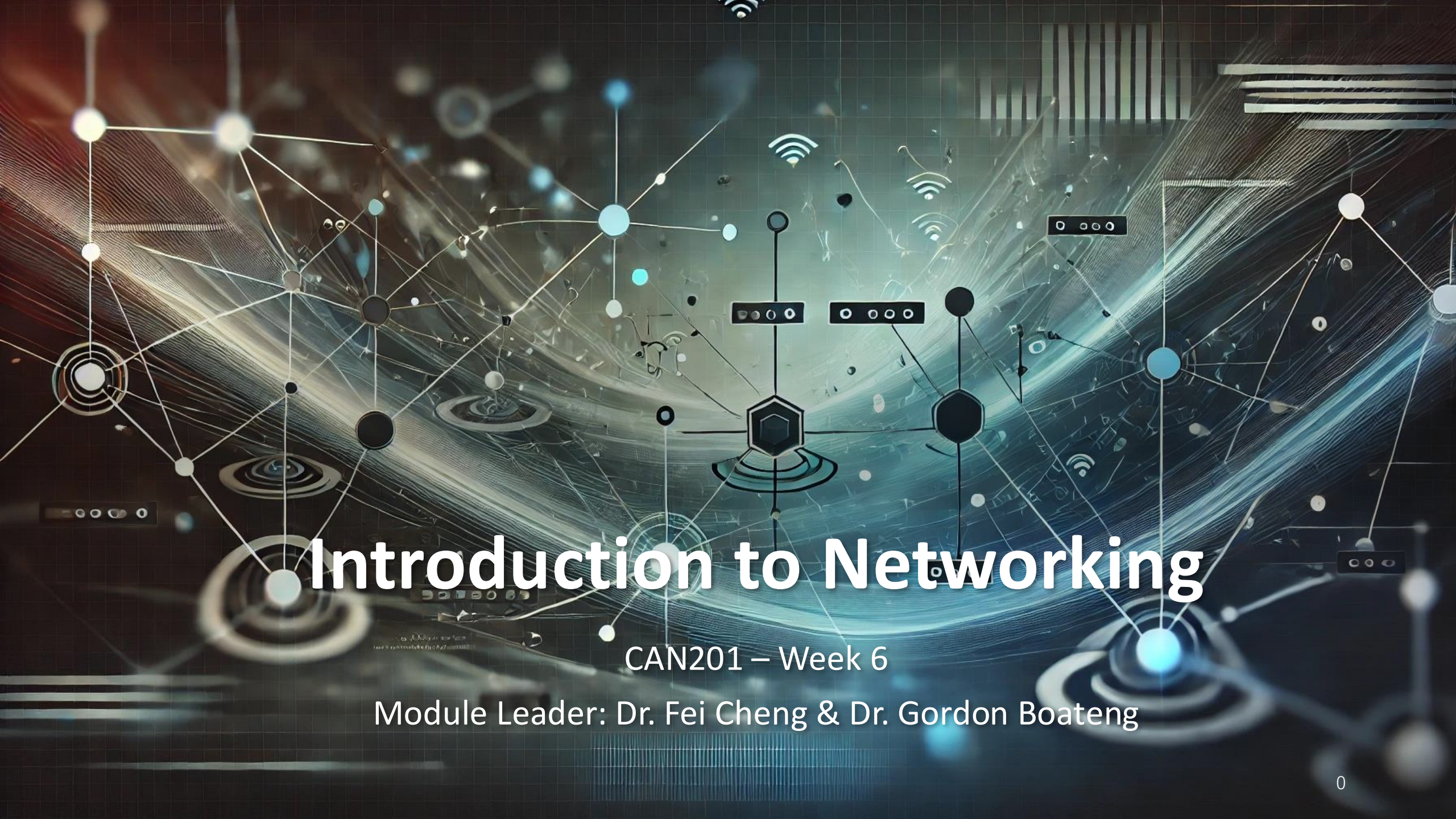


Introduction to Networking

CAN201 – Week 6

Module Leader: Dr. Fei Cheng & Dr. Gordon Boateng

Lecture 6 – Transport Layer (3)

- **Roadmap**
  1. TCP congestion control

Principles of congestion control

Congestion:

- "too many sources sending too much data too fast for *network* to handle"(informally)
- Manifestations:
  - long delays (queueing in router buffers)
  - packet loss (buffer overflow at routers)
- Different from flow control!

congestion control:
too many senders,
sending too fast

flow control: one sender
too fast for one receiver

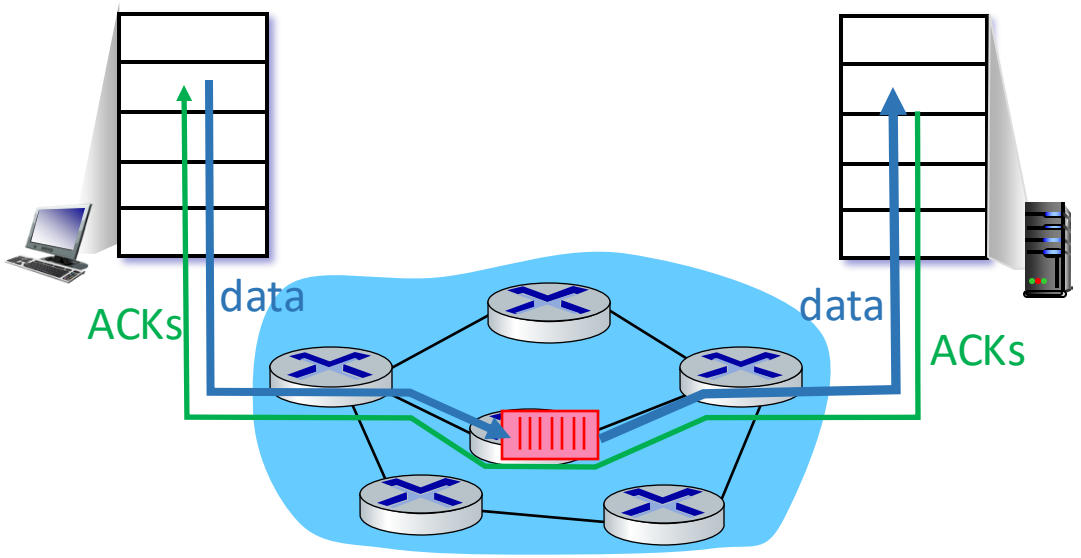
Causes and Costs of Network Congestion

- **Load increases:**
  - Many hosts send data at the same time, increasing the number of packets routers must process.
- **Queues build up:**
  - If packets arrive faster than they can be forwarded, they start to queue in router buffers.
- **Buffers overflow:**
  - When buffers are full, new packets are dropped.
- **Retransmissions add load:**
  - Dropped packets trigger retransmissions, adding more traffic.
- **=> Result: Higher delay, lower throughput, wasted bandwidth.**
- **Congestion occurs when the input rate exceeds the network's capacity.**

Approaches towards congestion control

End-end congestion control:

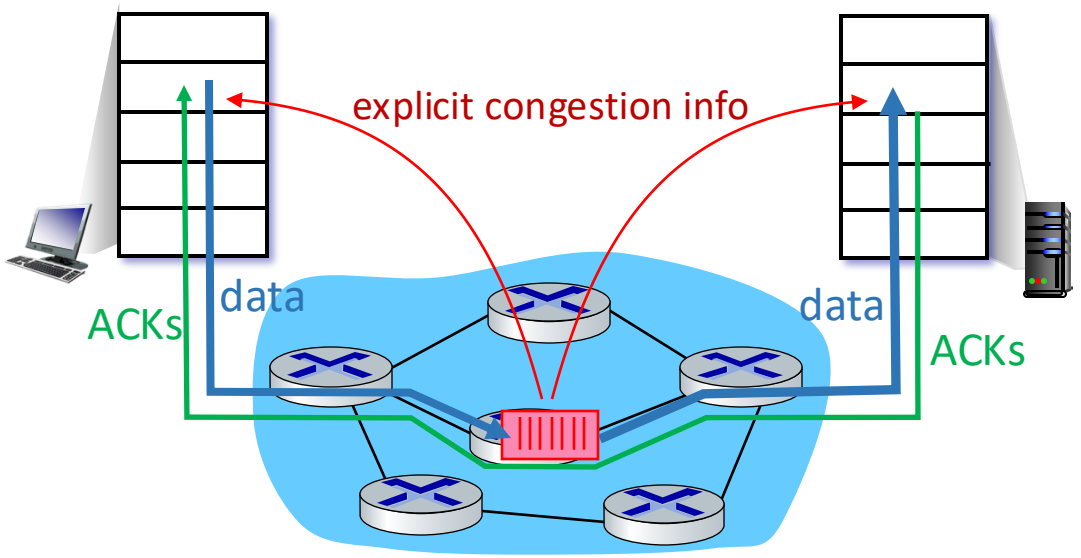
- no explicit feedback from network
- congestion *inferred* from observed loss, delay
  - approach taken by TCP



Approaches towards congestion control

Network-assisted congestion control:

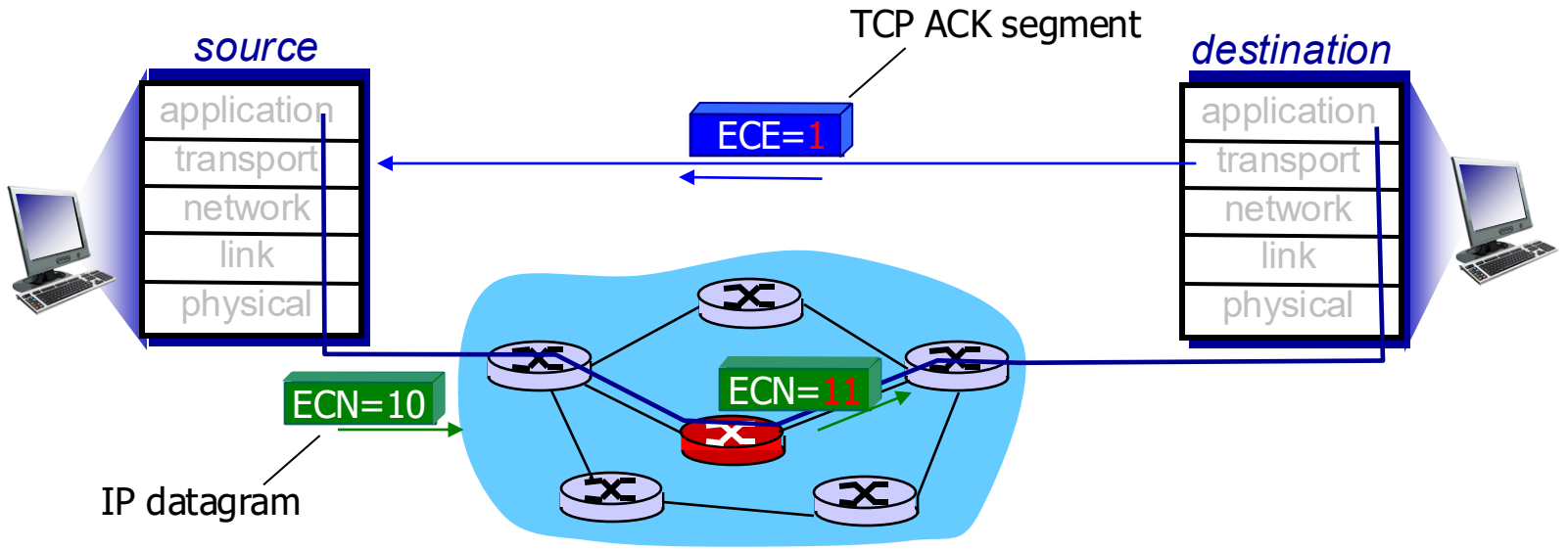
- routers provide *direct* feedback to sending/receiving hosts with flows passing through congested router
- may indicate congestion level or explicitly set sending rate
- TCP ECN …



Explicit Congestion Notification (ECN)

Network-assisted congestion control:

- Two bits in IP header (ToS field) marked *by network router* to indicate congestion
- Congestion indication carried to receiving host
- Receiver (seeing congestion indication in IP datagram)) sets ECE bit on receiver-to-sender ACK segment to notify sender of congestion



ECN:
00 Not-ECT*
01 ECT
10 ECT
11 CE**

*ECN-Capable Transport
** CE: Congestion Experienced

TCP congestion control: AIMD

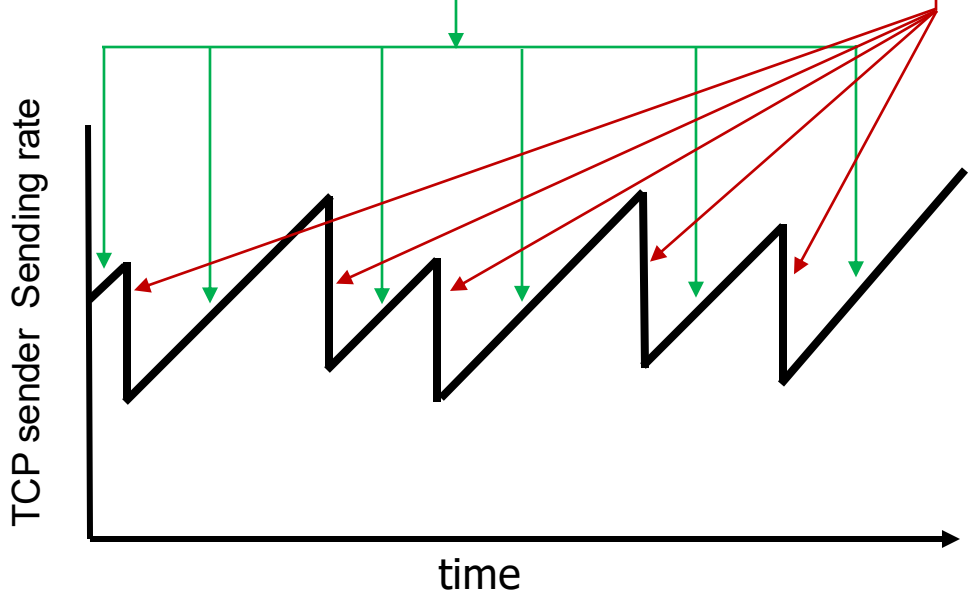
- *approach:* senders can increase sending rate until packet loss (congestion) occurs, then decrease sending rate on loss event

Additive Increase

increase sending rate by 1 maximum segment size every RTT until loss detected

Multiplicative Decrease

cut sending rate in half at each loss event



AIMD sawtooth behavior: *probing* for bandwidth

TCP AIMD: more

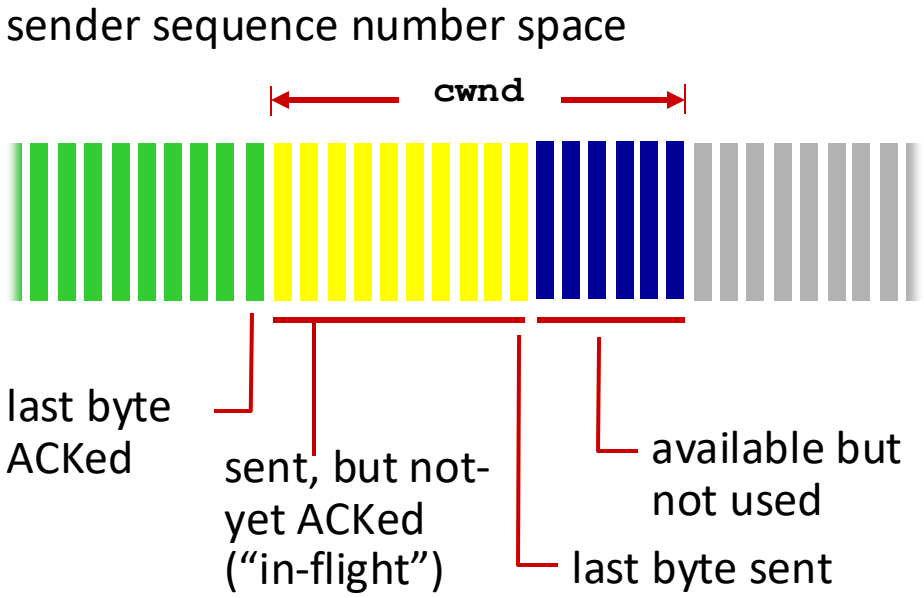
Multiplicative decrease detail: sending rate is

- Cut in half on loss detected by triple duplicate ACK (TCP Reno)
- Cut to 1 MSS (maximum segment size) when loss detected by timeout (TCP Tahoe)

Why AIMD?

- AIMD – a distributed, asynchronous algorithm – has been shown to:
  - optimize congested flow rates network wide
  - have desirable stability properties

TCP congestion control: details



TCP sending behavior:

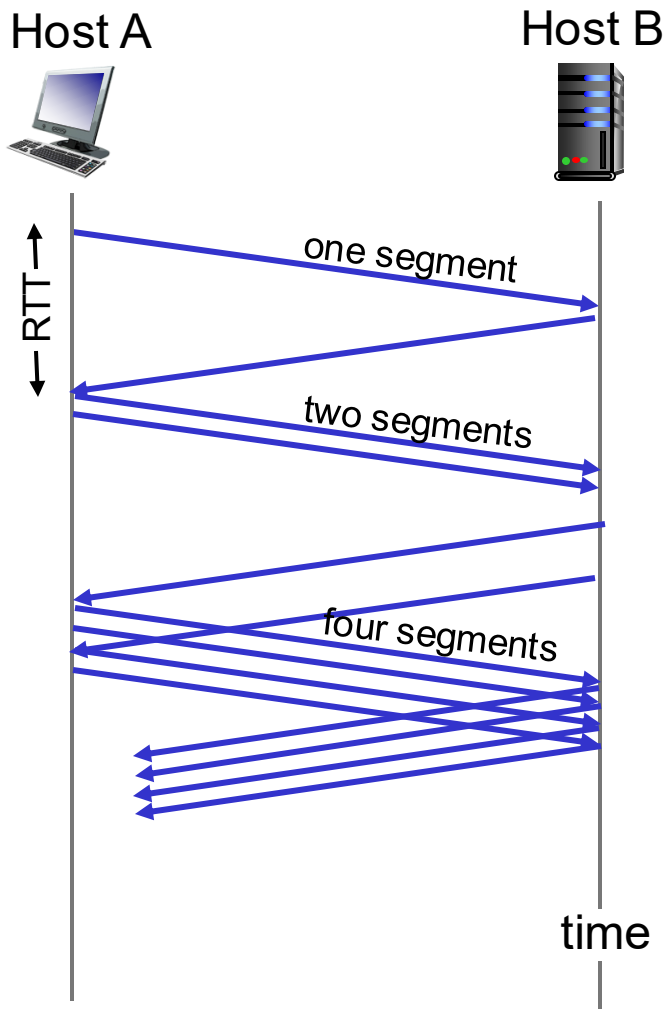
- *roughly:* send `cwnd` bytes, wait RTT for ACKS, then send more bytes

$$\text{TCP rate} \approx \frac{\texttt{cwnd}}{\text{RTT}} \text{ bytes/sec}$$

- TCP sender limits transmission: **`LastByteSent- LastByteAcked ≤ cwnd`**
- `cwnd` is dynamically adjusted in response to observed network congestion (implementing TCP congestion control)

TCP slow start

- when connection begins, increase rate exponentially until first loss event:
  - initially **`cwnd`** = 1 MSS
  - double **`cwnd`** every RTT
  - done by incrementing **`cwnd`** for every ACK received
- *summary:* initial rate is slow, but ramps up exponentially fast



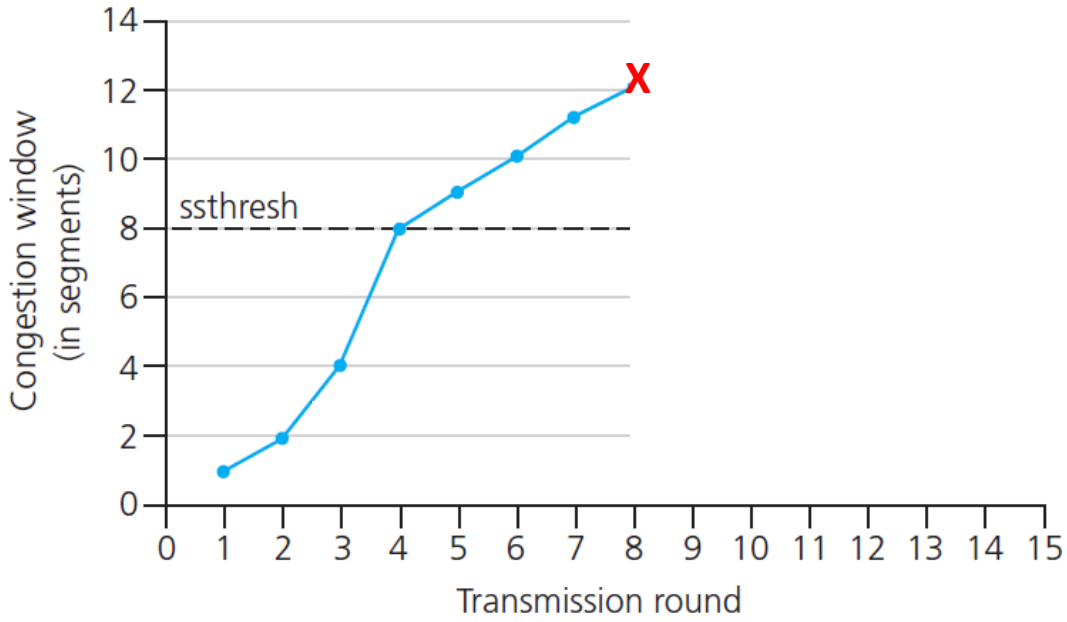
TCP: from slow start to congestion avoidance

Q: when should the exponential increase switch to linear?

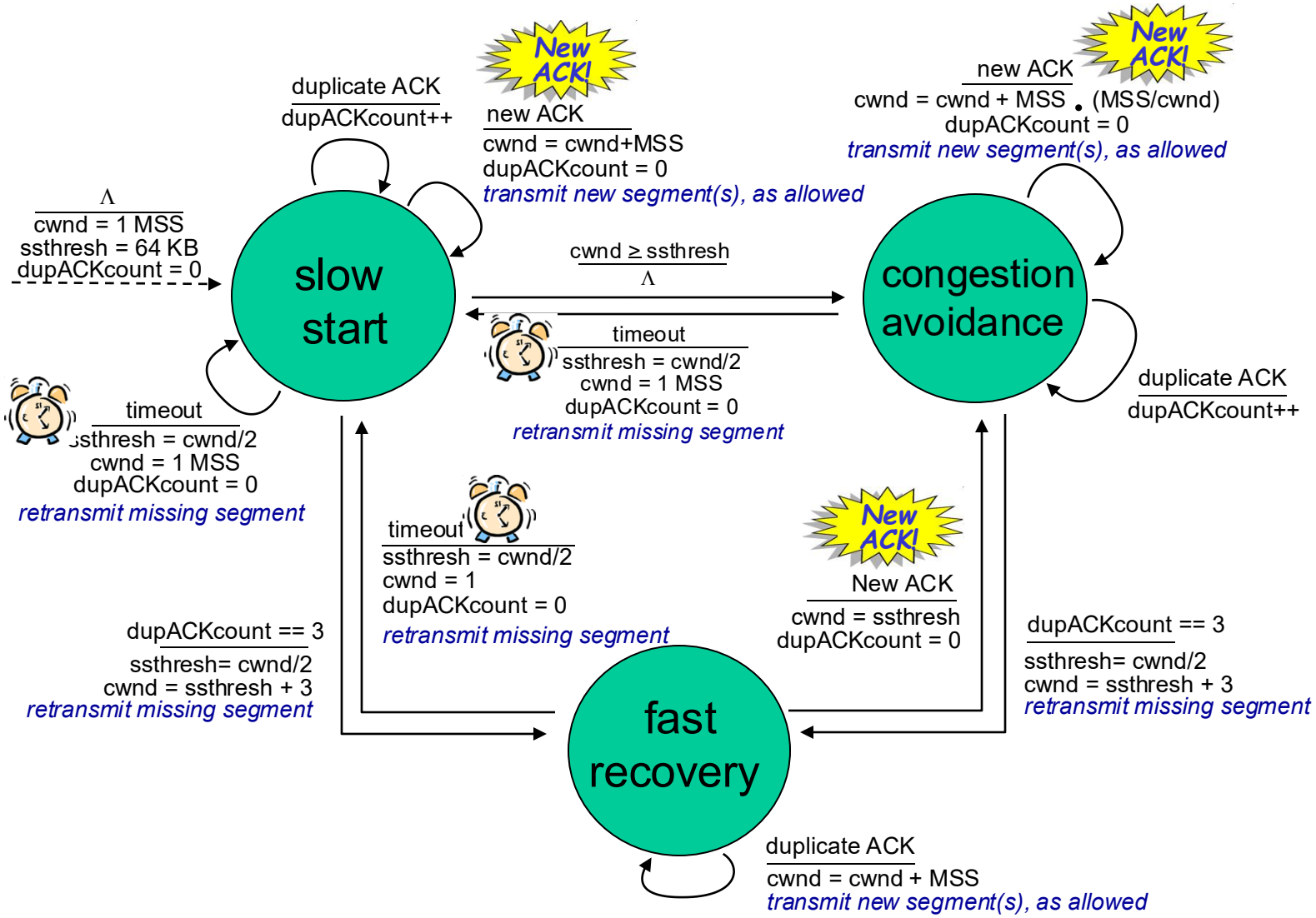
A: when **`cwnd`** gets to 1/2 of its value before timeout.

Implementation:

- variable **`ssthresh`**
- on loss event, **`ssthresh`** is set to 1/2 of **`cwnd`** just before loss event



Summary: TCP congestion control



More TCP Congestion Control Methods

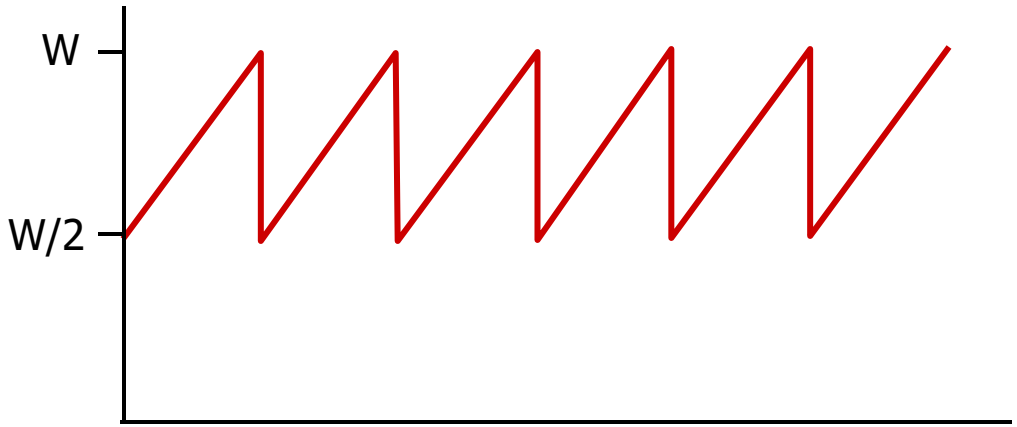
Variant	Feedback	Required changes	Benefits	Fairness
(New) Reno	Loss	—	—	Delay
Vegas	Delay	Sender	Less loss	Proportional
High Speed	Loss	Sender	High bandwidth	
BIC	Loss	Sender	High bandwidth	
CUBIC	Loss	Sender	High bandwidth	
C2TCP[9][10]	Loss/Delay	Sender	Ultra–low latency and high bandwidth	
NATCP[11]	Multi–bit signal	Sender	Near Optimal Performance	
Elastic–TCP	Loss and Delay	Sender	High bandwidth/short & long–distance	
Agile–TCP	Loss	Sender	High bandwidth/short–distance	
H–TCP	Loss	Sender	High bandwidth	
FAST	Delay	Sender	High bandwidth	Proportional
Compound TCP	Loss/Delay	Sender	High bandwidth	Proportional
Westwood	Loss/Delay	Sender	L	
Jersey	Loss/Delay	Sender	L	
BBR[12]	Delay	Sender	BLVC, Bufferbloat	
CLAMP	Multi–bit signal	Receiver, Router	V	Max–min
TFRC	Loss	Sender, Receiver	No Retransmission	Minimum delay
XCP	Multi–bit signal	Sender, Receiver, Router	BLFC	Max–min
VCP	2–bit signal	Sender, Receiver, Router	BLF	Proportional
MaxNet	Multi–bit signal	Sender, Receiver, Router	BLFSC	Max–min
JetMax	Multi–bit signal	Sender, Receiver, Router	High bandwidth	Max–min
RED	Loss	Router	Reduced delay	
ECN	Single–bit signal	Sender, Receiver, Router	Reduced loss	

https://en.wikipedia.org/wiki/TCP_congestion_control

TCP throughput

- **avg. TCP throughput as function of window size, RTT?**
  - ignore slow start, assume always data to send
- **W: window size (measured in bytes) where loss occurs**
  - avg. window size (# in-flight bytes) is ¾ W
  - avg. thruput is 3/4W per RTT

$$\text{avg TCP thruput} = \frac{3}{4}\frac{W}{RTT} \text{ bytes/sec}$$



Evolving transport-layer functionality

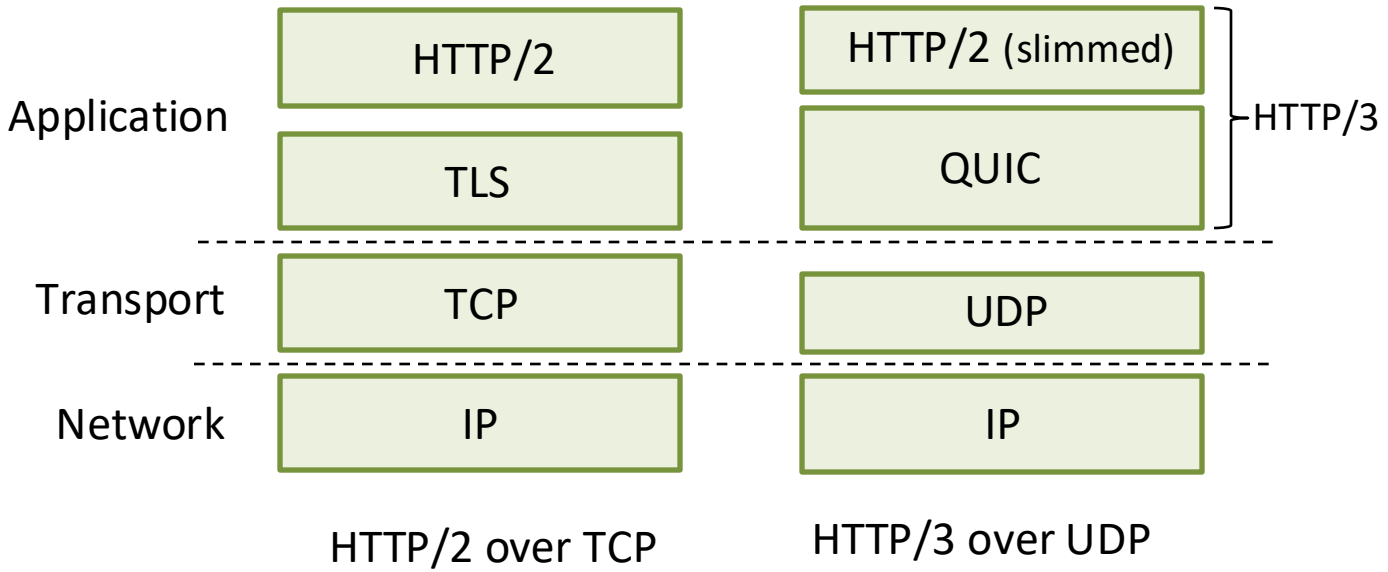
- TCP, UDP: principal transport protocols for 40 years
- different “flavors” of TCP developed, for specific scenarios:

Scenario	Challenges
Long, fat pipes (large data transfers)	Many packets “in flight”; loss shuts down pipeline
Wireless networks	Loss due to noisy wireless links, mobility; TCP treat this as congestion loss
Long-delay links	Extremely long RTTs
Data center networks	Latency sensitive
Background traffic flows	Low priority, “background” TCP flows

- moving transport–layer functions to application layer, on top of UDP
  - HTTP/3: QUIC

QUIC: Quick UDP Internet Connections

- application-layer protocol, on top of UDP
  - increase performance of HTTP
  - deployed on many Google servers, apps (Chrome, mobile YouTube app)



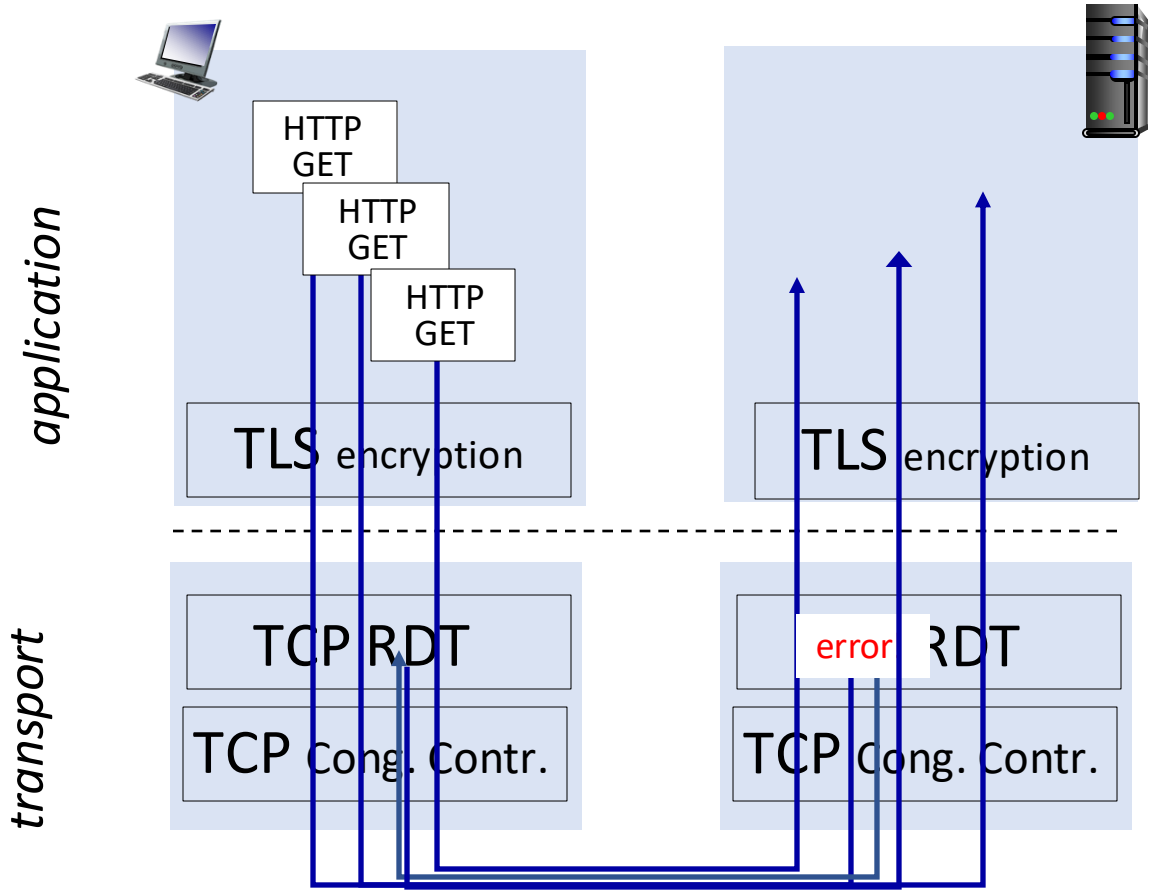
QUIC: Quick UDP Internet Connections

adopts approaches we've studied in this chapter for connection establishment, error control, congestion control

- **error and congestion control:** "Readers familiar with TCP's loss detection and congestion control will find algorithms here that parallel well-known TCP ones." [from QUIC specification]
- **connection establishment:** reliability, congestion control, authentication, encryption, state established in one (or even zero) RTT (recall HTTP3 discussion in Chapter 2)

- multiple application-level "streams" multiplexed over single QUIC connection
  - separate reliable data transfer, security
  - common congestion control

QUIC: streams: parallelism, no HOL blocking



(a) HTTP 1.1

Summary

- **Principles behind transport layer services:**
  - Multiplexing, demultiplexing
  - Reliable data transfer
  - Flow control
  - Congestion control
- **Implementation in the Internet**
  - UDP
  - TCP

Lecture 6 – Network Layer (1)

- **Roadmap**
  1. Overview of Network layer
  2. Router
  3. Internet Protocol

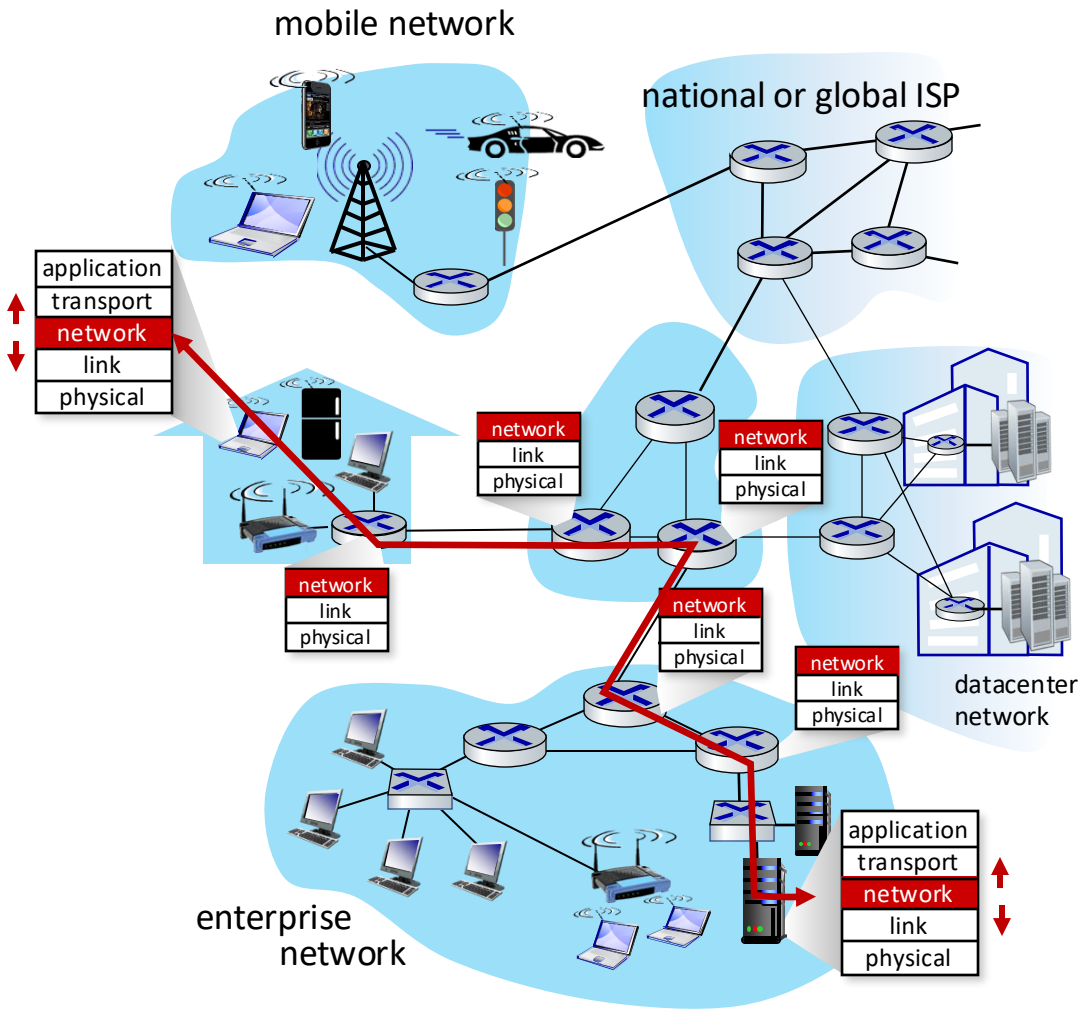
Network layer

Goal

- **Understand principles behind network layer services, focusing on data plane:**
  - Network layer service models
  - Forwarding versus routing
  - How a router works
  - Generalized forwarding
- **Instantiation, implementation in the Internet**

Network-layer services and protocols

- Transport segment from sending to receiving host
  - sender: encapsulates segments into datagrams, passes to link layer
  - receiver: delivers segments to transport layer protocol
- Network layer protocols in *every Internet device*: hosts, routers
- Routers:
  - examines header fields in all IP datagrams passing through it
  - moves datagrams from input ports to output ports to transfer datagrams along end-end path



Two key network-layer functions

Network-layer functions:

- **Forwarding: move packets from router's input to appropriate router output**
- **Routing: determine route taken by packets from source to destination**
  - Routing algorithms

Analogy: taking a trip

- forwarding: process of getting through single interchange
- routing: process of planning trip from source to destination

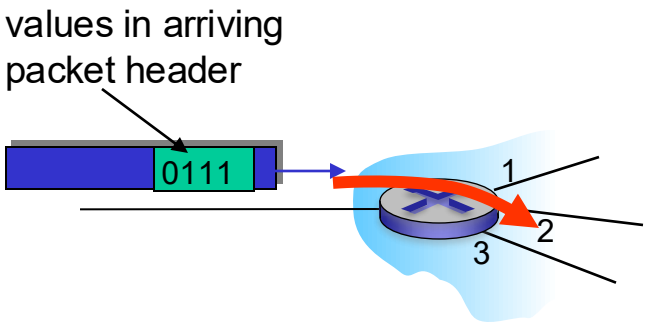
forwarding

routing

Network layer: data plane, control plane

Data plane

- Local, per-router function
- Determines how datagram arriving on router input port is forwarded to router output port
- Forwarding function

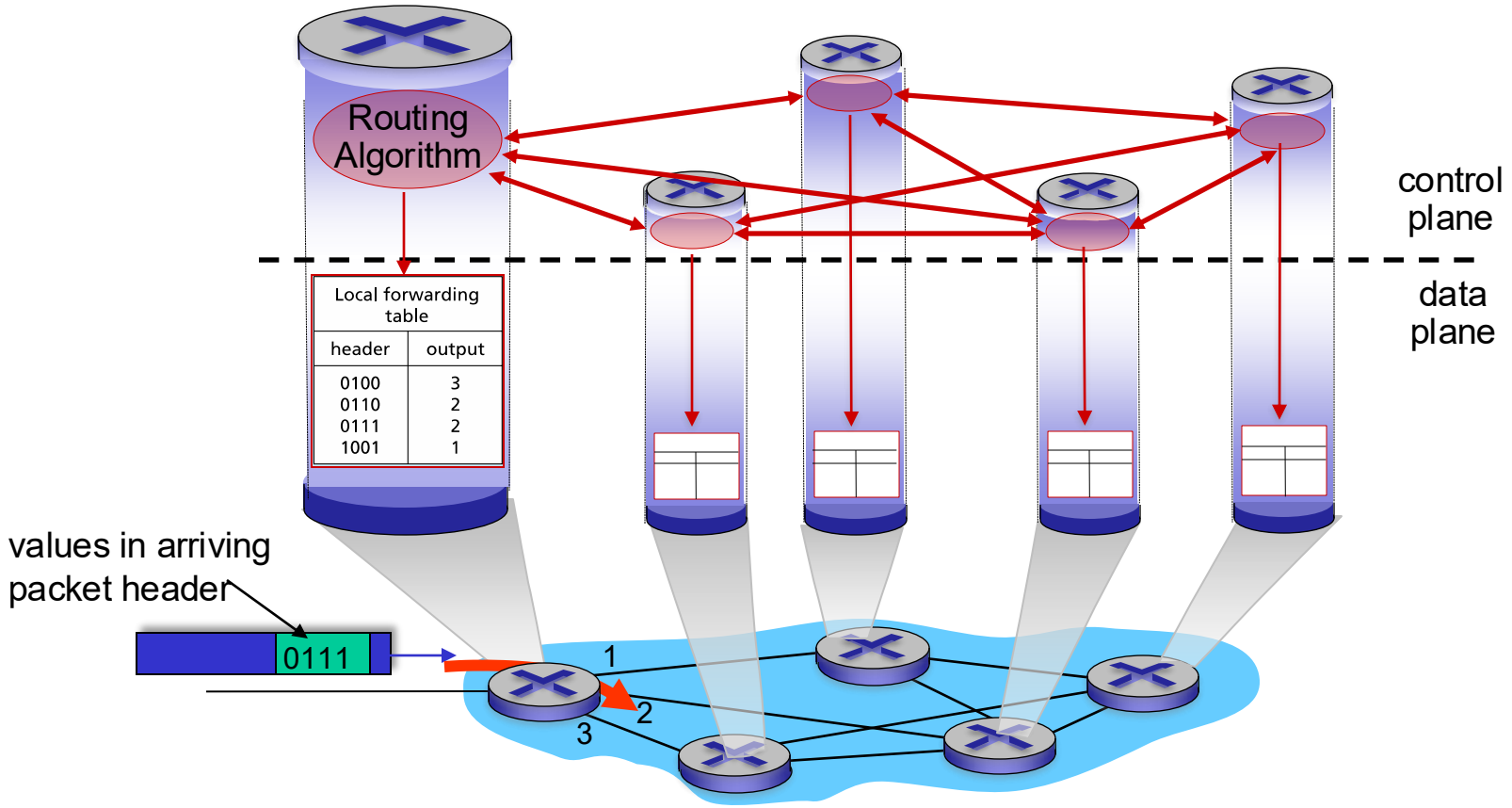


Control plane

- Network-wide logic
- Determines how datagram is routed among routers along end-end path from source host to destination host
- Two control-plane approaches:
  - *traditional routing algorithms:* implemented in routers
  - *software-defined networking (SDN)*: implemented in (remote) servers

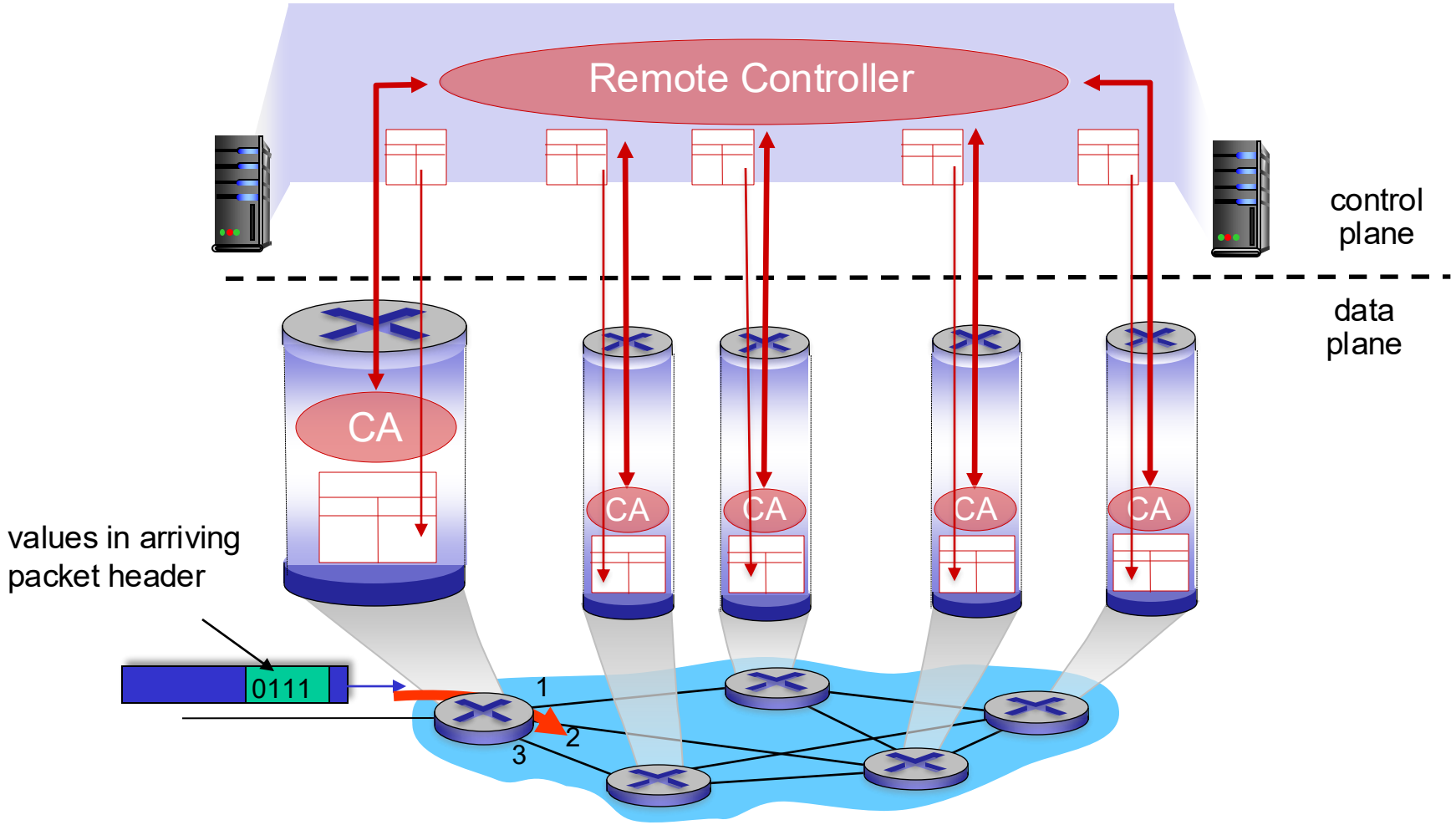
Per-router control plane

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



Software-Defined Networking (SDN) control plane

Remote controller computes, installs forwarding tables in routers



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 model

Network Architecture	Service Model	Quality of Service (QoS) Guarantees ?			
		Bandwidth	Loss	Order	Timing
Internet	best effort	none	no	no	no

Internet "best effort" service model

No guarantees on:

i. successful datagram delivery to destination
ii. timing or order of delivery
iii. bandwidth available to end-end flow

Network-layer service model

Network Architecture	Service Model	Quality of Service (QoS) Guarantees ?			
		Bandwidth	Loss	Order	Timing
Internet	best effort	none	no	no	no
ATM	Constant Bit Rate	Constant rate	yes	yes	yes
ATM	Available Bit Rate	Guaranteed min	no	yes	no
Internet	Intserv Guaranteed (RFC 1633)	yes	yes	yes	yes
Internet	Diffserv (RFC 2475)	possible	possibly	possibly	no

Reflections on best-effort service:

- simplicity of mechanism has allowed Internet to be widely deployed adopted
- sufficient provisioning of bandwidth allows performance of real-time applications (e.g., interactive voice, video) to be "good enough" for "most of the time"
- replicated, application-layer distributed services (datacenters, content distribution networks) connecting close to clients' networks, allow services to be provided from multiple locations
- congestion control of "elastic" services helps

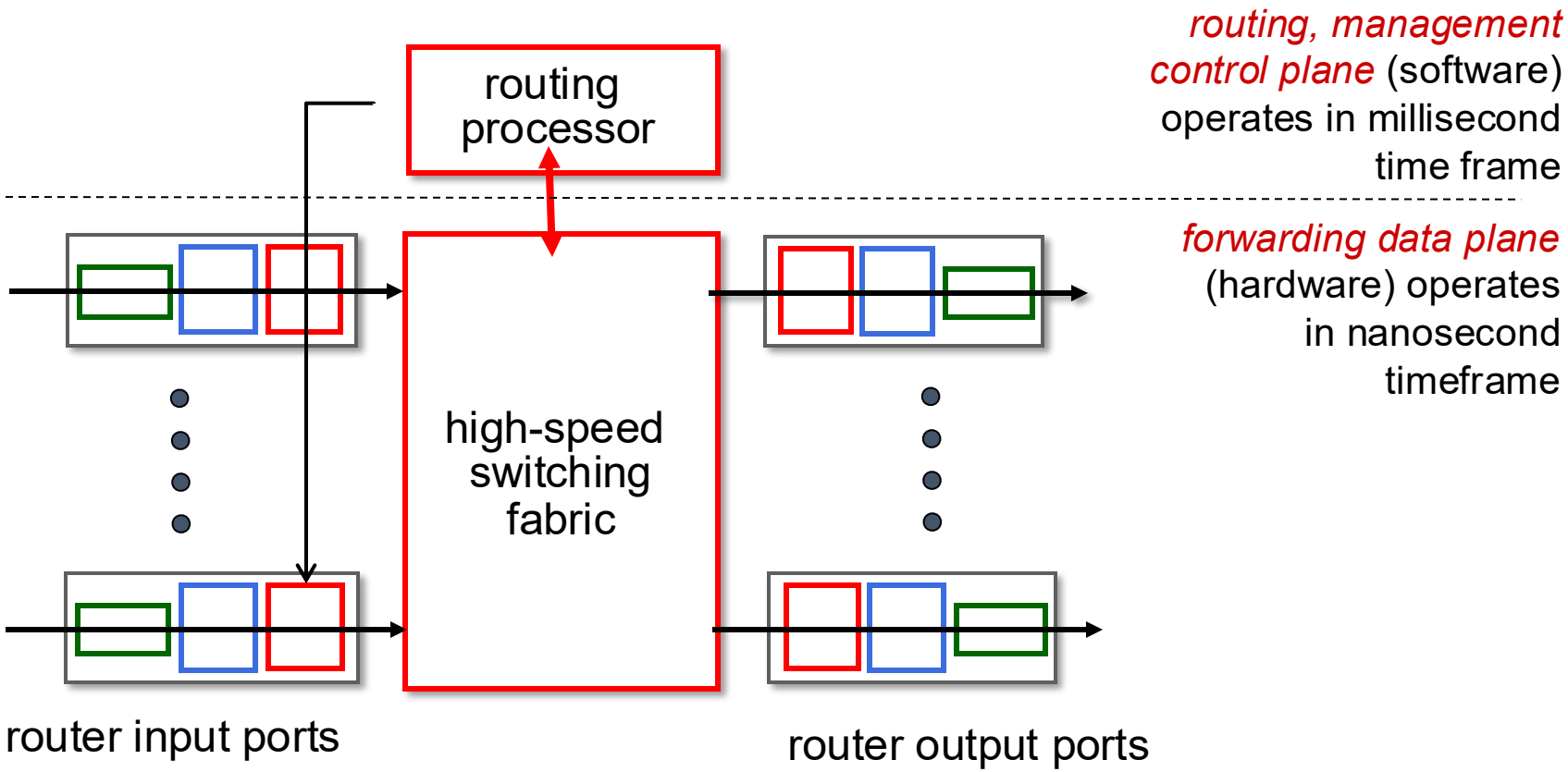
It's hard to argue with success of best-effort service model

Lecture 6 – Network Layer (1)

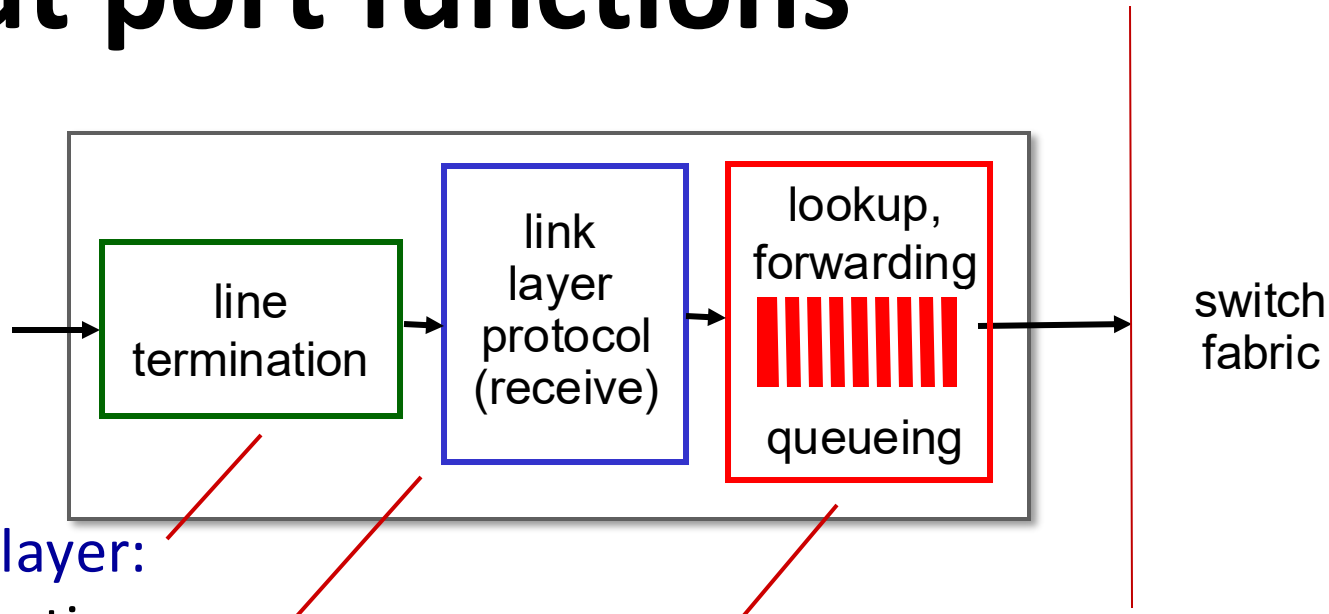
- **Roadmap**
  1. Overview of Network layer
  2. Router
  3. Internet Protocol

Router architecture overview

High-level view of generic router architecture:



Input port functions



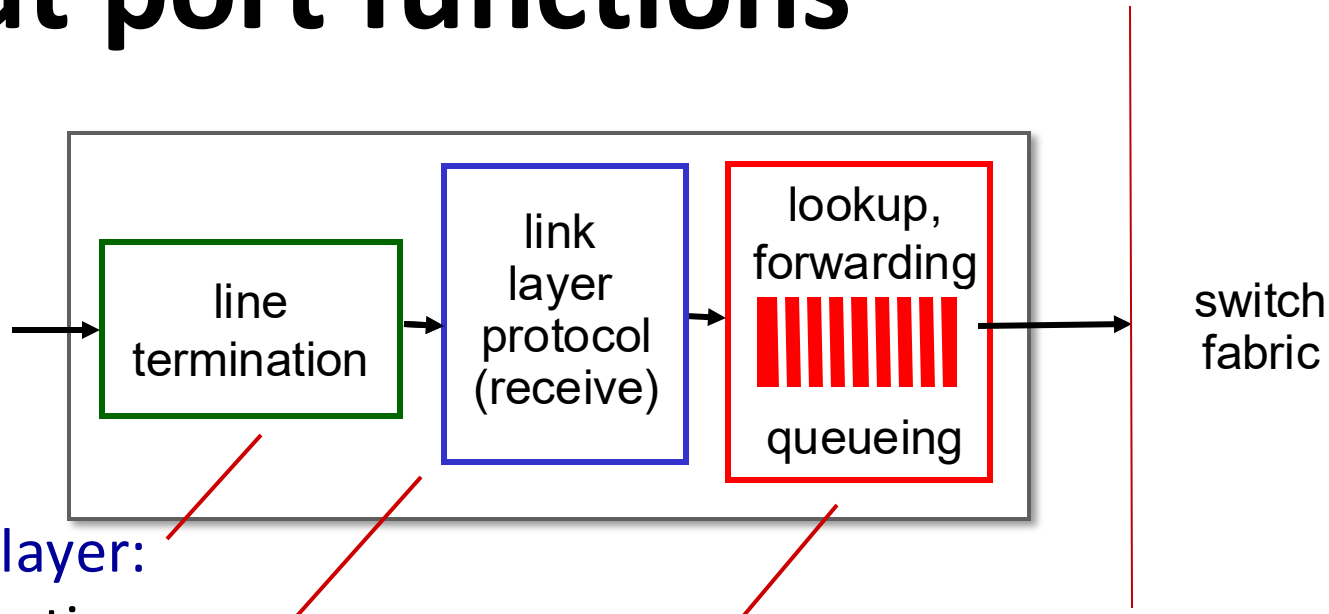
physical layer:
bit-level reception

link layer:
e.g., Ethernet
(chapter 6)

decentralized switching:

- using header field values, lookup output port using forwarding table in input port memory *("match plus action")*
- goal: complete input port processing at 'line speed'
- input port queuing: if datagrams arrive faster than forwarding rate into switch fabric

Input port functions



physical layer:
bit-level reception

link layer:
e.g., Ethernet (chapter 6)

decentralized switching:

- using 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

Destination-based forwarding

forwarding table

	Destination Address Range	Link Interface
200.23.16.0 200.23.23.255	**11001000 00010111 00010000 00000000** through **11001000 00010111 00010111 11111111**	0
200.23.24.0 200.23.24.255	**11001000 00010111 00011000 00000000** through **11001000 00010111 00011000 11111111**	1
200.23.25.0 200.23.31.255	**11001000 00010111 00011001 00000000** through **11001000 00010111 00011111 11111111**	2
	otherwise	3

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

Destination-based forwarding

forwarding table

Destination Address Range	Link Interface
11001000 00010111 00010000 00000000 through 11001000 00010111 00010000 00000100	0
11001000 00010111 00010000 00000101 through 11001000 00010111 00010000 00000111	3
11001000 00010111 00011000 11111111	
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 match

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:

11001000 00010111 00010110 10100001 which interface?

11001000 00010111 00011000 10101010 which interface?

Longest prefix matching

longest prefix match

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

match!

examples:

11001000 00010111 00010110 10100001 which interface?

11001000 00010111 00011000 10101010 which interface?

Longest prefix matching

longest prefix match

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

match!

examples:

11001000 00010111 00010110 10100001 which interface?

11001000 00010111 00011000 10101010 which interface?

Longest prefix matching

longest prefix match

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

match!

examples:

11001000 00010111 00010110 10100001 which interface?

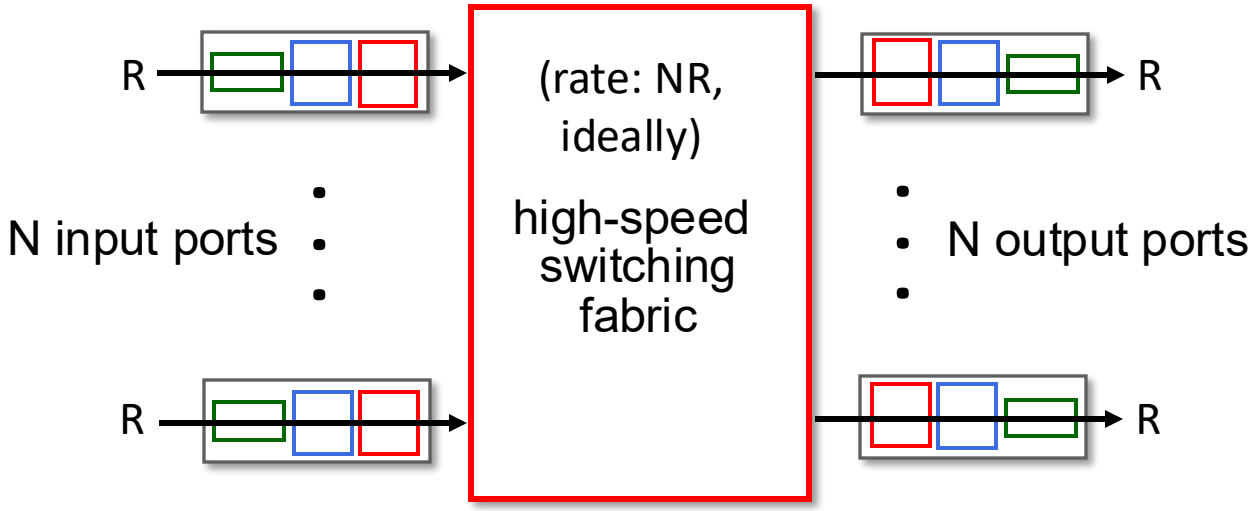
11001000 00010111 00011000 10101010 which interface?

Longest prefix matching

- we'll see *why* longest prefix matching is used shortly, when we study addressing
- longest prefix matching: often performed using ternary content addressable memories (TCAMs)
  - *content addressable:* present address to TCAM: retrieve address in one clock cycle, regardless of table size
  - Cisco Catalyst: ~1M routing table entries in TCAM

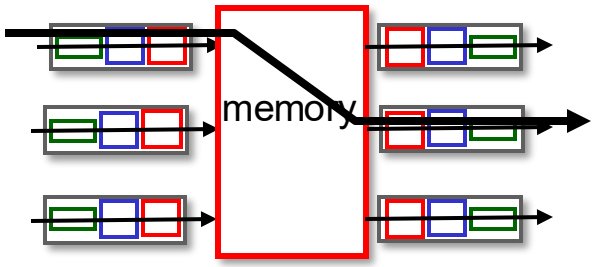
Switching fabrics

- transfer packet from input link to appropriate output link
- 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

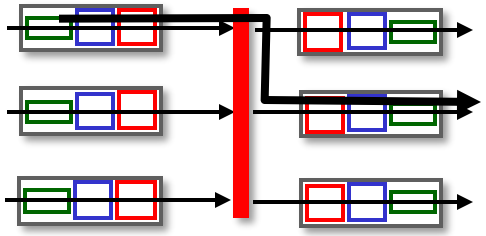


Switching fabrics

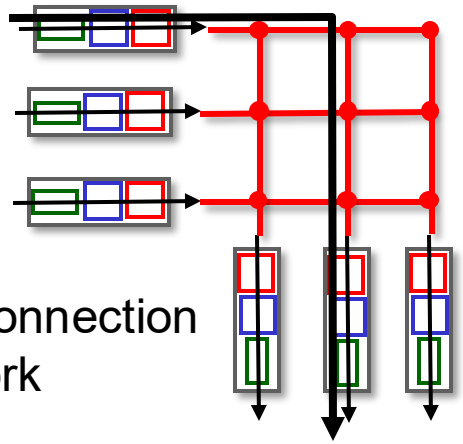
- transfer packet from input link to appropriate output link
- switching rate: rate at which packets can be transfer from inputs to outputs
  - often measured as multiple of input/output line rate
  - N inputs: switching rate N times line rate desirable
- three major types of switching fabrics:



memory



bus

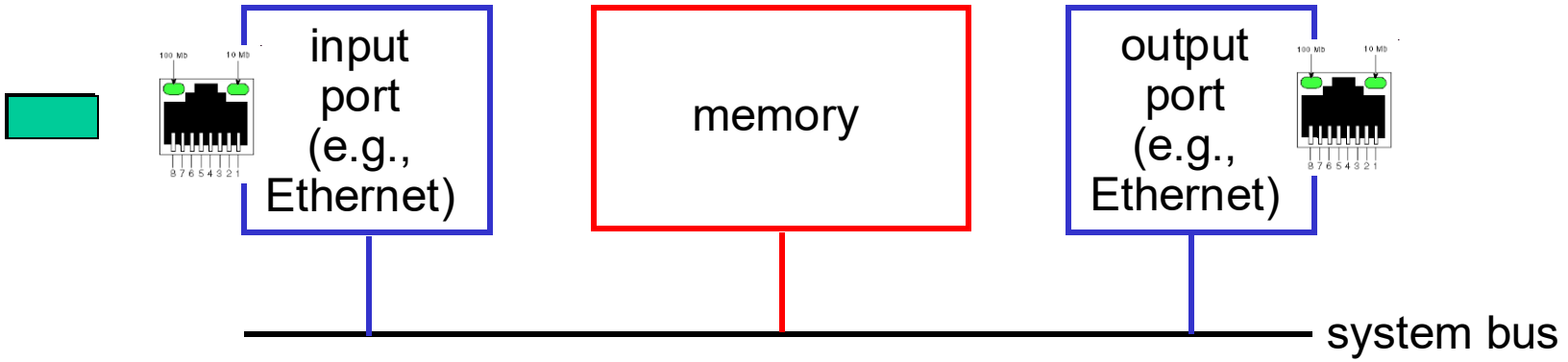


interconnection network

Switching via memory

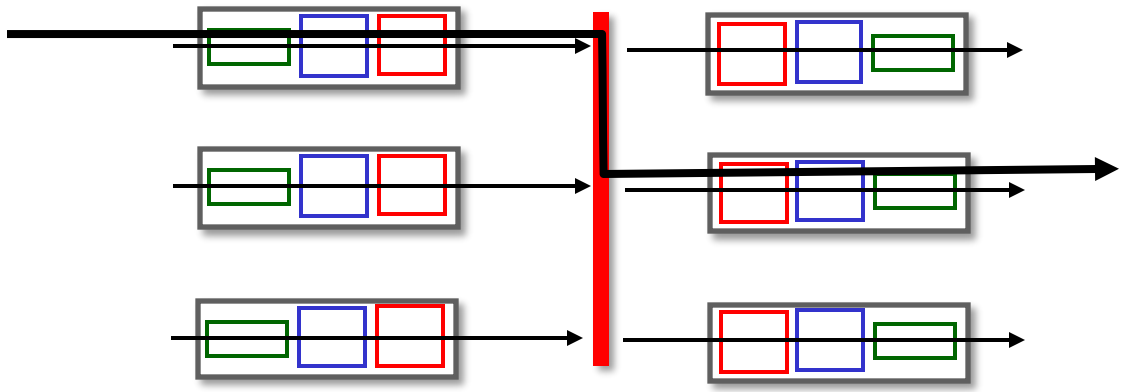
first generation routers:

- 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)



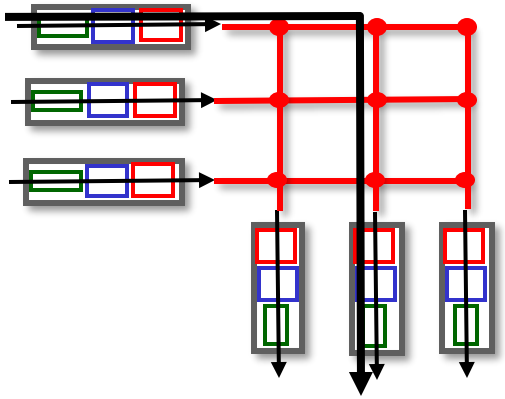
Switching via a bus

- 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 routers

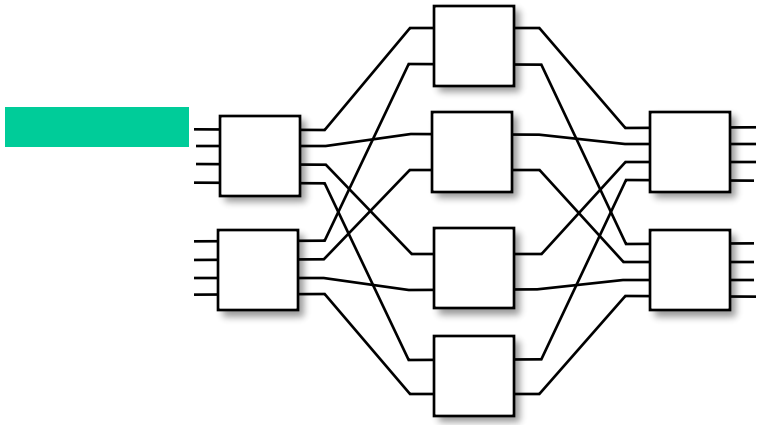


Switching via interconnection network

- Crossbar, Clos networks, other interconnection nets initially developed to connect processors in multiprocessor
- multistage switch: *nxn* switch from multiple stages of smaller switches
- exploiting parallelism:
  - fragment datagram into fixed length cells on entry
  - switch cells through the fabric, reassemble datagram at exit



3x3 crossbar

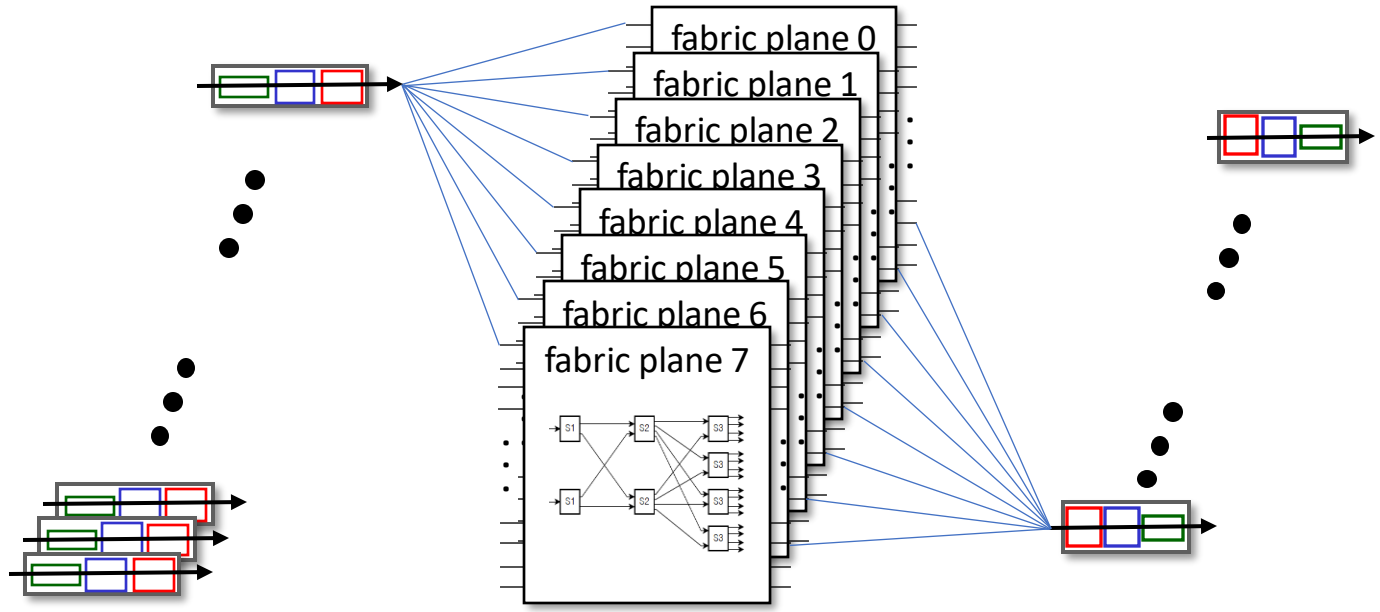


8x8 multistage switch
built from smaller-sized switches

Switching via interconnection network

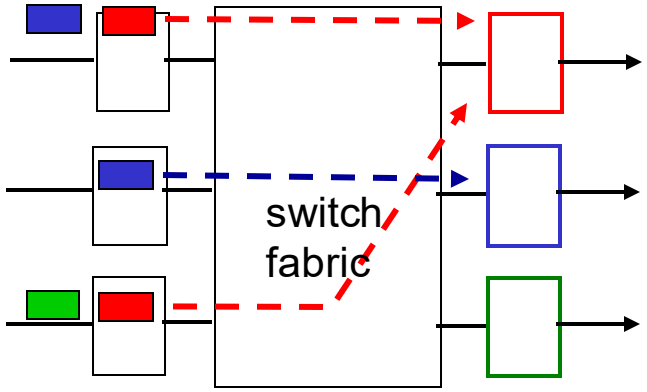
- scaling, using multiple switching “planes” in parallel:
  - speedup, scaleup via parallelism

- Cisco CRS router:
  - basic unit: 8 switching planes
  - each plane: 3-stage interconnection network
  - up to 100’s Tbps switching capacity

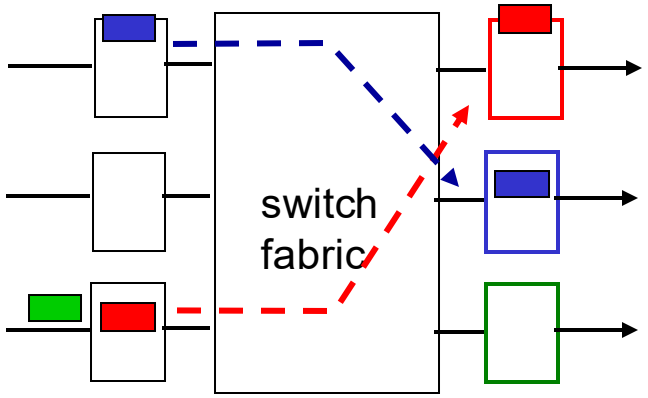


Input port queuing

- If switch 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

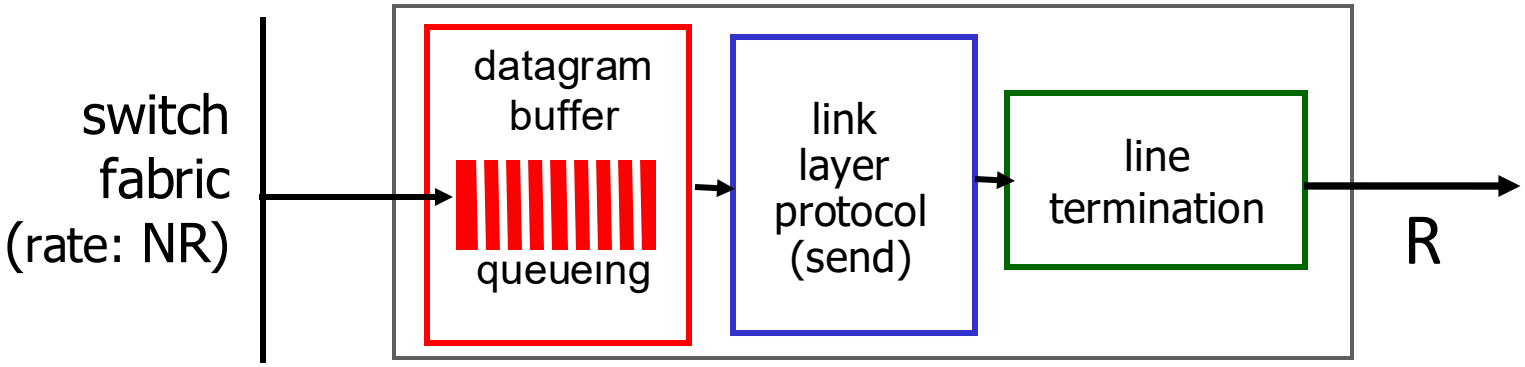


output port contention: only one red datagram can be transferred. lower red packet is *blocked*



one packet time later: green packet experiences HOL blocking

Output port queuing



This is a really important slide

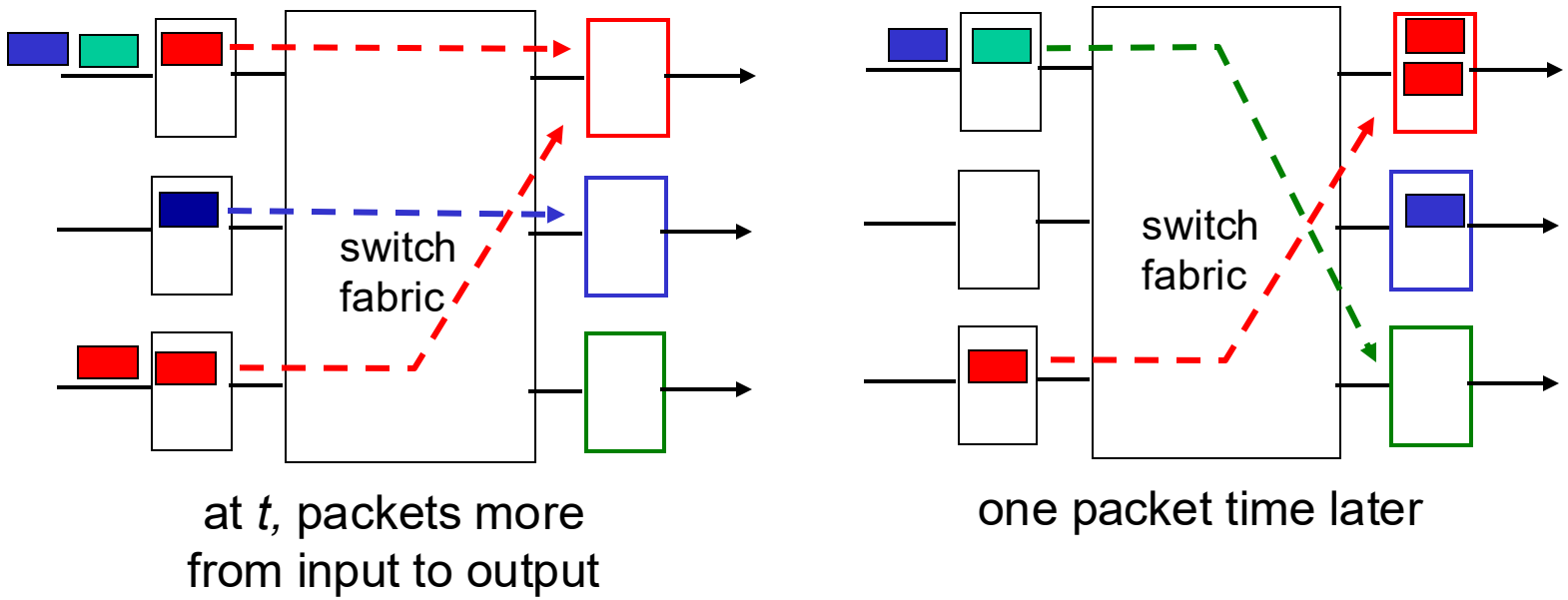
- *Buffering* required when datagrams arrive from fabric faster than link transmission rate. *Drop policy:* which datagrams to drop if no free buffers?

→ Datagrams can be lost due to congestion, lack of buffers

- *Scheduling discipline* chooses among queued datagrams for transmission

→ Priority scheduling – who gets best performance, network neutrality

Output port queuing

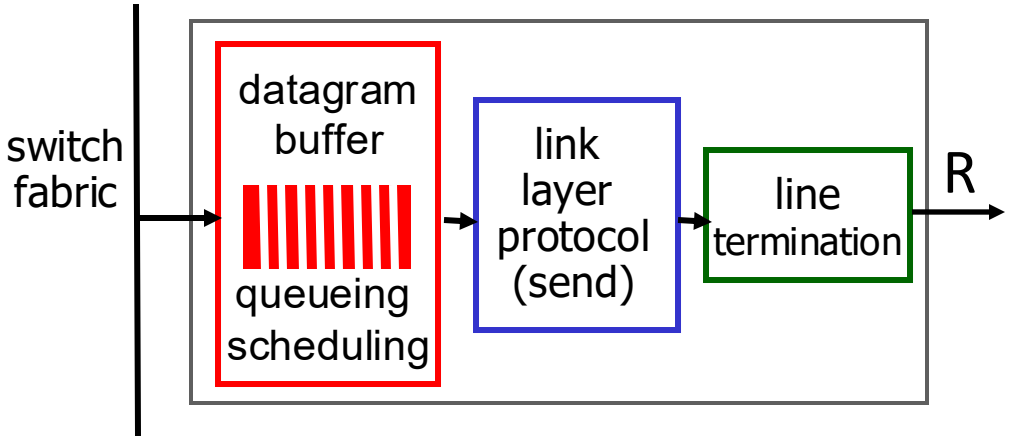


at *t*, packets more from input to output

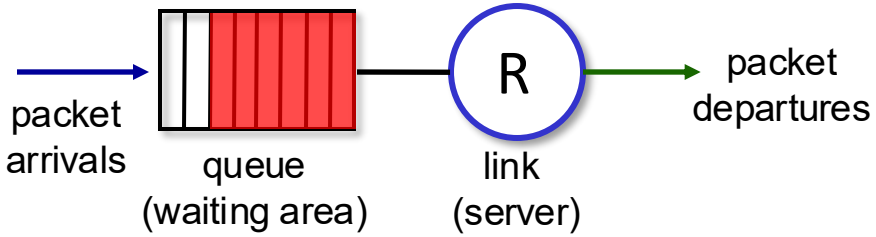
one packet time later

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

Buffer Management



Abstraction: queue



buffer management:

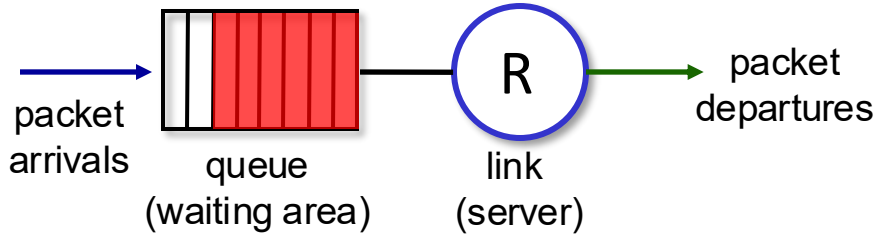
- drop: which packet to add, drop when buffers are full
  - tail drop: drop arriving packet
  - priority: drop/remove on priority basis
- marking: which packets to mark to signal congestion (ECN, RED)

Packet Scheduling: FCFS

packet scheduling: deciding which packet to send next on link

- first come, first served
- priority
- round robin
- weighted fair queueing

Abstraction: queue



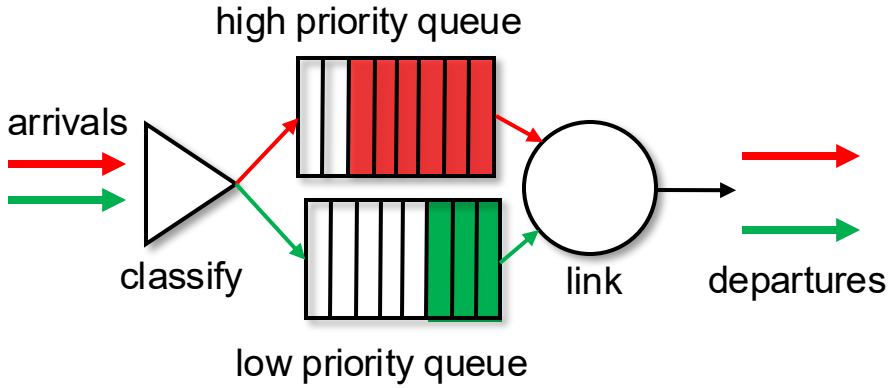
FCFS: packets transmitted in order of arrival to output port

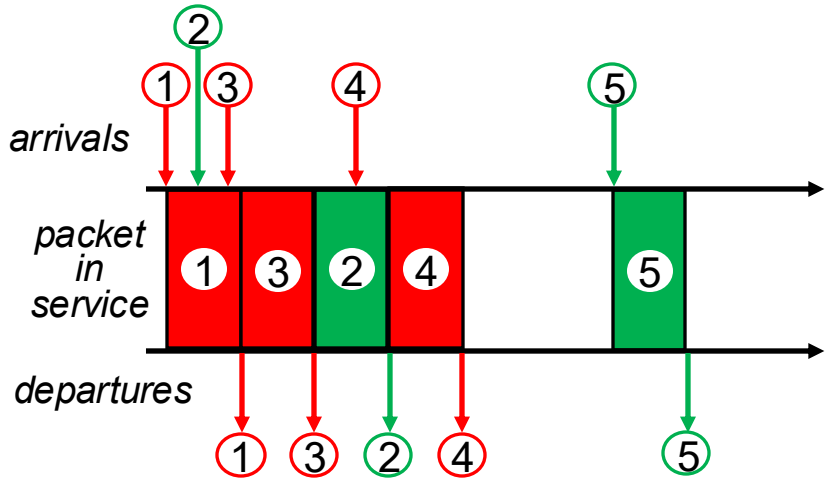
- also known as: First-in-first-out (FIFO)
- real world examples?

Scheduling policies: priority

Priority scheduling:

- arriving traffic classified, queued by class
  - any header fields can be used for classification
- send packet from highest priority queue that has buffered packets
  - FCFS within priority class

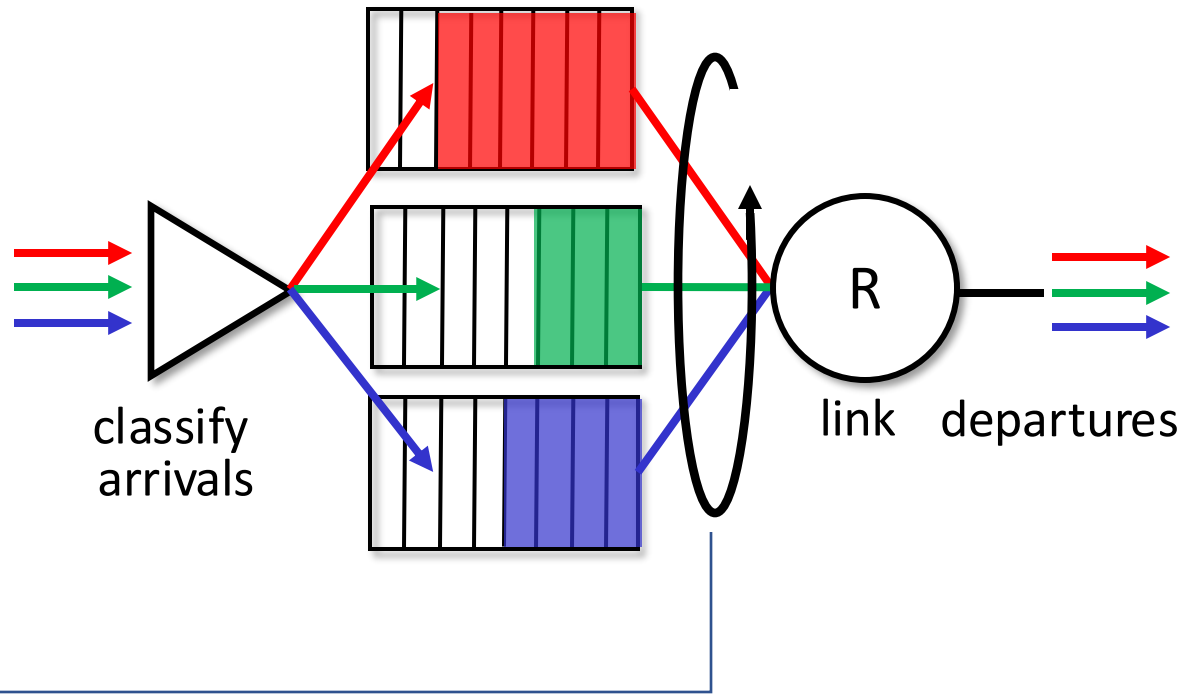




Scheduling policies: round robin

Round Robin (RR) scheduling:

- arriving traffic classified, queued by class
  - any header fields can be used for classification
- server cyclically, repeatedly scans class queues, sending one complete packet from each class (if available) in turn



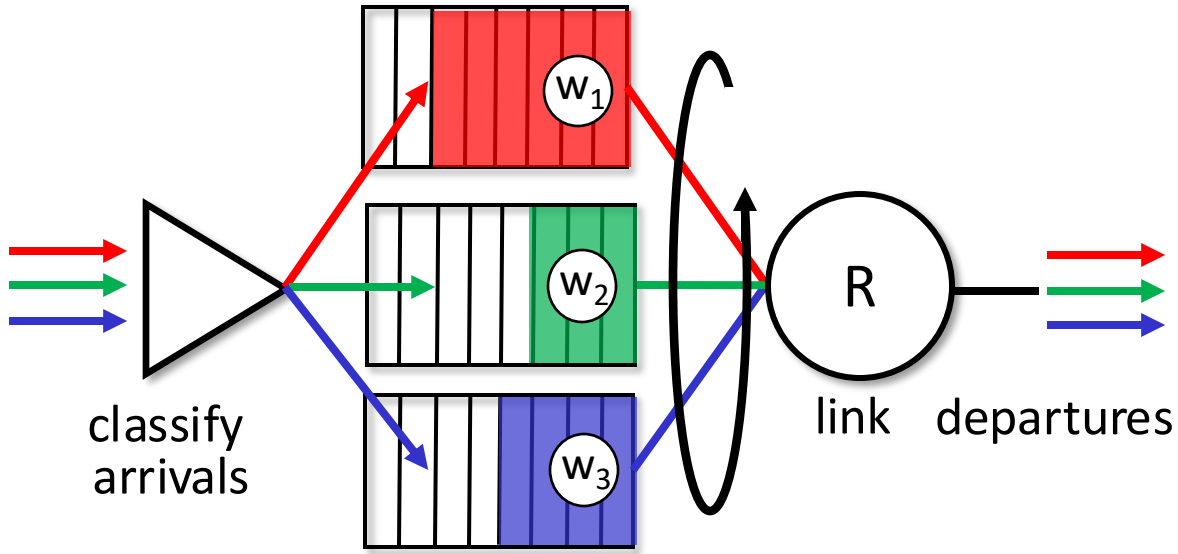
Scheduling policies: weighted fair queueing

Weighted Fair Queuing (WFQ):

- generalized Round Robin
- each class, *i*, has weight, w_i, and gets weighted amount of service in each cycle:

$$\frac{w_i}{\sum_j w_j}$$

- minimum bandwidth guarantee (per-traffic-class)

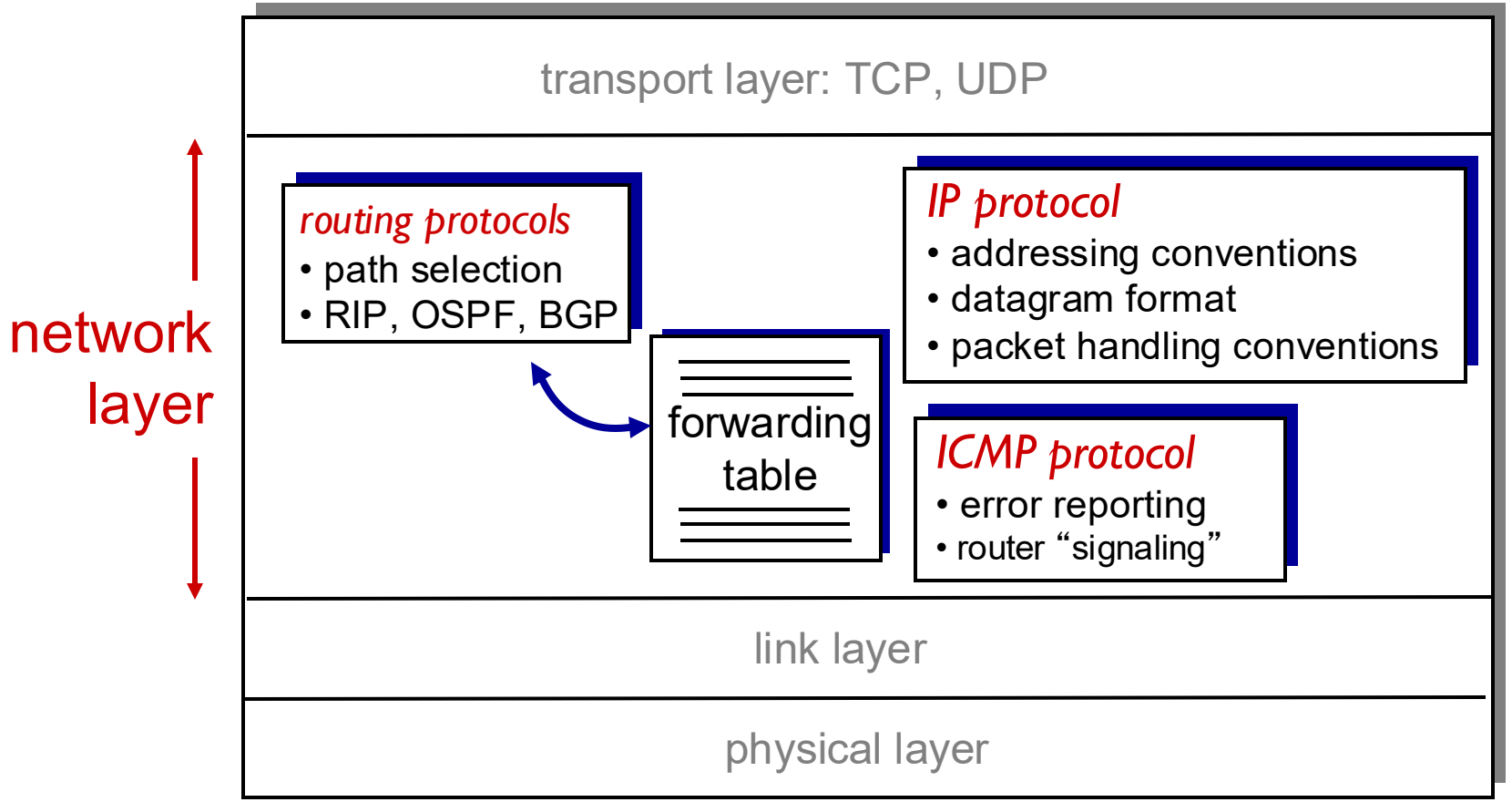


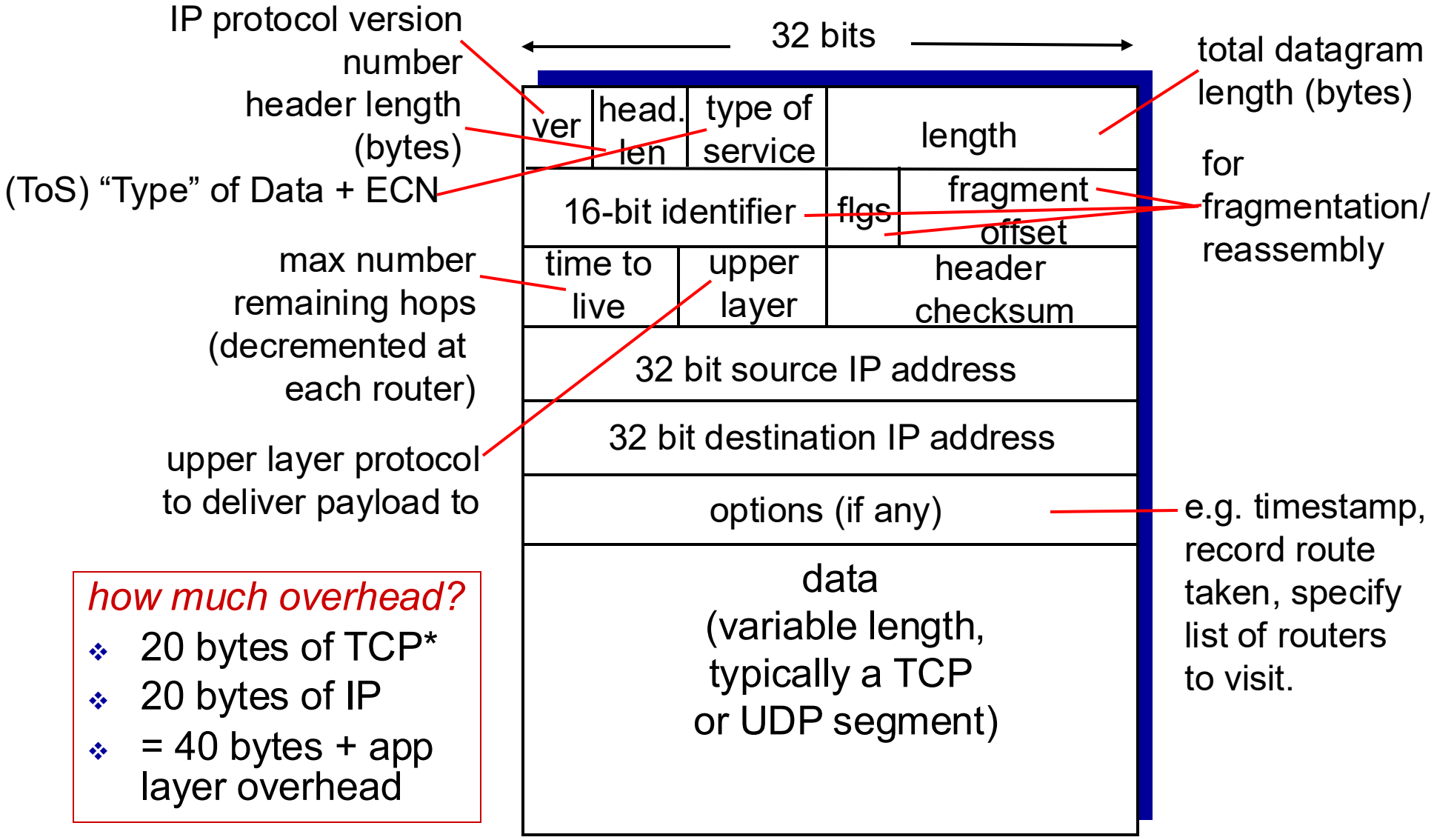
Lecture 6 – Network Layer (1)

- **Roadmap**
  1. Overview of Network layer
  2. Router
  3. Internet Protocol

The Internet network layer

Host, router network layer functions:





how much overhead?

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

IP datagram format

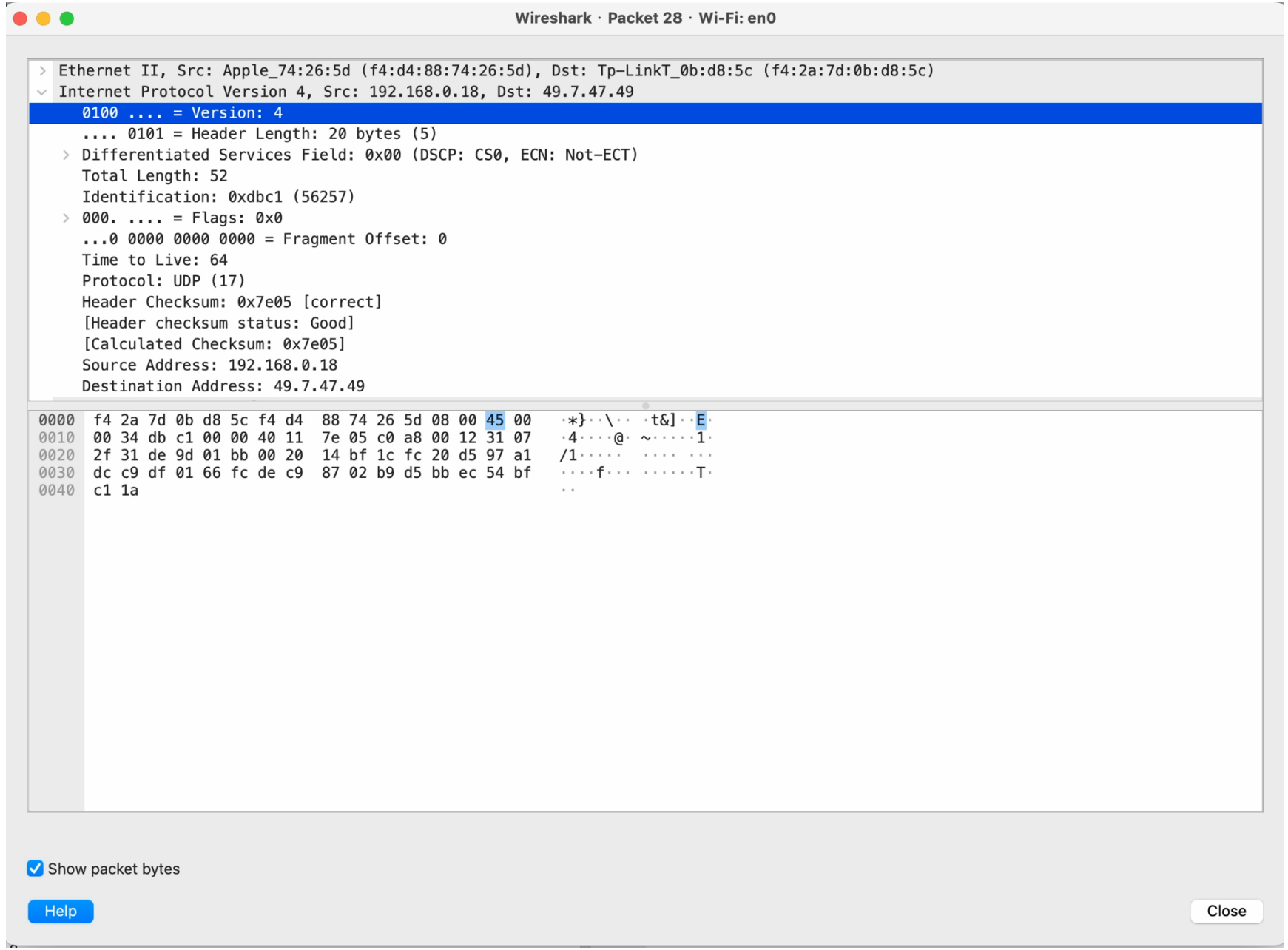
*sometimes the optional header will be used, the overhead of TCP will be more.

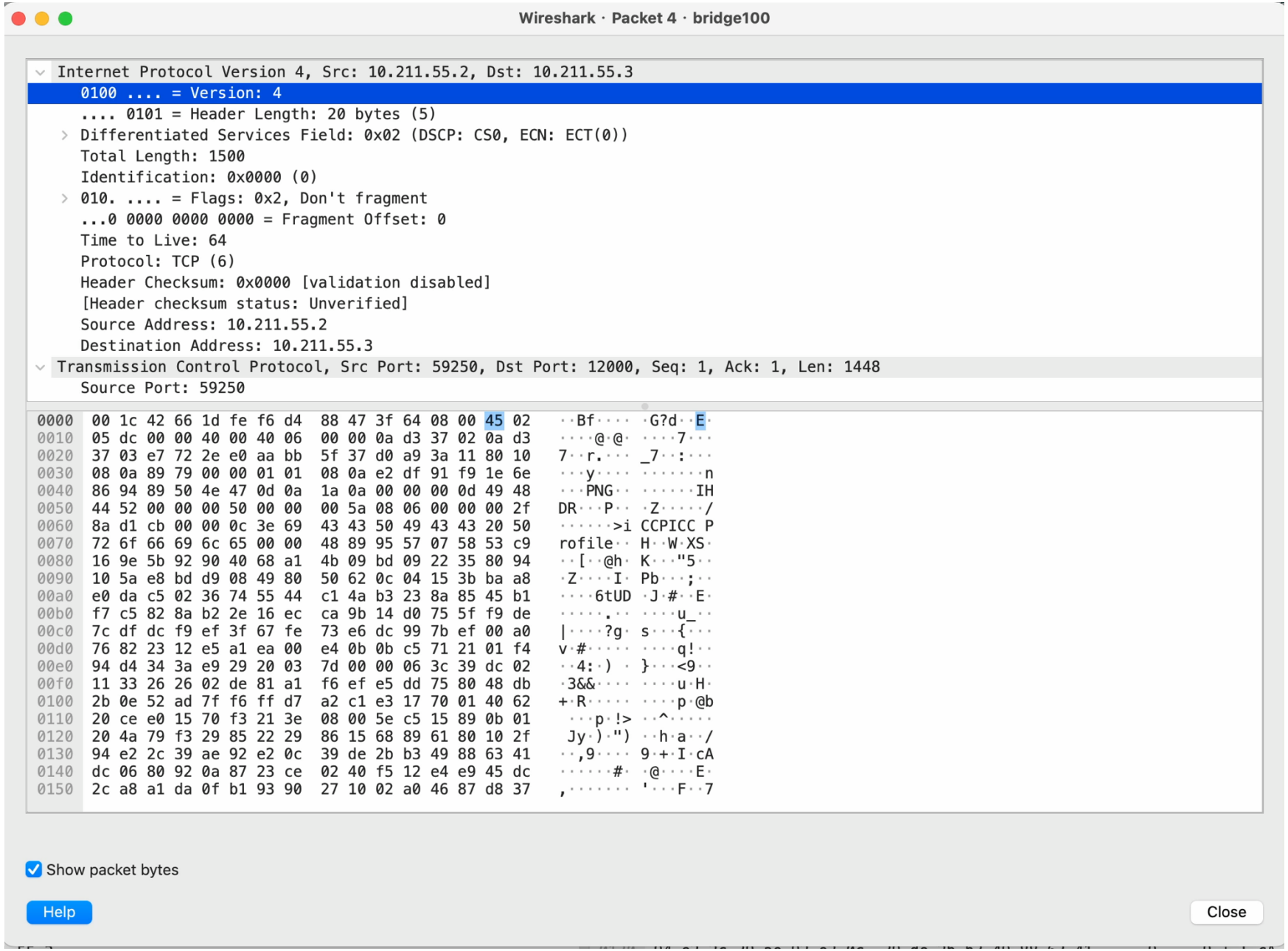
ToS: Type of service

7	6	5	4	3	2	1	0
DSCP						ECN	

Differentiated Services Code Point (DSCP)

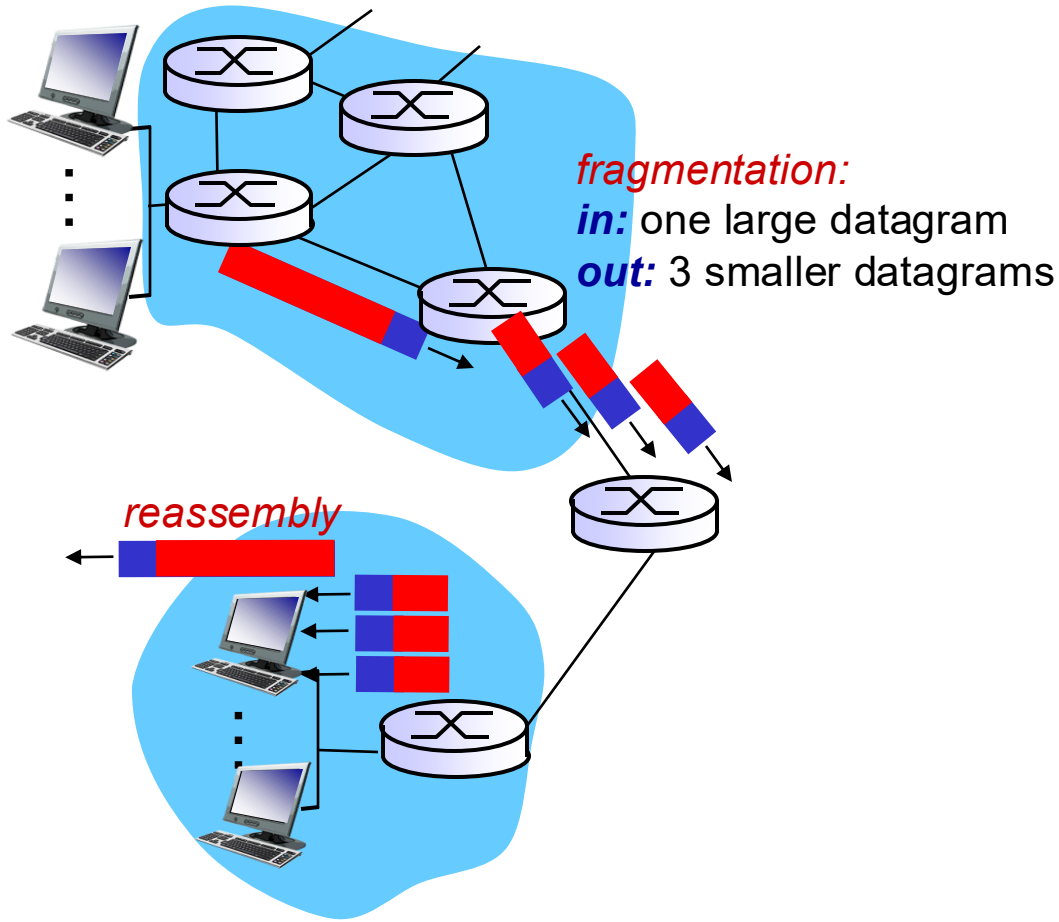
Service class	DSCP Name	DSCP Value	Examples of application
Standard	CS0 (DF)	0	
Low–priority data	CS1	8	File transfer (FTP, SMB)
Network operations, administration and management (OAM)	CS2	16	SNMP, SSH, Ping, Telnet, syslog
Broadcast video	CS3	24	RTSP broadcast TV streaming of live audio and video events video surveillance video–on–demand
Real–time interactive	CS4	32	Gaming, low priority video conferencing
Signaling	CS5	40	Peer–to–peer (SIP, H.323, H.248), NTP
Network control	CS6	48	Routing protocols (OSPF, BGP, ISIS, RIP)
Reserved for future use	CS7	56	


Wireshark · Packet 28 · Wi-Fi: en0
Ethernet II, Src: Apple_74:26:5d (f4:d4:88:74:26:5d), Dst: Tp-LinkT_0b:d8:5c (f4:2a:7d:0b:d8:5c)
Internet Protocol Version 4, Src: 192.168.0.18, Dst: 49.7.47.49
0100 .... = Version: 4
.... 0101 = Header Length: 20 bytes (5)
Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
Total Length: 52
Identification: 0xdbc1 (56257)
000. .... = Flags: 0x0
...0 0000 0000 0000 = Fragment Offset: 0
Time to Live: 64
Protocol: UDP (17)
Header Checksum: 0x7e05 [correct]
[Header checksum status: Good]
[Calculated Checksum: 0x7e05]
Source Address: 192.168.0.18
Destination Address: 49.7.47.49
0000 f4 2a 7d 0b d8 5c f4 d4 88 74 26 5d 08 00 45 00
0010 00 34 db c1 00 00 40 11 7e 05 c0 a8 00 12 31 07
0020 2f 31 de 9d 01 bb 00 20 14 bf 1c fc 20 d5 97 a1
0030 dc c9 df 01 66 fc de c9 87 02 b9 d5 bb ec 54 bf
0040 c1 1a
Show packet bytes
Help
Close


Wireshark · Packet 4 · bridge100
Internet Protocol Version 4, Src: 10.211.55.2, Dst: 10.211.55.3
0100 .... = Version: 4
.... 0101 = Header Length: 20 bytes (5)
Differentiated Services Field: 0x02 (DSCP: CS0, ECN: ECT(0))
Total Length: 1500
Identification: 0x0000 (0)
010. .... = Flags: 0x2, Don't fragment
...0 0000 0000 0000 = Fragment Offset: 0
Time to Live: 64
Protocol: TCP (6)
Header Checksum: 0x0000 [validation disabled]
[Header checksum status: Unverified]
Source Address: 10.211.55.2
Destination Address: 10.211.55.3
Transmission Control Protocol, Src Port: 59250, Dst Port: 12000, Seq: 1, Ack: 1, Len: 1448
Source Port: 59250
0000 00 1c 42 66 1d fe f6 d4 88 47 3f 64 08 00 45 02
0010 05 dc 00 00 40 00 40 06 00 00 0a d3 37 02 0a d3
0020 37 03 e7 72 2e e0 aa bb 5f 37 d0 a9 3a 11 80 10
0030 08 0a 89 79 00 00 01 01 08 0a e2 df 91 f9 1e 6e
0040 86 94 89 50 4e 47 0d 0a 1a 0a 00 00 00 0d 49 48
0050 44 52 00 00 00 50 00 00 00 5a 08 06 00 00 00 2f
0060 8a d1 cb 00 00 0c 3e 69 43 43 50 49 43 43 20 50
0070 72 6f 66 69 6c 65 00 00 48 89 95 57 07 58 53 c9
0080 16 9e 5b 92 90 40 68 a1 4b 09 bd 09 22 35 80 94
0090 10 5a e8 bd d9 08 49 80 50 62 0c 04 15 3b ba a8
00a0 e0 da c5 02 36 74 55 44 c1 4a b3 23 8a 85 45 b1
00b0 f7 c5 82 8a b2 2e 16 ec ca 9b 14 d0 75 5f f9 de
00c0 7c df dc f9 ef 3f 67 fe 73 e6 dc 99 7b ef 00 a0
00d0 76 82 23 12 e5 a1 ea 00 e4 0b 0b c5 71 21 01 f4
00e0 94 d4 34 3a e9 29 20 03 7d 00 00 06 3c 39 dc 02
00f0 11 33 26 26 02 de 81 a1 f6 ef e5 dd 75 80 48 db
0100 2b 0e 52 ad 7f f6 ff d7 a2 c1 e3 17 70 01 40 62
0110 20 ce e0 15 70 f3 21 3e 08 00 5e c5 15 89 0b 01
0120 20 4a 79 f3 29 85 22 29 86 15 68 89 61 80 10 2f
0130 94 e2 2c 39 ae 92 e2 0c 39 de 2b b3 49 88 63 41
0140 dc 06 80 92 0a 87 23 ce 02 40 f5 12 e4 e9 45 dc
0150 2c a8 a1 da 0f b1 93 90 27 10 02 a0 46 87 d8 37
Show packet bytes
Help
Close

IP fragmentation, reassembly

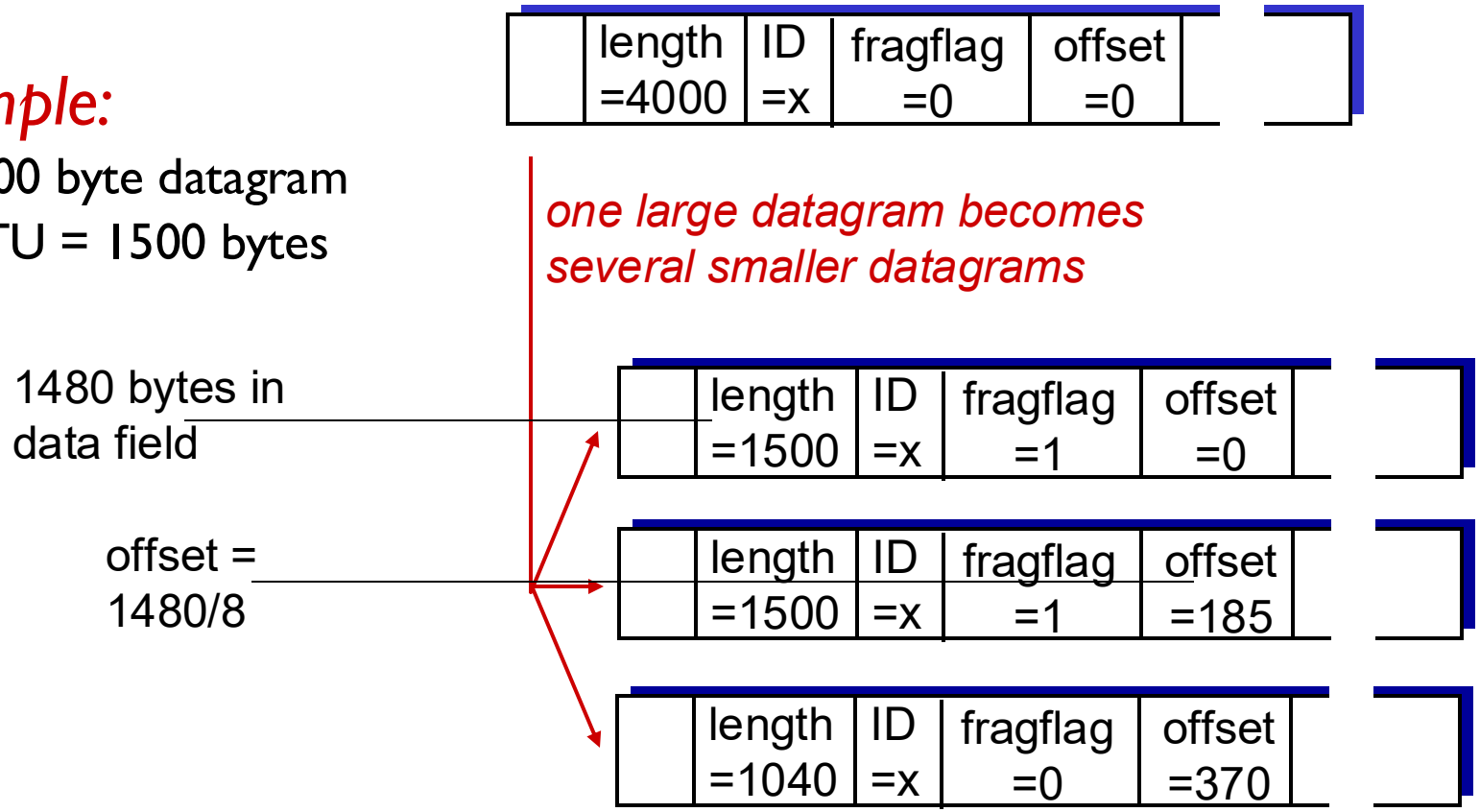
- **Network links have MTU (max.transfer size) - largest possible link-level frame**
  - Different link types, different MTUs
- **Large IP datagram divided ("fragmented") within net**
  - One datagram becomes several datagrams
  - "Reassembled" only at final destination
  - IP header bits used to identify, order related fragments



IP fragmentation, reassembly

example:

- 4000 byte datagram
- MTU = 1500 bytes



Thanks.