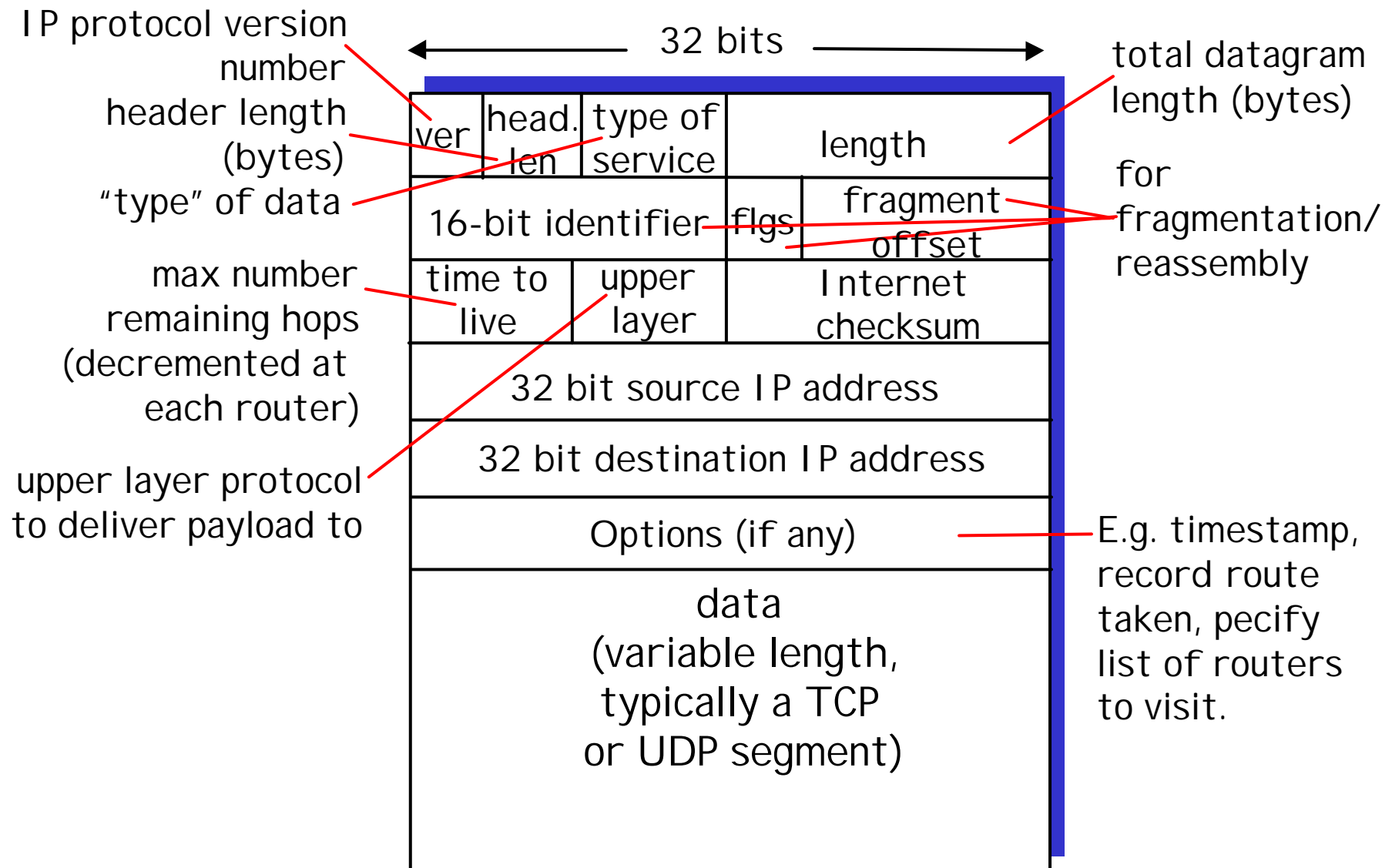
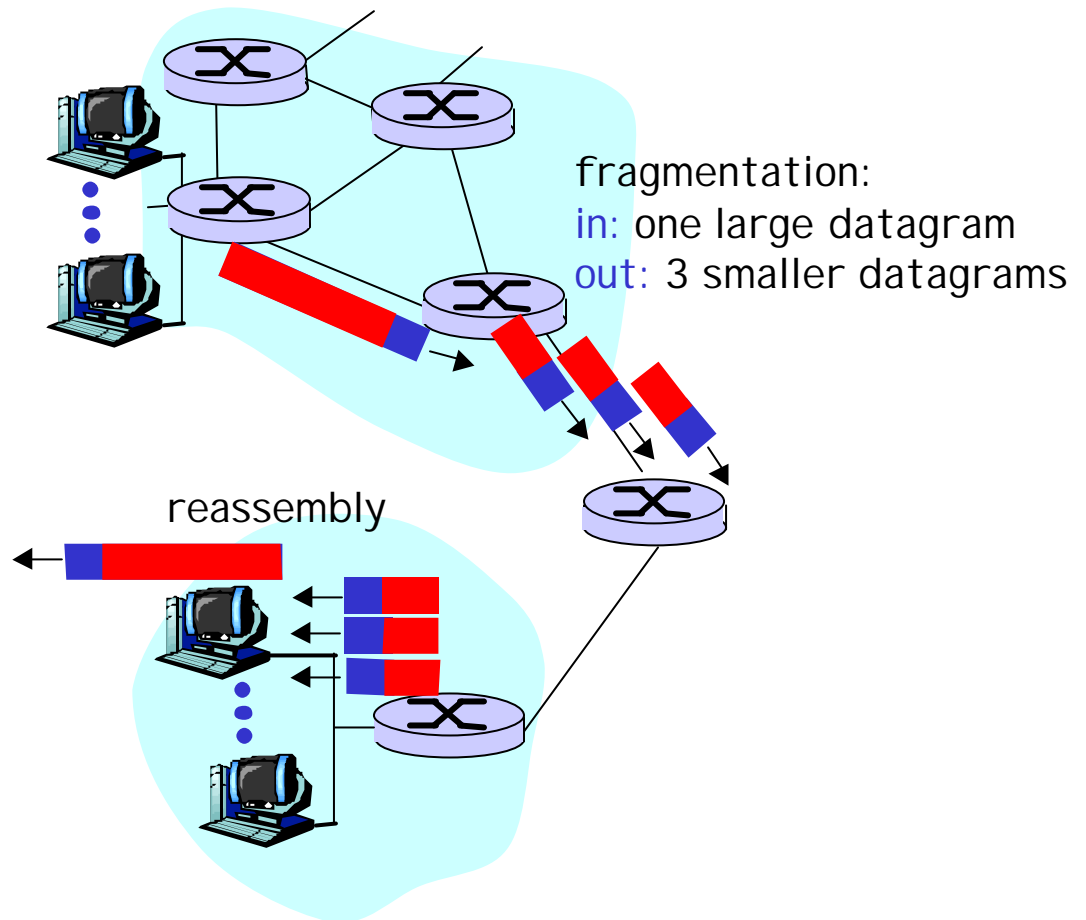


IP datagram format



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	ID	fragflag	offset	
	=4000	=x	=0	=0	

One large datagram becomes
several smaller datagrams

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

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

	length	ID	fragflag	offset	
	=1040	=x	=0	=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

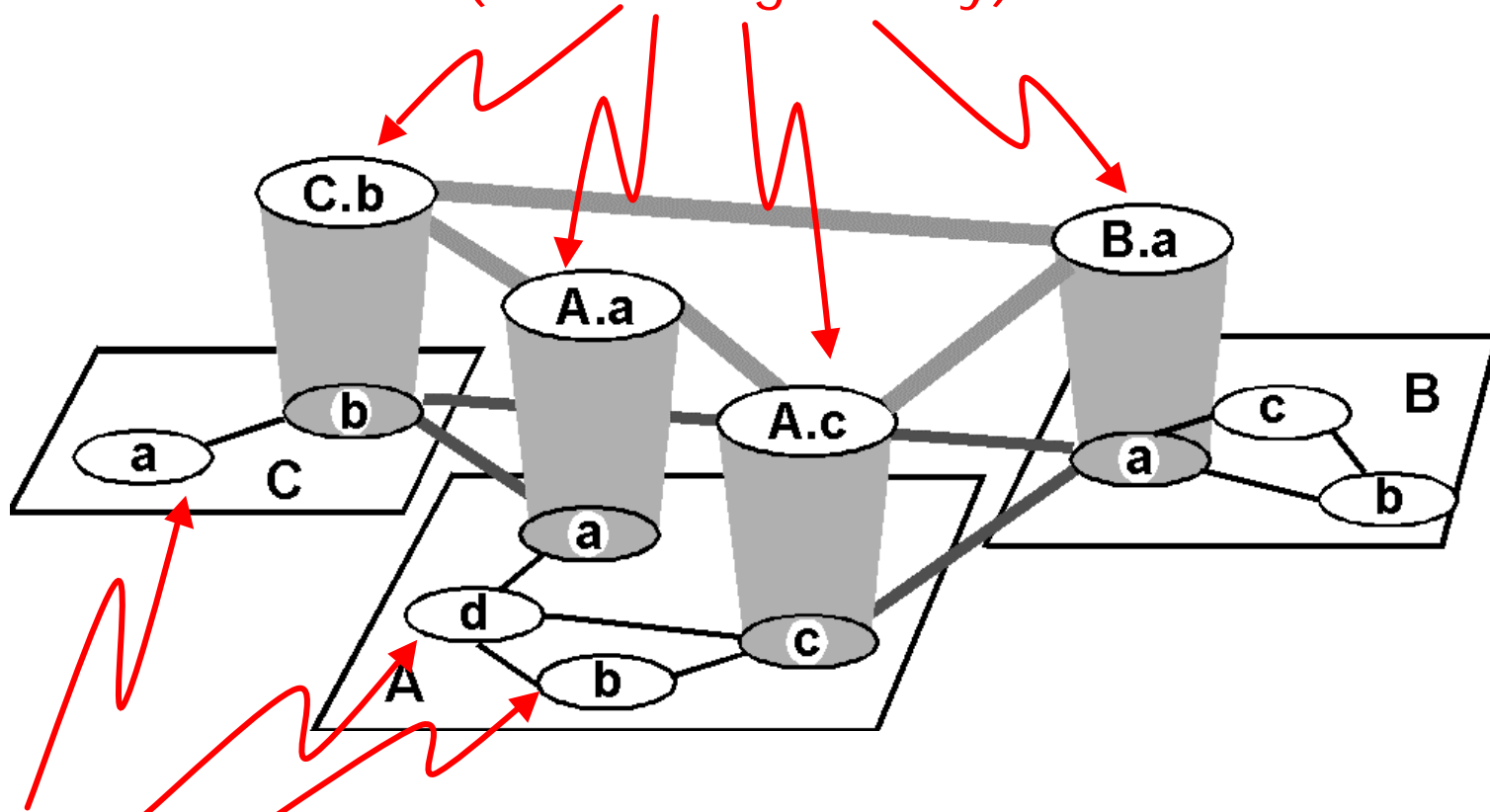
Routing in the Internet

- ❑ 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

Intra-AS border (exterior gateway) routers



Inter-AS interior (gateway) routers

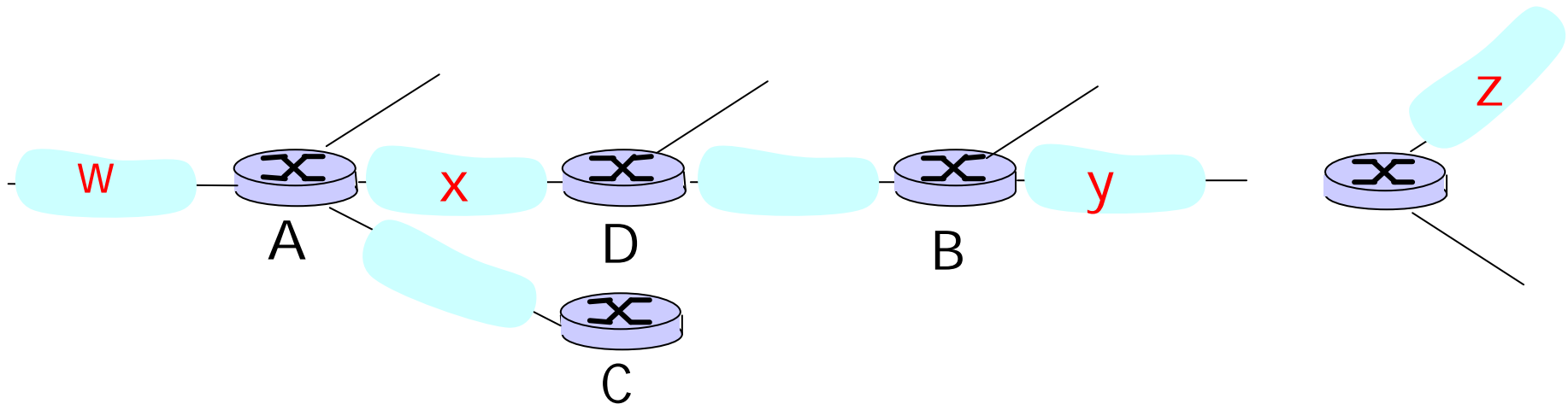
Intra-AS Routing

- ❑ 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	Next Router	Num. of hops to dest.
W	A	2
y	B	2
z	B	7
x	--	1
....

Routing table in D

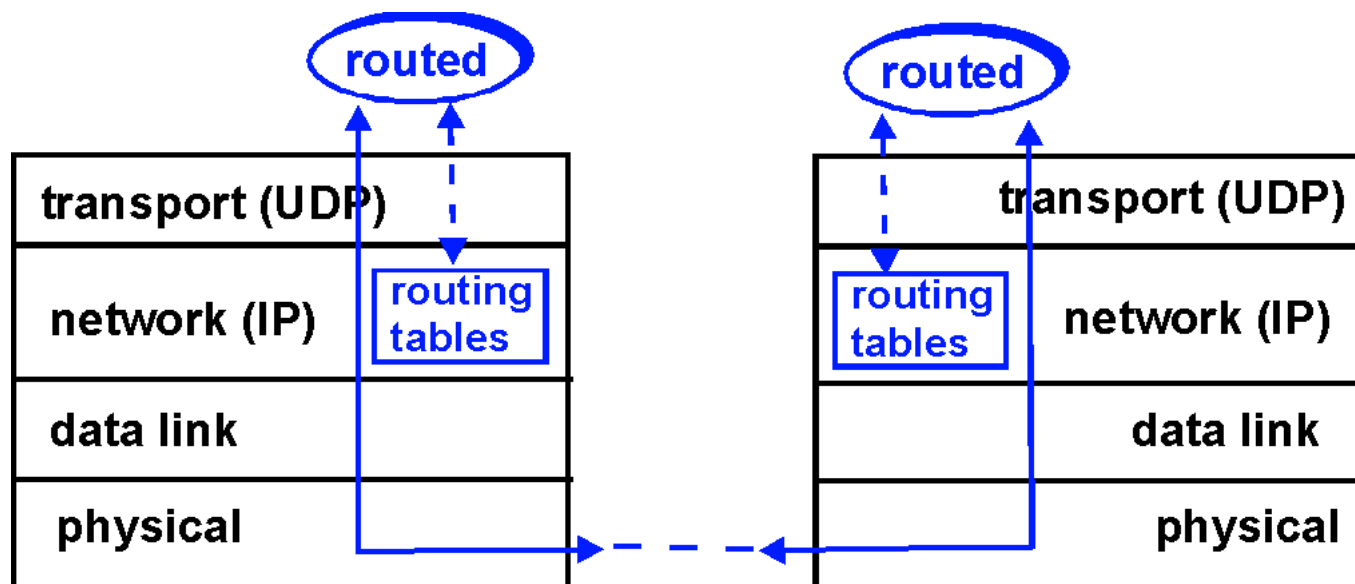
RI P: Link Failure and Recovery

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)

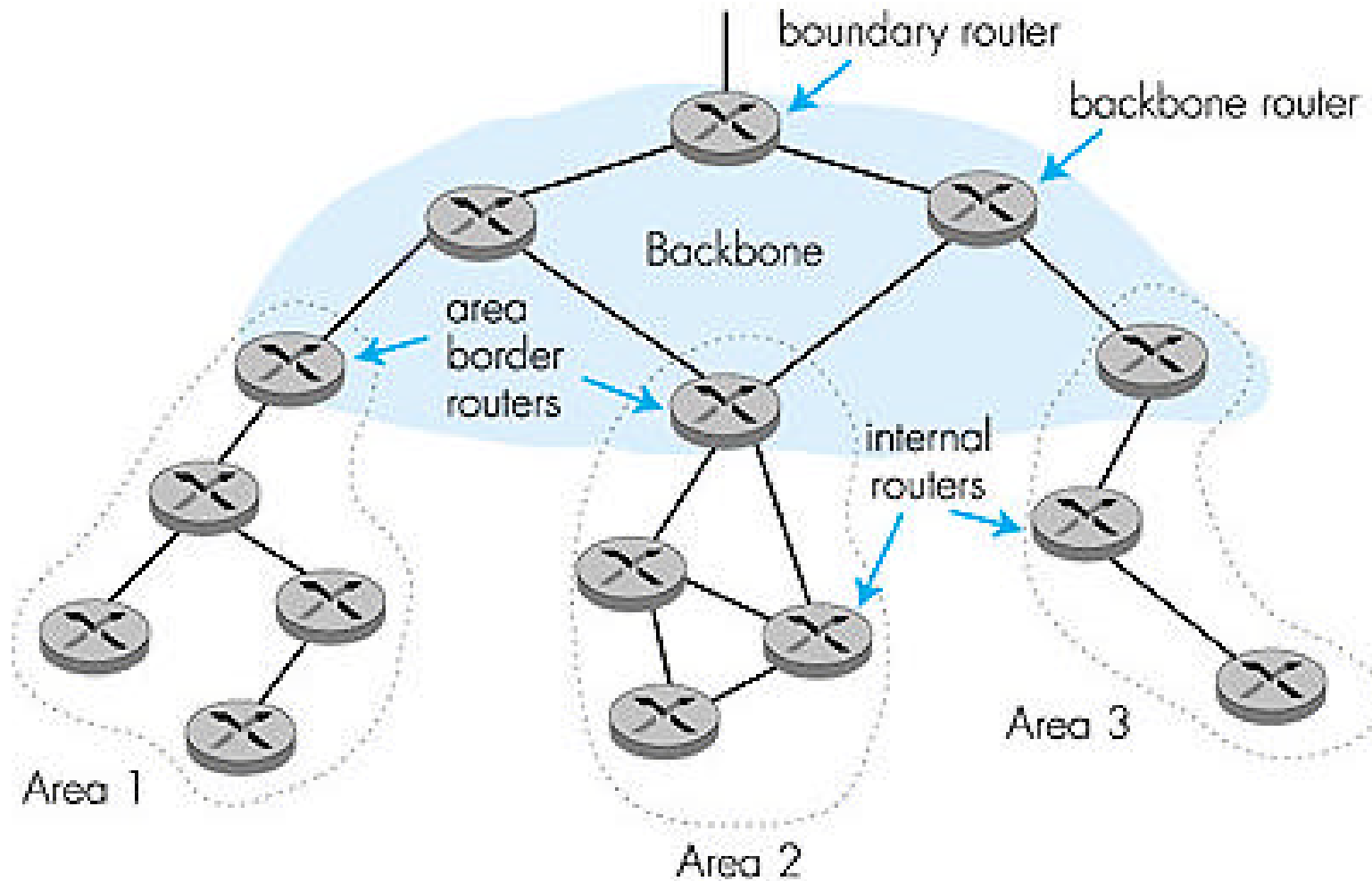
OSPF (Open Shortest Path First)

- ❑ “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 R I P)

- ❑ **Security**: all OSPF messages authenticated (to prevent malicious intrusion); TCP connections used
- ❑ **Multiple** same-cost **paths** allowed (only one path in R I P)
- ❑ 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



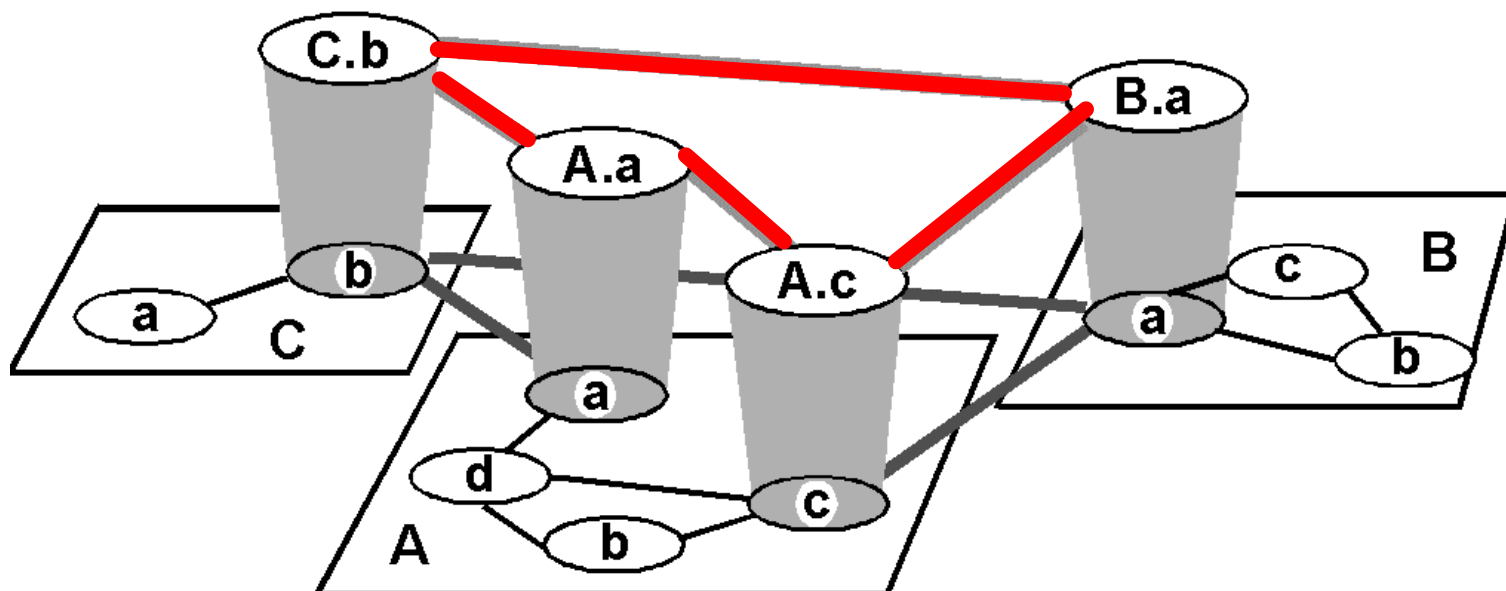
Hierarchical OSPF

- ❑ **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.

I GRP (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*

Inter-AS routing



Internet inter-AS routing: BGP

- ❑ **BGP (Border Gateway Protocol):** *the de facto standard*
- ❑ **Path Vector** protocol:
 - similar to Distance Vector protocol
 - each Border Gateway broadcast to neighbors (peers) *entire path* (I.e, sequence of ASs) to destination
 - E.g., Gateway X may send its path to dest. Z:

$\text{Path (X,Z)} = X, Y_1, Y_2, Y_3, \dots, Z$

Internet inter-AS routing: BGP

Suppose: gateway X send its path to peer gateway W

- ❑ W may or may not select path offered by X
 - cost, policy (don't route via competitors AS), loop prevention reasons.
- ❑ If W selects path advertised by X, then:
$$\text{Path (W,Z)} = w, \text{Path (X,Z)}$$
- ❑ Note: X can control incoming traffic by controlling its route advertisements to peers:
 - e.g., don't want to route traffic to Z -> don't advertise any routes to Z

Internet inter-AS routing: BGP

- ❑ BGP messages exchanged using TCP.
- ❑ BGP messages:
 - **OPEN**: opens TCP connection to peer and authenticates sender
 - **UPDATE**: advertises new path (or withdraws old)
 - **KEEPALIVE** keeps connection alive in absence of UPDATES; also ACKs OPEN request
 - **NOTIFICATION**: reports errors in previous msg; also used to close connection

Why different Intra- and Inter-AS routing ?

Policy:

- ❑ Inter-AS: admin wants control over how its traffic routed, who routes through its net.
- ❑ Intra-AS: single admin, so no policy decisions needed

Scale:

- ❑ hierarchical routing saves table size, reduced update traffic

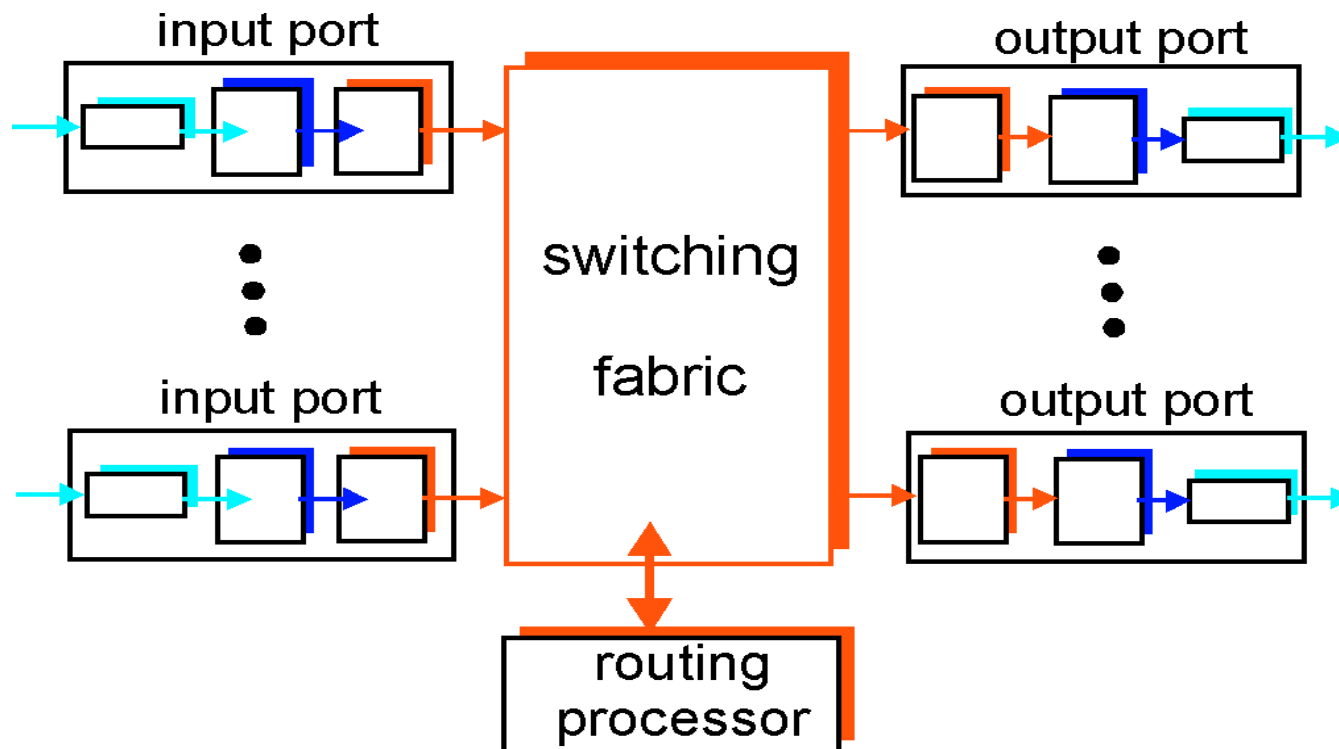
Performance:

- ❑ Intra-AS: can focus on performance
- ❑ Inter-AS: policy may dominate over performance

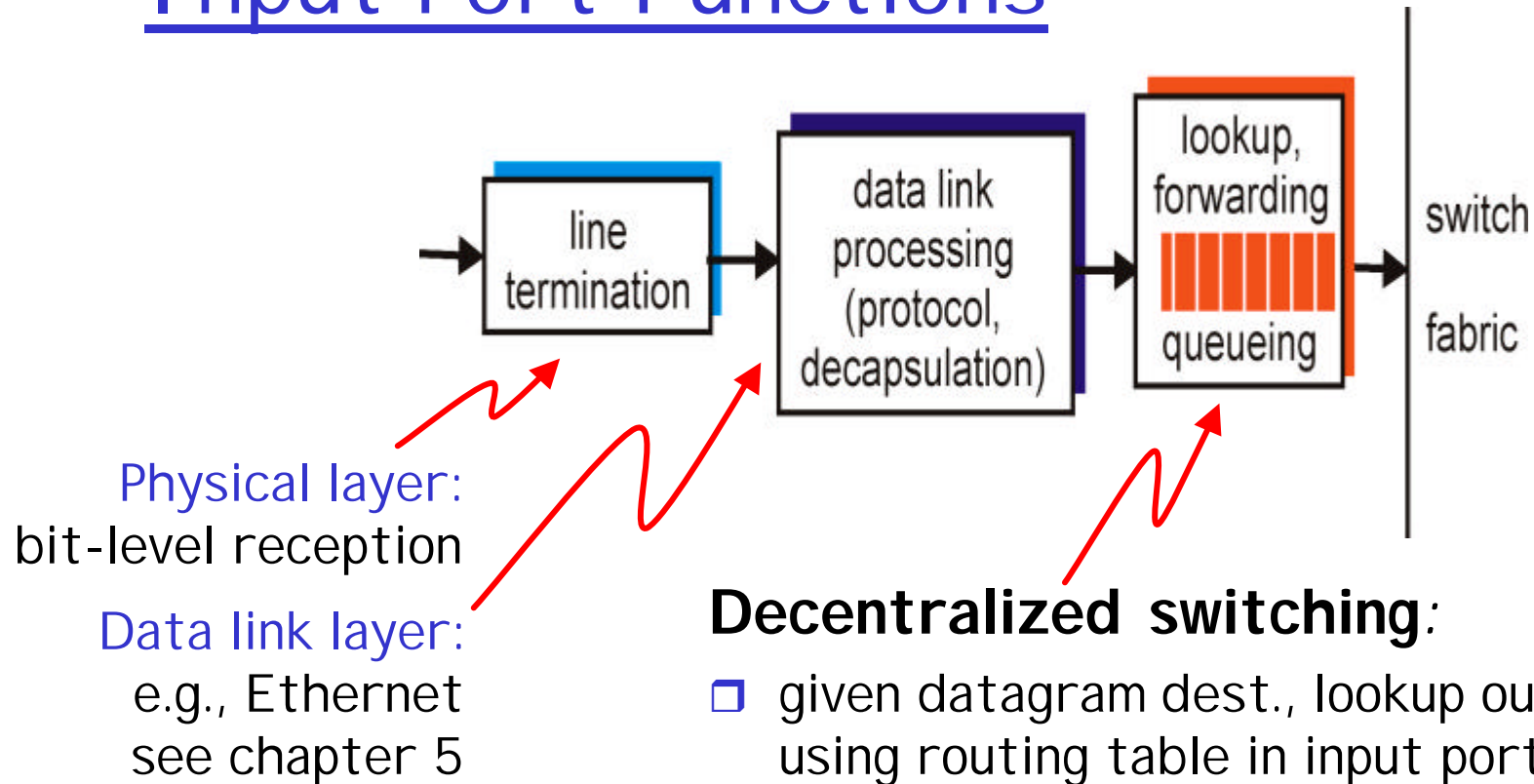
Router Architecture Overview

Two key router functions:

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



Input Port Functions

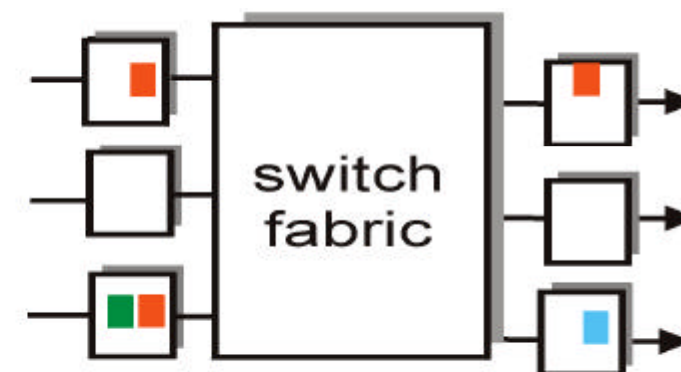
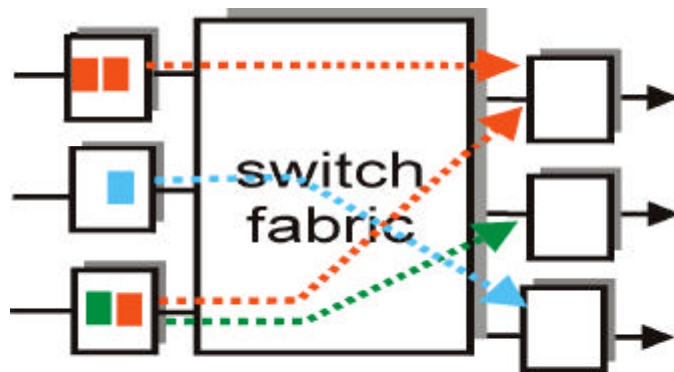


Decentralized switching:

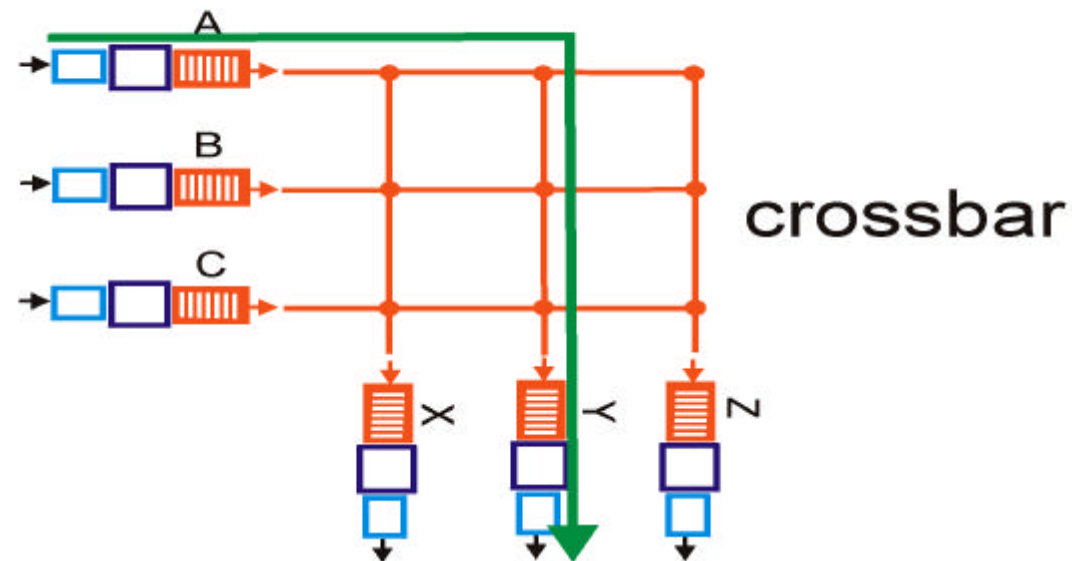
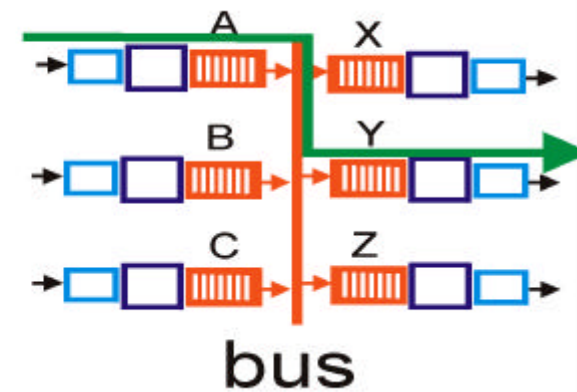
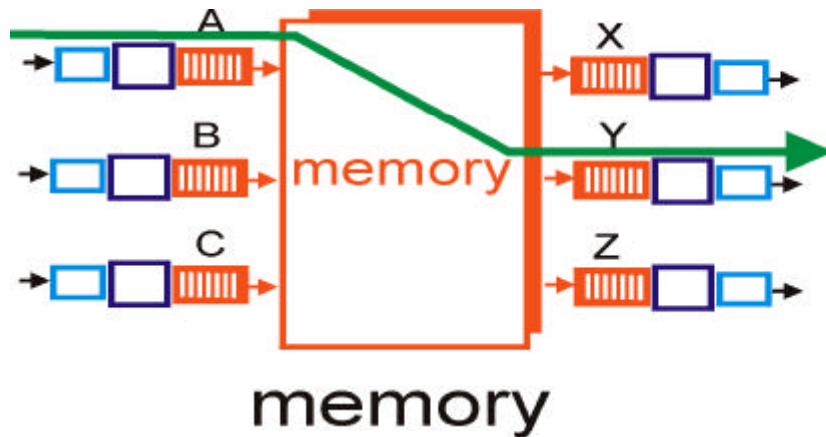
- ❑ given datagram dest., lookup output port using routing table in input port memory
- ❑ goal: complete input port processing at 'line speed'
- ❑ queuing: if datagrams arrive faster than forwarding rate into switch fabric

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!*



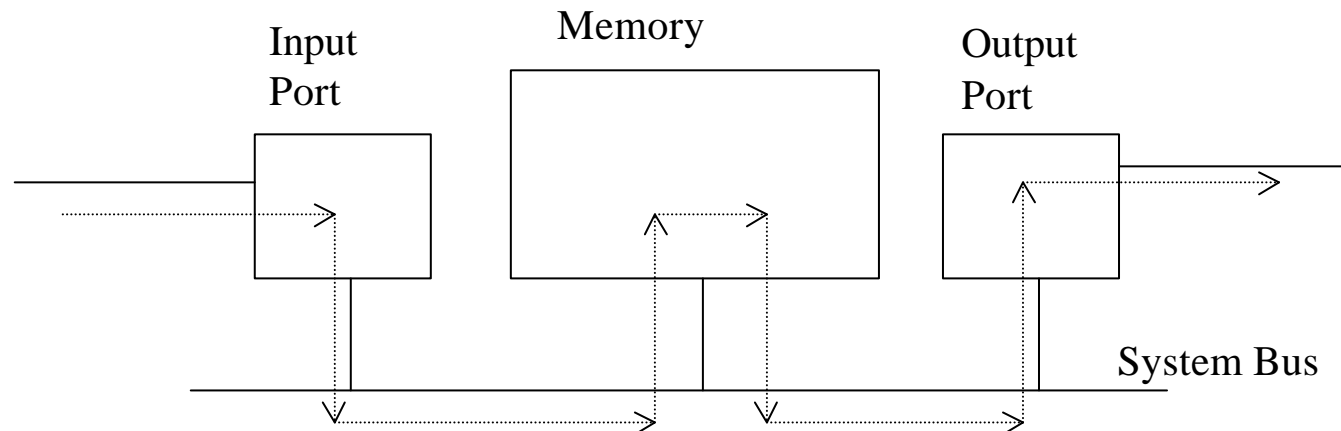
Three types of switching fabrics



Switching Via Memory

First generation routers:

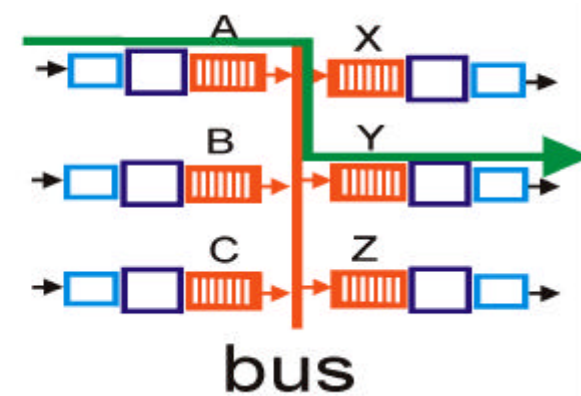
- ❑ packet copied by system's (single) CPU
- ❑ speed limited by memory bandwidth (2 bus crossings per datagram)



Modern routers:

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

Switching Via Bus

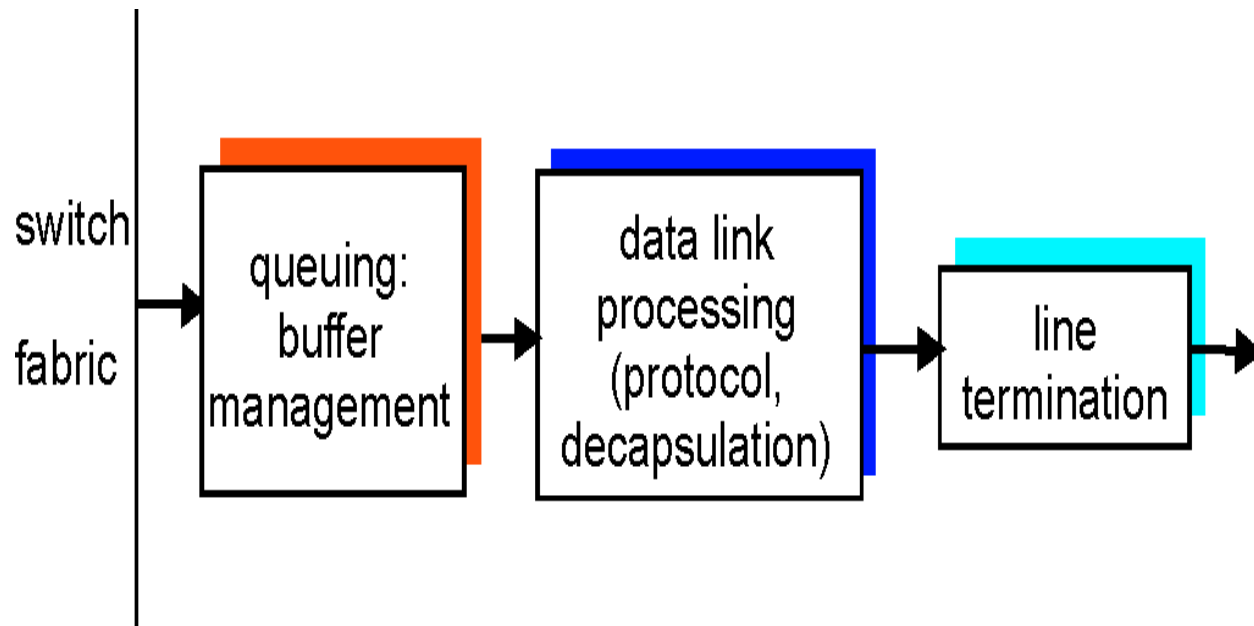


- ❑ datagram from input port memory to output port memory via a shared bus
- ❑ **bus contention:** switching speed limited by bus bandwidth
- ❑ 1 Gbps bus, Cisco 1900: sufficient speed for access and enterprise routers (not regional or backbone)

Switching Via An Interconnection Network

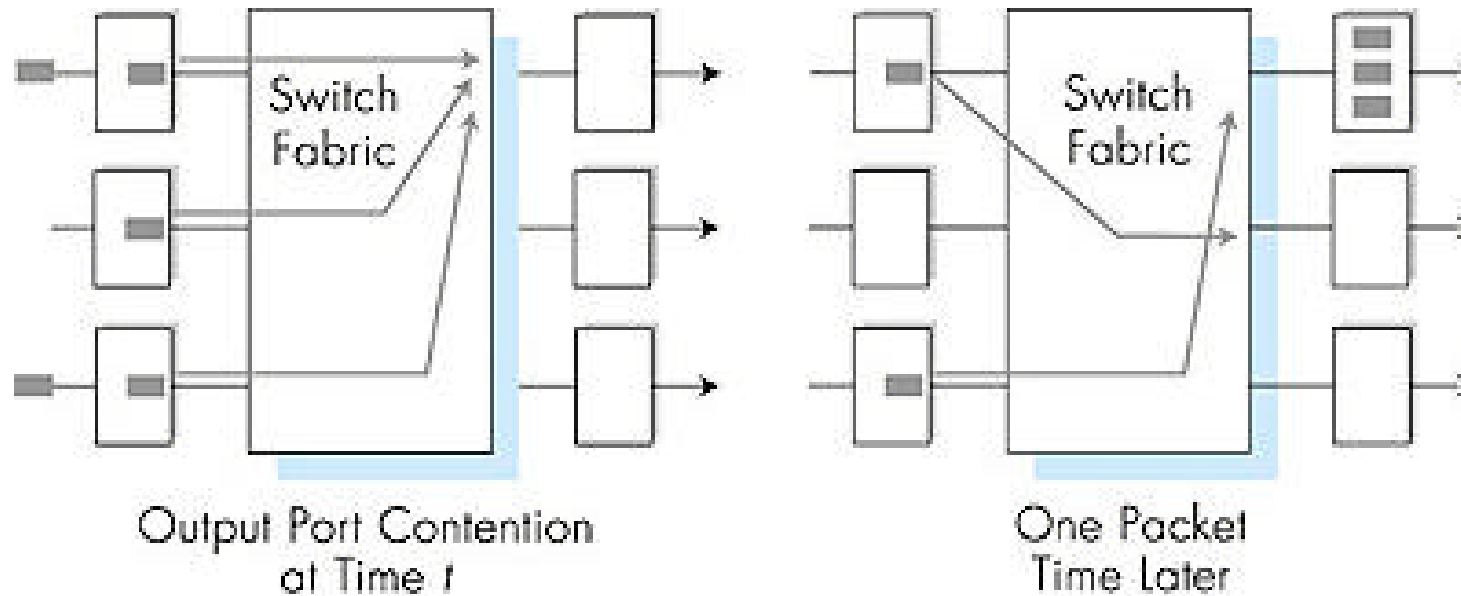
- ❑ overcome bus bandwidth limitations
- ❑ Banyan networks, other interconnection nets initially developed to connect processors in multiprocessor
- ❑ Advanced design: fragmenting datagram into fixed length cells, switch cells through the fabric.
- ❑ Cisco 12000: switches Gbps through the interconnection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

I Pv6

- ❑ **I nitial motivation:** 32-bit address space completely allocated by 2008.
- ❑ **A**dditional motivation:
 - header format helps speed processing/forwarding
 - header changes to facilitate QoS
 - new “anycast” address: route to “best” of several replicated servers
- ❑ **I Pv6 datagram format:**
 - fixed-length 40 byte header
 - no fragmentation allowed

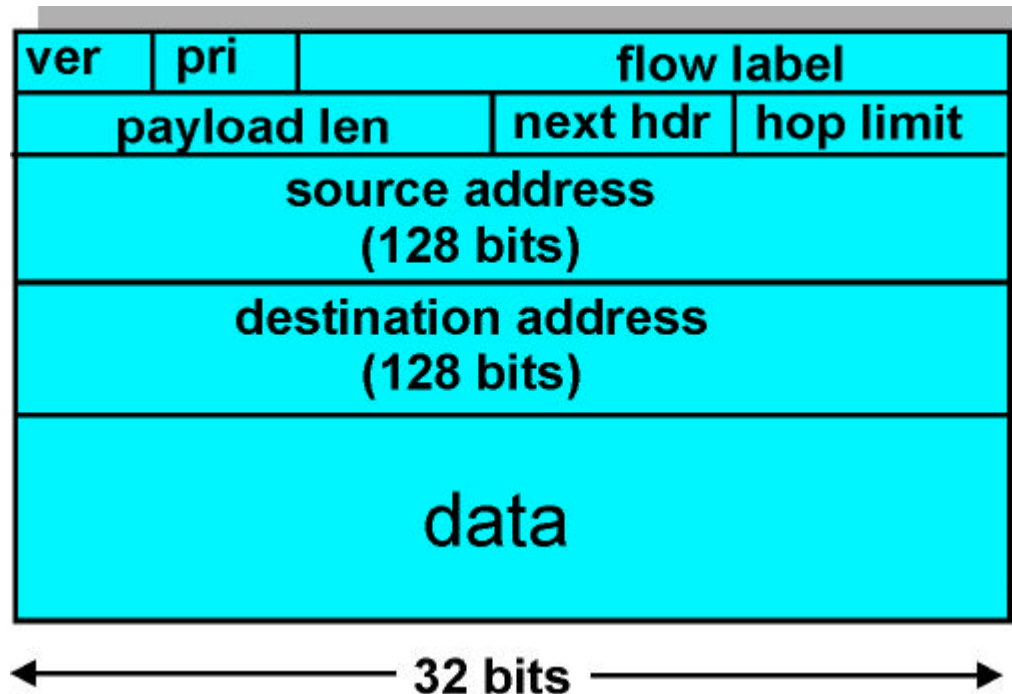
I Pv6 Header (Cont)

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



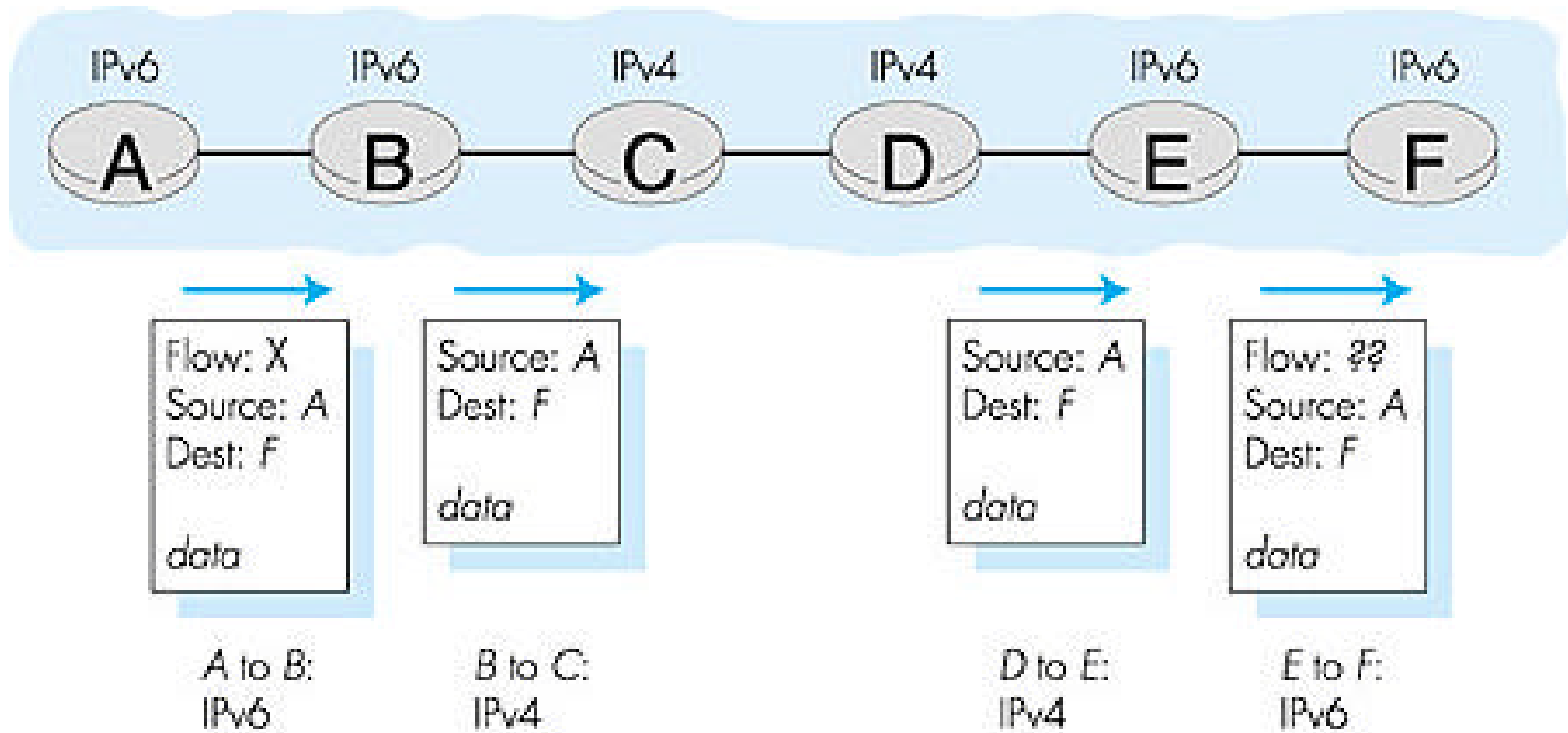
Other Changes from IPv4

- ❑ *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 I Pv4 To I Pv6

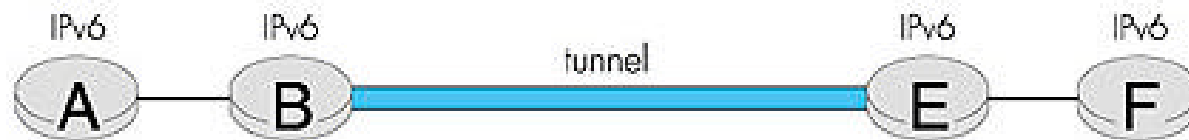
- ❑ Not all routers can be upgraded simultaneous
 - no “flag days”
 - How will the network operate with mixed I Pv4 and I Pv6 routers?
- ❑ Two proposed approaches:
 - *Dual Stack*: some routers with dual stack (v6, v4) can “translate” between formats
 - *Tunneling*: I Pv6 carried as payload in I Pv4 datagram among I Pv4 routers

Dual Stack Approach



Tunneling

Logical view



Physical view

