C-F	1	E-F	3
B-C	3	D-A	2	F-C	1
B-D	2	D-B	2	F-E	3
C-A	5	D-C	3		



Routing Table

Dest	OutLink
B	B
C	D
D	D
E	D
F	D

Sample AS

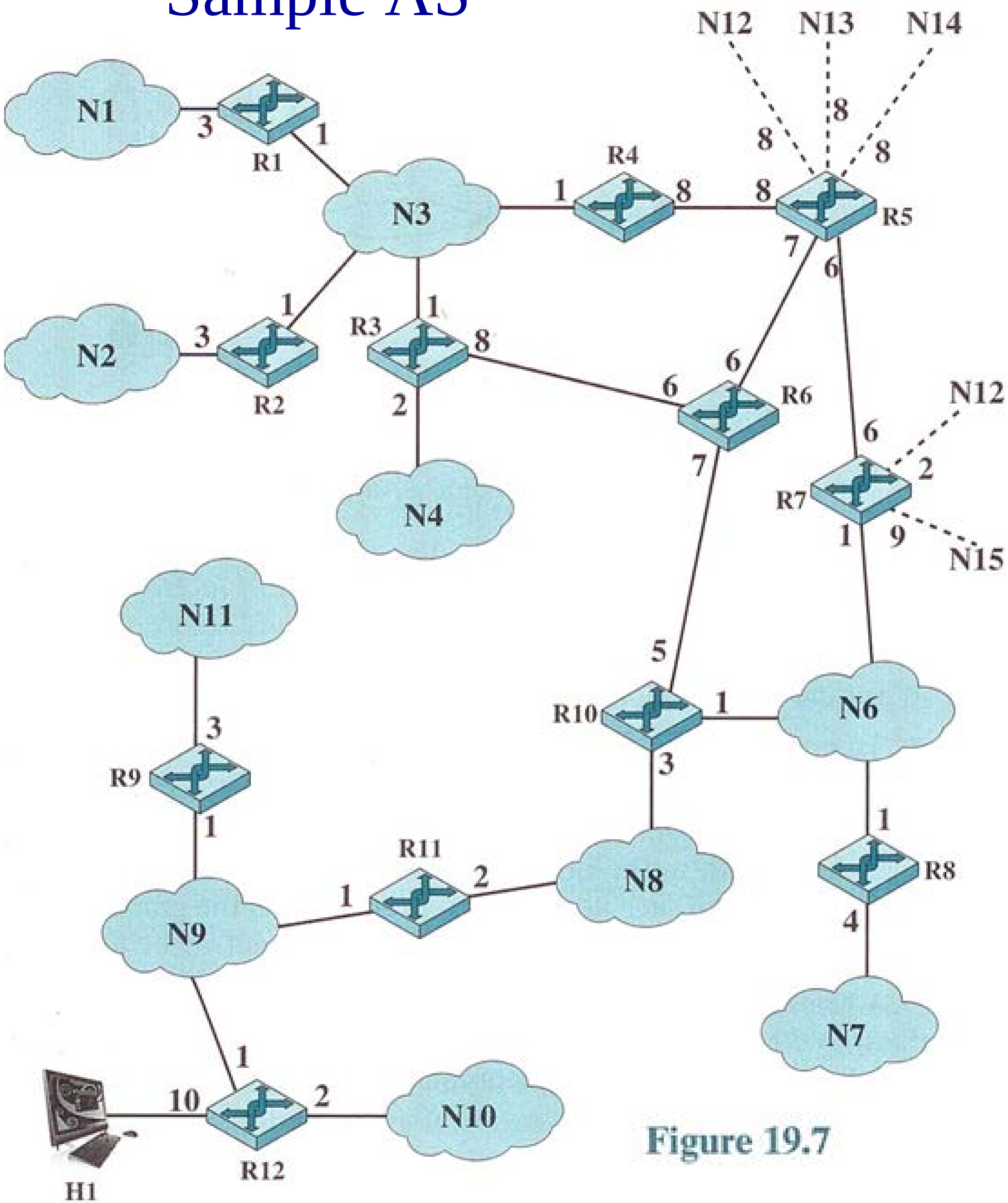


Figure 19.7

Open Shortest Path First

- Link State Routing
- Link State
  - Throughput
  - Delay
  - Cost
  - Packet error rate

OSPF: Graph Modeling

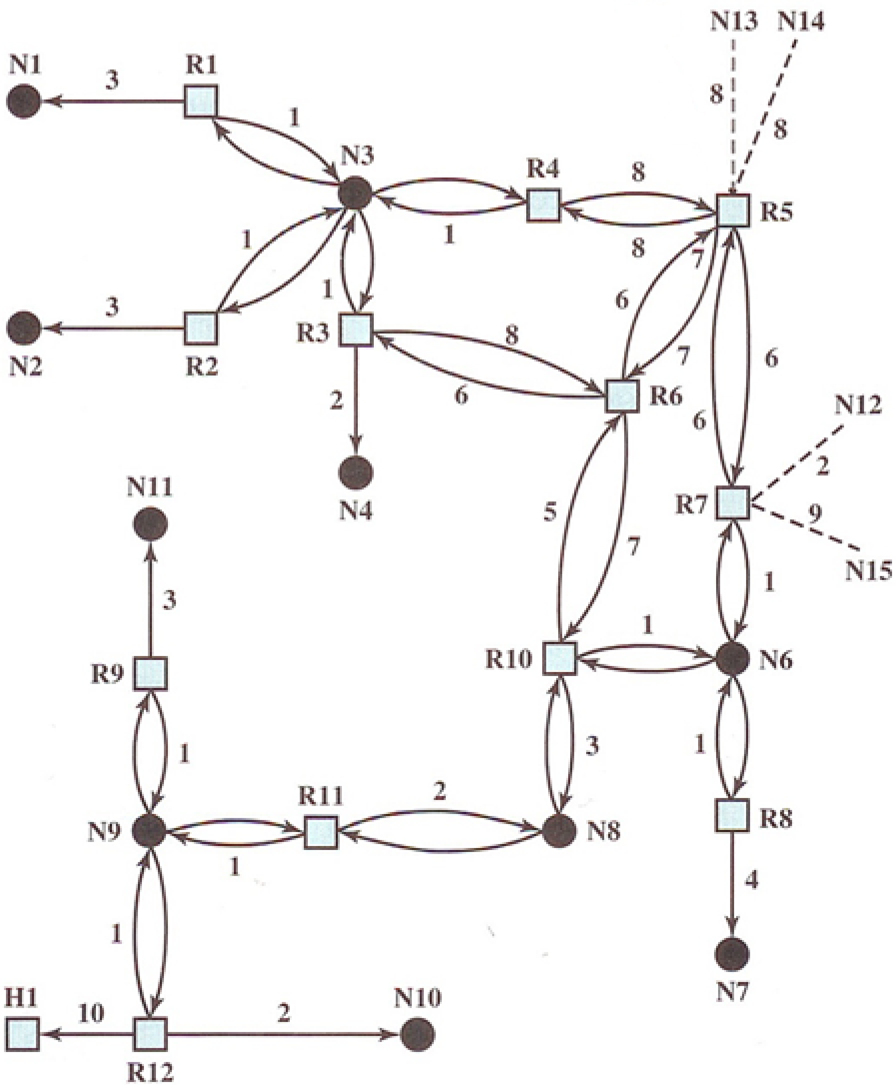


Figure 19.8 Directed Graph of Autonomous System of Figure 19.7

- Vertex
  - Router
  - Network
- A pair of directional edges
  - between two routers
  - between a transit network and router
- Directional edge from router to a stub net.
- Directional edge from router to a host

OSPF: Routing Table at R6

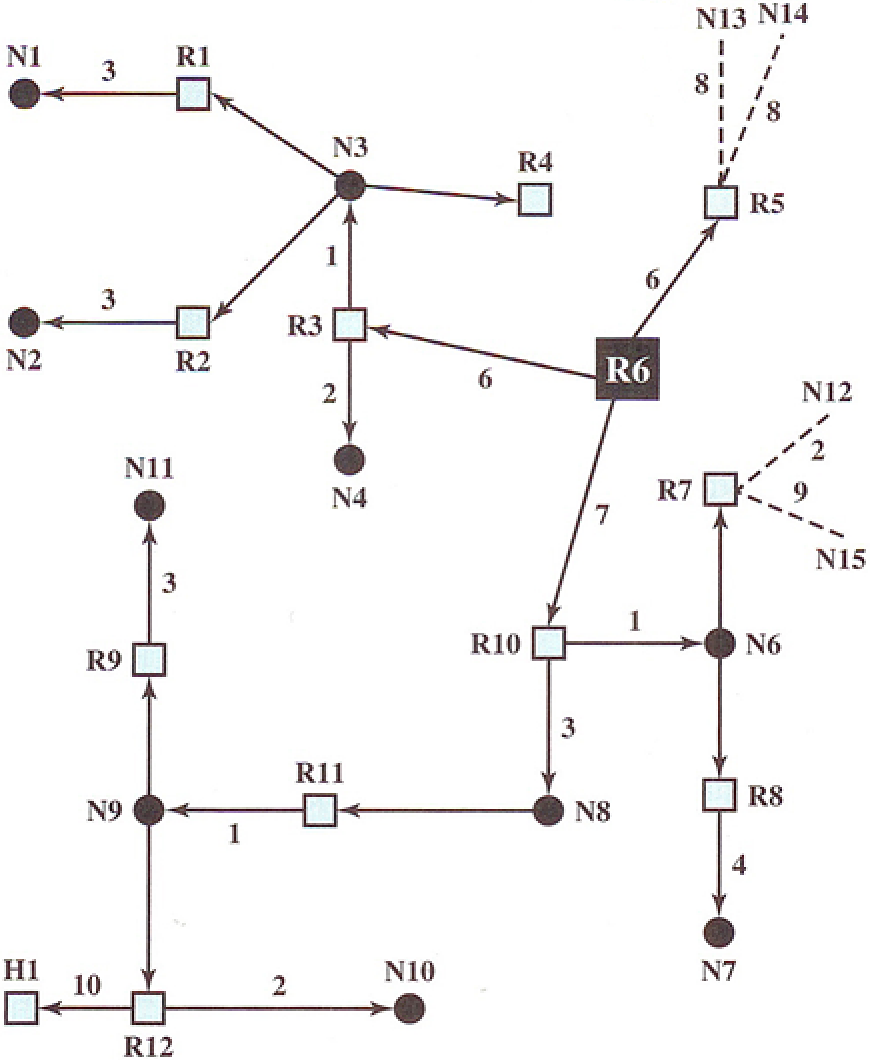


Figure 19.9 The SPF Tree for Router R6

Destination	Next Hop	Distance
N1	R3	10
N2	R3	10
N3	R3	7
N4	R3	8
N6	R10	8
N7	R10	12
N8	R10	10
N9	R10	11
N10	R10	13
N11	R10	14
H1	R10	21
R5	R5	6
R7	R10	8
N12	R10	10
N13	R5	14
N14	R5	14
N15	R10	17

ICMP
(Internet Control Message Protocol)

ICMP

- used by hosts & routers to exchange network-level information
  - Error reporting: unreachable host, network, port, protocol
  - Simple query (echo request/reply)
- network-layer "above" IP:
  - ICMP messages are carried in IP datagrams
- ICMP message format:
  - type (1 byte) and code (1 byte)
  - IP header + first 8 bytes of IP datagram payload, 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

0	8	16	31
Type	Code	Checksum	
Unused			
IP header + 8 bytes of original datagram			

dest unreachable, source quench, TTL expired

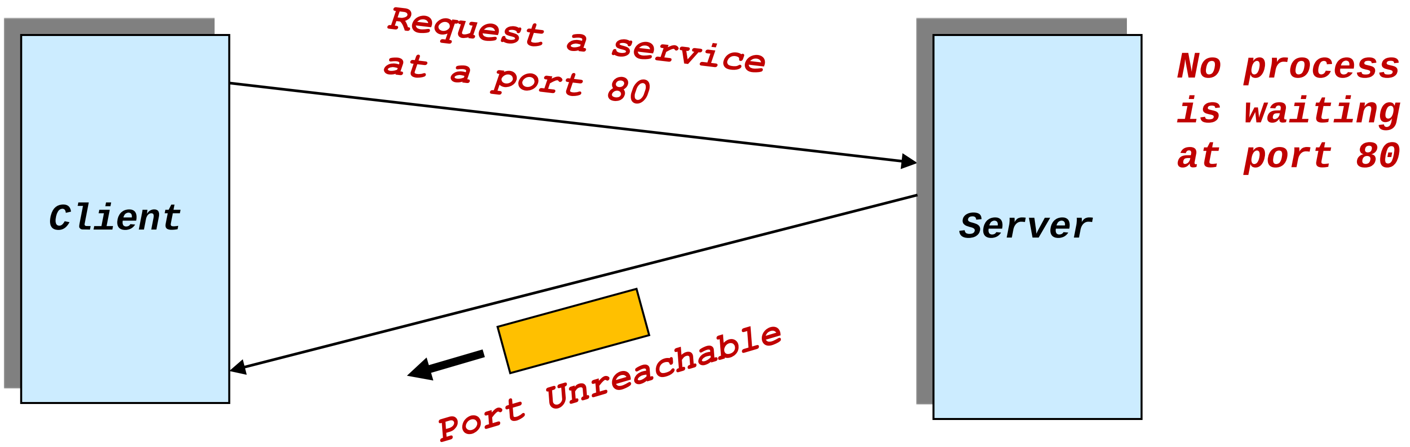
0	8	16	31
Type	Code	Checksum	
Identifier		Sequence number	
Optional data			

echo request, echo reply

Example

Destination Process/Port Unreachable

- If, in the destination host, the IP module cannot deliver the datagram because the indicated protocol module or process port is not active, the destination host may send a destination unreachable message to the source host.
- Scenario

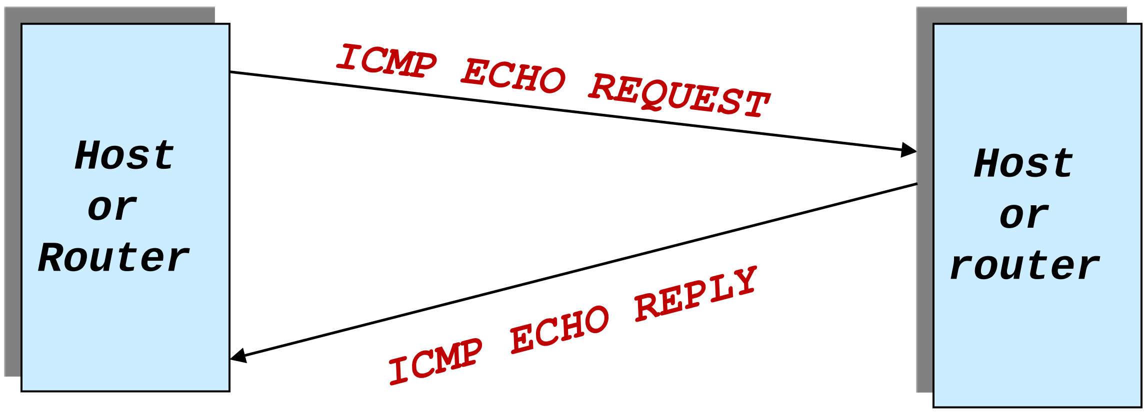


port 80: well-known http port

Example

Ping: Echo Request/Echo Reply

- Each Ping is translated into an Echo Request
- The Ping'ed host responds with an Echo Reply



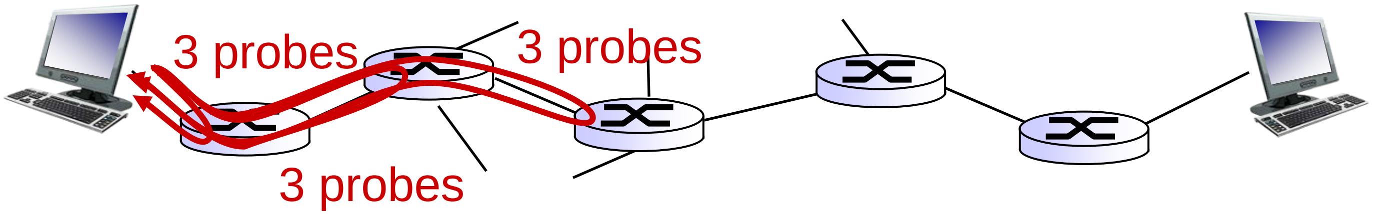
Example: Traceroute

- The source host sends a series of UDP segments to destination
  - first set has TTL =1
  - second set has TTL=2, etc.
  - unlikely port number
- when datagram in *n*th set arrives to *n*th router:
  - router discards datagram and sends ICMP message (type 11, code 0: "TTL expired") to source host
  - ICMP message include name of router & IP address of router

- when ICMP message arrives, source records RTTs

stopping criteria:

- UDP segment eventually arrives at destination host
- destination returns ICMP "port unreachable" message (type 3, code 3)
- source stops



System Congestion

- Situation that a system is loaded beyond its capacity
- Effects:
  Performance degradation

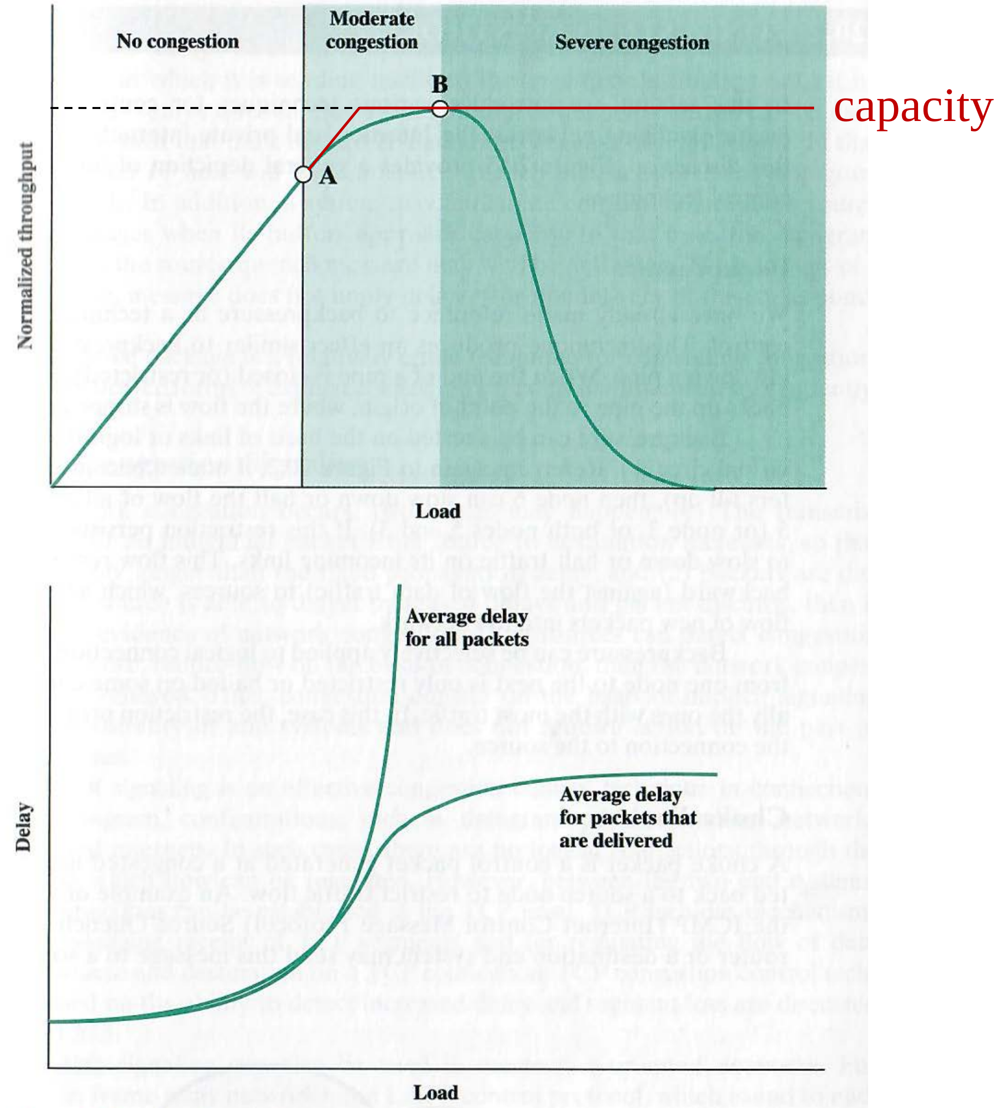


Figure 20.4 The Effects of Congestion

Solution:
Resource Allocation, Congestion Control

- Two ways to solve the network congestion
  - Resource Allocation
    - Request resource before sending packets
    - Limits the sending rate to the agreed amount
  - Congestion Control
    - Send packets
    - If congestion occurs, reduce the data rate
- General Principles of Congestion Control
  - Monitor the system for detecting when and where congestion occurs.
  - Congestion notification to places where action can be taken.
  - Adjust the data rate to solve the problem.

Congestion Signaling

- Explicit signaling
  - The network alerts end systems of growing congestion
  - End systems take steps to reduce the offered load
  - Example: ICMP source quench, ECN(IP)+ECE(TCP) flag
- Implicit signaling
  - A packet loss may occur
  - Transmission delay seriously increases.
  - The source detects this an implicit congestion indication
  - Less accurate but smaller overhead
  - Example: TCP congestion control (lecture-13)

Choke Packet

- A router monitors the utilization of each output line
- The output line: a warning state if utilization > threshold.
- Each arriving packet is checked if its output line is in warning state.
- If so, the router sends a choke packet back to source host
- When the source gets a choke packet, it reduces the traffic sent to the destination by *x* percent
- The source host ignores choke packets referring to that destination for a predefined time interval
- If no choke packets arrive during listen interval, the host may increases the flow rate.

- ICMP source quench
  (a kind of choke packet)
  - from router or destination to source
    - when it must discard IP packets because of full buffer
    - when its buffers approach capacity
  - source cuts back the sending rate until it no longer receives source quench datagrams
  - little used owing to overhead

Internet: Network-assisted congestion control

- Explicit Congestion Notification (ECN)
  - Two bits in IP header marked by router to indicate congestion
  - Congestion indication carried to receiving host
  - TCP receiver (seeing congestion indication in IP datagram) sets ECE b it (in TCP header) of receiver-to-sender TCP ACK segment to notify s ender of congestion
  - TCP sender: after reducing the date rate, CWR flag setting