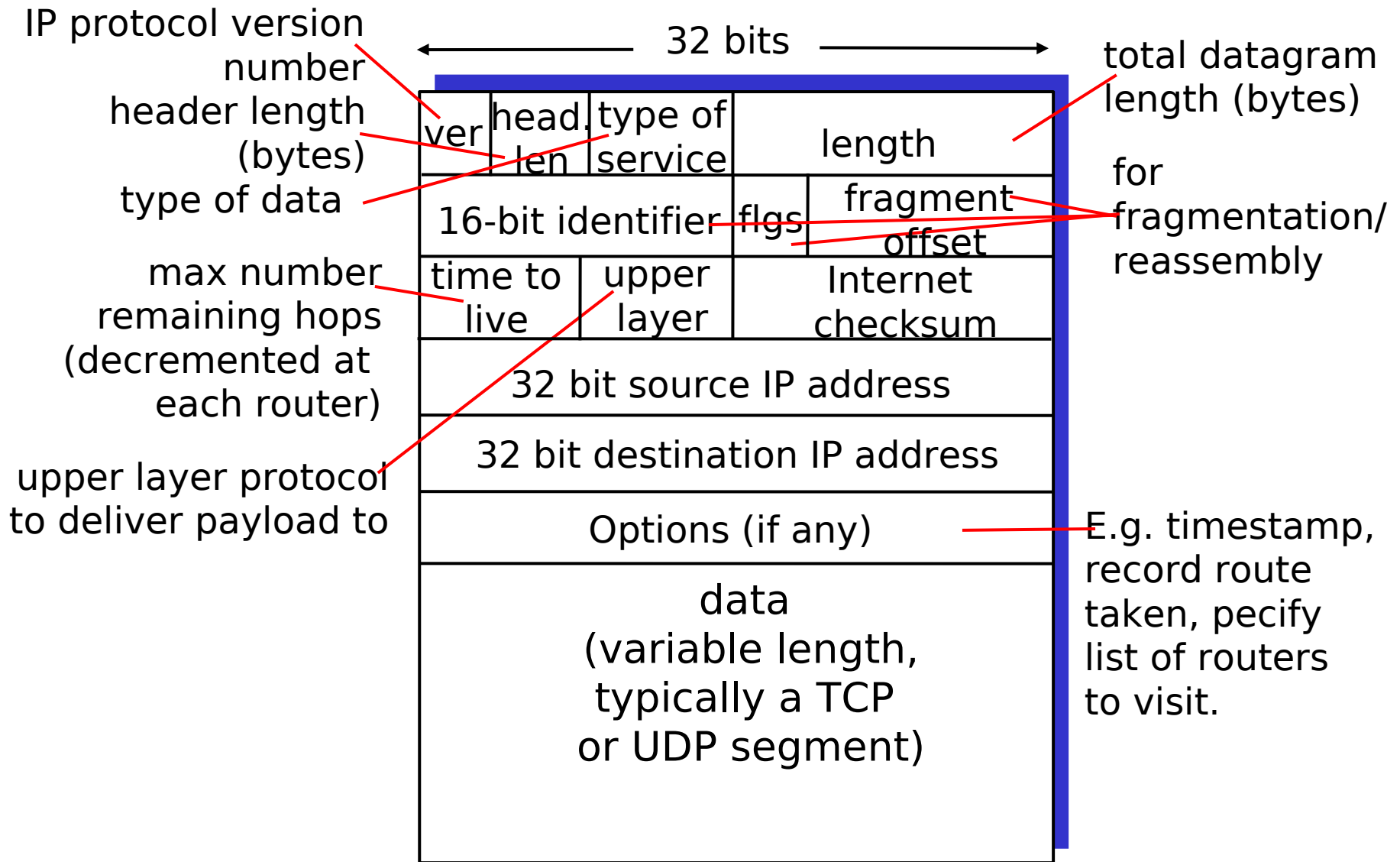
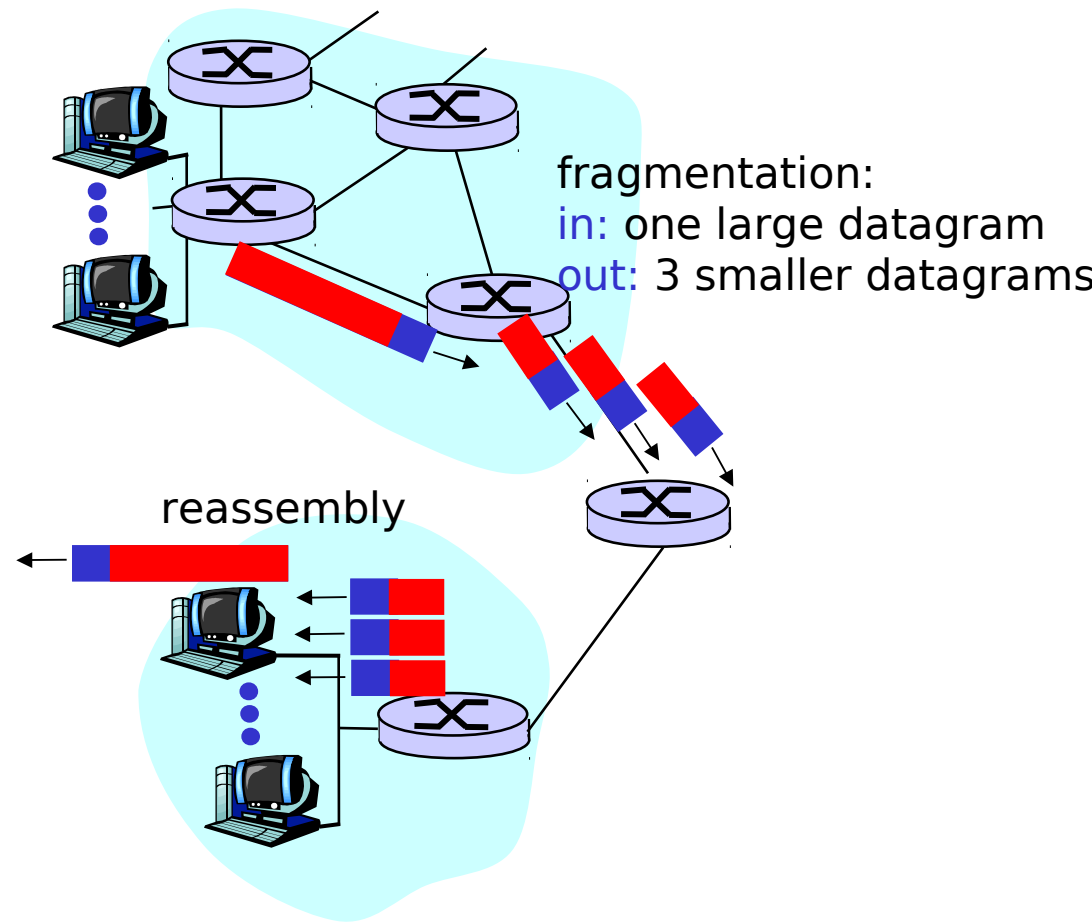


IP datagram format



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

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

One large datagram becomes
several smaller datagrams



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

	length =1500	ID =x	fragflag =1	offset =1480	
--	-----------------	----------	----------------	-----------------	--

	length =1040	ID =x	fragflag =0	offset =2960	
--	-----------------	----------	----------------	-----------------	--

ICMP: Internet Control Message Protocol

- used by hosts, routers, gateways 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

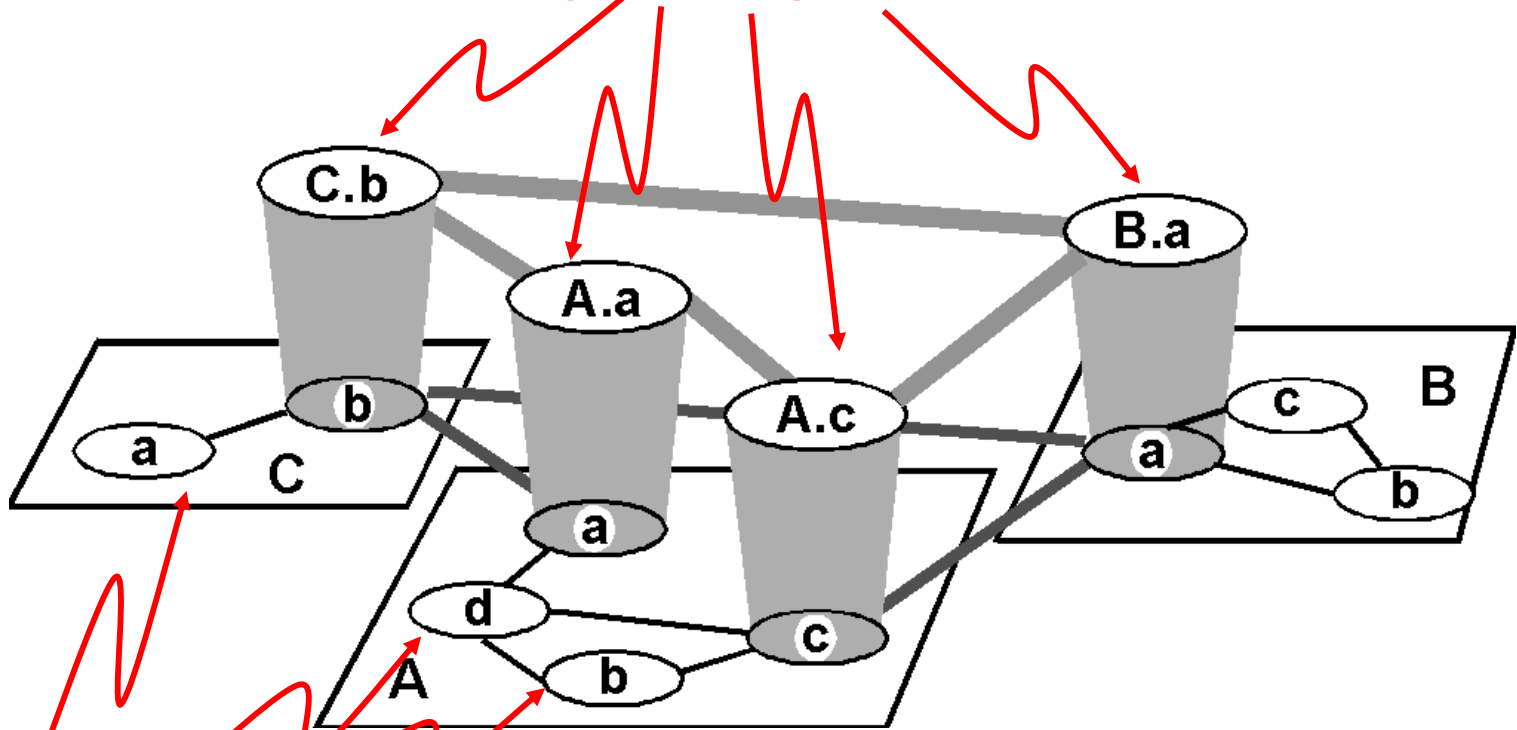
Routing in the Internet

- The Global Internet consists of **Autonomous Systems (AS)** interconnected with each other:
 - **Stub AS**: small corporation
 - **Multihomed AS**: large corporation (no transit)
 - **Transit AS**: provider

- Two-level routing:
 - **Intra-AS**: administrator is responsible for choice
 - **Inter-AS**: unique standard

Internet AS Hierarchy

AS border (exterior gateway) routers



AS interior (gateway) routers

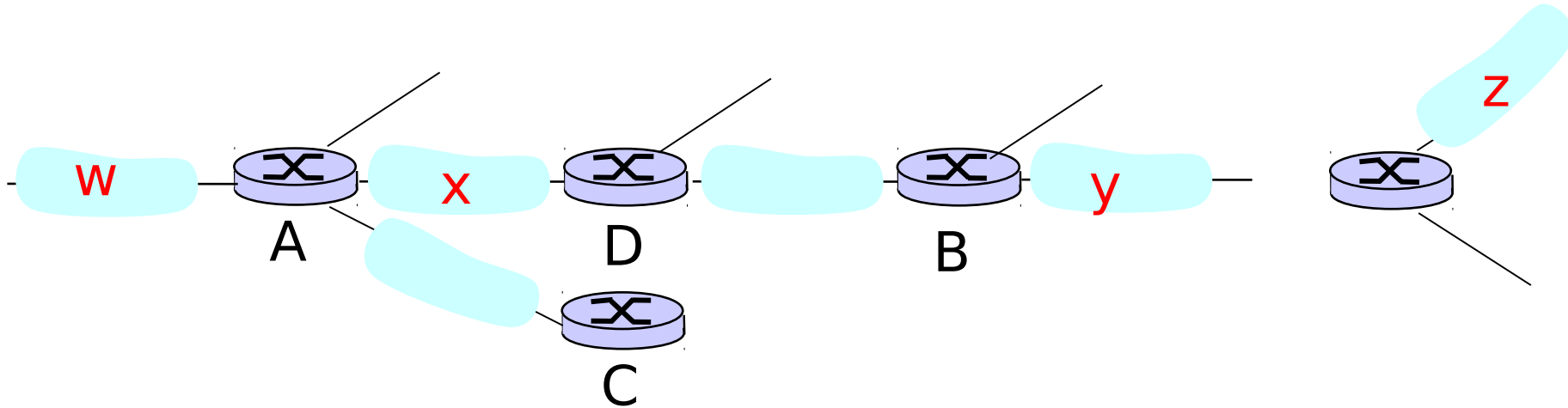
Intra-AS Routing

- Also known as **Interior Gateway Protocols (IGP)**
- Most common IGPs:
 - RIP: Routing Information Protocol
 - OSPF: Open Shortest Path First
 - IGRP: Interior Gateway Routing Protocol (Cisco propr.)

RIP (Routing Information Protocol)

- Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops)
 - *Can you guess why?*
- Distance vectors: exchanged every 30 sec via Response Message (also called **advertisement**)
- Each advertisement: route to up to 25 destination nets

RIP (Routing Information Protocol)



Destination Network to dest.	Next Router	Num. of hops
w	A	2
y	B	2
z	B	7
x	--	1

...

Routing table in D

....

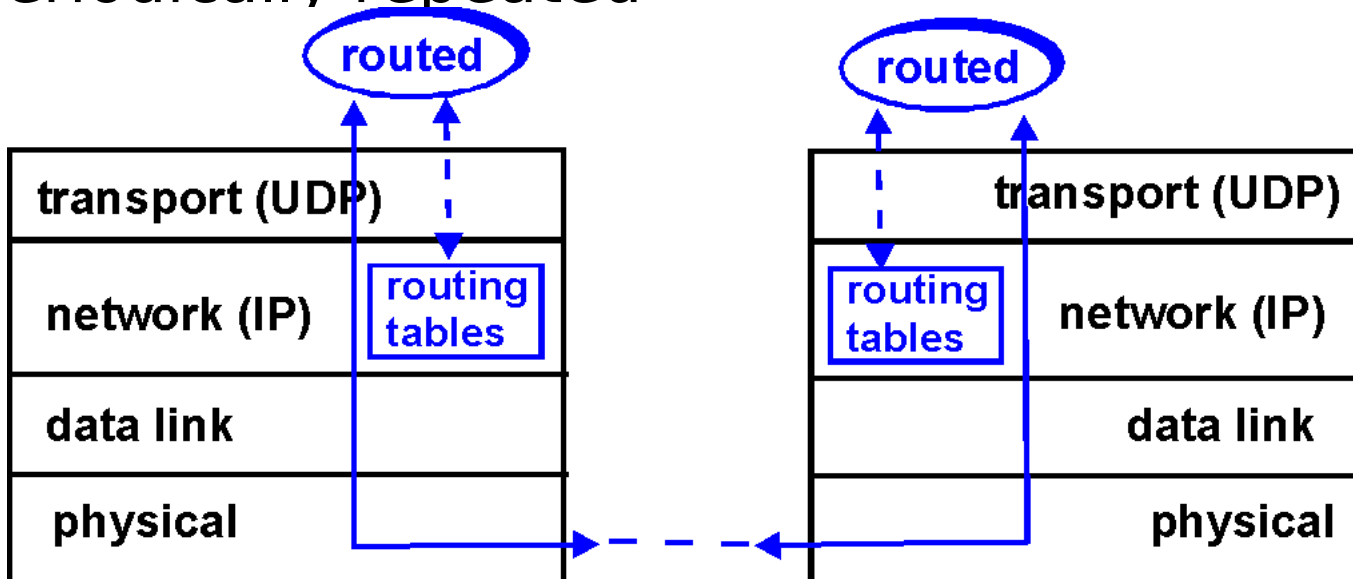
RIP: Link Failure and Recovery

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

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

RIP Table_processing

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



RIP Table example (continued)

Router: *girofflee.eurocom.fr*

Destination	Gateway	Flags	Ref	Use	Interface
-----	-----	-----	-----	-----	-----
127.0.0.1	127.0.0.1	UH	0	26492	lo0
192.168.2.	192.168.2.5	U	2	13	fa0
193.55.114.	193.55.114.6	U	3	58503	le0
192.168.3.	192.168.3.5	U	2	25	qaa0
224.0.0.0	193.55.114.6	U	3	0	le0
default	193.55.114.129	UG	0	143454	

- Three attached class C networks (LANs)
- Router only knows routes to attached LANs
- Default router used to go up
- Route multicast address: 224.0.0.0
- Loopback interface (for debugging)

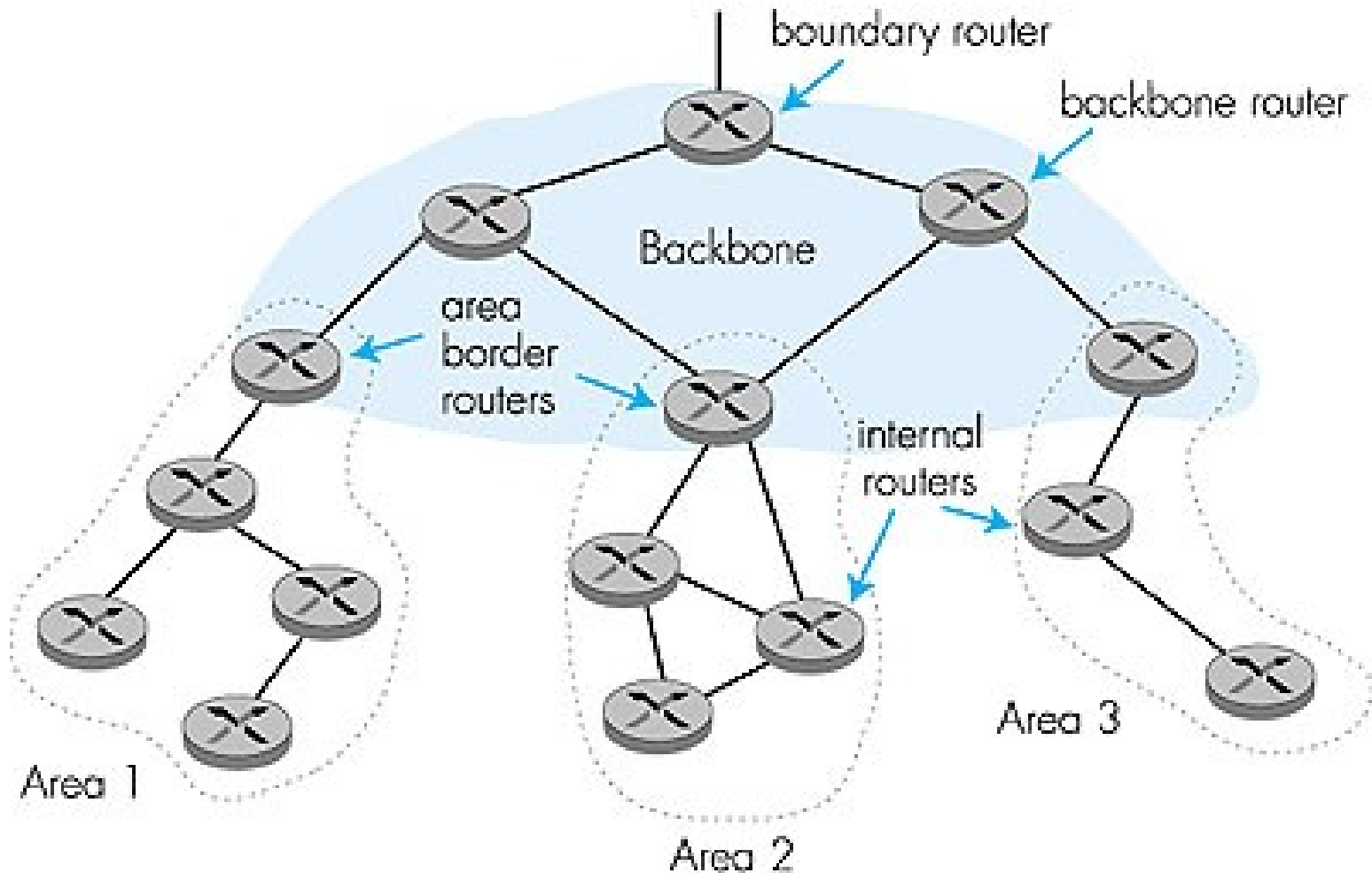
OSPF (Open Shortest Path First)

- open: publicly available
- Uses Link State algorithm
 - LS packet dissemination
 - Topology map at each node
 - Route computation using Dijkstra's algorithm
- OSPF advertisement carries one entry per neighbor router
- Advertisements disseminated to **entire** AS (via flooding)

OSPF advanced features (not in RIP)

- ▢ **Security:** all OSPF messages authenticated (to prevent malicious intrusion); TCP connections used
- ▢ **Multiple** same-cost **paths** allowed (only one path in RIP)
- ▢ For each link, multiple cost metrics for different **TOS** (eg., satellite link cost set low for best effort; high for real time)
- ▢ Integrated uni- and **multicast** support:
 - ▢ Multicast OSPF (MOSPF) uses same topology data base as OSPF
- ▢ **Hierarchical** OSPF in large domains.

Hierarchical OSPF



Hierarchical OSPF

- ▮ **Two-level hierarchy:** local area, backbone.
 - ▮ Link-state advertisements only in area
 - ▮ each nodes has detailed area topology; only know direction (shortest path) to nets in other areas.
- ▮ **Area border routers:** summarize distances to nets in own area, advertise to other Area Border routers.
- ▮ **Backbone routers:** run OSPF routing limited to backbone.
- ▮ **Boundary routers:** connect to other ASs.

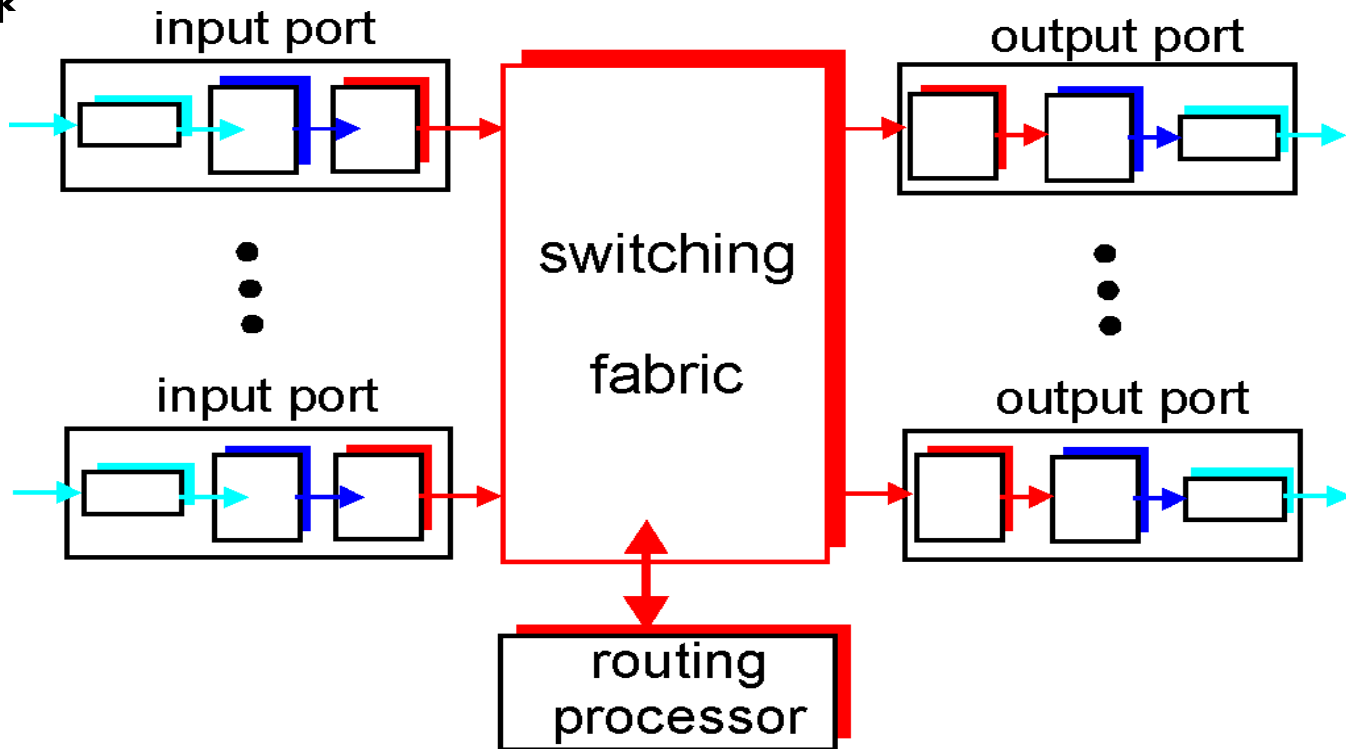
IGRP (Interior Gateway Routing Protocol)

- ❑ CISCO proprietary; successor of RIP (mid 80s)
- ❑ Distance Vector, like RIP
- ❑ several cost metrics (delay, bandwidth, reliability, load etc)
- ❑ uses TCP to exchange routing updates
- ❑ Loop-free routing via Distributed Updating Alg. (DUAL) based on *diffused computation*

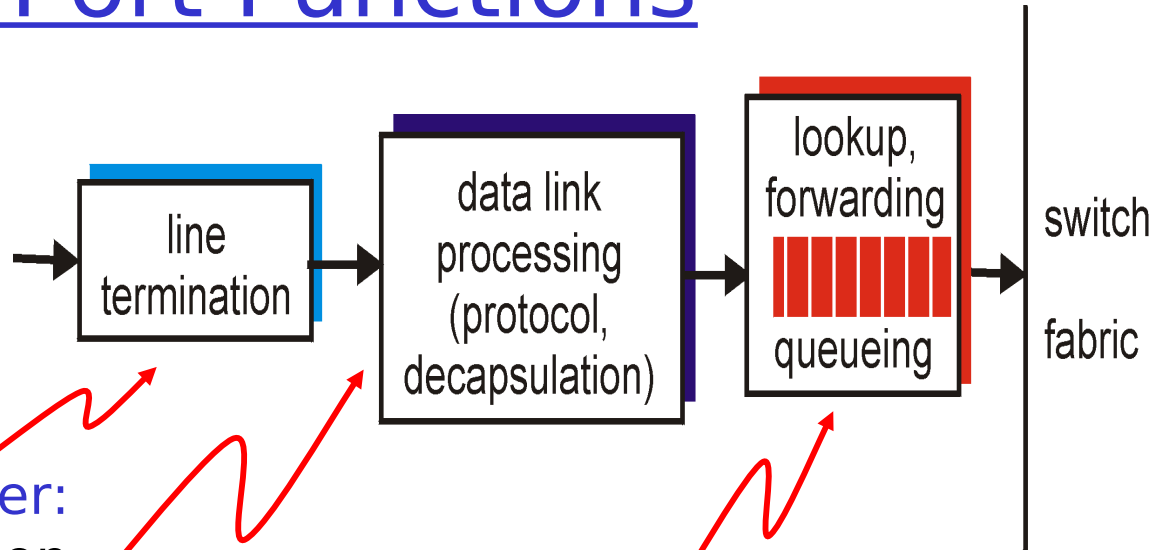
Router Architecture Overview

Two key router functions:

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



Input Port Functions



Physical layer:
bit-level reception

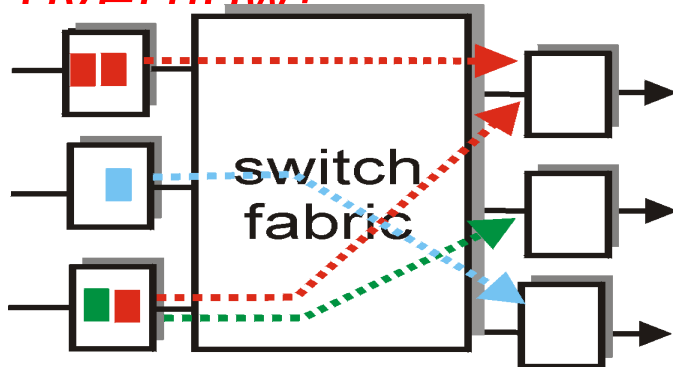
Data link layer:
e.g., Ethernet
see chapter 5

Decentralized switching:

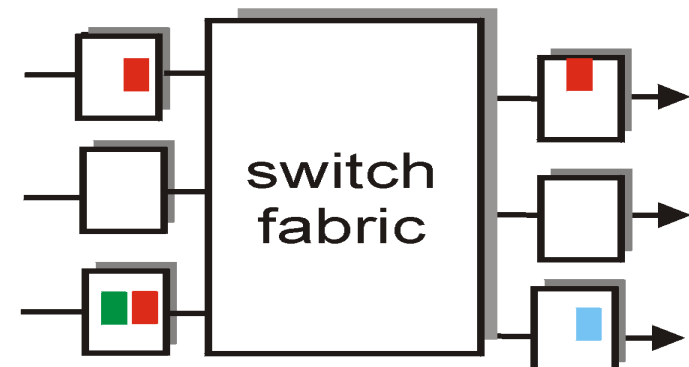
- given datagram dest., lookup output port using routing table in input port memory
- goal: complete input port processing at 'line speed'
- queuing: if datagrams arrive faster than forwarding rate into switch fabric

Input Port Queuing

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

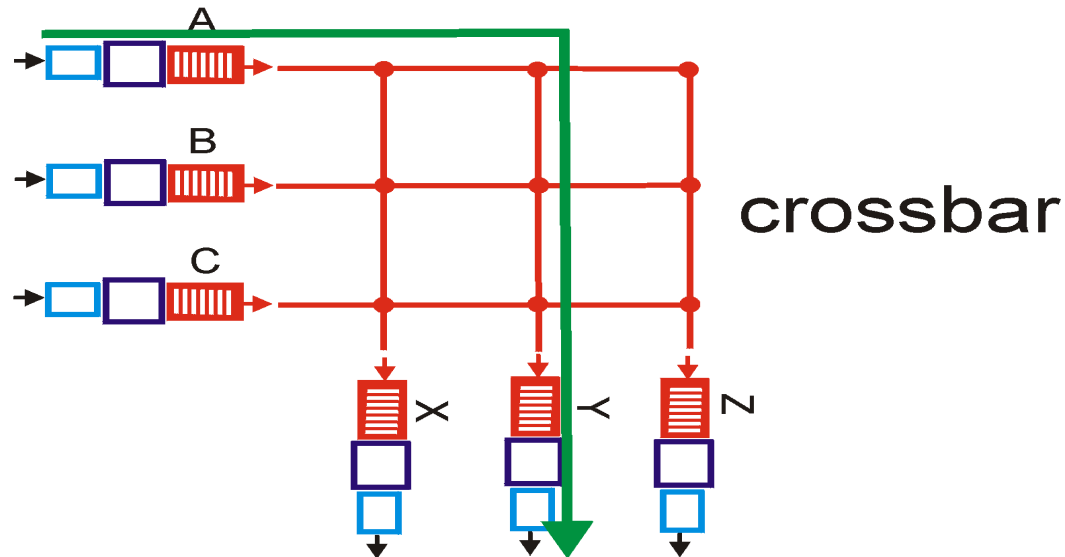
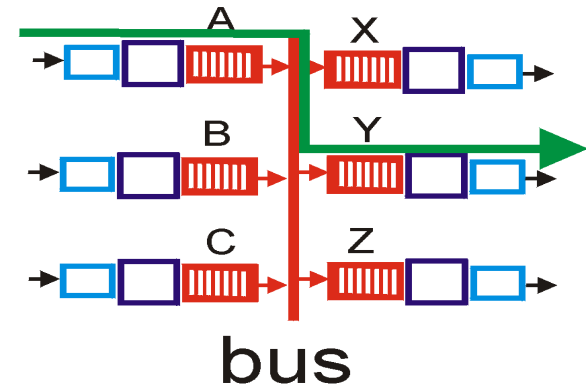
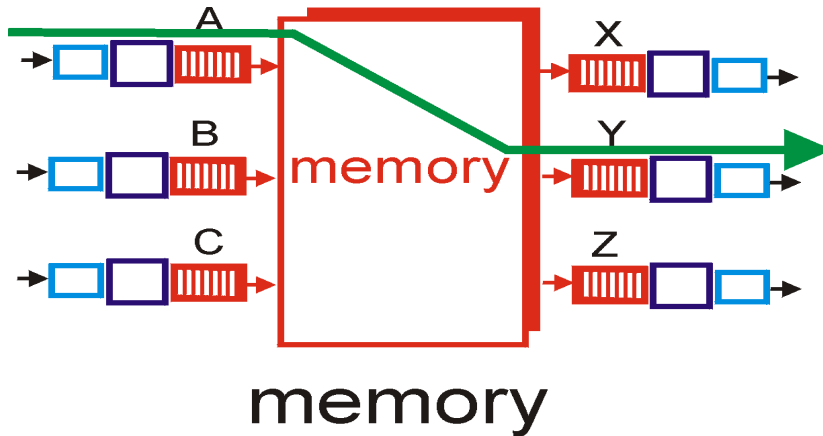


output port contention
at time t - only one red
packet can be transferred



green packet
experiences HOL blocking

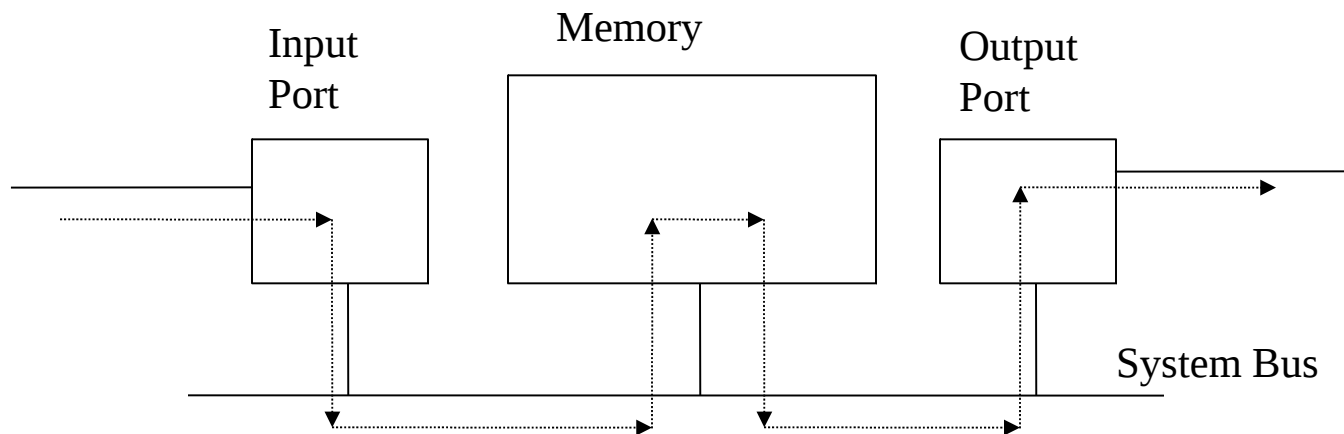
Three types of switching fabrics



Switching Via Memory

First generation routers:

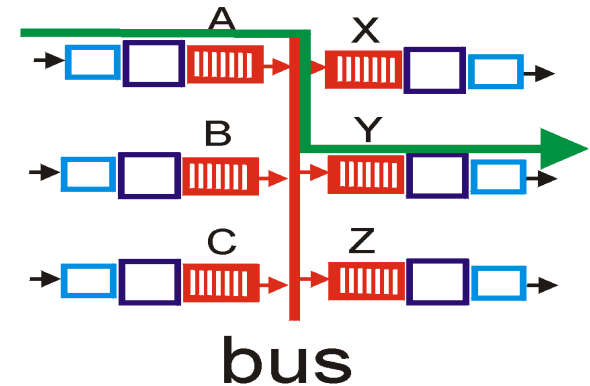
- packet copied by system's (single) CPU
- speed limited by memory bandwidth (2 bus crossings per datagram)



Modern routers:

- input port processor performs lookup, copy into memory
- Cisco Catalyst 8500

Switching Via Bus

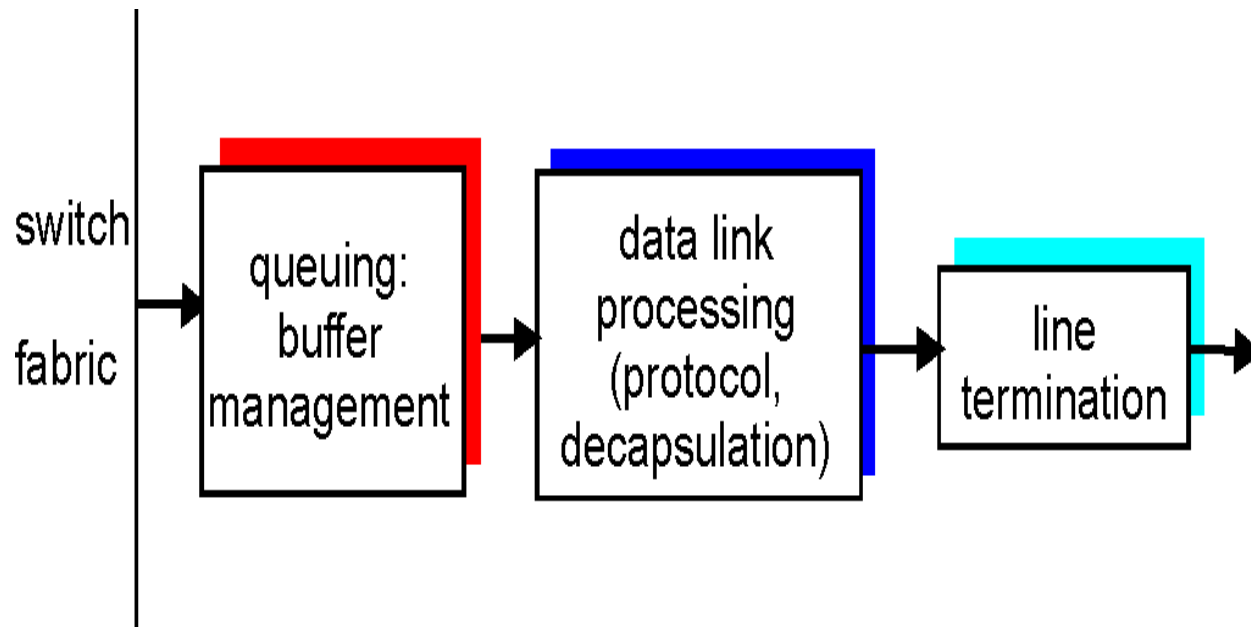


- datagram from input port memory
to output port memory via a
shared bus
- **bus contention:** switching speed
limited by bus bandwidth
- 1 Gbps bus, Cisco 1900:
sufficient speed for access and
enterprise routers (not regional
or backbone)

Switching Via An Interconnection Network

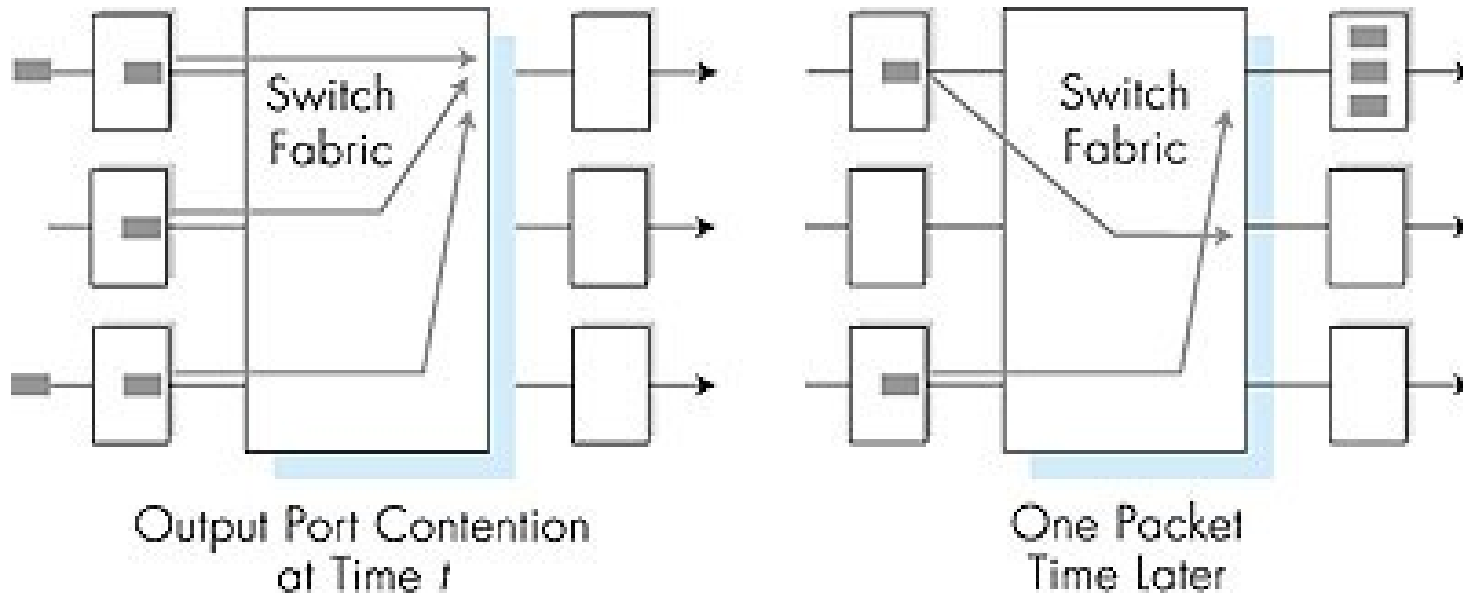
- overcome bus bandwidth limitations
- Banyan networks, other interconnection nets initially developed to connect processors in multiprocessor
- Advanced design: fragmenting datagram into fixed length cells, switch cells through the fabric.
- Cisco 12000: switches Gbps through the interconnection network

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



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

IPv6

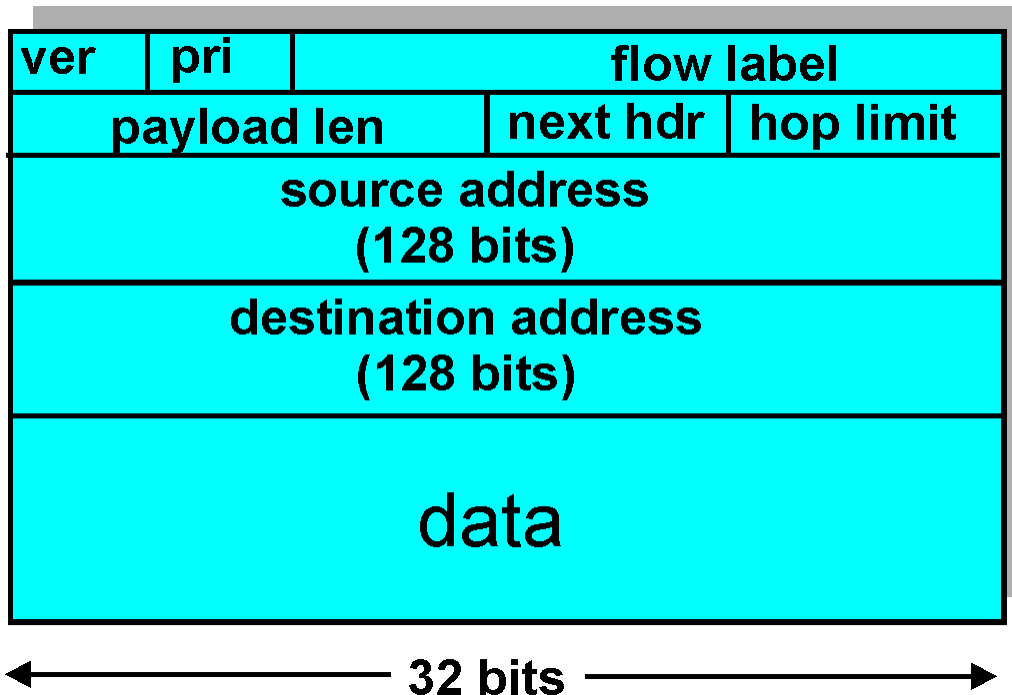
- **Initial motivation:** 32-bit address space completely allocated by 2008.
- **Additional motivation:**
 - header format helps speed processing/forwarding
 - header changes to facilitate QoS
 - new anycast address: route to best of several replicated servers
- **IPv6 datagram format:**
 - fixed-length 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



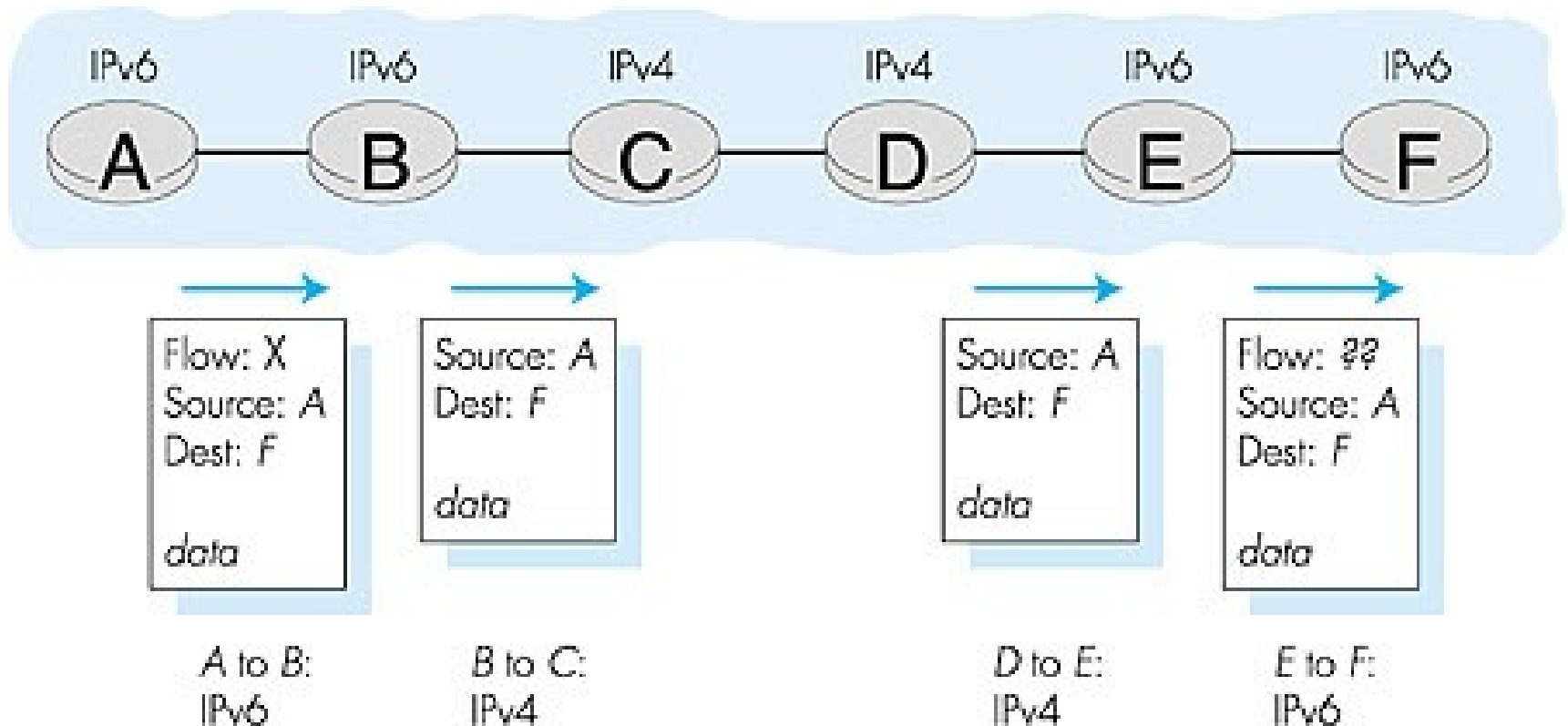
Other Changes from IPv4

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

Transition From IPv4 To IPv6

- Not all routers can be upgraded simultaneously
 - no flag days
 - How will the network operate with mixed IPv4 and IPv6 routers?
- Two proposed approaches:
 - *Dual Stack*: some routers with dual stack (v6, v4) can translate between formats
 - *Tunneling*: IPv6 carried as payload in IPv4 datagram among IPv4 routers

Dual Stack Approach

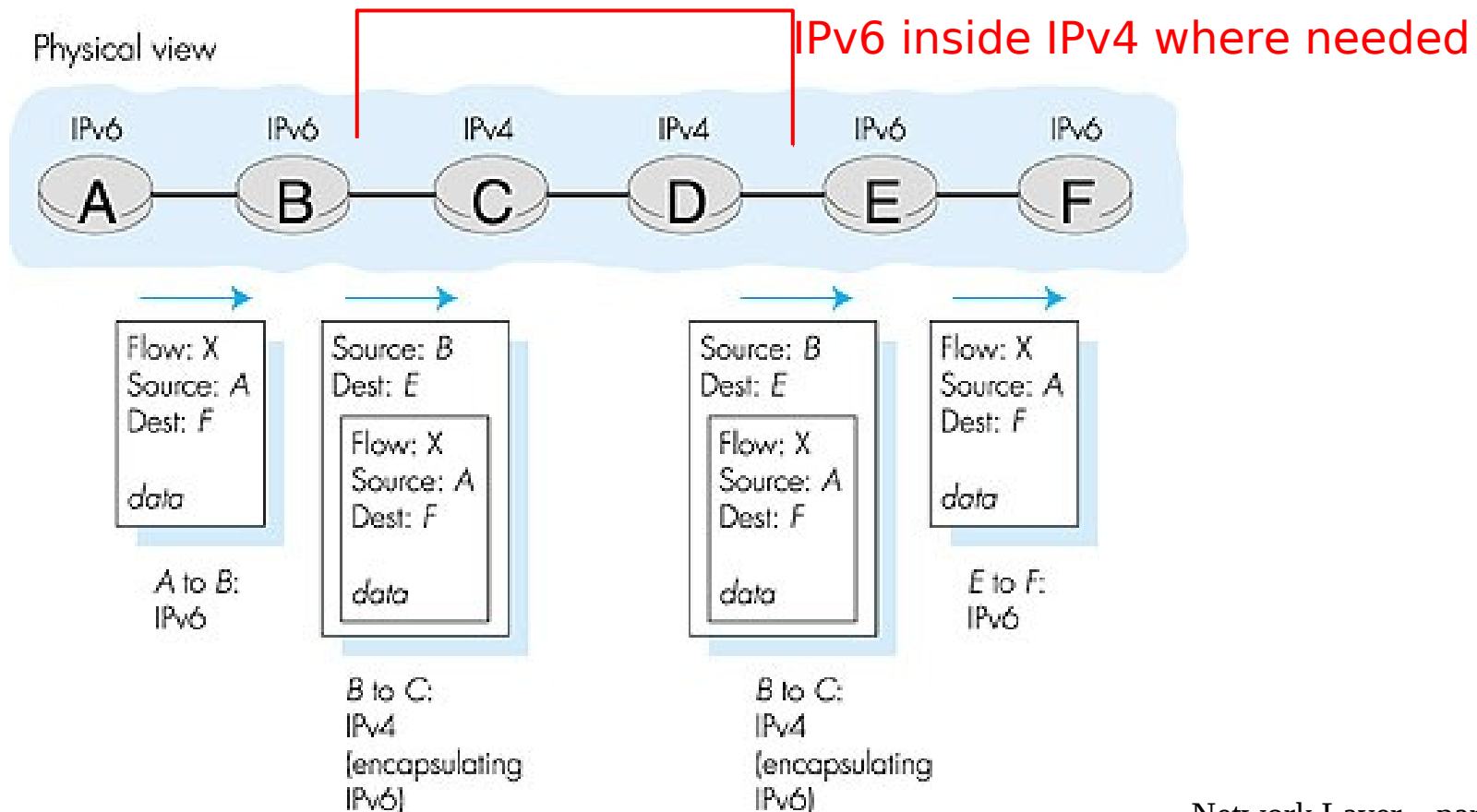


Tunneling

Logical view



Physical view



Extra exercises:

- ▮ `tracert` command (or `tracert` on Windows):
 - ▮ Useful to track the path between machines
 - ▮ Tells something about the time between hops
 - ▮ Can indicate bottlenecks in the path between your host and the destination
 - ▮ However:
 - Not 100% reliable
 - Some routers may not respond (no ICMP or blocked by firewalls)
 - Paths may change rapidly

Traceroute

- ▮ Use ifconfig (ipconfig) to find your IP address
- ▮ Ping a neighbour address until finding a live one
- ▮ traceroute (tracert) that neighbour
 - ▮ How many hops?

Traceroute

▮ traceroute (tracert) IT046600

▮ e.g.

```
prompt$ traceroute it046600 <enter>
```

```
traceroute to it046600 (130.123.248.237), 30 hops max, 40 byte packets
```

```
 1  it051752-vlan205.massey.ac.nz (130.123.246.129)  1.556 ms  1.736 ms  1.931 ms
 2  it050005-vlan907.massey.ac.nz (10.100.250.57)   1.616 ms  1.810 ms  1.952 ms
 3  it051761-vlan911.massey.ac.nz (10.100.250.74)   1.674 ms  1.875 ms  2.066 ms
 4  it046600.massey.ac.nz (130.123.248.237)  0.114 ms  0.113 ms  0.104 ms
```

Each line indicates the routers between this host and it046600 (a linux machine in a laboratory at Massey).

The traceroute works by sending an IP protocol packet with a small TTL to try to get an ICMP message back. The message is warning that the time has exceeded. The RTT can be estimated. In the listing above, the command sends three probes by default.

Traceroute

□ traceroute (tracert) www.google.com

□ e.g.

```
prompt$ traceroute www.google.com <enter>
```

```
traceroute to www.google.com (74.125.237.144), 30 hops max, 40 byte packets
```

```
 1  it051752-vlan205.massey.ac.nz (130.123.246.129)  1.810 ms  2.014 ms  2.220 ms
 2  it050006-vlan908.massey.ac.nz (10.100.250.61)   1.909 ms  2.115 ms  2.320 ms
 3  it032224-virt801.massey.ac.nz (10.100.3.225)    0.600 ms  0.489 ms  0.737 ms
 4  it028215-ge0-0.massey.ac.nz (130.123.3.236)    1.434 ms  1.492 ms  2.248 ms
 5  it028216-vlan2100.massey.ac.nz (130.123.1.2)   9.366 ms  9.265 ms  9.265 ms
 6  it028230-vlan802.massey.ac.nz (130.123.3.185)  10.066 ms 10.468 ms  9.206 ms
 7  anr2.karen.net.nz (210.7.32.2)  23.003 ms 22.505 ms 22.754 ms
 8  ani1.karen.net.nz (210.7.36.227)  23.172 ms 23.538 ms 23.521 ms
 9  ani9.karen.net.nz (210.7.36.182)  45.976 ms 45.003 ms 45.355 ms
10  202.167.228.73 (202.167.228.73)  42.734 ms 43.821 ms 44.022 ms
11  66.249.95.234 (66.249.95.234)  44.661 ms 44.727 ms 44.765 ms
12  72.14.237.137 (72.14.237.137)  44.694 ms 43.361 ms 44.203 ms
13  syd01s13-in-f16.1e100.net (74.125.237.144)  51.013 ms 50.586 ms 50.587 ms
```

Traceroute

- traceroute (tracert) www.google.com
- Note that the times are not accurate

