Fundamentals of Packet-Switched Networks

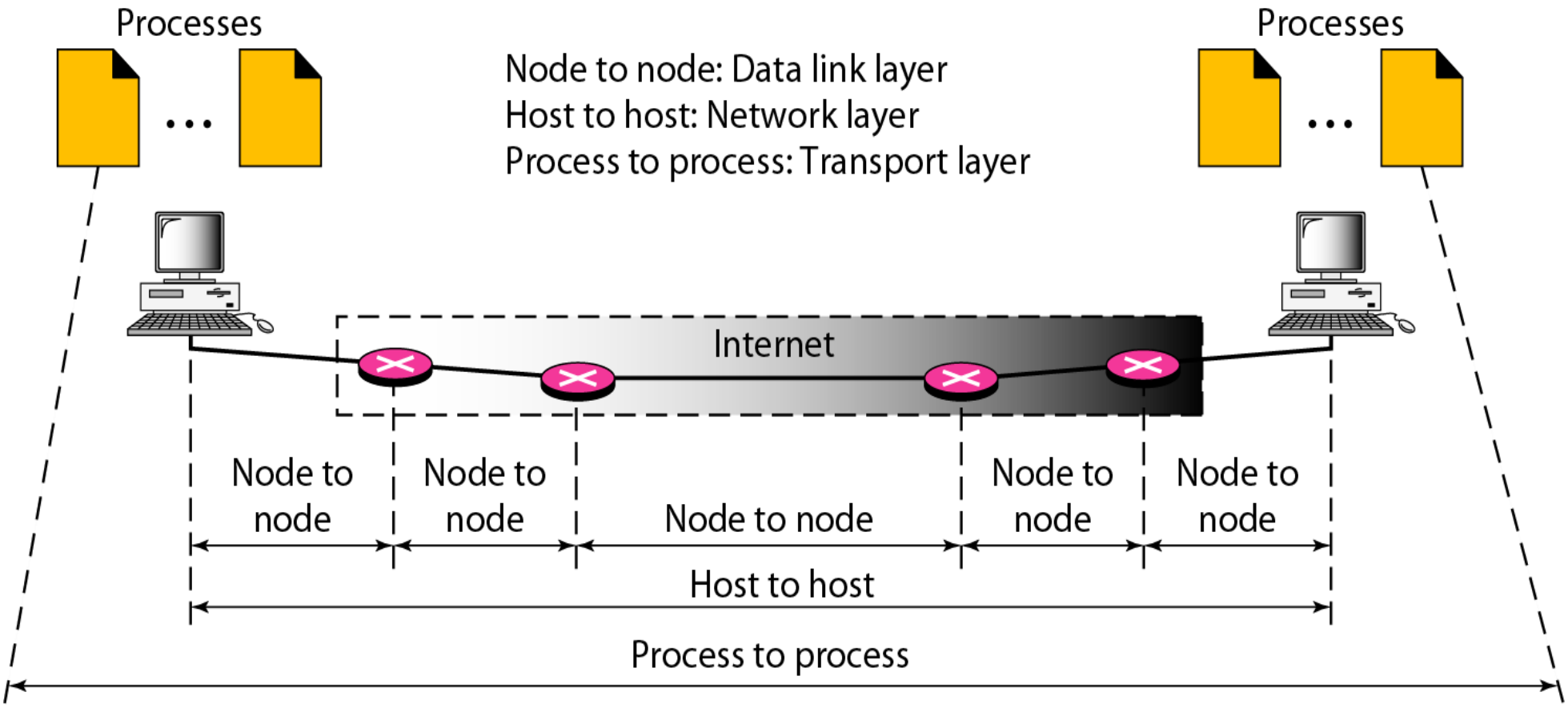
Peerapon S.

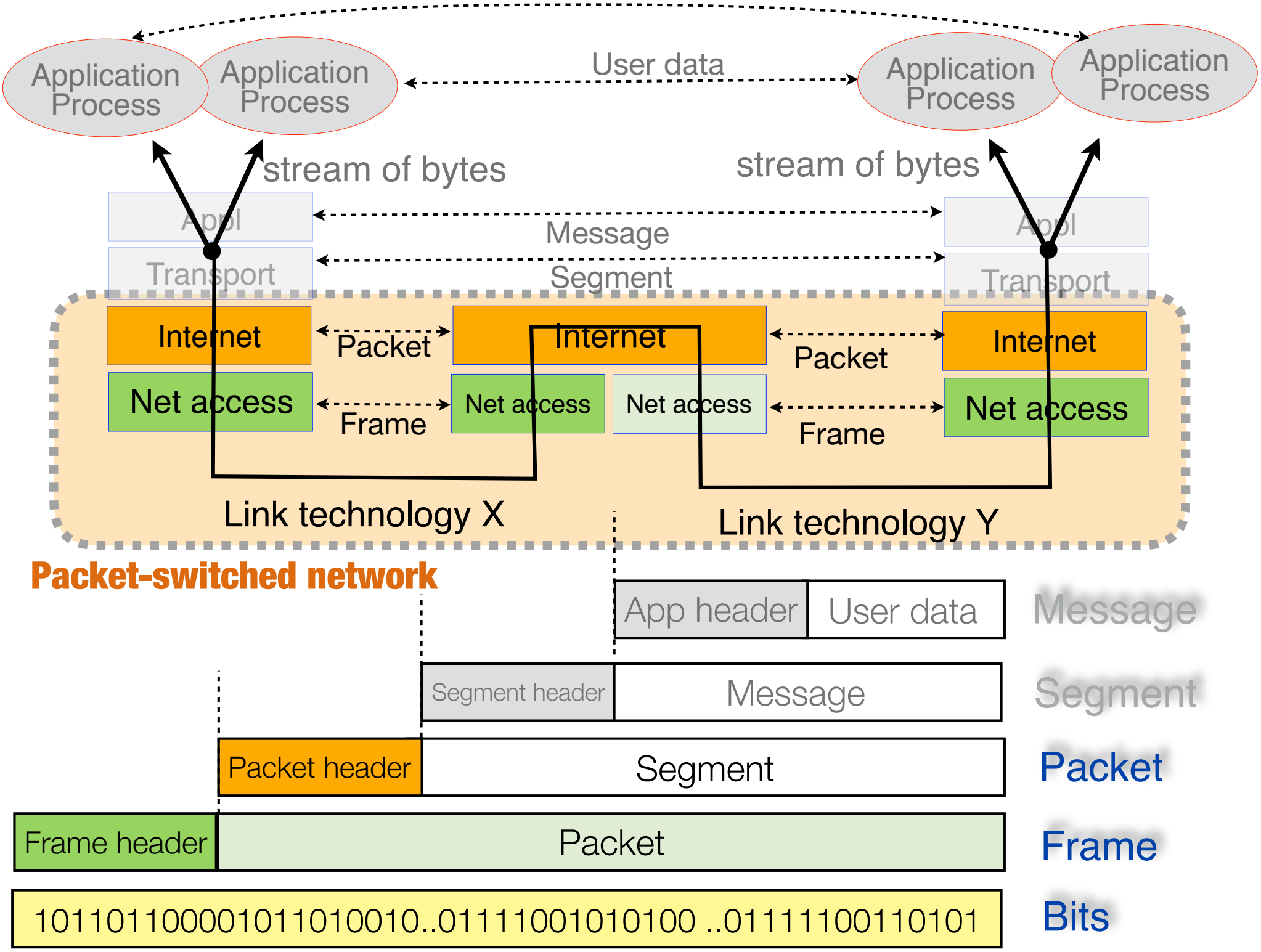
CPE 314: Computer Networks (2/61)

Topics

- Principles of packet switching
- Router architecture
- Types of packet-switched networks
- Network performance measures
- Network congestion and TCP congestion control

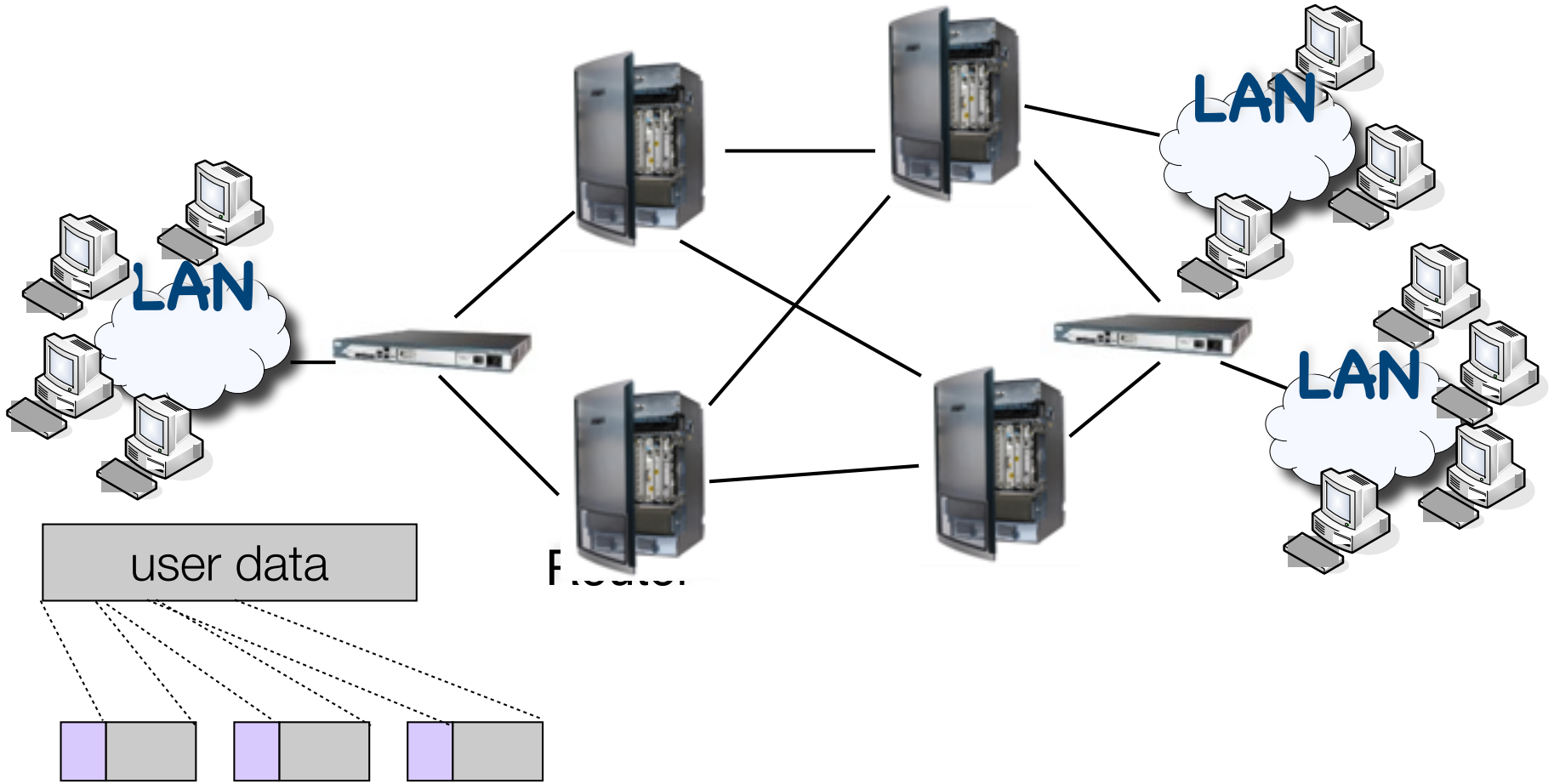
- Reading
  - Forouzan text, Ch 3.4, 4.1, 4.2

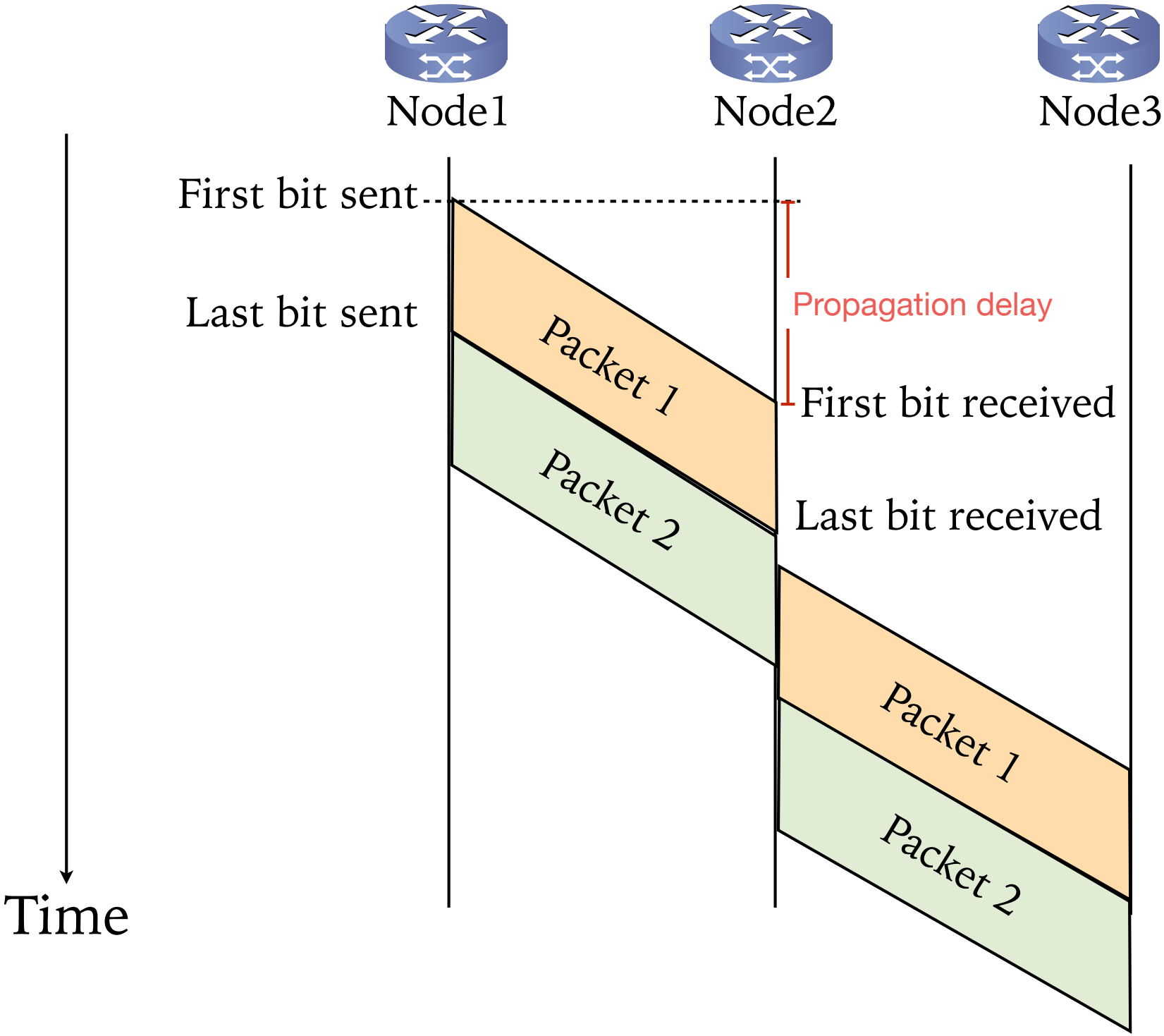

Processes
Processes
Node to node: Data link layer
Host to host: Network layer
Process to process: Transport layer
Internet
Node to node
Node to node
Node to node
Node to node
Node to node
Host to host
Process to process

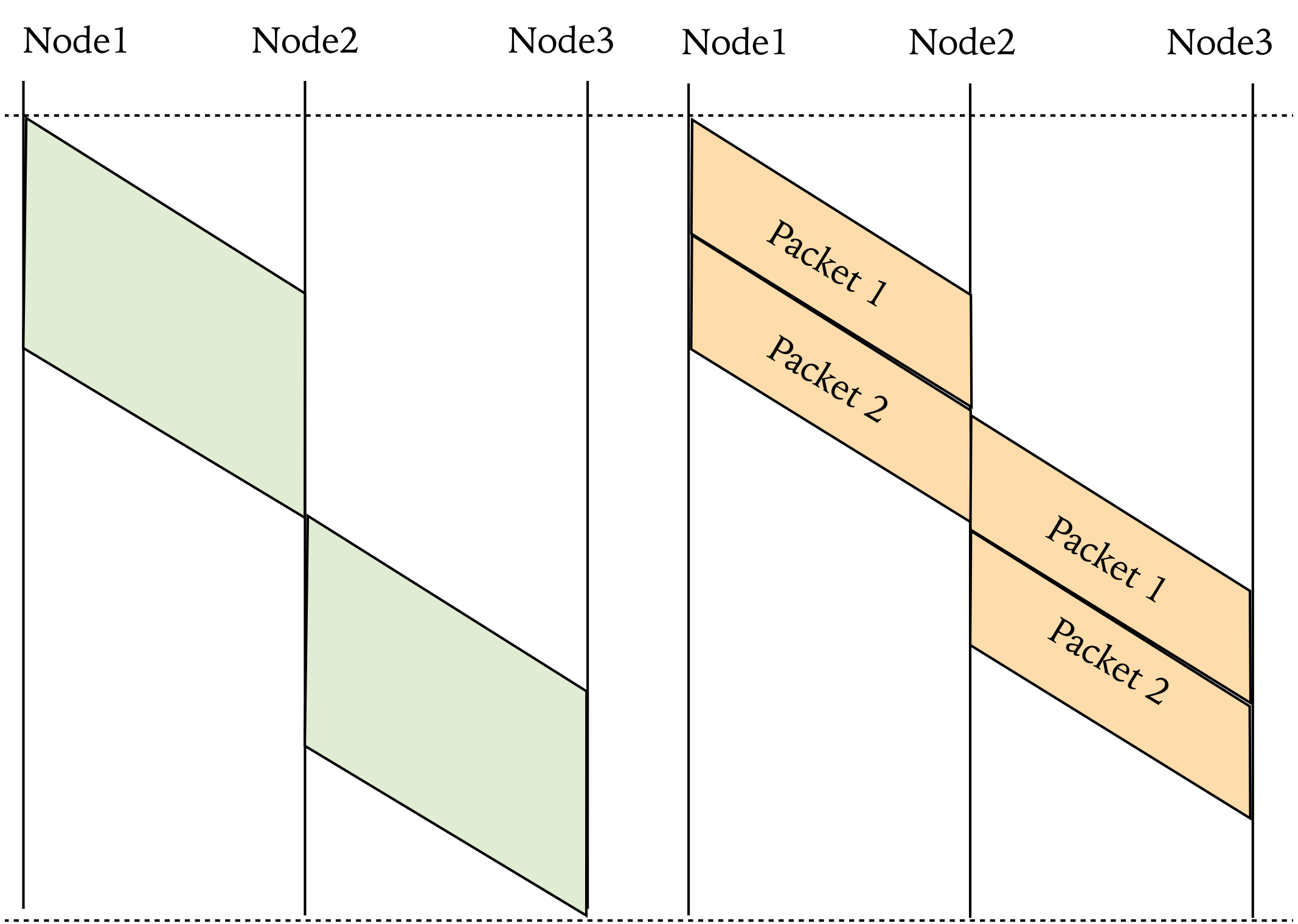

Application Process
Application Process
User data
Application Process
Application Process
stream of bytes
stream of bytes
Appl
Transport
Message
Segment
Appl
Transport
Internet
Packet
Internet
Packet
Internet
Net access
Frame
Net access
Net access
Frame
Net access
Link technology X
Link technology Y
Packet-switched network
App header
User data
Message
Segment header
Message
Segment
Packet header
Segment
Packet
Frame header
Packet
Frame
1011011000010110100 10..01111001010100 ..01111100110101
Bits

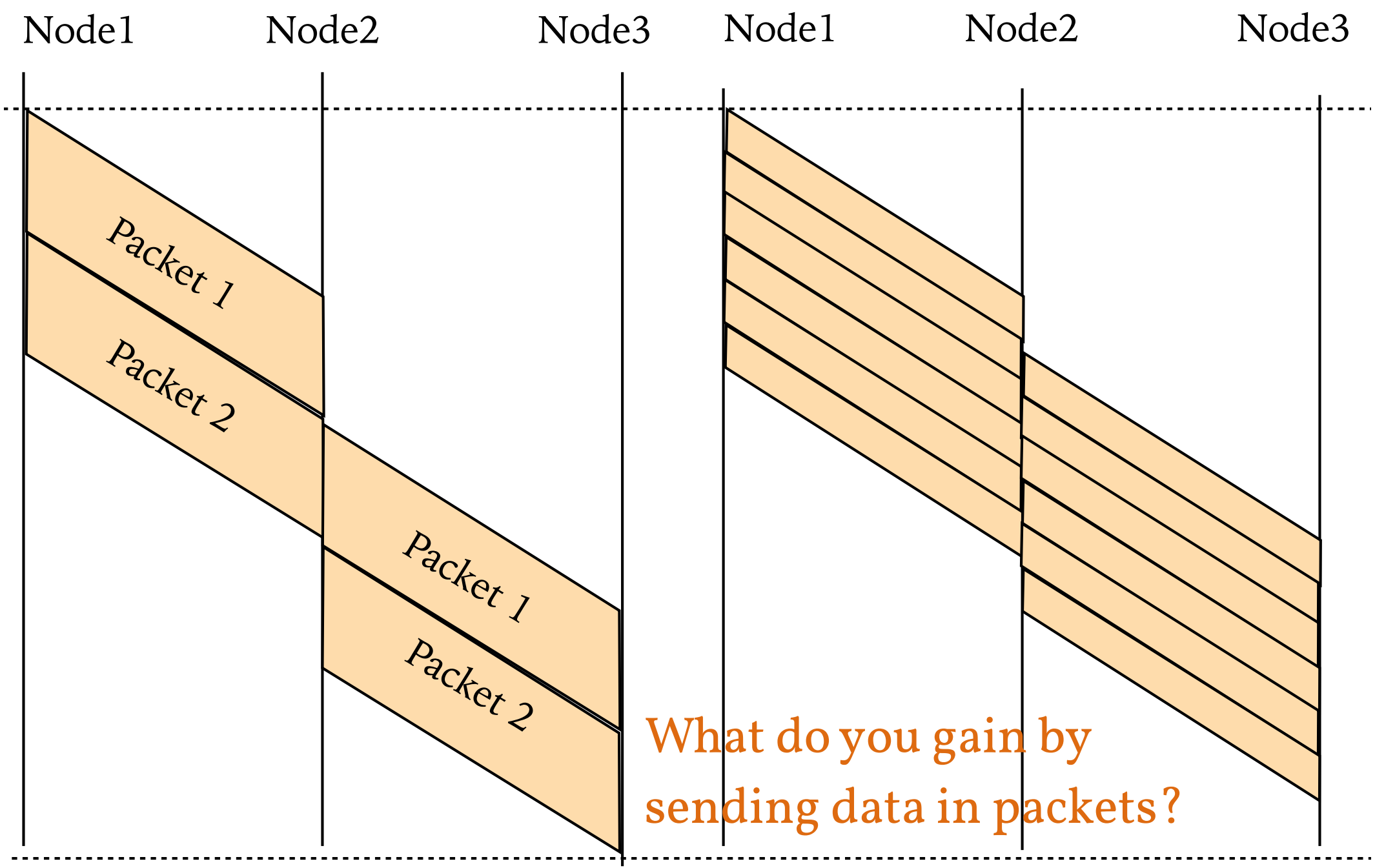
Packet-Switched Network

- Store-and-forward operation
- Link bandwidth sharing

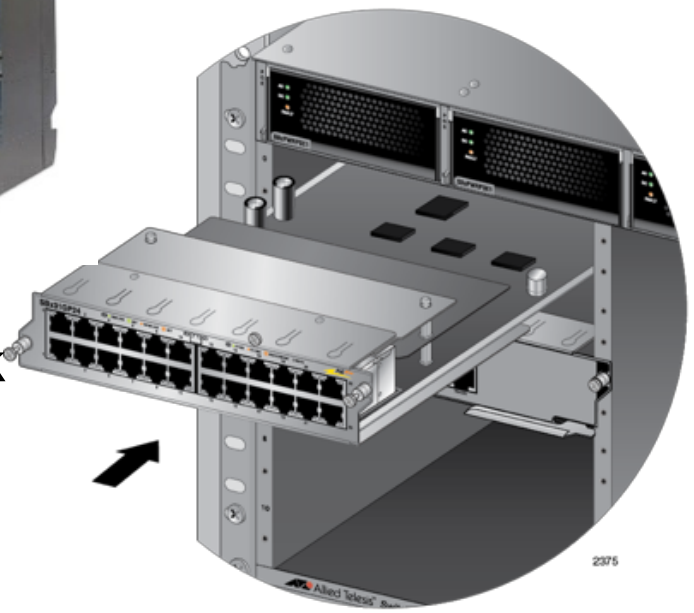


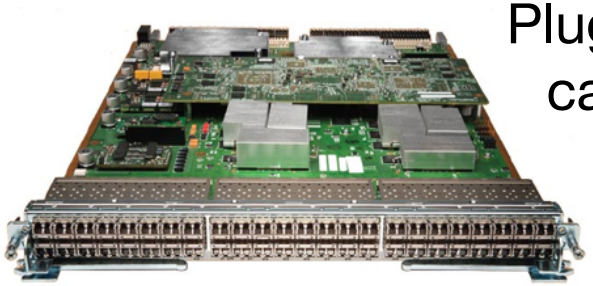

Node1
Node2
Node3
First bit sent
Last bit sent
Packet 1
Packet 2
Propagation delay
First bit received
Last bit received
Packet 1
Packet 2
Time


Node1
Node2
Node3
Node1
Node2
Node3
Packet 1
Packet 2
Packet 1
Packet 2


Node1
Node2
Node3
Packet 1
Packet 2
Packet 1
Packet 2
Node1
Node2
Node3
What do you gain by sending data in packets?

Non-modular routers used in local/access networks

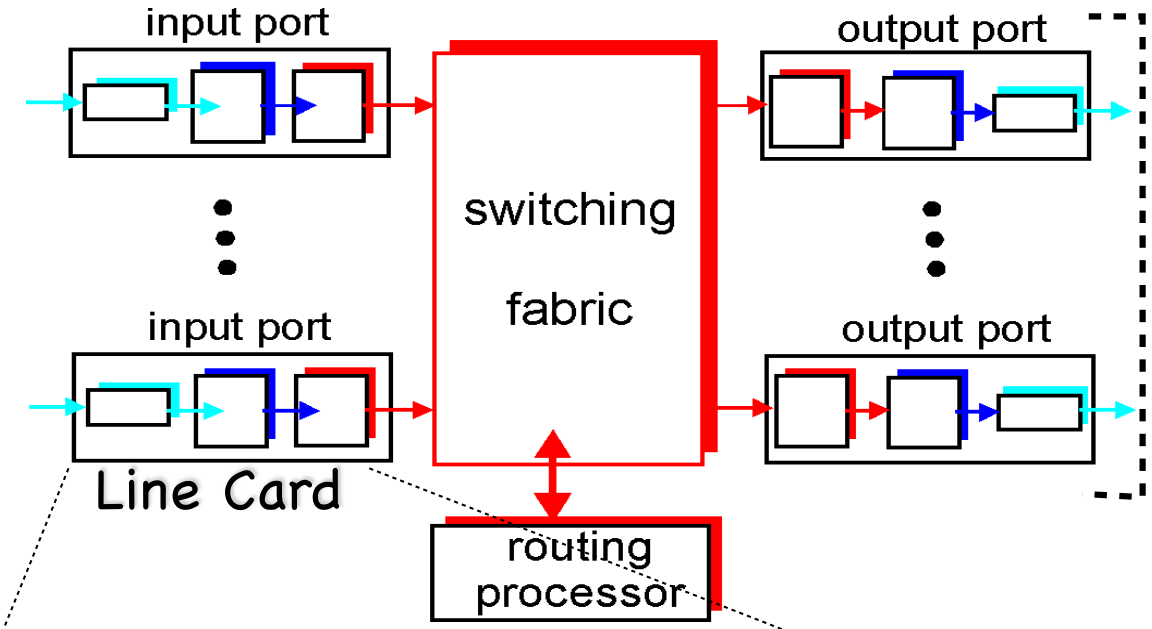




Plug-in network card module

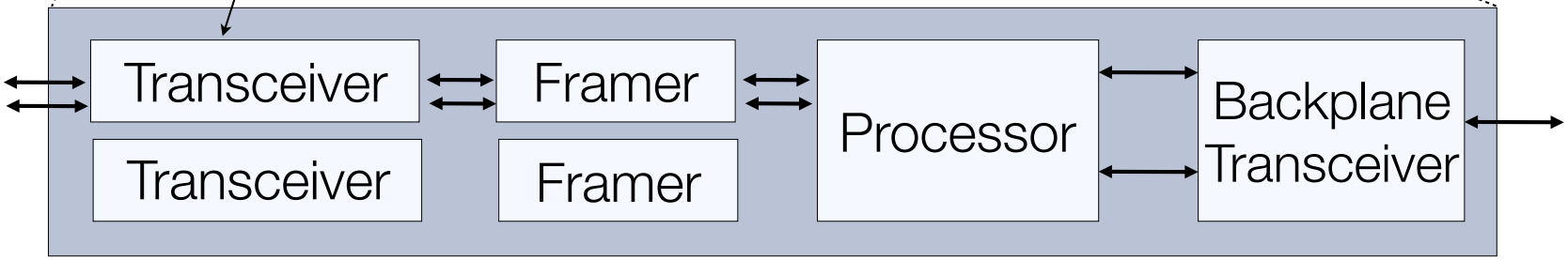
Modular routers used in backbone networks

Basic Router/Switch HW Architecture



Different card types, link speeds

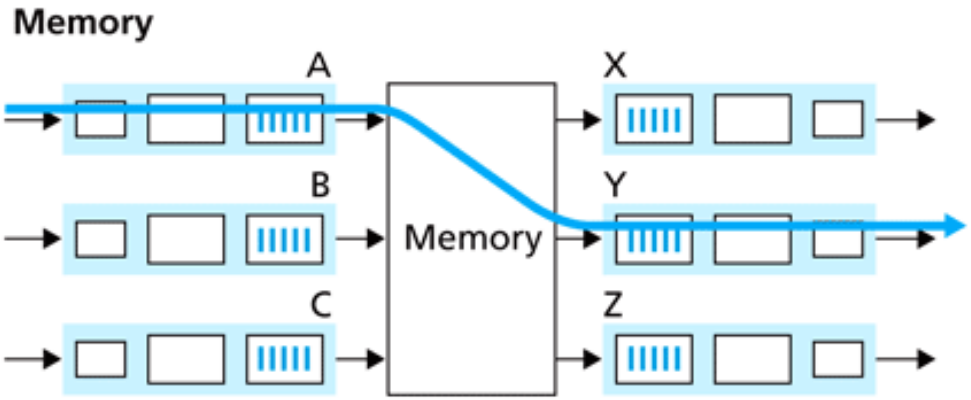
Input+output ports

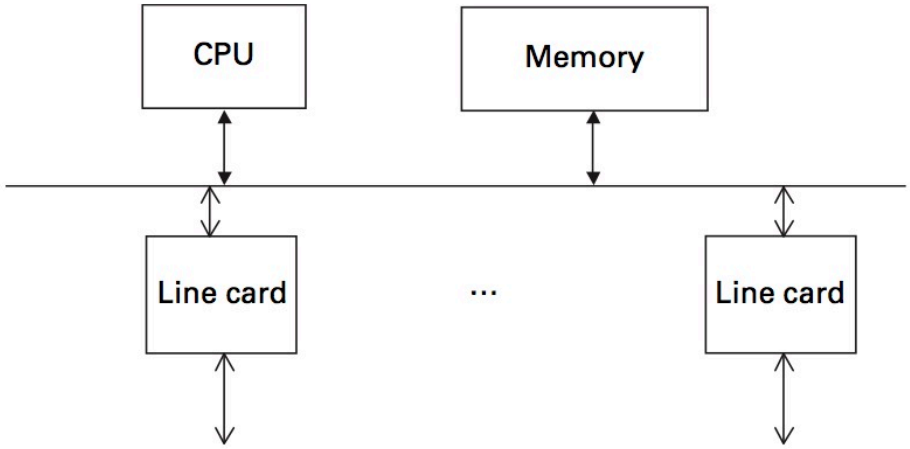


To switch fabric

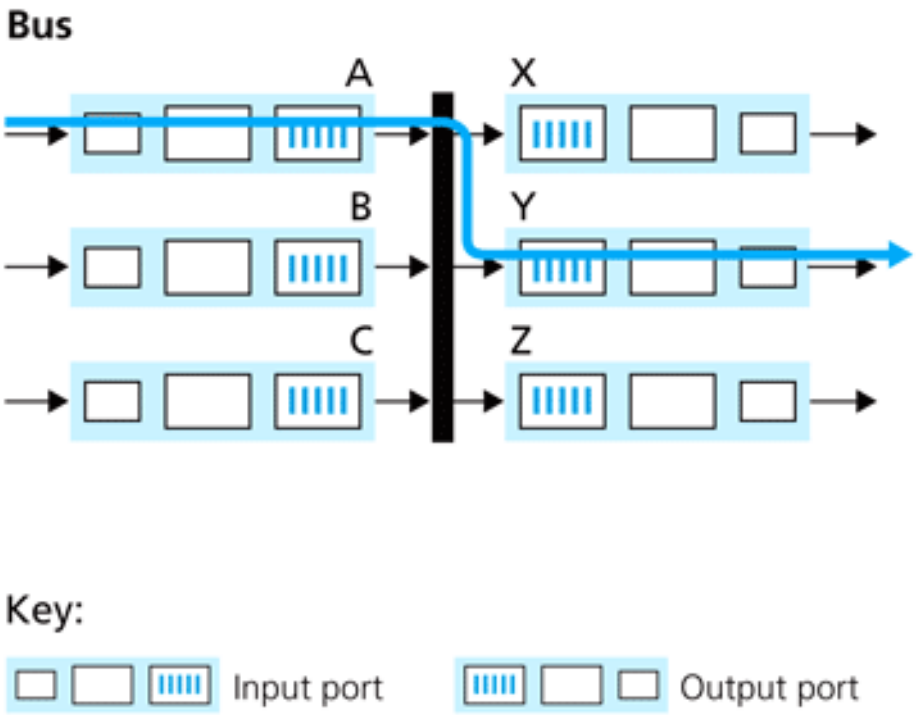
Switching Fabrics

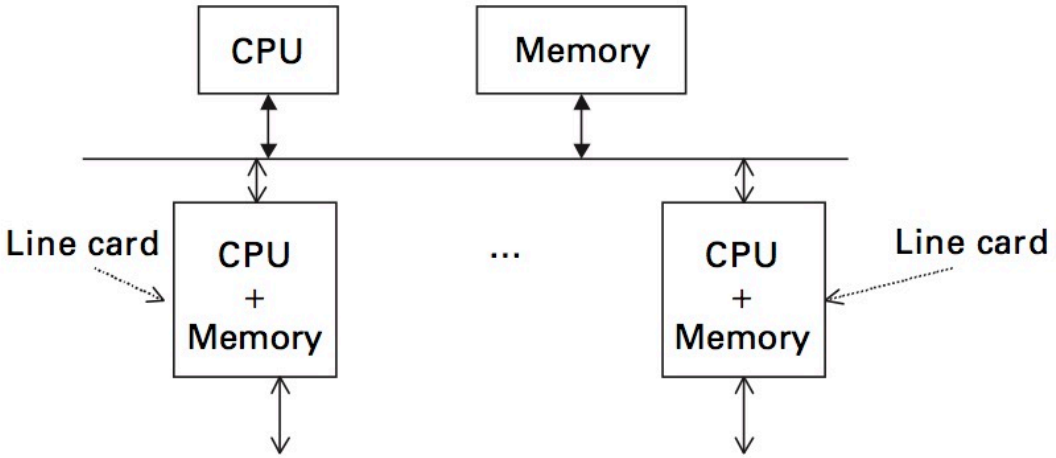
- **Gen 1**: All packets processed by Router CPU
  - Incoming packets from line card copied to memory
  - Outgoing packets sent to line cards
- Ex: Home-grade routers, PC routers.

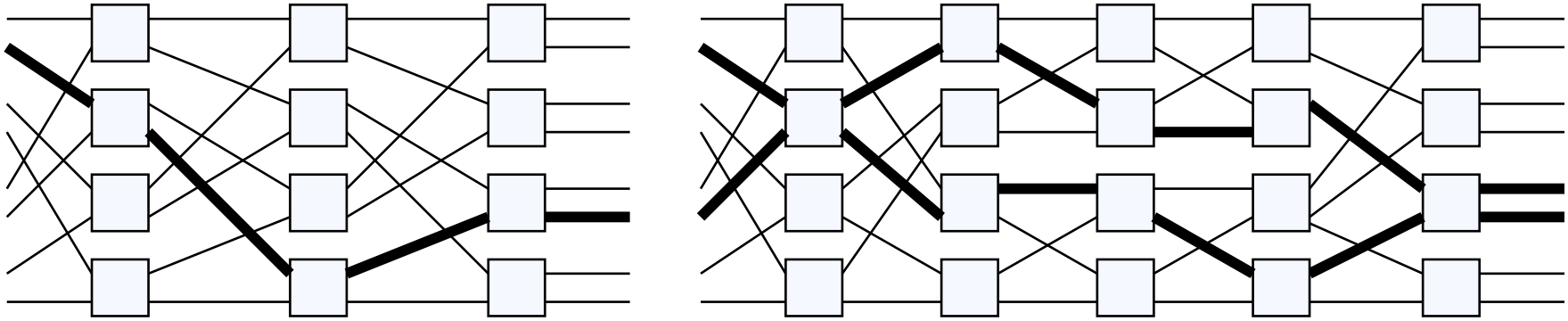




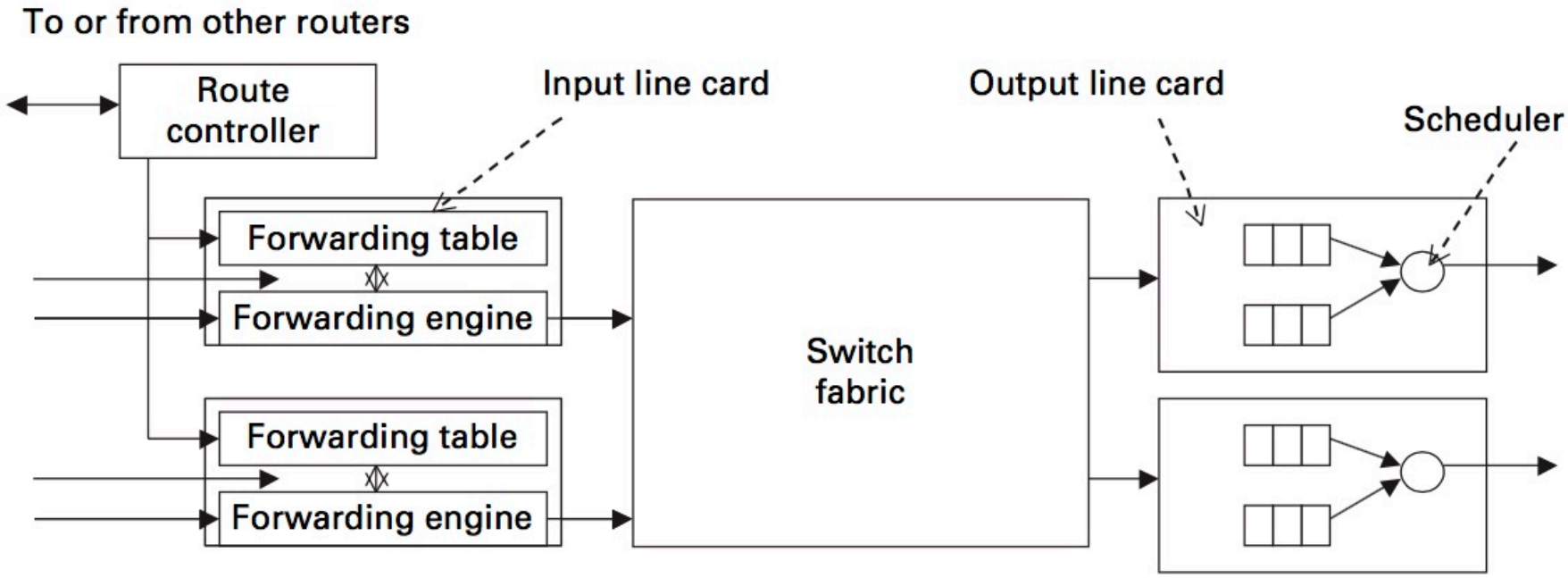
- **Gen 2**: Packets processed by line card CPUs
  - Forwarding engine and table inside line card
  - Packets transferred directly among ports over bus
- Ex: small to medium (non-modular) routers, e.g., Cisco 1xxx, 2xxx



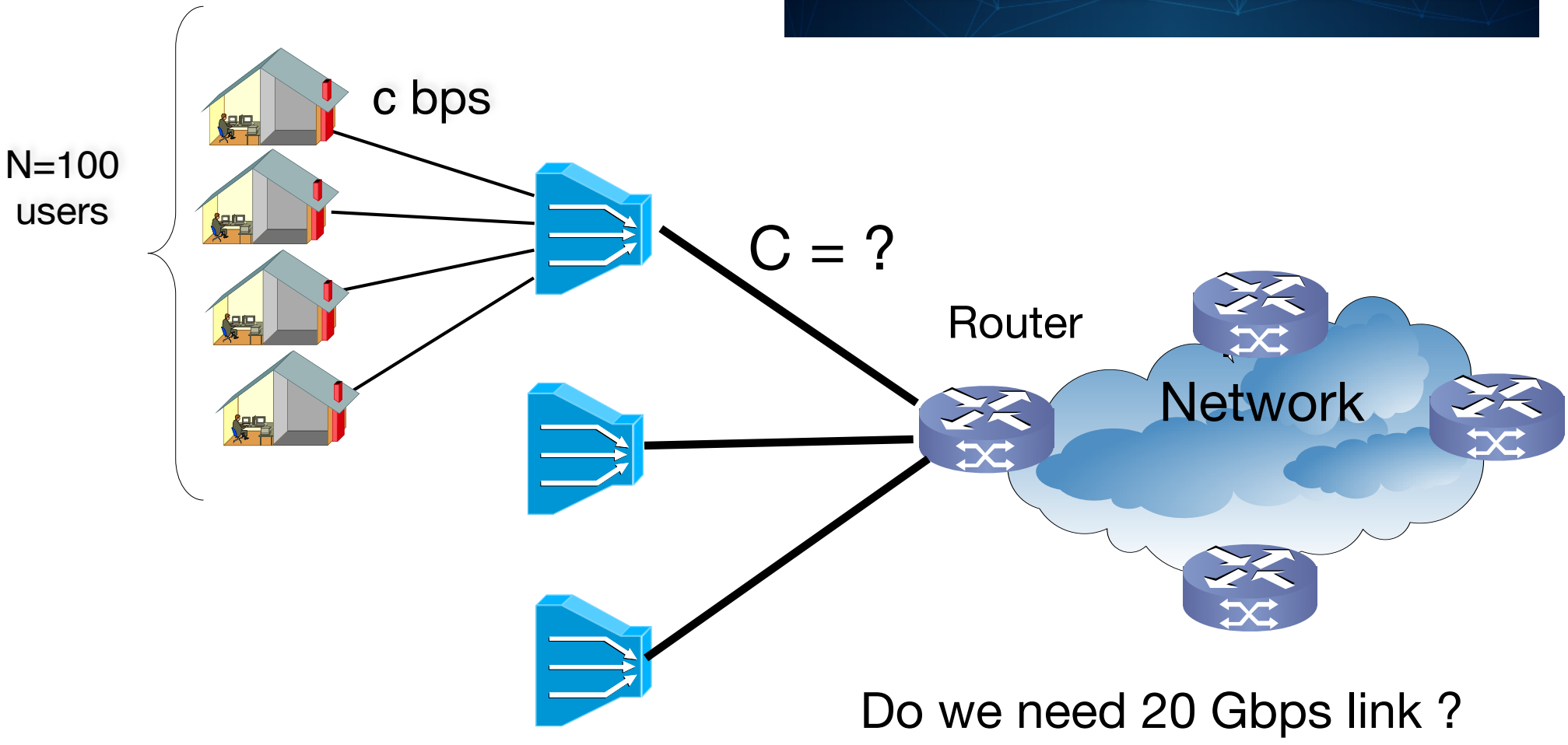




- **Gen 3**: Replace bus by interconnection network
  - Much faster speed with several forwardings in paralllel
  - Core network (modular) routers, e.g., Cisco 1xxxx




เลือกแพ็กเกจที่เป็นคุณ
เกมเมอร์ และนักกีฬาอีสปอร์ต
ทั้งเล่นเกม และไลฟ์สด
ไฟเบอร์ความเร็วสูงสุด
200/200
Mbps
เน็ตมือถือ 4G|3G
แรงเต็มสปีด
10GB
899 บาท/เดือน
ฮาร์ดคอร์เกมเมอร์ ทั้งเกม ไลฟ์สด
ดูหนังพร้อมกัน หลายจอได้
ไฟเบอร์ความเร็วสูงสุด
300/300
Mbps
เน็ตมือถือ 4G|3G
แรงเต็มสปีด
10GB
999 บาท/เดือน
โปรแกรมเมอร์
และนักกีฬาอีสปอร์ตตัวจริง
ไฟเบอร์ความเร็วสูงสุด
1000/500
Mbps
เน็ตมือถือ 4G|3G
แรงเต็มสปีด
10GB
1,999 บาท/เดือน


N=100
users
c bps
C = ?
Router
Network
Do we need 20 Gbps link ?

Statistical Multiplexing Gain

- Assume that P(User active) = 0.2
  - Download speed c=200 Mbps when active
  - Chance to see download speed less than adverstised ~ 5%

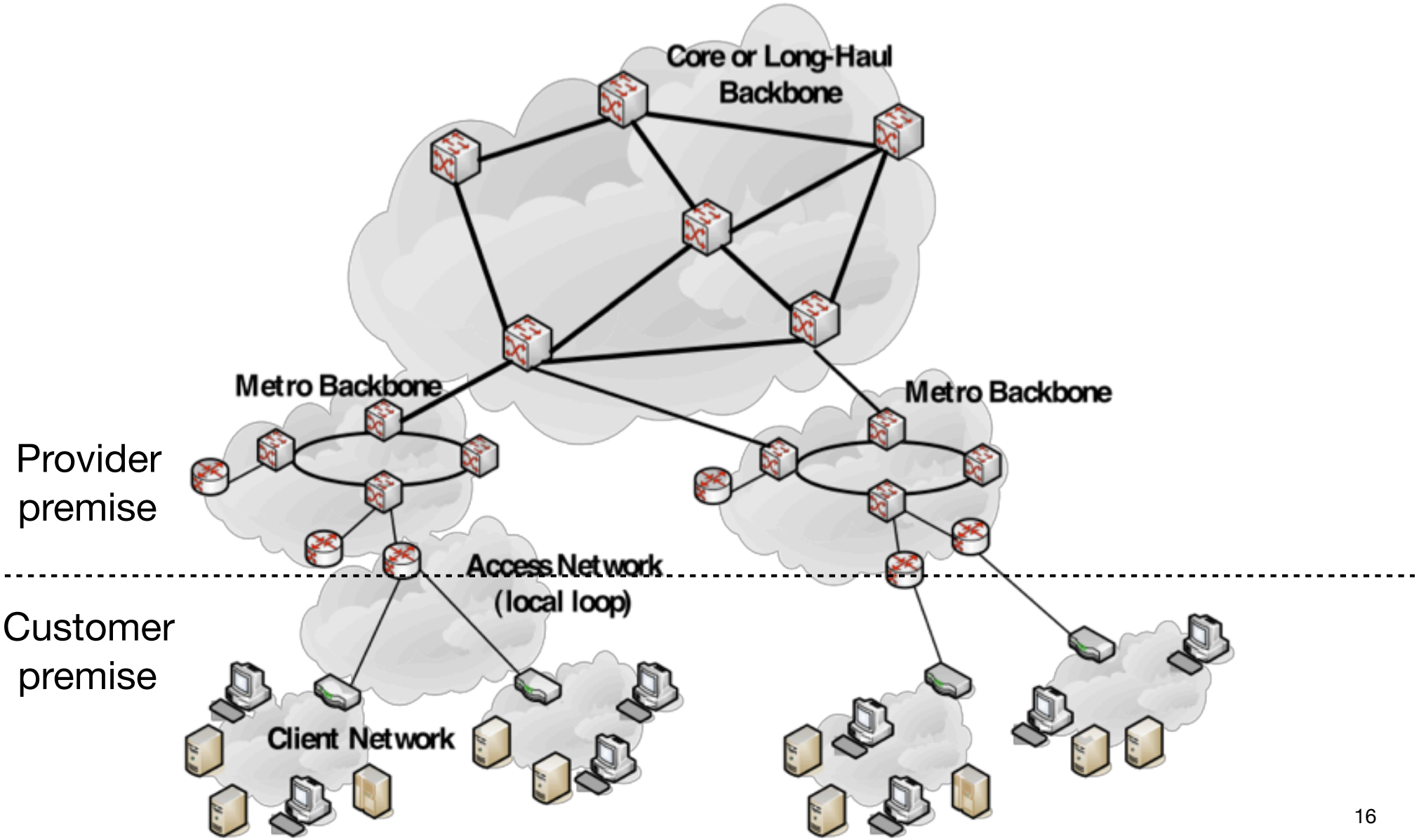
$$\mathbb{P}(\text{Total download rate} > C) = 0.05$$

$$\mathbb{P}(\text{Number of active users} > C/c) = 0.05$$

$$C = 5 \quad \text{Gbps}$$

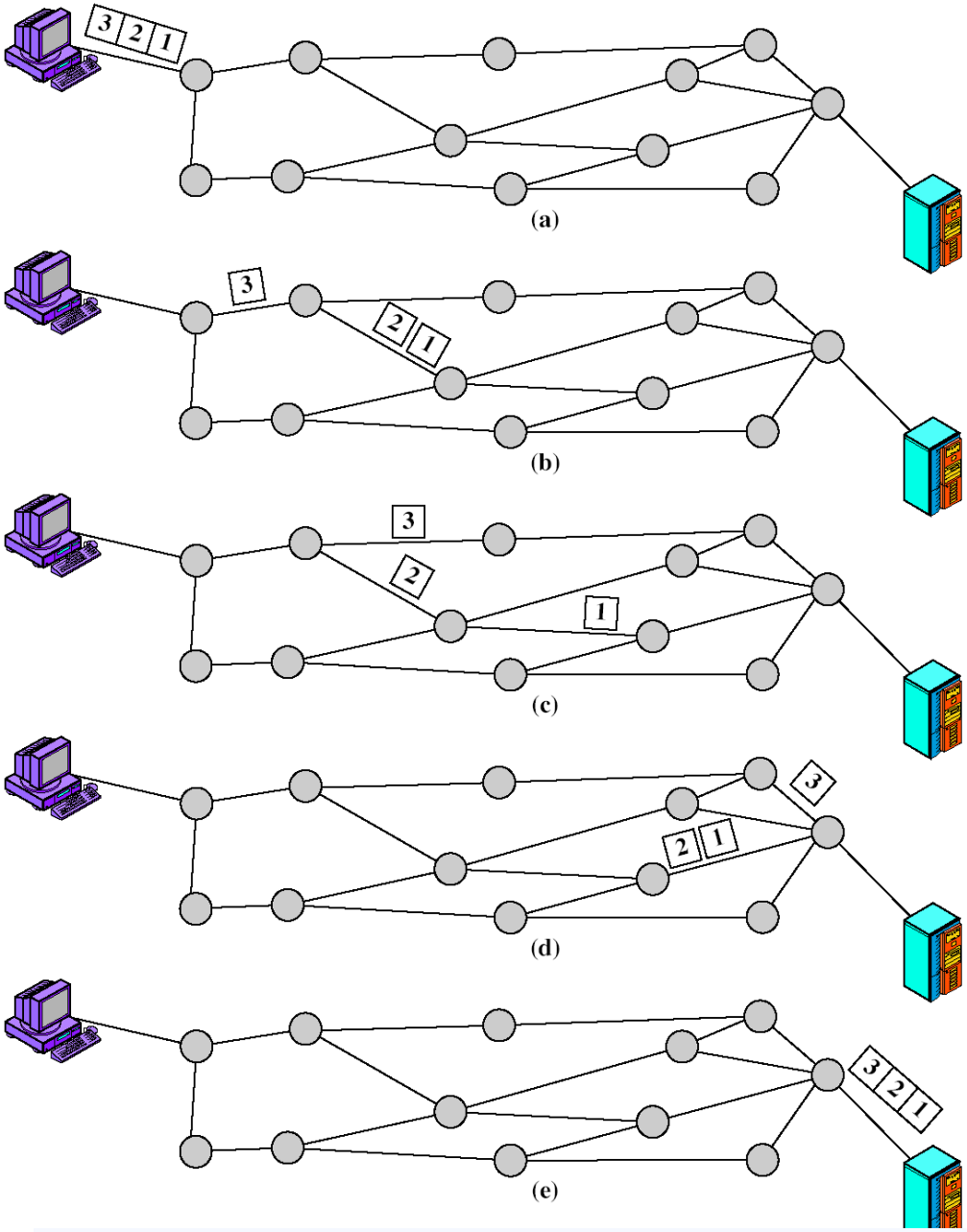
- Multiplexing gain = Total peak rate/Actual link rate =100*200/5000 = 4

Typical Large-scale Network Infrastructure

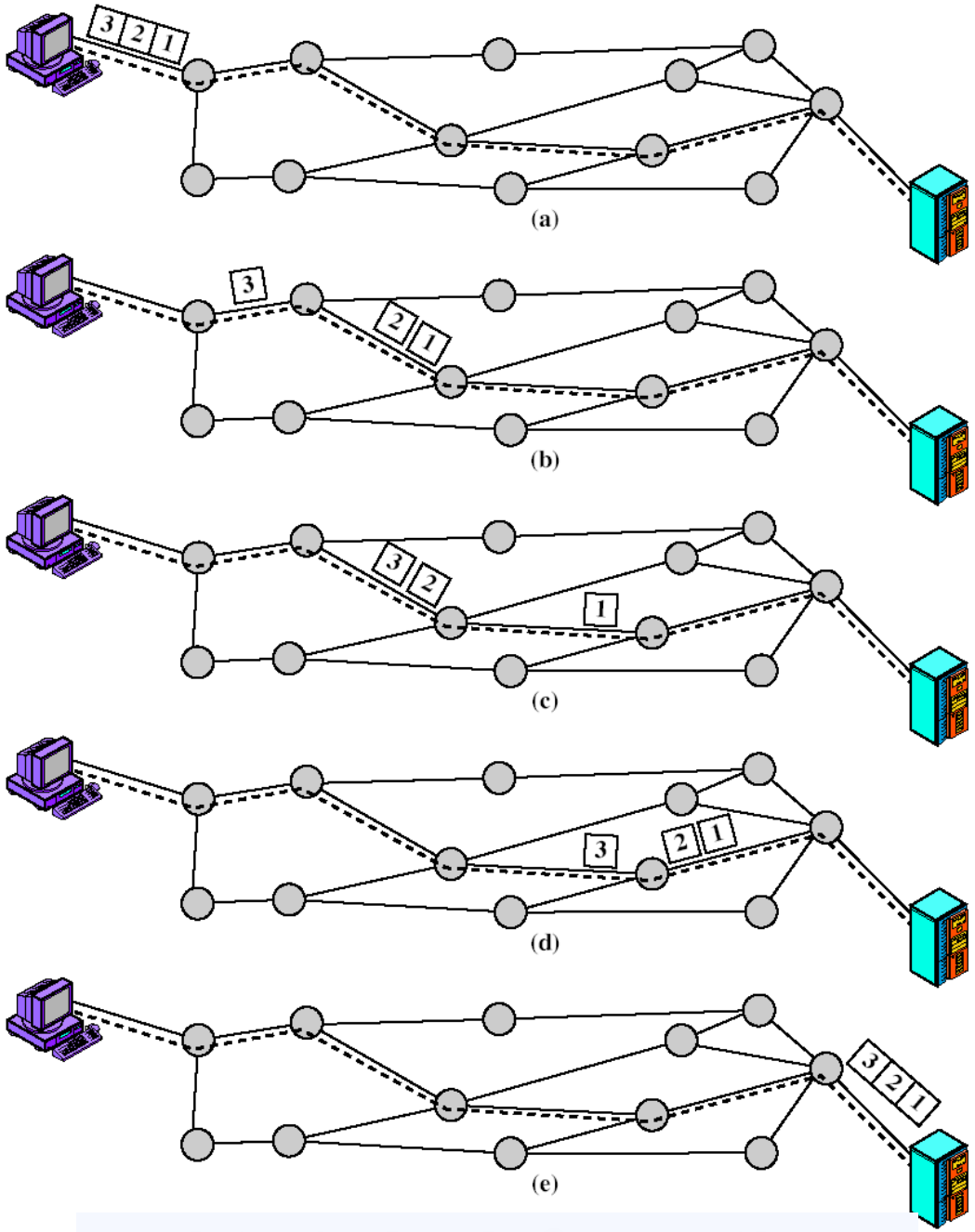


Who are in the path? Traceroute

```
$ tracert 85.17.172.180
Tracing route to 85.17.172.180 over a maximum of 30 hops
  1     2 ms     1 ms     1 ms  . [192.168.2.1]
  2    28 ms    27 ms    27 ms  ppp-61-90-69-1.revip.asianet.co.th [61.90.69.1]
  3    27 ms    27 ms    27 ms  ppp-210-86-189-12.revip.asianet.co.th [210.86.189.12]
  4    33 ms    30 ms    30 ms  10.169.12.1
  5    30 ms    27 ms    27 ms  58-97-4-46.static.asianet.co.th [58.97.4.46]
  6    30 ms    30 ms    29 ms  58-97-4-41.static.asianet.co.th [58.97.4.41]
  7    29 ms    29 ms    29 ms  61-91-210-50.static.asianet.co.th [61.91.210.50]
  8    28 ms    29 ms    28 ms  203-144-144-28.static.asianet.co.th [203.144.144.28]
  9    29 ms    29 ms    29 ms  61-91-210-5.static.asianet.co.th [61.91.210.5]
 10    63 ms    63 ms    66 ms  tig-net28-157.trueinternetgateway.com [122.144.28.157]
 11    65 ms    67 ms    65 ms  SG-ICR-ANC2-26-214.trueinternetgateway.com
    [122.144.26.214]
 12   270 ms   267 ms   270 ms  TIG-Net26-94.trueinternetgateway.com [122.144.26.94]
 13   347 ms   347 ms   353 ms  lon.tc2.leaseweb.net [195.66.225.100]
 14   351 ms   347 ms   349 ms  te5-4.evo-hv1.leaseweb.net [85.17.191.250]
 15   352 ms   350 ms   351 ms  85.17.172.180
Trace complete.
```

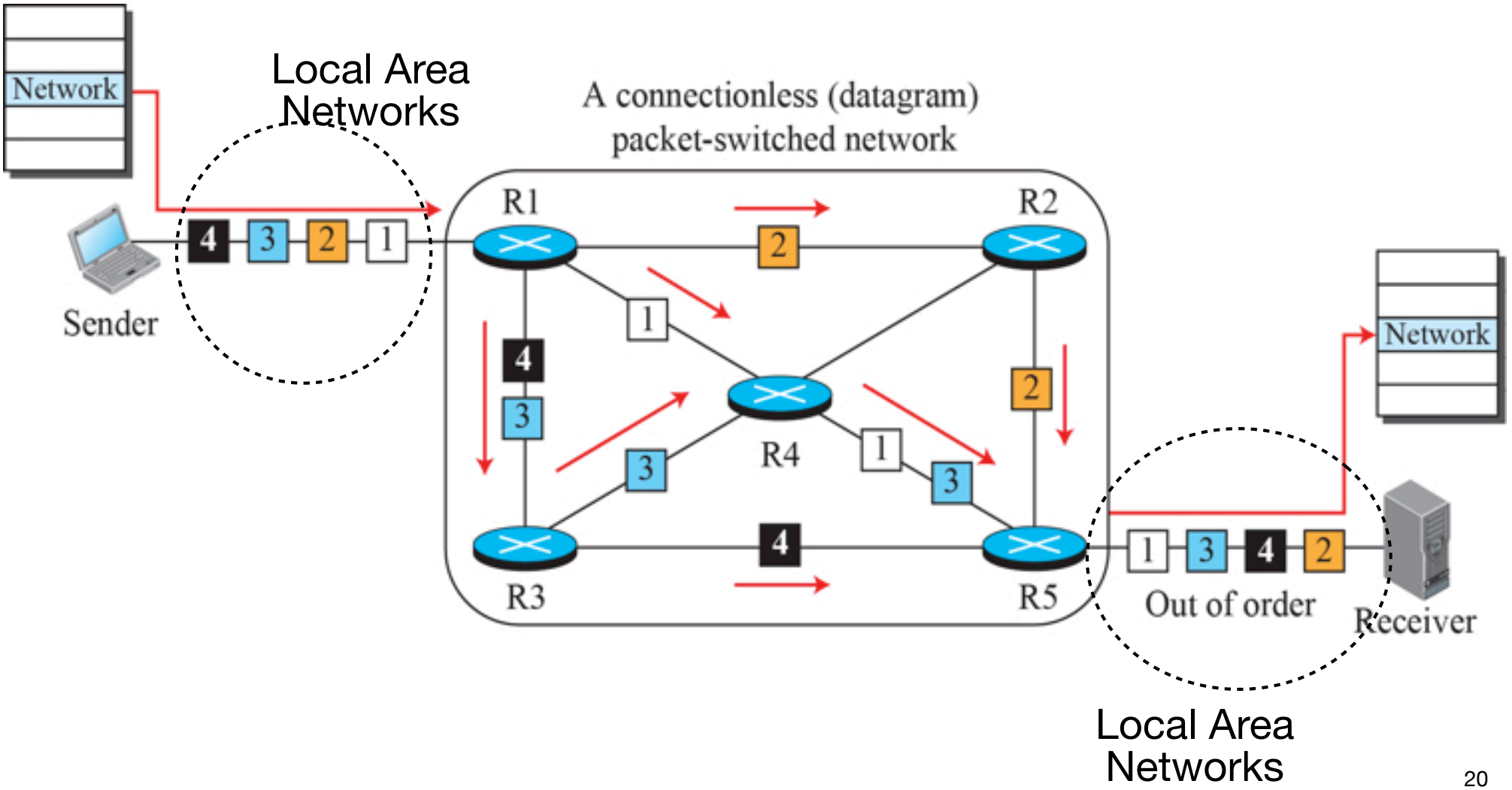


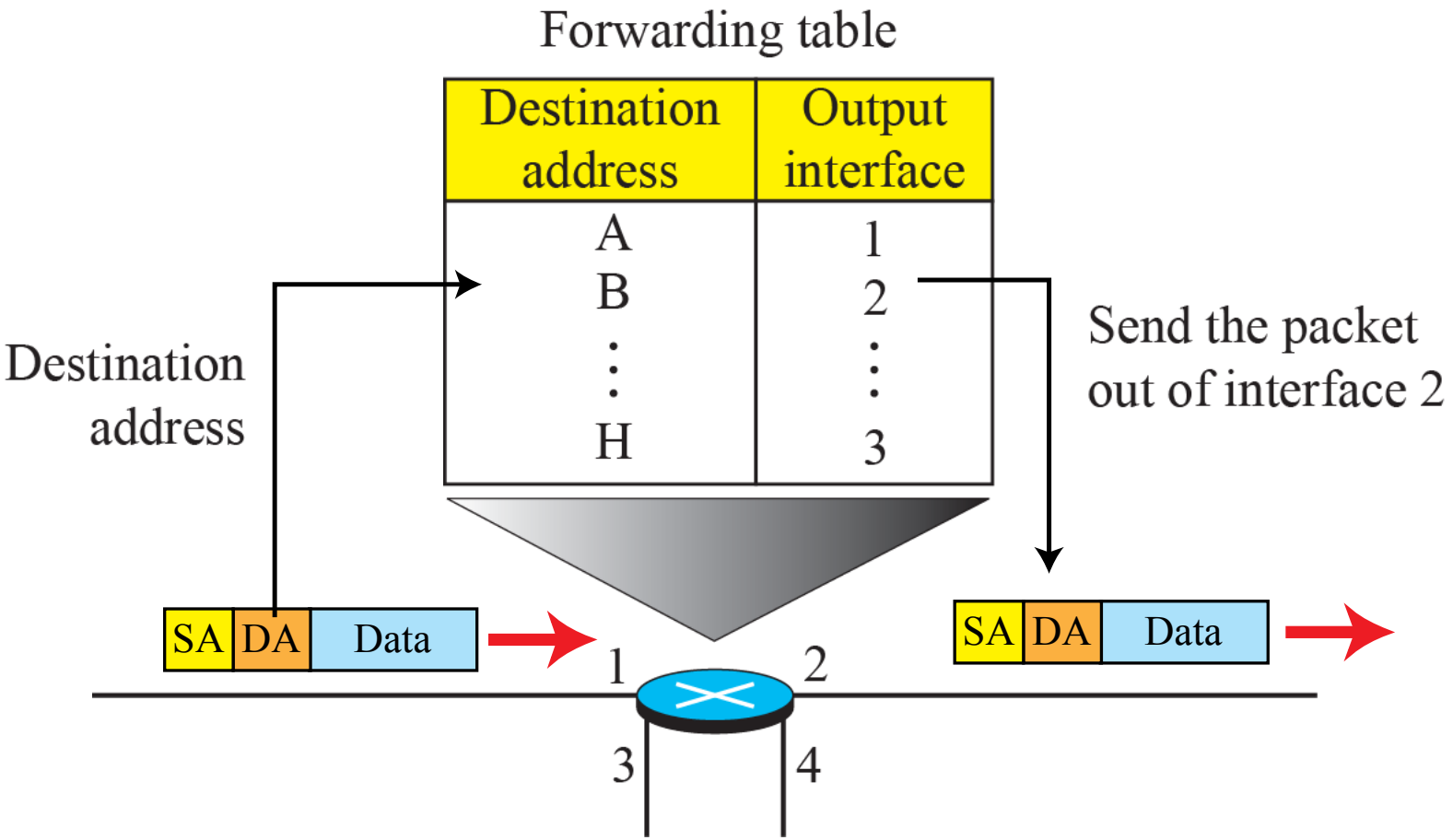
Datagram (connectionless) network



Virtual Circuit (connection-oriented) network

Datagram (Connectionless) Network

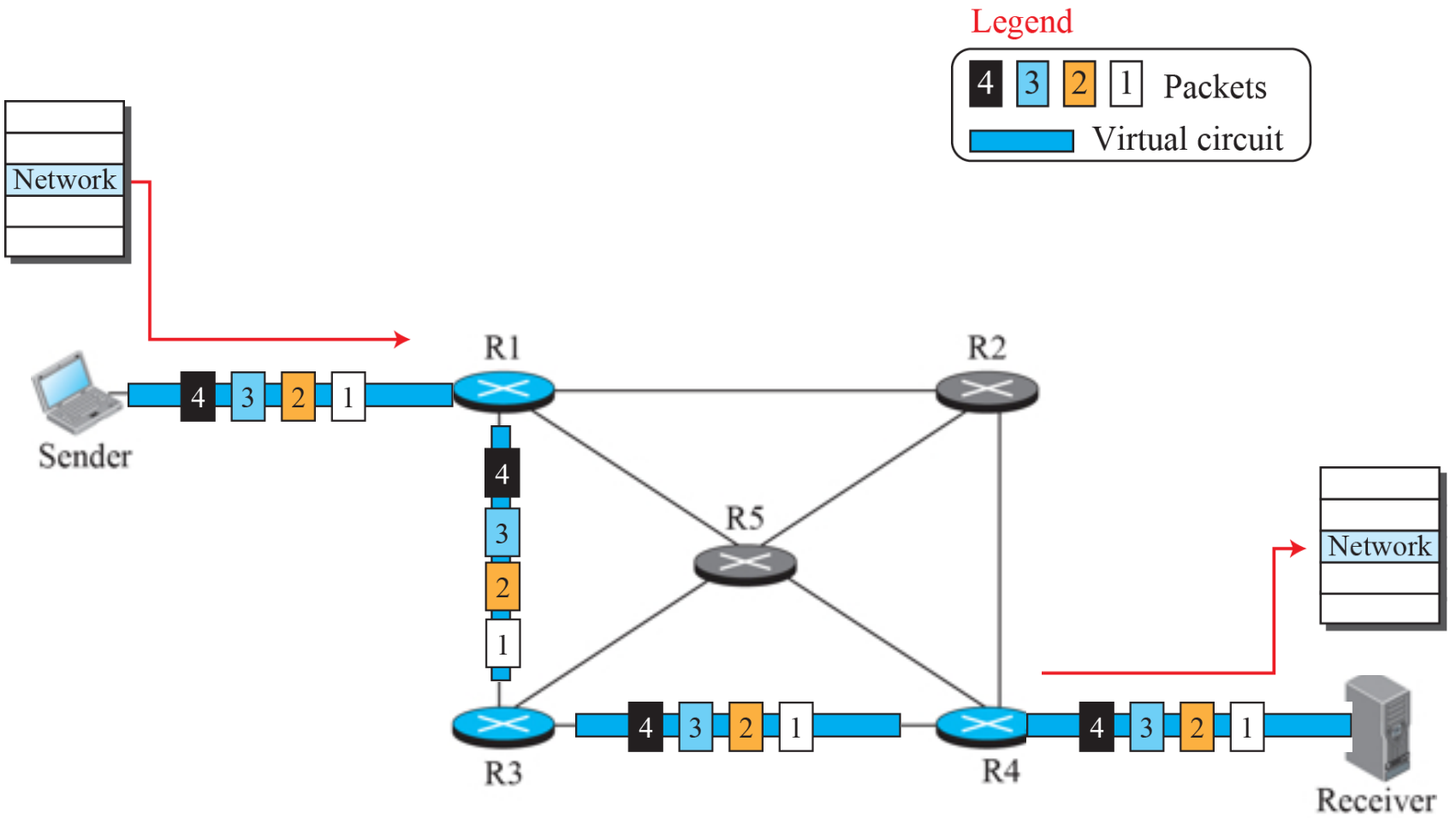




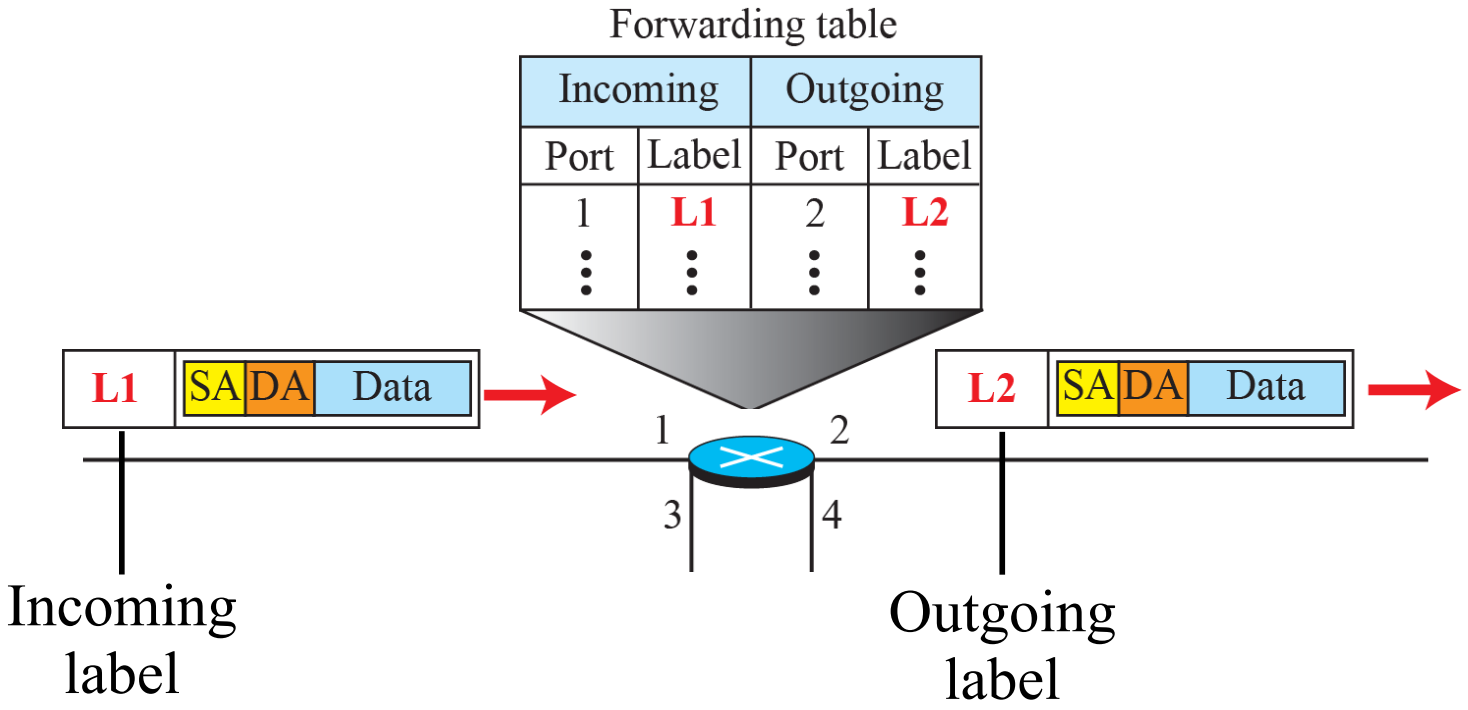
SA: Source Address

DA: Destination Address

Virtual Circuit (Connection-Oriented) Network

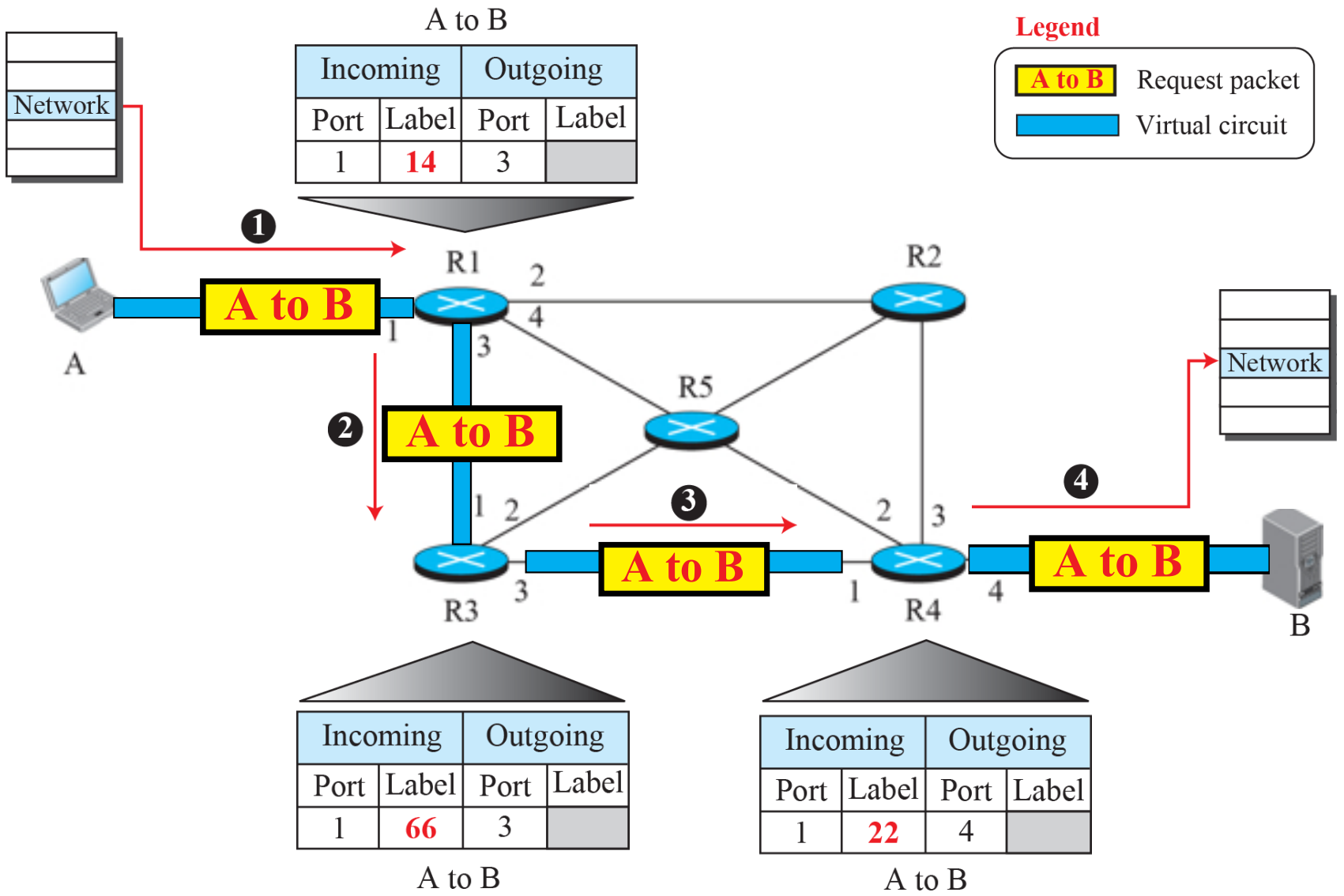


Ex: Multi-Protocol Label Switching (MPLS), X.25, Frame Relay

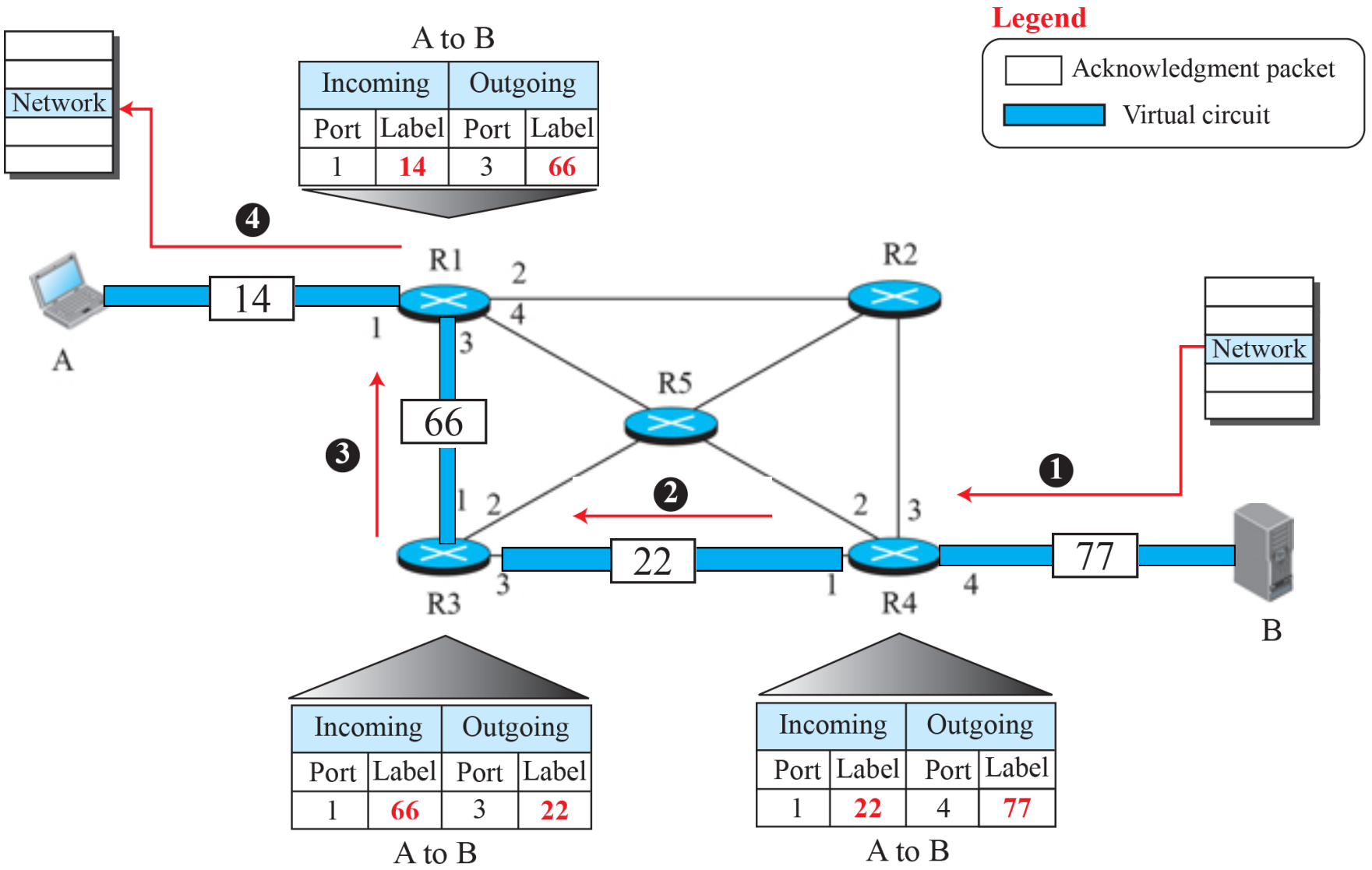

Forwarding table
Incoming
Outgoing
Port
Label
Port
Label
1
L1
2
L2
L1
SA
DA
Data
L2
SA
DA
Data
1
2
3
4
Incoming label
Outgoing label

Connection Setup Phase

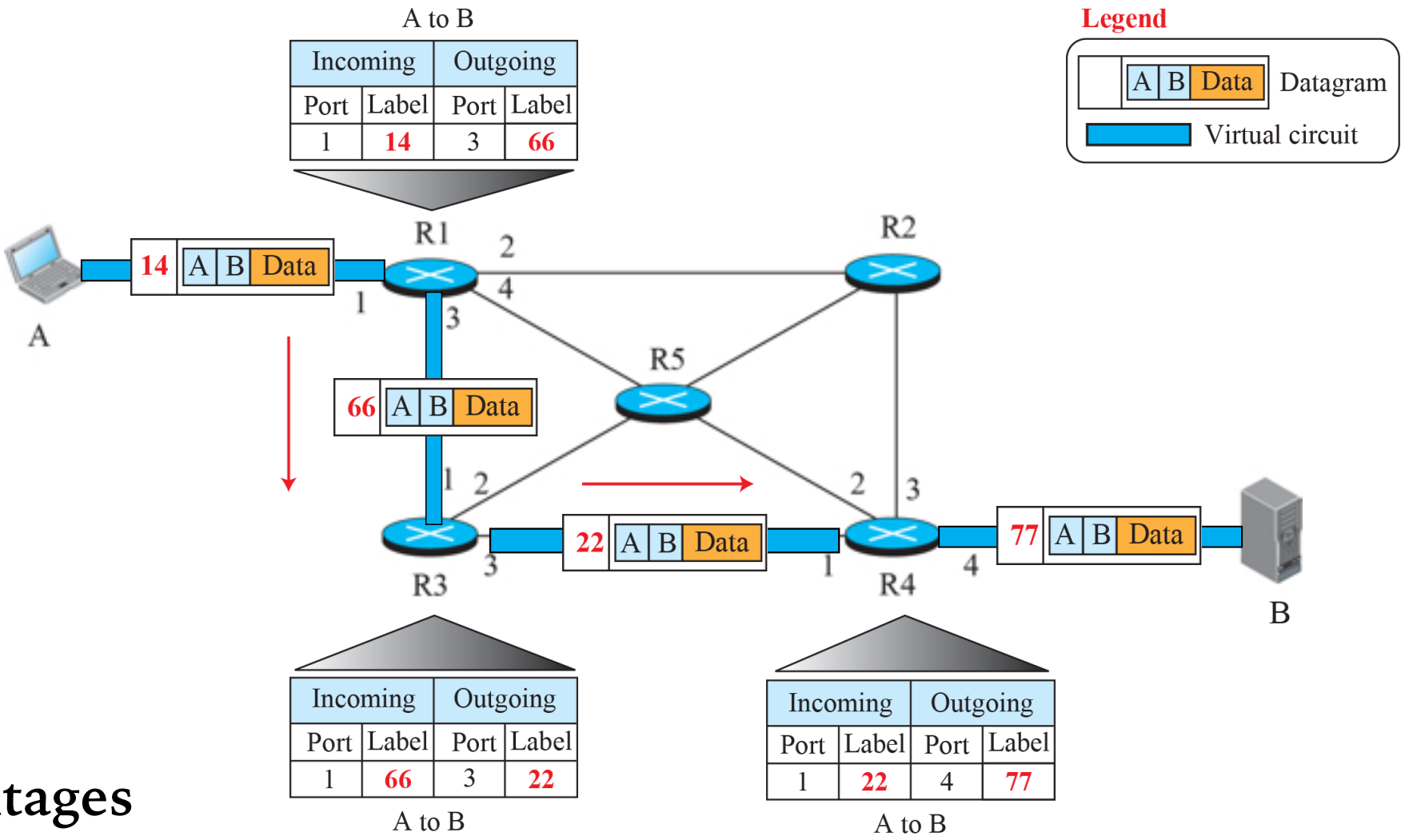
Upstream direction



Downstream direction



Packet Forwarding Process



Advantages

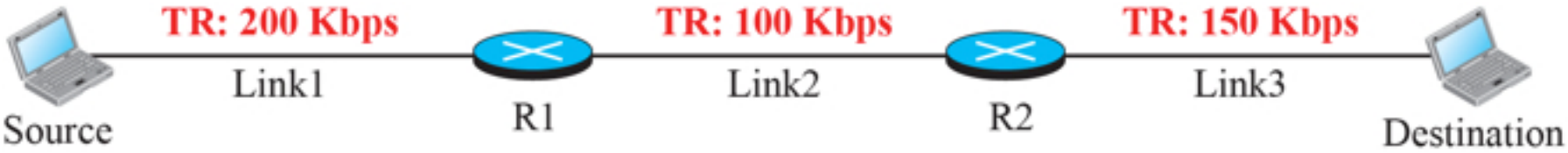
Resource management

Quality of services

Network Performance Measures

Application	Requirements		
	Loss	Throughput	Response time/Delay
File transfer, WWW	No loss	Elastic ~ 50 kbps	few secs to mins
E-mail	No loss	None	N/A
Remote access, Interactive chat	No loss	Elastic ~ few kbps	100s msec
VoIP	< 0.5%	≤ 64 kbps	≤ 300 msec
Real-time audio/video	< 1%	Audio: 5k-1Mbps Video: 10k-5Mbps	100s msec
Stored audio/video			few secs
Interactive games	< 1%	Few kbps	100s msec
Telemetry	< 5%	Few kbps	few secs

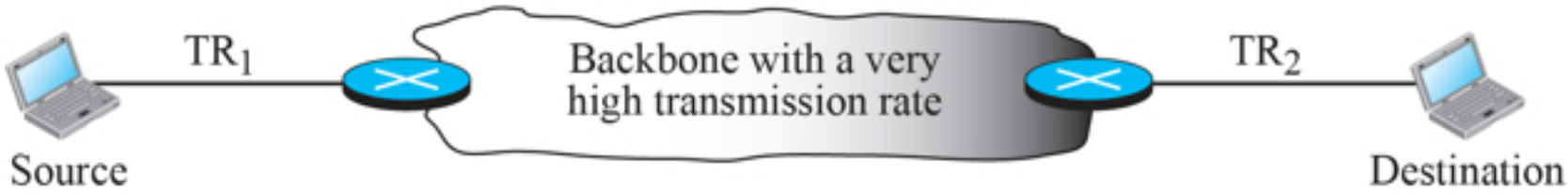
Network Throughput = End-to-End Data Rate



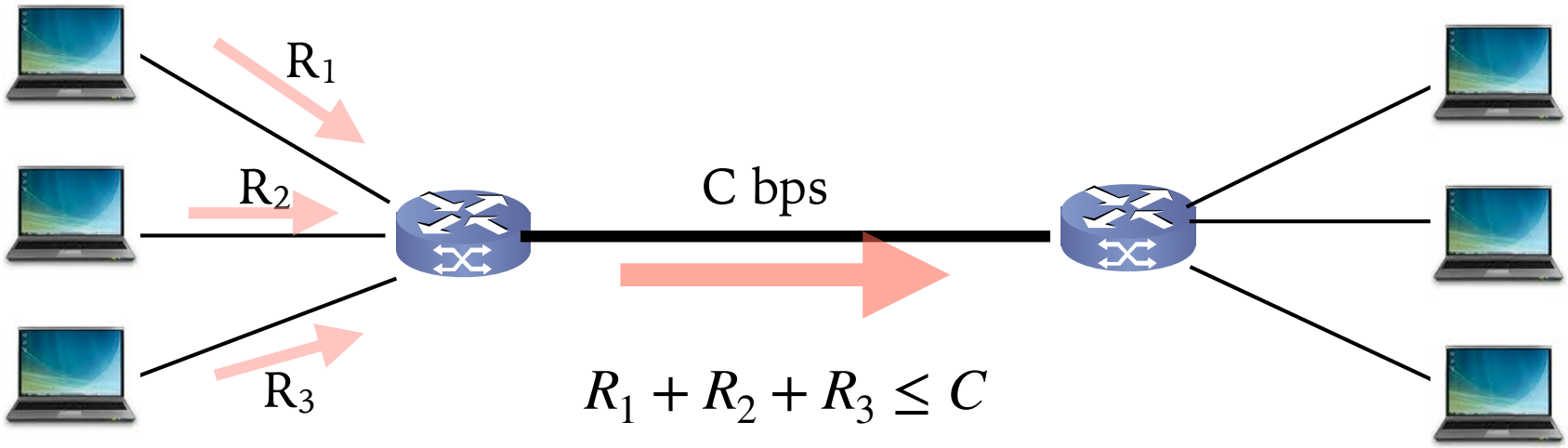
a. A path through three links

b. Simulation using pipes

TR: Transmission rate



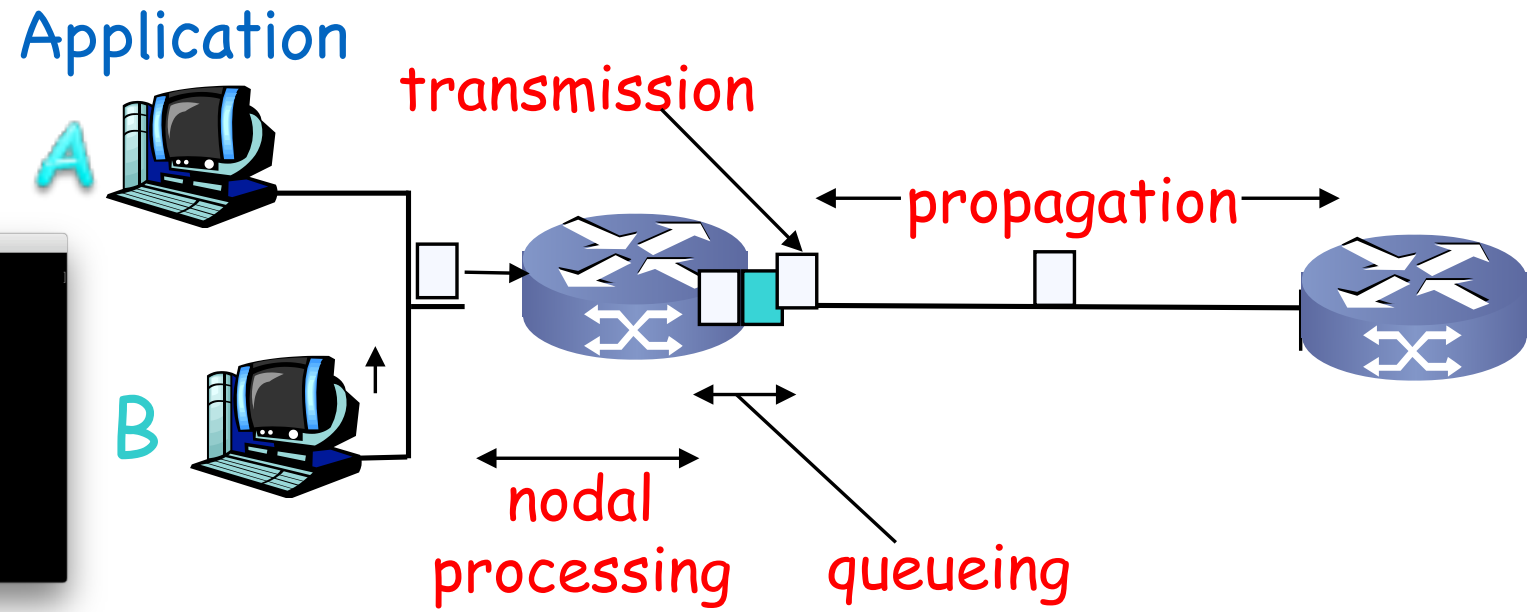
Network Throughtput ≤ Bottleneck Link Capacity



Network Delay Components (One-way)

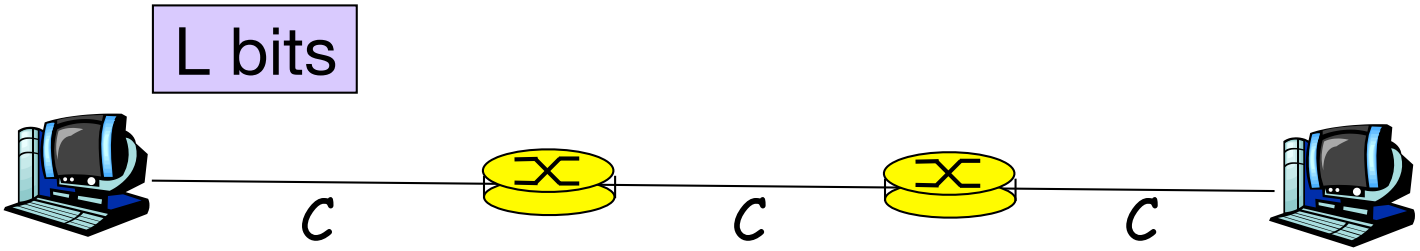
See delay variation in ping output

```
peerapon@metis ~
$ ping -c 10 8.8.8.8
PING 8.8.8.8 (8.8.8.8): 56 data bytes
64 bytes from 8.8.8.8: icmp_seq=0 ttl=54 time=30.699 ms
64 bytes from 8.8.8.8: icmp_seq=1 ttl=54 time=31.952 ms
64 bytes from 8.8.8.8: icmp_seq=2 ttl=54 time=32.457 ms
64 bytes from 8.8.8.8: icmp_seq=3 ttl=54 time=30.924 ms
64 bytes from 8.8.8.8: icmp_seq=4 ttl=54 time=30.888 ms
64 bytes from 8.8.8.8: icmp_seq=5 ttl=54 time=30.973 ms
64 bytes from 8.8.8.8: icmp_seq=6 ttl=54 time=30.766 ms
64 bytes from 8.8.8.8: icmp_seq=7 ttl=54 time=30.815 ms
64 bytes from 8.8.8.8: icmp_seq=8 ttl=54 time=37.182 ms
64 bytes from 8.8.8.8: icmp_seq=9 ttl=54 time=31.118 ms

--- 8.8.8.8 ping statistics ---
10 packets transmitted, 10 packets received, 0.0% packet loss
round-trip min/avg/max/stddev = 30.699/31.777/37.182/1.882 ms

peerapon@metis ~
$
```

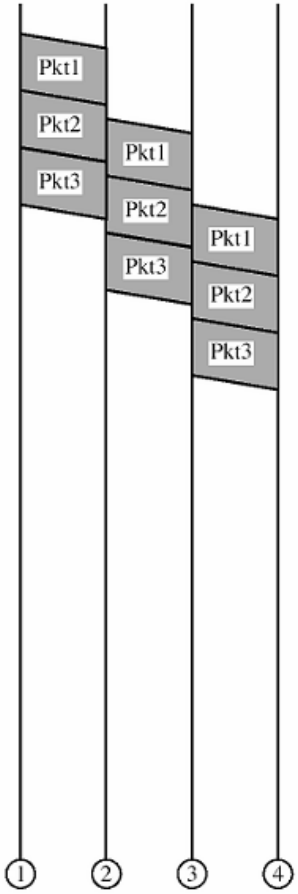


- Application delay
  - ≤ 1 – 3s Web page retrieval
  - 10 ms DB record retrieval.
  - ≤ 50ms VoIP coding/ decoding delay
- Nodal delay. Negligible if low CPU utilization
  - Header processing in L3/L4
  - Make forwarding decision

Transmission and Propagation Delays (d_{tran} and d_{prop})

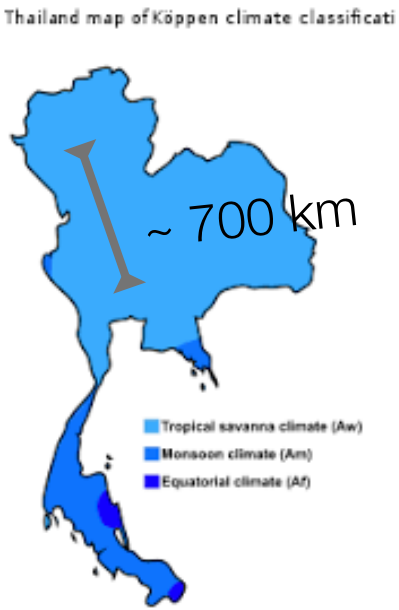


Effect of packet size on delay and error recovery ?

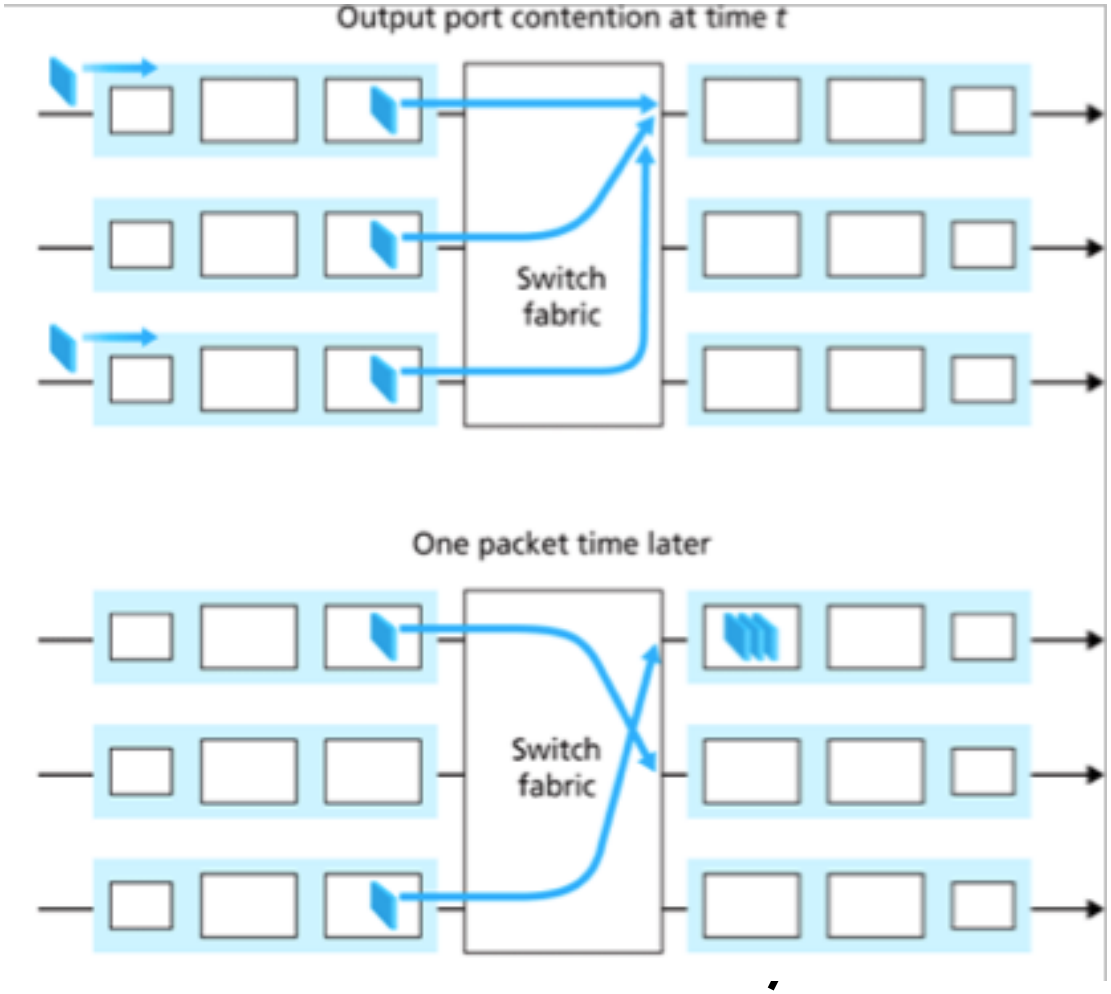


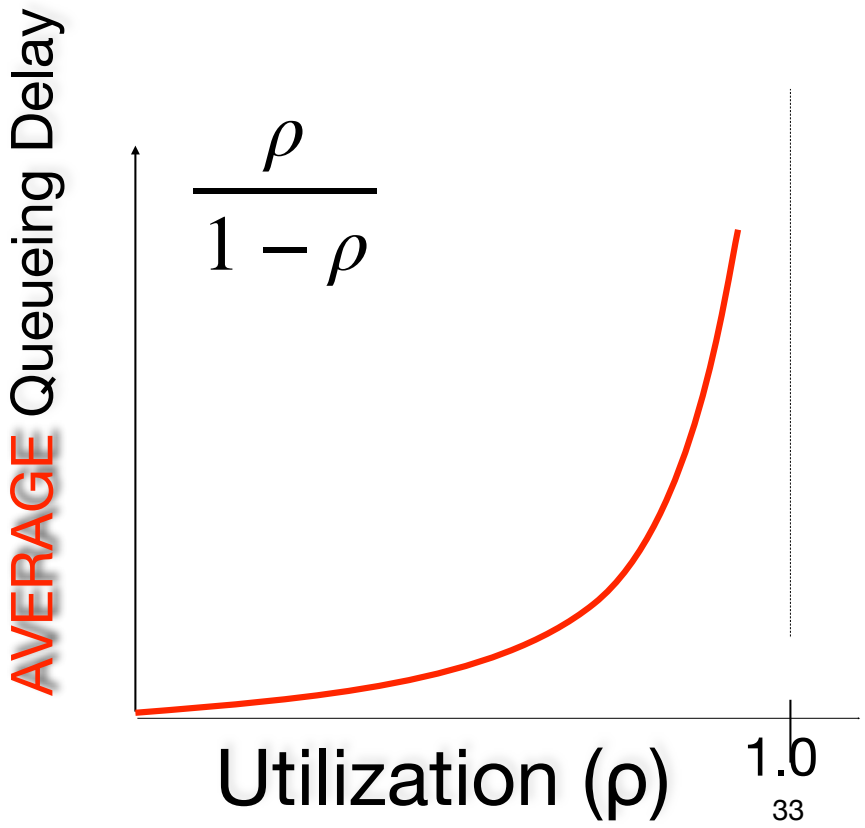
- Take L/C to send (push out) L-bit packet on to C-bps link
- Entire packet must arrive at router before it can be transmitted on next link: *store and forward*
- Transmission delay (d_{tran}) = 3L/C
- Does the packet size matter?

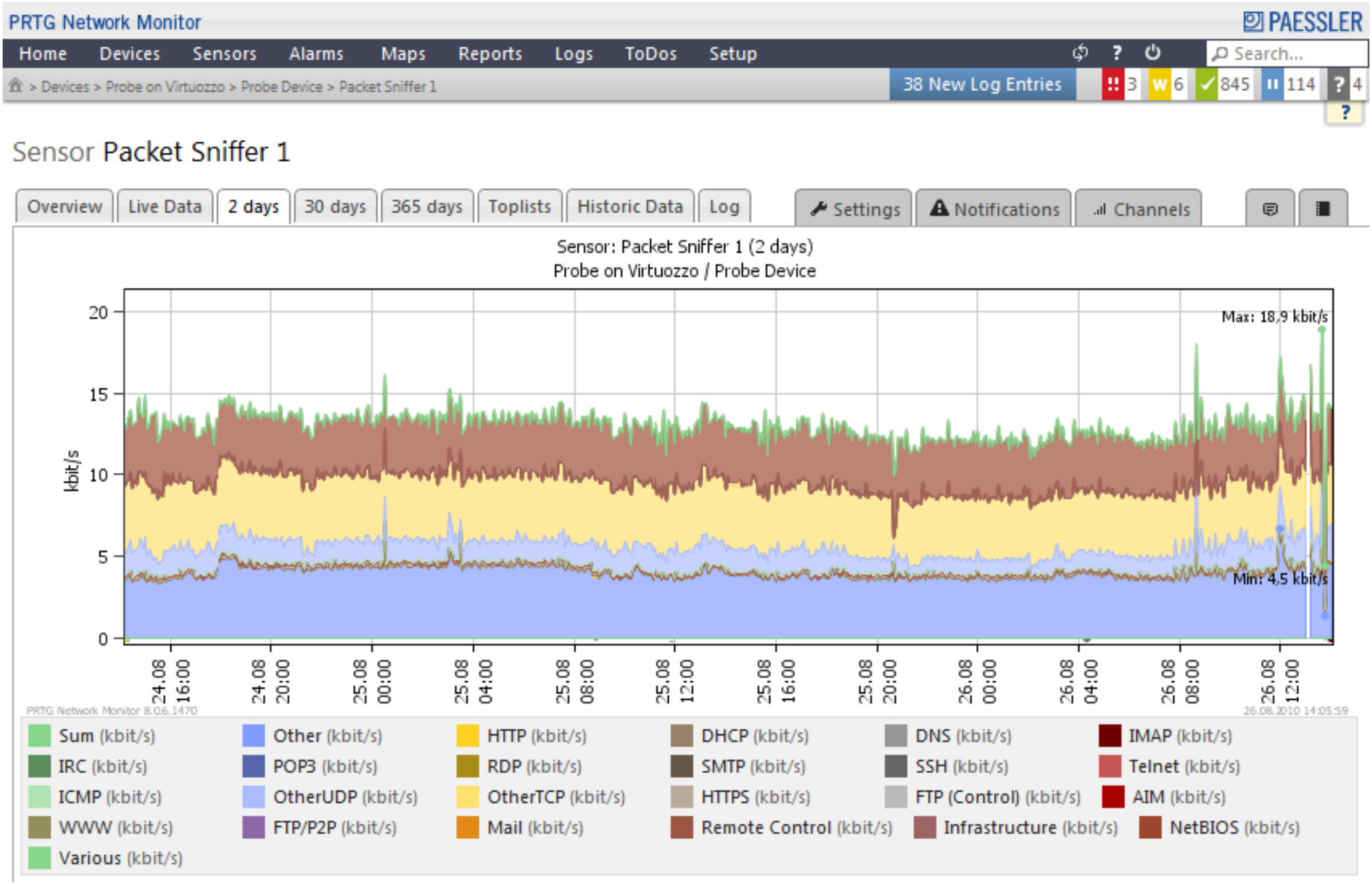
- Propagation delay directly computed from media type and length.
- Fiber optic cable propagates at roughly 2/3 speed of light (3e8 m/s)
  - ~ 20 ms for US transcontinental link.
  - ~ 3.5 ms for a 700-km link
  - Small enough on short fiber links to ignore.
- ~ 180 ms for Satellite link



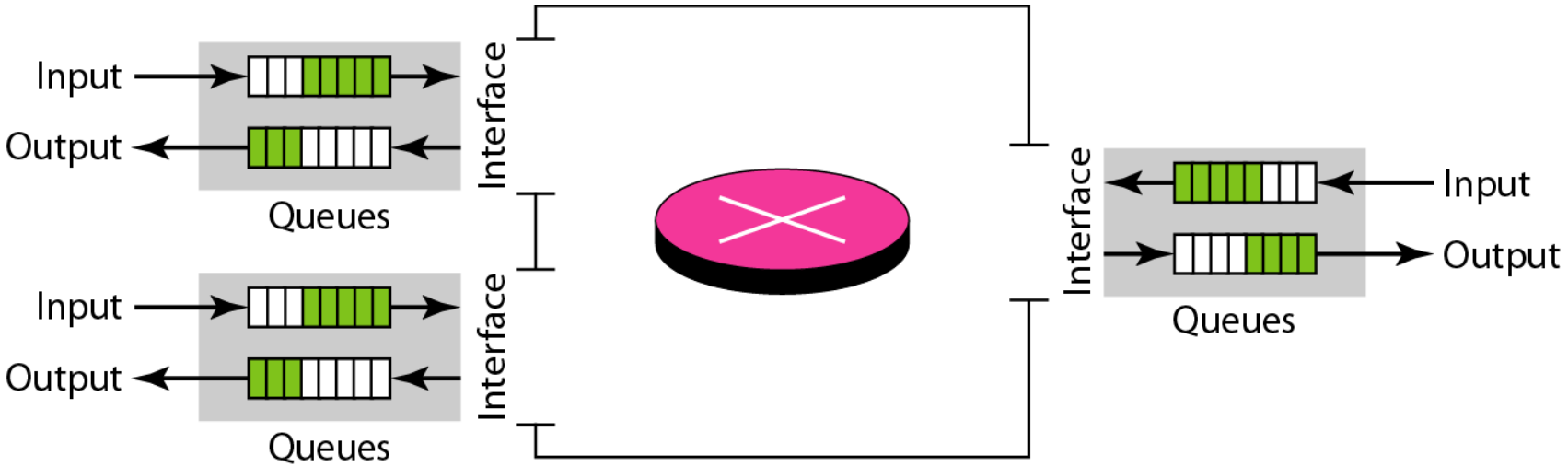
Queueing Delay (d_{queue})

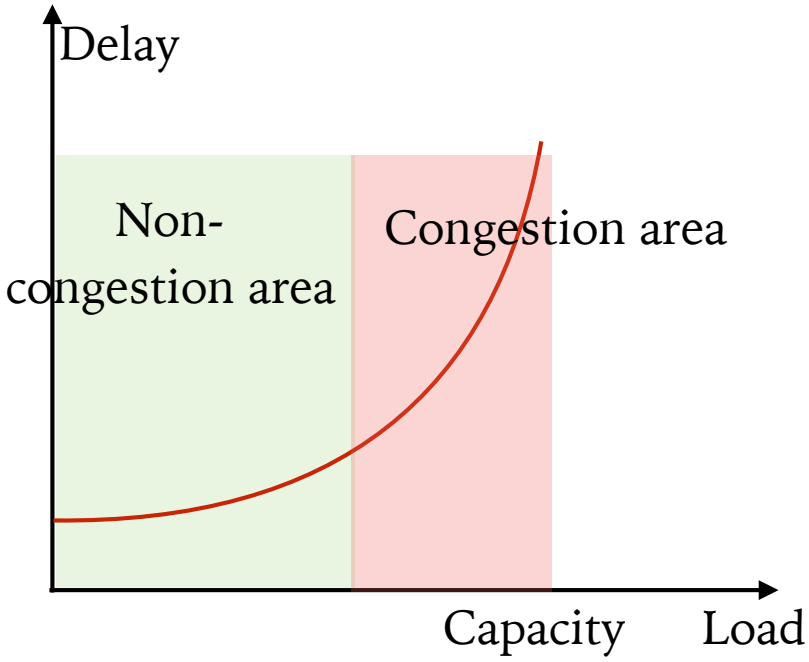


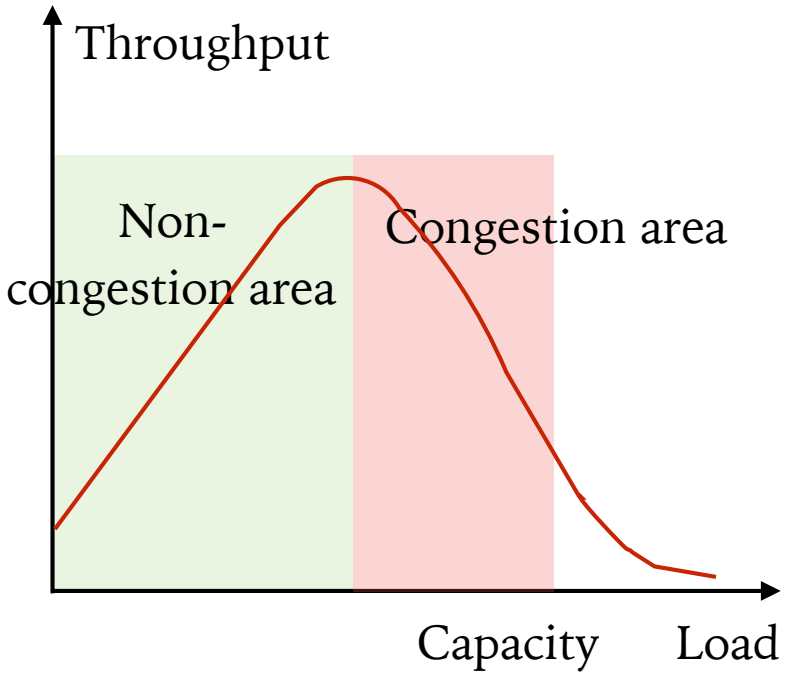



PRTG Network Monitor
PAESSLER
Home
Devices
Sensors
Alarms
Maps
Reports
Logs
ToDos
Setup
Search...
> Devices > Probe on Virtuozzo > Probe Device > Packet Sniffer 1
38 New Log Entries
3
6
845
114
4
Sensor Packet Sniffer 1
Overview
Live Data
2 days
30 days
365 days
Toplists
Historic Data
Log
Settings
Notifications
Channels
Sensor: Packet Sniffer 1 (2 days)
Probe on Virtuozzo / Probe Device
Max: 18,9 kbit/s
Min: 4,5 kbit/s
kbit/s
20
15
10
5
0
24.08 16:00
24.08 20:00
25.08 00:00
25.08 04:00
25.08 08:00
25.08 12:00
25.08 16:00
25.08 20:00
26.08 00:00
26.08 04:00
26.08 08:00
26.08 12:00
PRTG Network Monitor 8.0.6.1470
26.08.2010 14:05:59
Sum (kbit/s)
Other (kbit/s)
HTTP (kbit/s)
DHCP (kbit/s)
DNS (kbit/s)
IMAP (kbit/s)
IRC (kbit/s)
POP3 (kbit/s)
RDP (kbit/s)
SMTP (kbit/s)
SSH (kbit/s)
Telnet (kbit/s)
ICMP (kbit/s)
OtherUDP (kbit/s)
OtherTCP (kbit/s)
HTTPS (kbit/s)
FTP (Control) (kbit/s)
AIM (kbit/s)
WWW (kbit/s)
FTP/P2P (kbit/s)
Mail (kbit/s)
Remote Control (kbit/s)
Infrastructure (kbit/s)
NetBIOS (kbit/s)
Various (kbit/s)

Network Congestion







TCP Congestion Control RFC 2581

- Sender limits transmission by maintaining that:

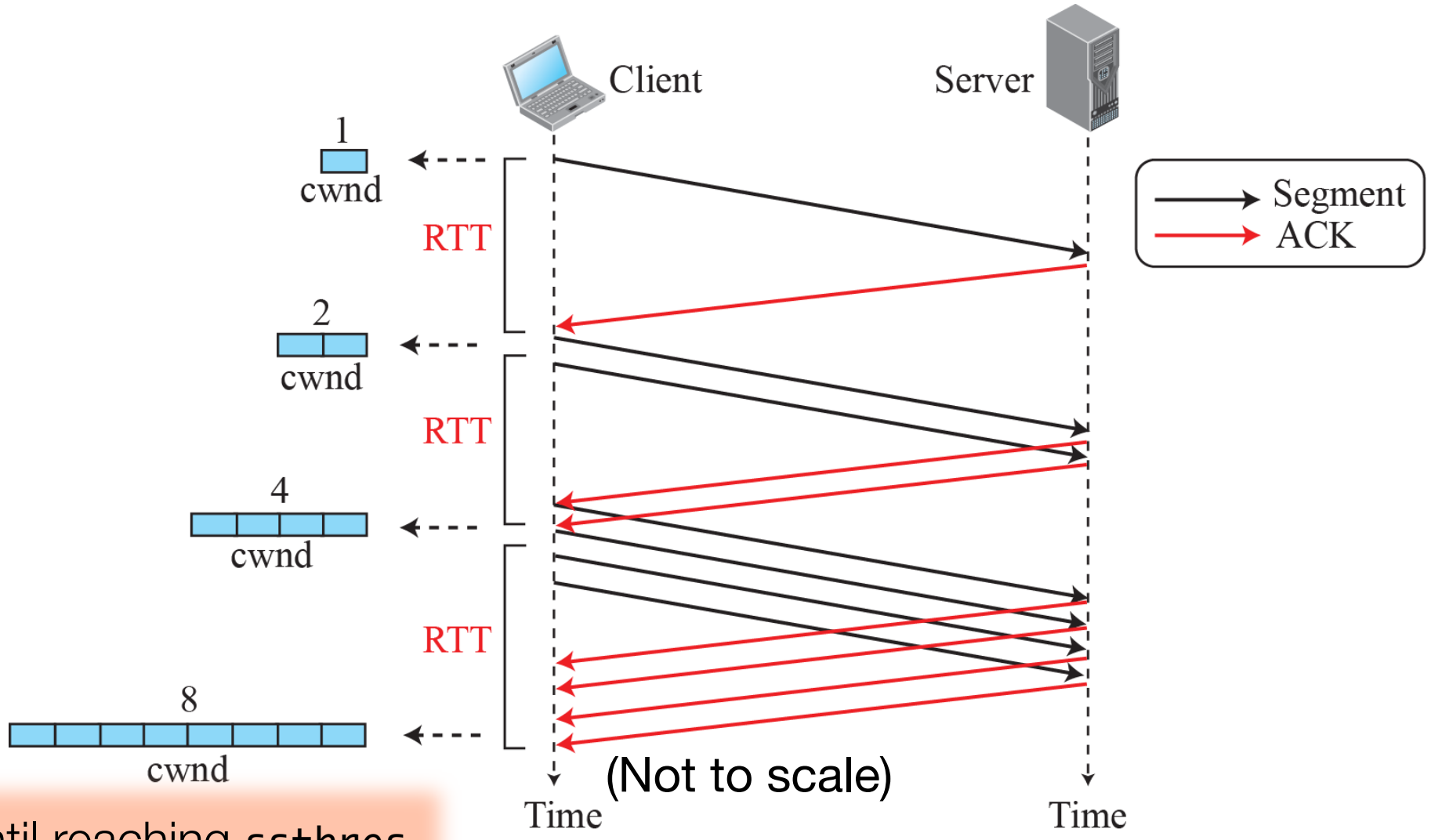
  `LastByteSent - LastByteAcked ≤ min (cwnd*MSS, rwnd)`

  - `MSS` : maximum segment size (in bytes)
  - `cwnd` (Congestion Window) : Dynamic fn. of perceived network congestion (unit MSS).

- Assuming a very large receiver buffer,

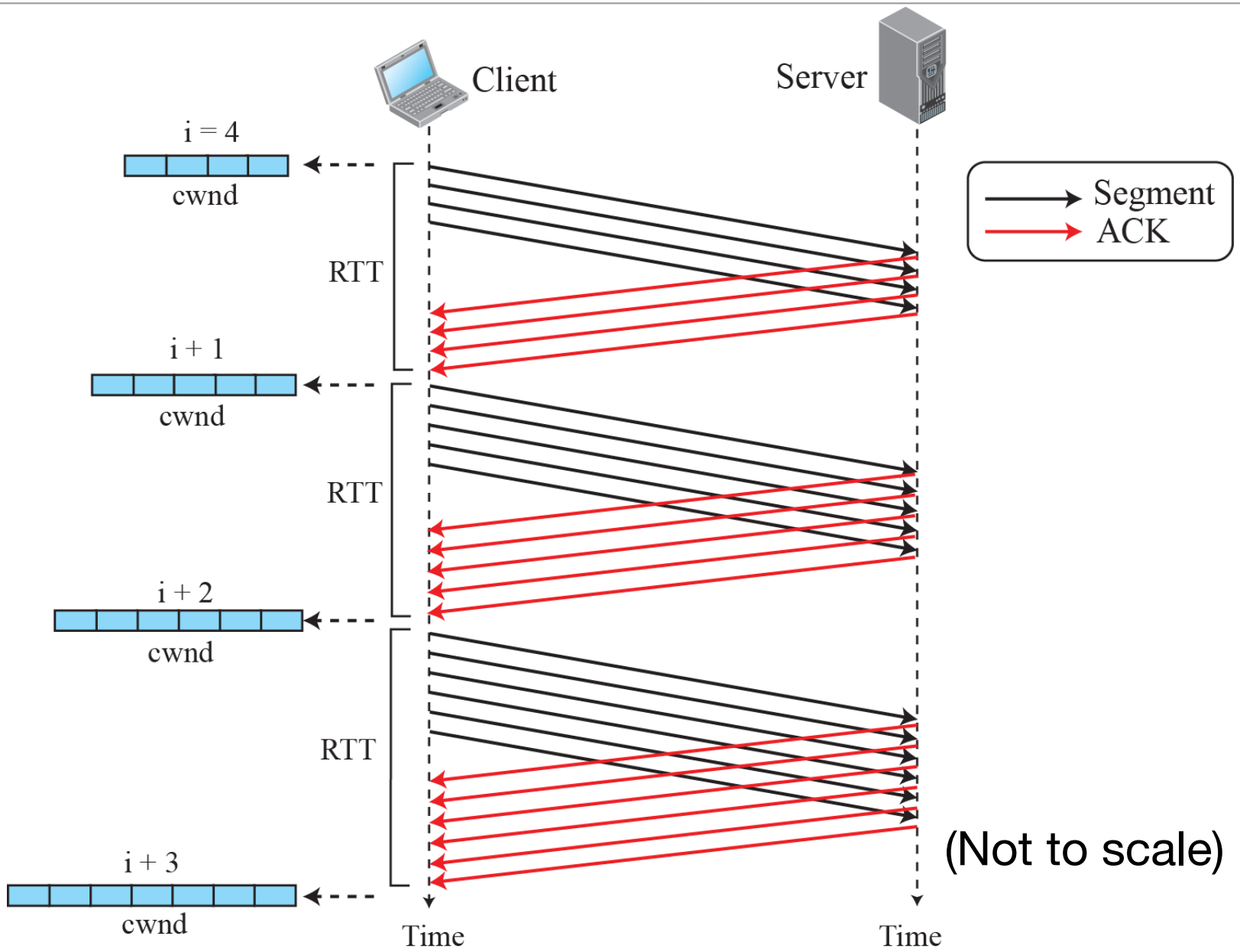
$$\textbf{Avg. Rate} \sim \frac{\text{Avg cwnd} * \text{MSS}}{\text{RTT}} \text{ Bytes/sec}$$

TCP Slow Start (Bandwidth Probing)



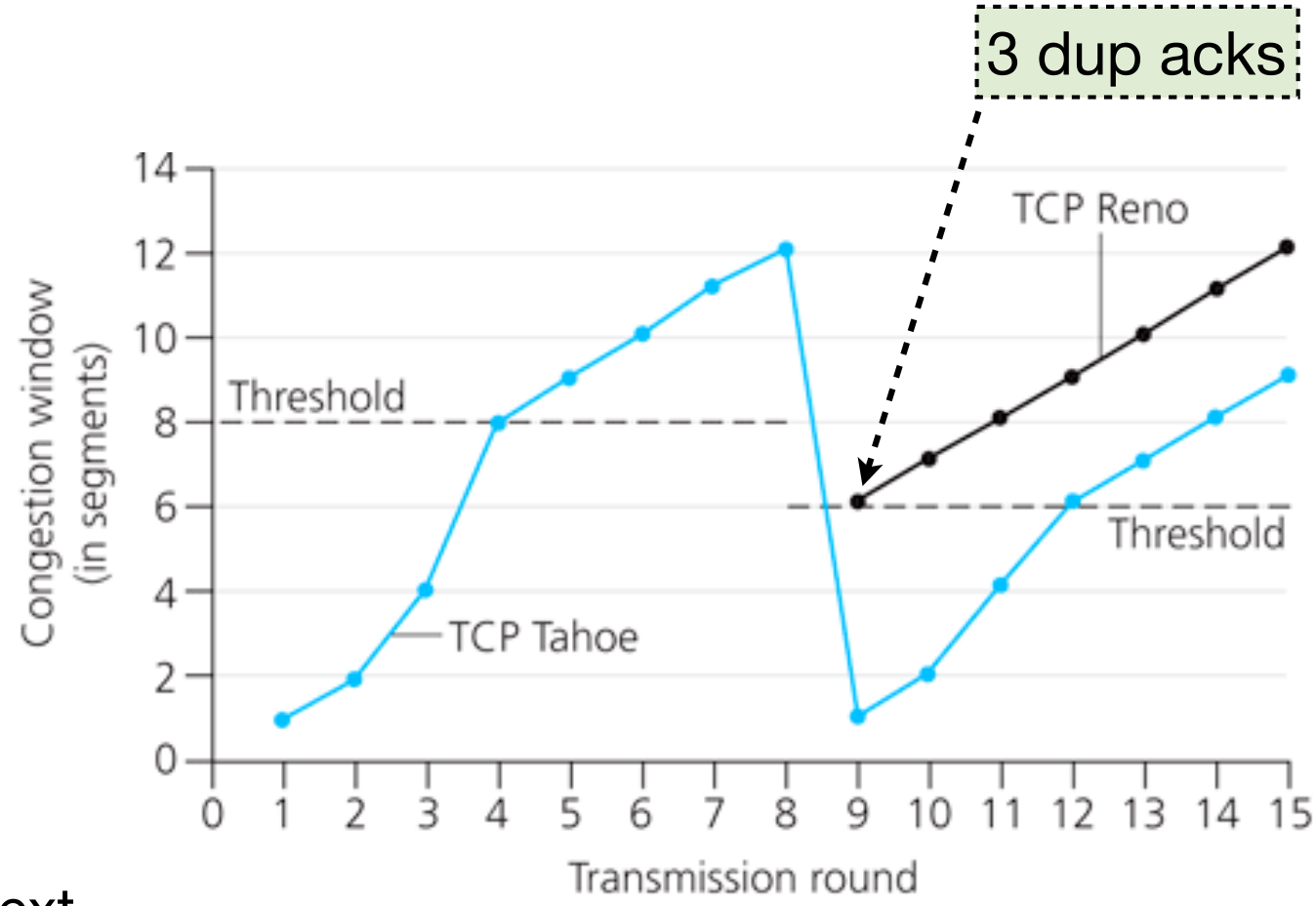
Until reaching `ssthres`
(Slow start threshold)

Congestion Avoidance : Additive Increase (AI)



Congestion Avoidance : Multiplicative Decrease (MD)

- Decrease cwnd when a "loss event" occurs.
  - Two versions: Tahoe and Reno

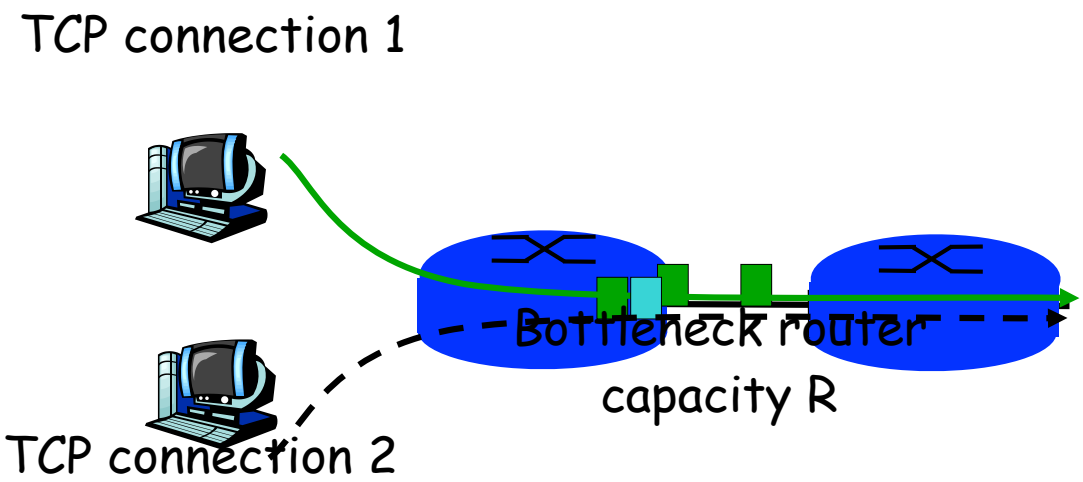


Kurose text

Figure 3.53 ♦ Evolution of TCP's congestion window (Tahoe and Reno)

TCP Fairness

- Fairness goal: if K TCP sessions share the same bottleneck link of bandwidth R, each should get an average rate of R/K.



Why is TCP fair?

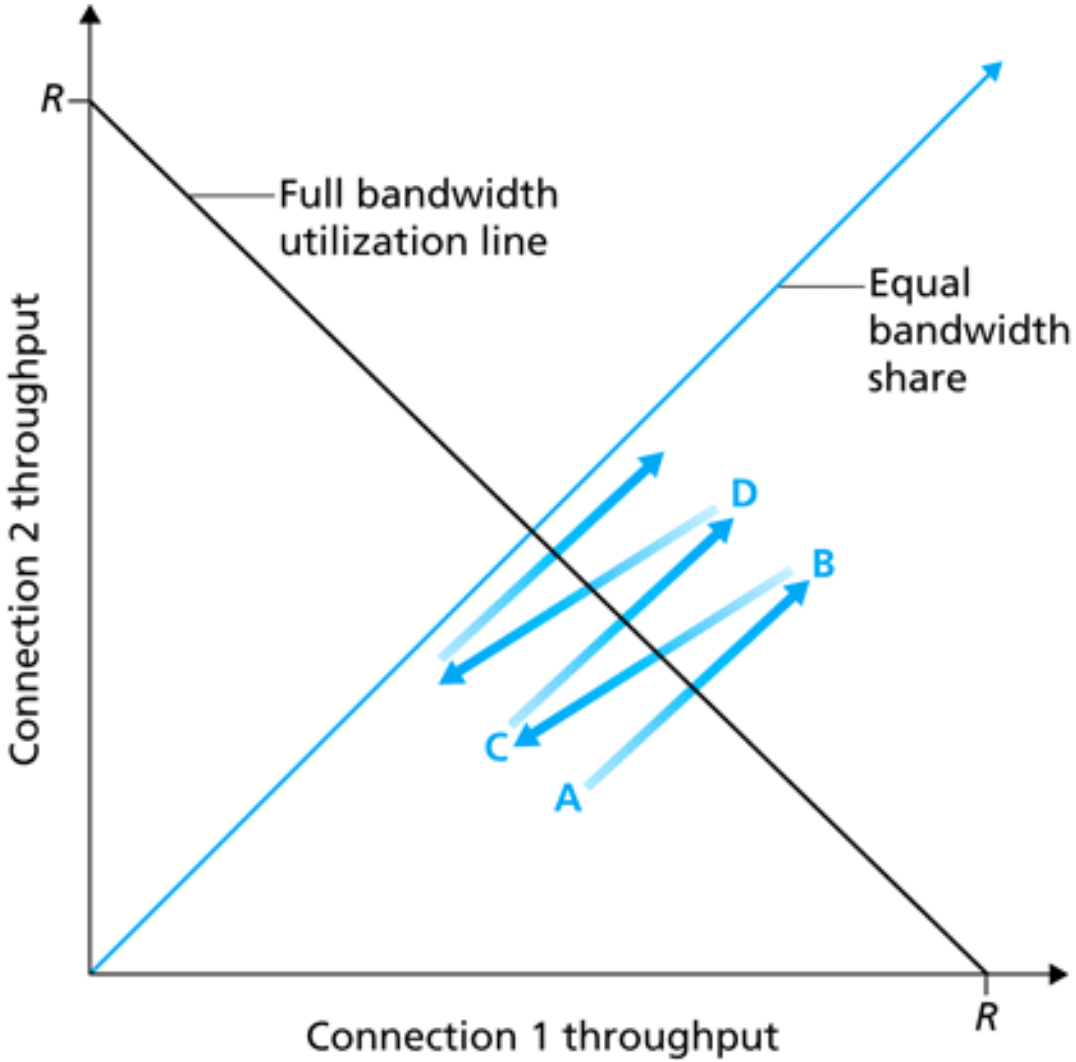


Figure 3.55 ♦ Throughput realized by TCP connections 1 and 2

Highly Simplified TCP throughput

- Macroscopic observation
  - Long-run approximate behavior
  - Ignore slow start (very short)

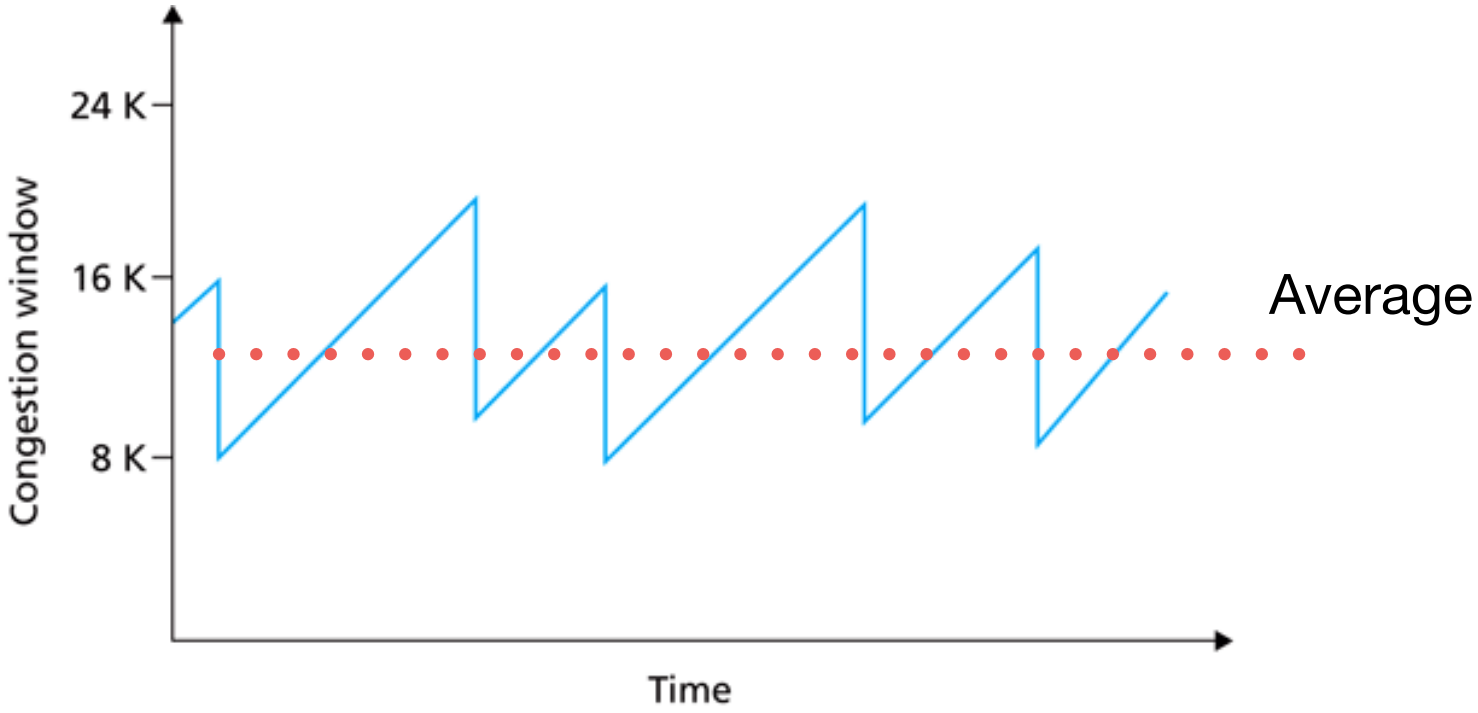
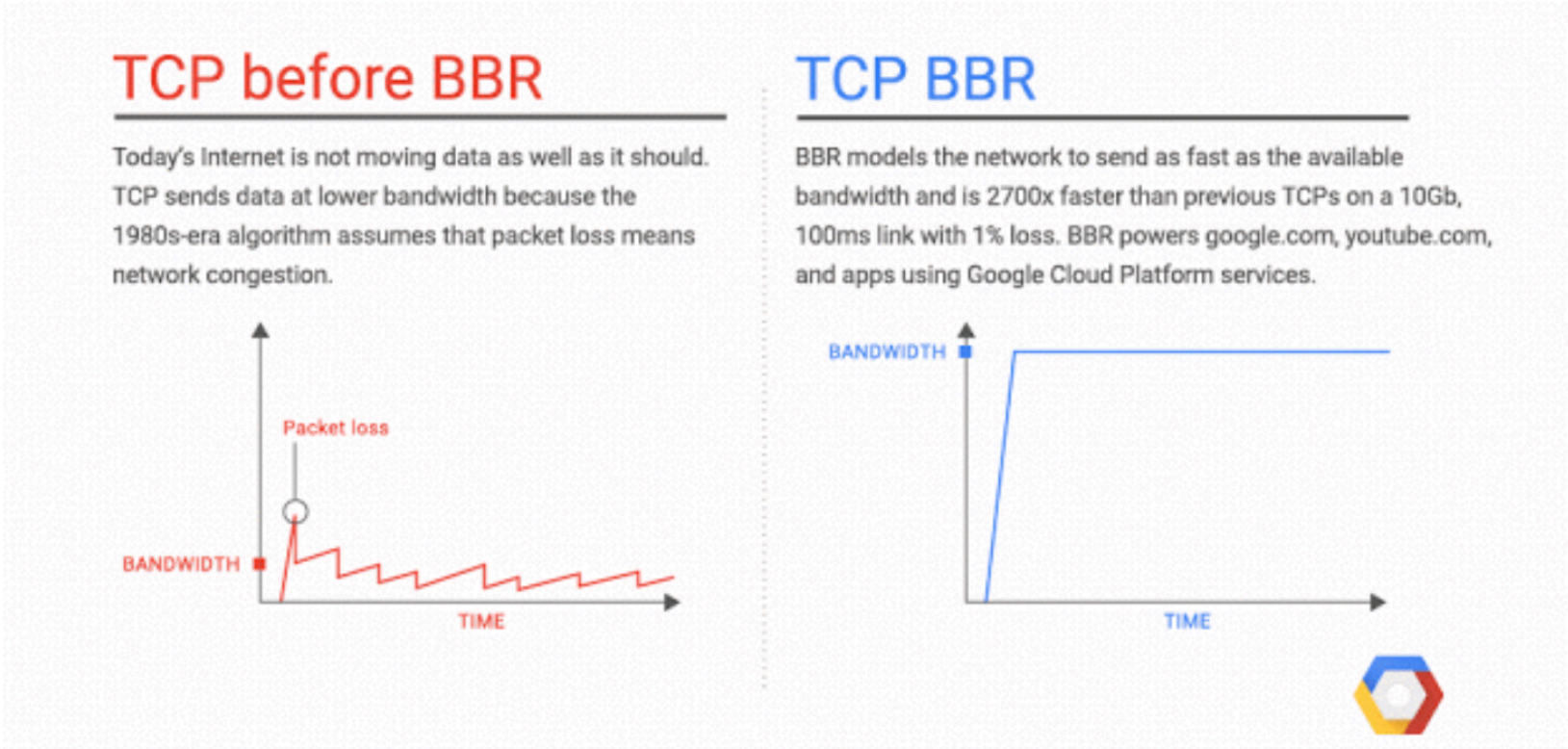


Figure 3.51 ♦ Additive-increase, multiplicative-decrease congestion control

GOOGLE ประกาศใช้ TCP BBR ในระบบเครือข่ายของ GCP เพิ่ม THROUGHPUT สูงสุด 2,700 เท่า

July 21, 2017
Cloud and Systems, Cloud Services, Google, IT Knowledge, IT Researches, Networking, Products, Switch and Router

Google ออกมาประกาศถึงการนำ TCP BBR ซึ่งเป็น Congestion Control Algorithm ใหม่สำหรับระบบเครือข่ายมาใช้งานภายใน Google Cloud Platform (GCP) ซึ่งช่วยให้ประสิทธิภาพการทำงานของระบบในบางแง่มุมสูงขึ้นถึง 2,700 เท่าเลยทีเดียวเมื่อเทียบกับ Algorithm แบบก่อนๆ



https://cloudplatform.googleblog.com/2017/07/TCP-BBR-congestion-control-comes-to-GCP-your-Internet-just-got-faster.html

Summary

- Basic principles of packet-switching networks: Store-and-Forward and Bandwidth sharing
  - Parallelization and efficient error recovery
  - Trade-off with overheads
- Router architecture -- Line cards, Fabric, Processors
  - Different designs for better performance
- Two types of packet-switched networks
  - Datagram (connectionless) vs. Virtual-circuit (connection-oriented)
  - Trade-off cost, simplicity, and performance.

- Key network performance measures
  - Throughput, delay, loss are closely related
  - Affected by lots of factors: Topology, link capacities/distance/types, flow pattern, # users, input traffic rate (congestion level)
- End-to-end solution to network congestion problem
  - Congestion avoidance
  - React to “loss events”
  - Additive Increase Multiplicative Decrease
  - Also result in throughput fairness
- Issues of TCP congestion control
  - Favor small RTT connections
  - Not suit multimedia applications, e.g., voice/video streaming
  - Parallel TCP connections exploited