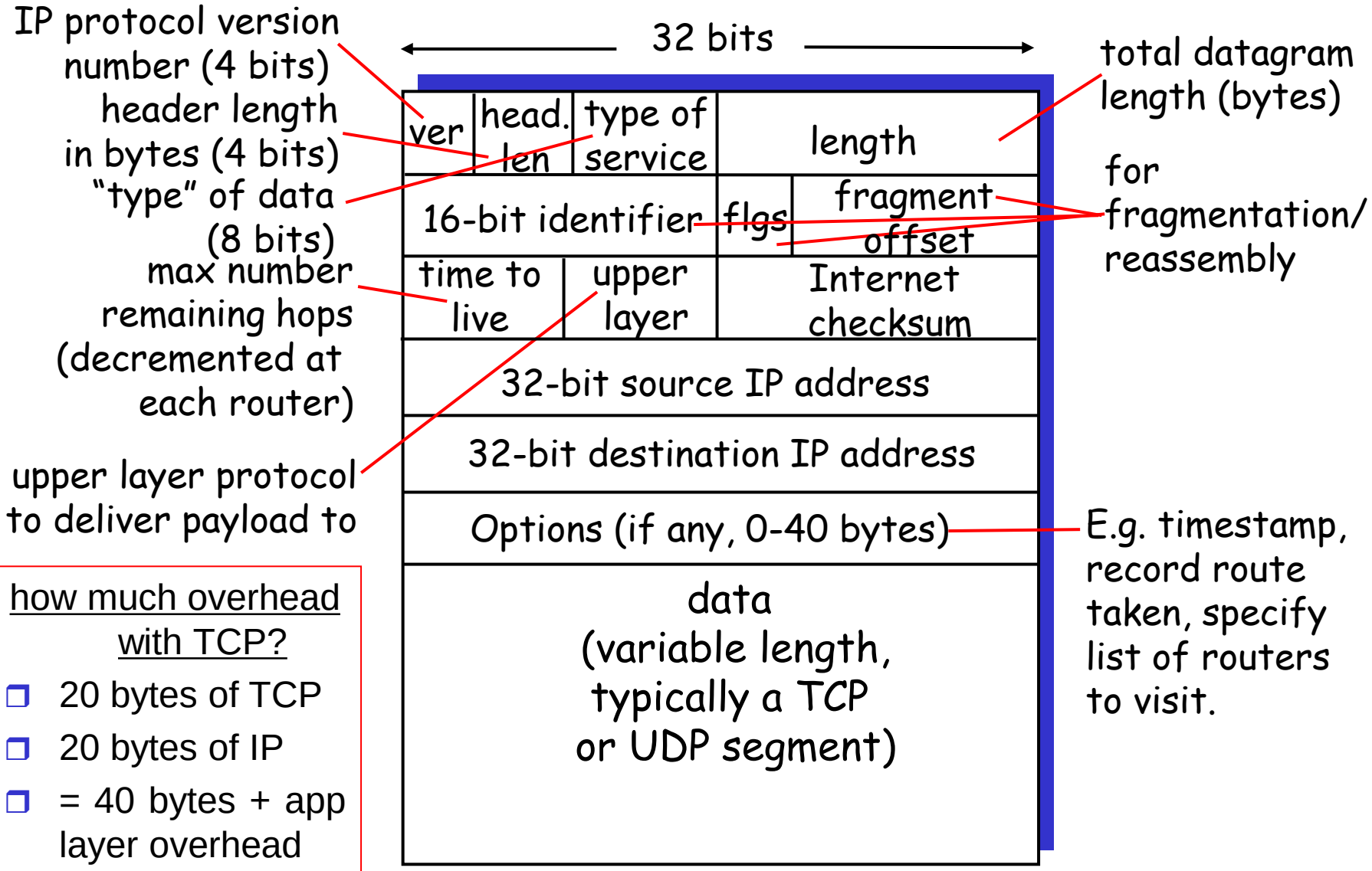
IP datagram format



how much overhead with TCP?

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

Upper layer Protocol Numbers

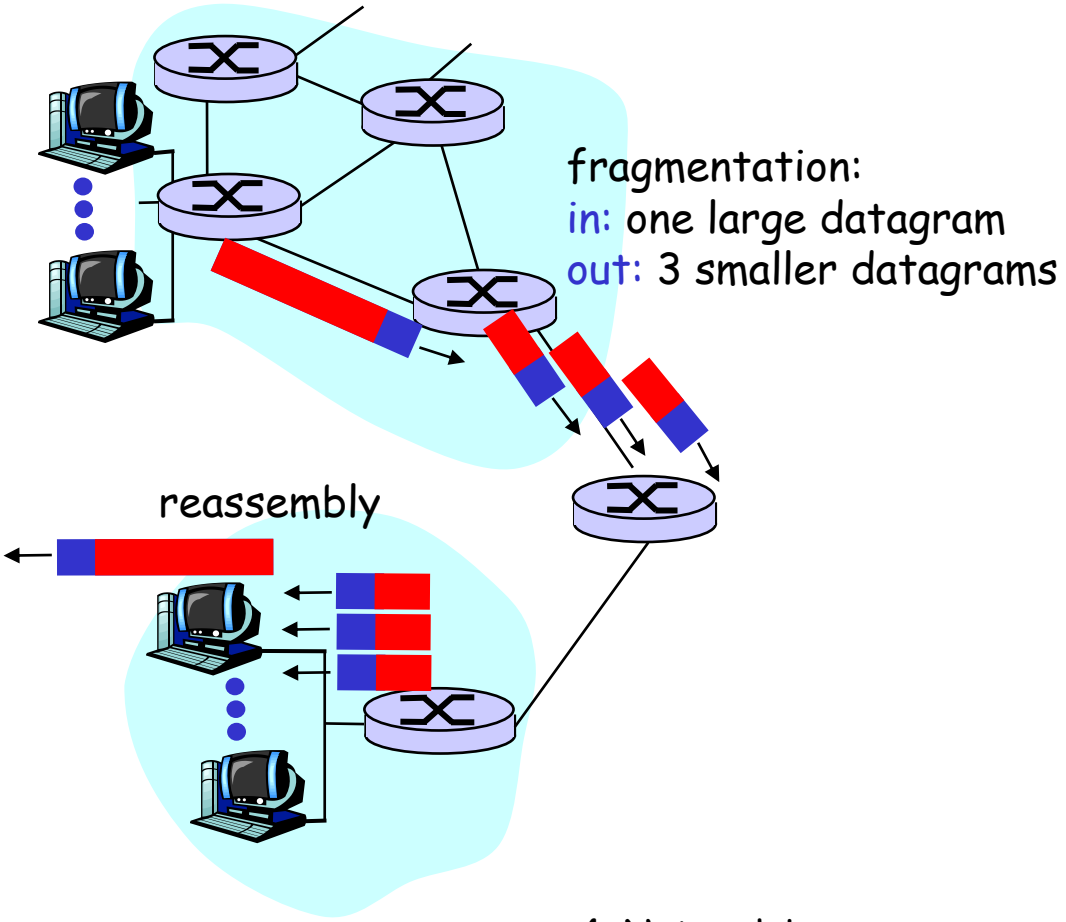
Decimal	Keyword	Protocol
0		Reserved
1	**ICMP**	Internet Control Message Protocol
2	IGMP	Internet Group Management Protocol
4	SCTP	Stream Control Transport Protocol
5	**TCP**	Transmission Control Protocol
8	EGP	Exterior Gateway Protocol
9	IGP	Interior Gateway Protocol
17	**UDP**	User Datagram Protocol
41	**IPv6**	IPv6 Encapsulation

Μέγιστη Μονάδα Μεταφοράς Δεδομένων (Maximum Transmission Unit – MTU)

- Κάθε φυσικό δίκτυο μεταδίδει πακέτα IP που έχουν ένα μέγιστο μήκος
- Τυπικές τιμές των IP πακέτων: 68 bytes μέχρι 8192 bytes
- IP MTU = IP Datagram (data + header)
- Το (μέγιστο) μήκος πλαισίου (frame) κάθε δικτύου είναι καθορισμένο
  - Ethernet MTU: 1518 bytes (= 18 bytes header + **1500 bytes data**).
  - Fiber Distributed Data Interface (FDDI – token ring) MTU: 4500 bytes.
  - Wireless LAN WiFi MTU: 2304 bytes (2312 bytes με κρυπτογράφηση).
- Η MTU οποιωνδήποτε πρωτοκόλλων πάνω από τα πρωτόκολλα του στρώματος ζεύξης, πρέπει να ταιριάζει στην MTU του στρώματος ζεύξης.
- Η IP MTU μπορεί να καθοριστεί ανεξάρτητα από την Ethernet MTU, αλλά αν το IP χρησιμοποιεί το Ethernet, η IP MTU πρέπει να είναι μικρότερη από την Ethernet MTU.

IP Fragmentation & Reassembly

- Κατακερματισμός (Fragmentation):
- 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



Κατακερματισμός (Fragmentation)

- Όλα τα επιμέρους πακέτα πλην του τελευταίου θα πρέπει να είναι πολλαπλάσια των 8 bytes.
  - Η τιμή στο fragment offset field σε κάθε IP πακέτο υπολογίζεται με βάση τη τρέχουσα αρχική θέση στο πακέτο σε σχέση με το αρχικό πακέτο. Μετράται σε μονάδες των 8-byte units.
  - Υπολογισμός του μήκους για την επικεφαλίδα του κάθε διασπασμένου πακέτου
  - Συνολικό μήκος του διασπασμένου πακέτου.
  - Νέος Υπολογισμός του CRC

IP Fragmentation and Reassembly

4000 bytes datagram

⇒ 20 bytes IP header +

3980 data (segment size)

3980 : 0 … 3979

MTU = 1500

Τότε σε κάθε fragment έχουμε 1480 bytes in data field (το πολύ)

1ο fragment: **0** – 1479

offset = **0** (1480/8 = 185)

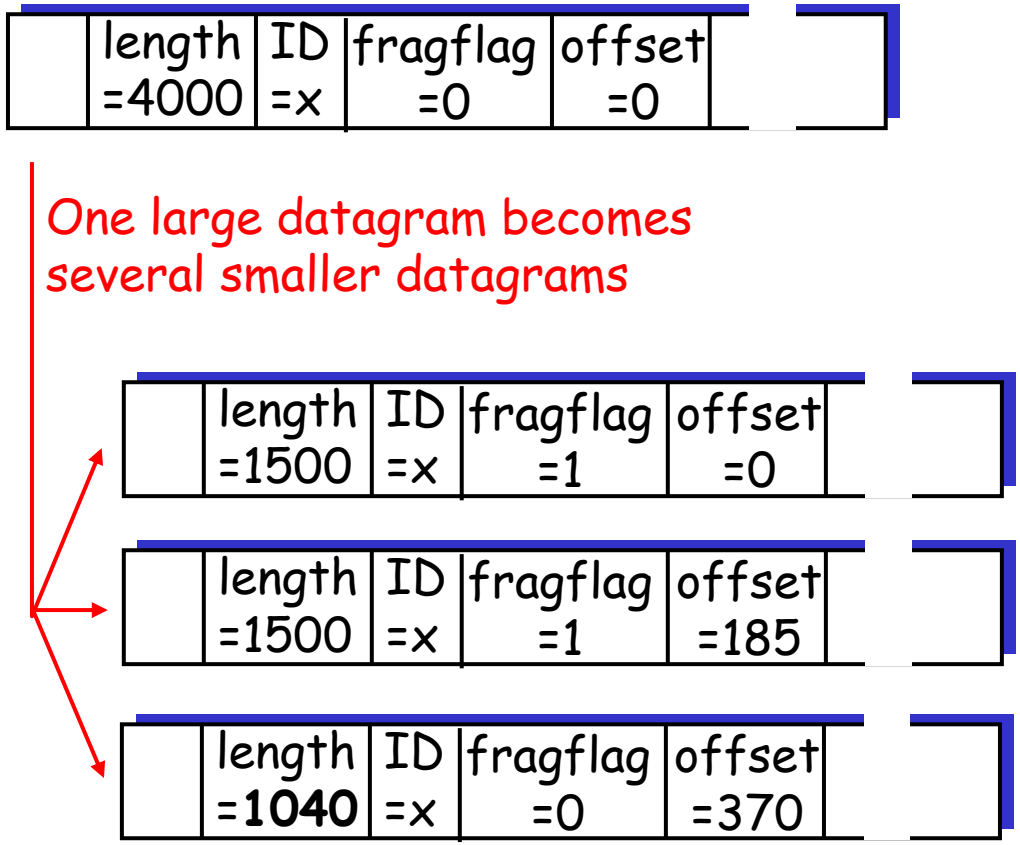
2 ο fragment: **1480** – 2959

offset = 1480 ή **1480/8 =185**

3 ο fragment: **2960** – 3979

offset = 2960 ή **2960/8 =370**

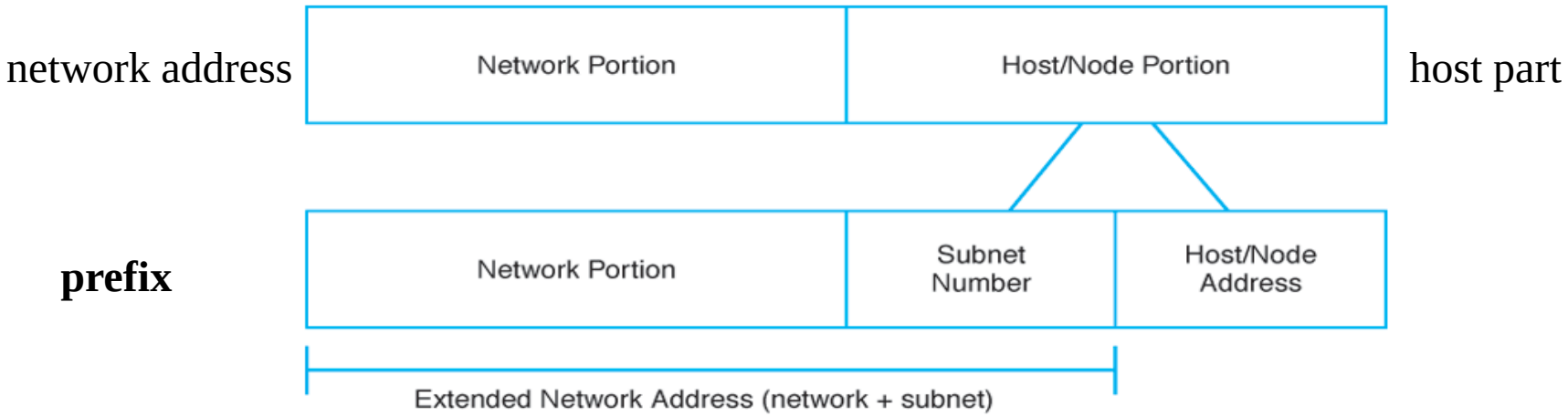
3979 – 2960 + 1 + 20 = **1040**



Χωρισμός σε υποδίκτυα (subnetting)

- Για την αποφυγή χρήσης επιπρόσθετων IP-διεύθυνσεων
- Για την **σμίκρυνση των πινάκων δρομολόγησης** – ιεραρχική δρομολόγηση
- Το 'host part' της IP-διεύθυνσης υποδιαιρείται σε *subnet number* και *host number*
  - Η IP διεύθυνση είναι της μορφής:

    <**network address**> < **subnet number**> <**host number**>
- Ο συνδυασμός του subnet number και host number ονομάζεται **local address**



- http://www.subnetmask.info/

Σμίκρυνση πινάκων δρομολόγησης – Ιεραρχική δρομολόγηση

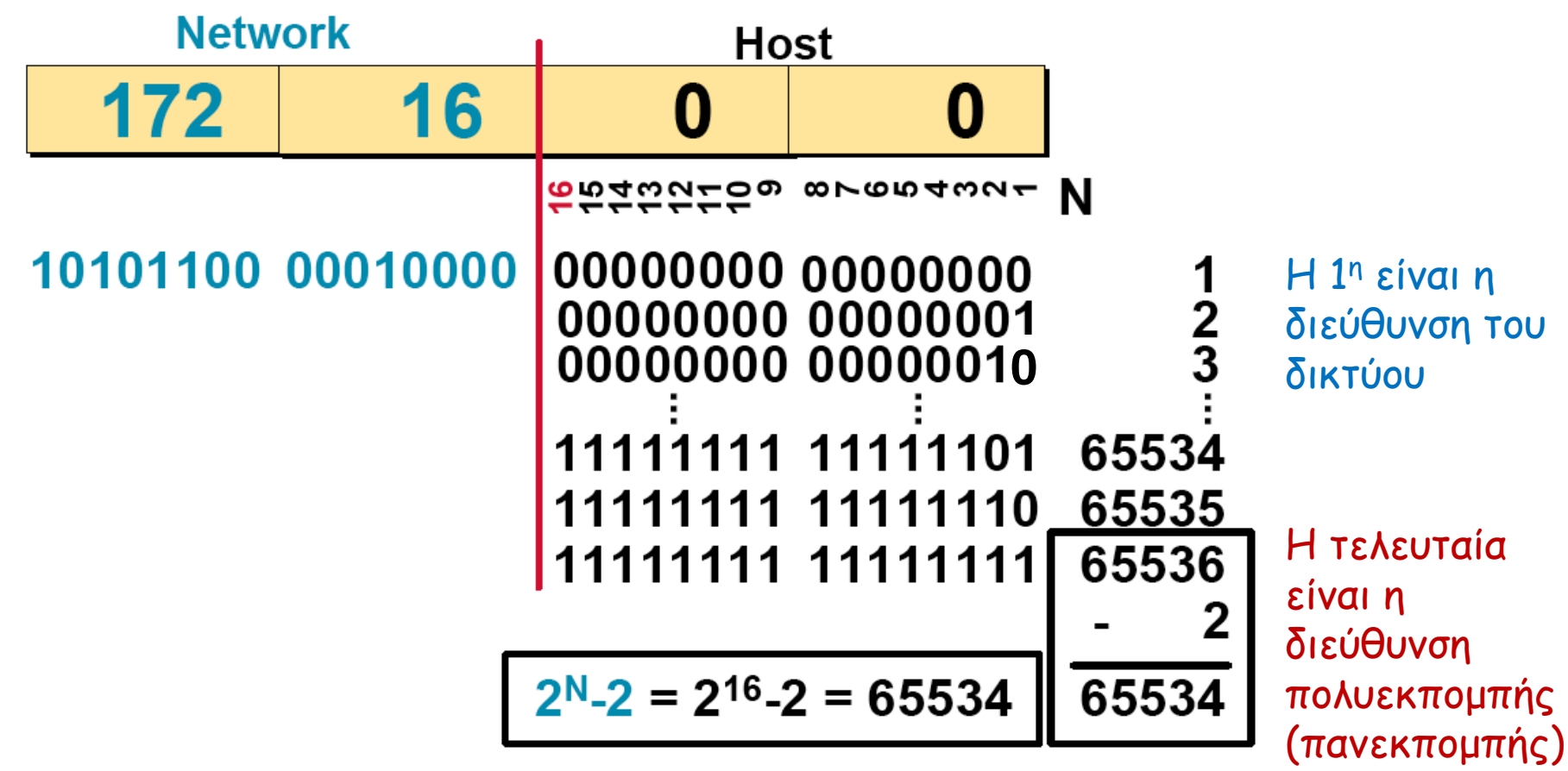
Destination Address	Interface
0000	…
0001	…
0010	…
0011	…
0100	…
0101	…
0110	…
0111	…
1000	…
1001	…
1010	…
1011	…
1100	xxx
1101	…
1110	…
1111	…

(13)

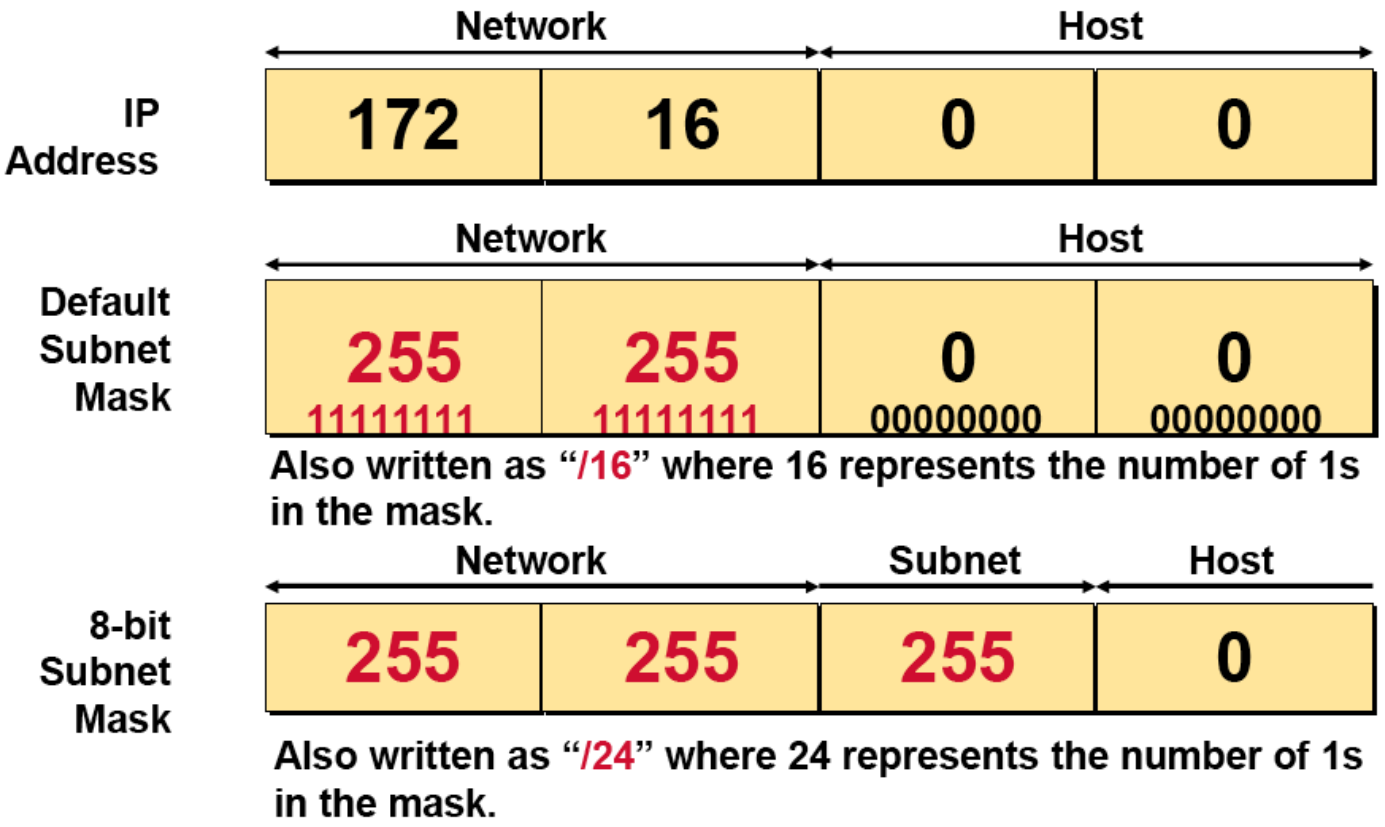
Subnet	Destination Address	Interface
00	0000 0001 0010 0011	… … … …
01	0100 0101 0110 0111	… … … …
10	1000 1001 1010 1011	… … … …
11	1100 1101 1110 1111	xxx … … …

(5)

Διαθέσιμες διευθύνσεις IP σε χρήστες



Η χρήση μάσκας (subnetmask)



Παράδειγμα χωρισμού σε υποδίκτυα - subnetting

- Host στο Διαδίκτυο έχει διεύθυνση IPv4: 144.128.128.16.

(α) Σε ποια κλάση ανήκει η διεύθυνση αυτή;

(β) Σύμφωνα με την κλάση, να ευρεθεί κατάλληλη μάσκα υποδικτύου (subnetmask), ώστε να χωριστεί το δίκτυο στο οποίο ανήκει σε 4 ίσα υποδίκτυα (subnetworks).

(γ) Να ορισθούν οι διευθύνσεις IP των 4 υποδικτύων και οι διευθύνσεις IP των hosts σε κάθε υποδίκτυο, καθώς και η διεύθυνση πολυεκπομπής κάθε υποδικτύου. Οι απαντήσεις να καταχωρηθούν στον ακόλουθο πίνακα.

IP address του Sub-Network	Hosts (IP address)		Broadcast IP (Πολυεκπομπή)
	Από	Μέχρι	
...	...	...	...

(δ) Χωρίς να βασίζεστε στον πίνακα, αλλά μέσω της μάσκας υποδικτύου, βρείτε σε ποιο υποδίκτυο ανήκει ο host με τη διεύθυνση 144.128.128.16.

- Άλλες λειτουργίες της subnetmask.

Παράδειγμα χωρισμού σε υποδίκτυα – subnetting (συν. 1)

- (α) 144.128.128.16 → 10010000 10000000 10000000 00010000

  Το πρώτο byte της IP διεύθυνσης, το 144, σε δυαδική μορφή γράφεται: 10010000. Οπότε βλέπουμε ότι το πρώτο «0» βρίσκεται στην $2^η$ θέση, άρα έχουμε **class B**.

- (β) Για να διακρίνουμε 4 υποδίκτυα χρειαζόμαστε 2 bits ($2^2 = 4$).

  Αυτά τα bits θα τα πάρουμε από τα πρώτα bits του host part της διεύθυνσης.

  Αφού η (**default** όπως λέμε) **subnetmask** της class B είναι 255.255.0.0,

  Ή, 11111111 11111111 00000000 00000000

  Τα 2 bits που θα μας δώσουν το subnet number είναι τα bits #17 και #18 (κόκκινα). Άρα η subnetmask θα γίνει:

  11111111 11111111 11000000 00000000 → **255.255.192.0**.

  - Η CIDR μορφή της μάσκας υποδικτύων είναι /18, δηλ. **144.128.128.16/18**

Παράδειγμα χωρισμού σε υποδίκτυα – subnetting (συν.2)

- (γ) Για να βρούμε τις διευθύνσεις των subnets παίρνουμε την network address και κολλάμε τους συνδυασμούς των 2 bits (17, 18) ως εξής:

10010000 10000000 **00**000000 00000000 → **144.128.0.0** **subnet**
10010000 10000000 **00**111111 11111111 → 144.128.63.255 broadcast

10010000 10000000 **01**000000 00000000 → **144.128.64.0** **subnet**
10010000 10000000 **01**111111 11111111 → 144.128.127.255 broadcast

10010000 10000000 **10**000000 00000000 → **144.128.128.0** **subnet**
10010000 10000000 **10**111111 11111111 → 144.128.191.255 broadcast

10010000 10000000 **11**000000 00000000 → **144.128.192.0** **subnet**
10010000 10000000 **11**111111 11111111 → 144.128.255.255 broadcast

(3ο)

IP address του Sub-Network	Hosts (IP address)		Broadcast IP (Πολυεκπομπή)
	Από	Μέχρι	
144.128.0.0	144.128.0.1	144.128.63.254	144.128.63.255
144.128.64.0	144.128.64.1	144.128.127.254	144.128.127.255
144.128.128.0	144.128.128.1	144.128.191.254	144.128.191.255
144.128.192.0	144.128.192.1	144.128.255.254	144.128.255.255

Παράδειγμα χωρισμού σε υποδίκτυα – subnetting (συν.3)

- (δ) Για να βρούμε το υποδίκτυο στο οποίο ανήκει μια διεύθυνση IP που έχει δοθεί σε host, εκτελούμε την πράξη AND μεταξύ της IP address και της subnetmask:

IP address: 10010000 10000000 10000000 00010000
subnetmask: 11111111 11111111 11000000 00000000
----------AND
subnet address: 10010000 10000000 10000000 00000000 → **144.128.128.0** (3ο)

- Άλλες λειτουργίες της subnetmask

Subnet address: 10010000 10000000 10000000 00000000 (144.128.128.0)
Inverted subnetmask: 00000000 00000000 00111111 11111111
----------XOR
Broadcast address: 10010000 10000000 10111111 11111111 → **144.128.191.255**

ip address: 10010000.10000000.10000000.00010000
inverted subnet mask: 00000000.00000000.00111111.11111111
----------AND
Host number: 00000000.00000000.00000000.00010000 → **16**

Συσσωμάτωση διευθύνσεων IP - Aggregation

- Router πρέπει να προωθεί πακέτα στις εξόδους του ως εξής:

Destination address Range	Interface (port)
Από 11001000 00010111 00010000 00000000 Μέχρι 11001000 00010111 00010111 11111111	I0
Από 11001000 00010111 00011000 00000000 Μέχρι 11001000 00010111 00011000 11111111	I1
Από 11001000 00010111 00011001 00000000 Μέχρι 11001000 00010111 00011111 11111111	I2
Άλλες διευθύνσεις (default)	I3

- Βρείτε τον πίνακα δρομολόγησης με τις μικρότερες εγγραφές.

Συσσωμάτωση διευθύνσεων IP - Aggregation

Routing Table

Destination address Range	Destination address	Interface (port)
Από **11001000 00010111 00010**000 00000000 Μέχρι 11001000 00010111 00010111 11111111	200.23.16.0/21	I0
Από **11001000 00010111 00011000** 00000000 Μέχρι 11001000 00010111 00011000 11111111	200.23.24.0/24	I1
Από **11001000 00010111 00011**001 00000000 Μέχρι 11001000 00010111 00011111 11111111	200.23.25.0/24 200.23.26.0/23 200.23.28.0/22	I2 I2 I2
Άλλες διευθύνσεις (default)	default	I3

Destination Address	Interface
192.168.0.0/24	I1
192.168.1.0/24	I1
192.168.2.0/24	I1
192.168.3.0/24	I1
192.168.4.0/24	I1

⇒

Longest prefix matching

Destination Address	Interface
192.168.0.0/**22**	I1
192.168.4.0/**24**	I1

ICMP: Internet Control Message Protocol

Upper layer Protol Number = 1

- used by hosts, routers, gateways to communication 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

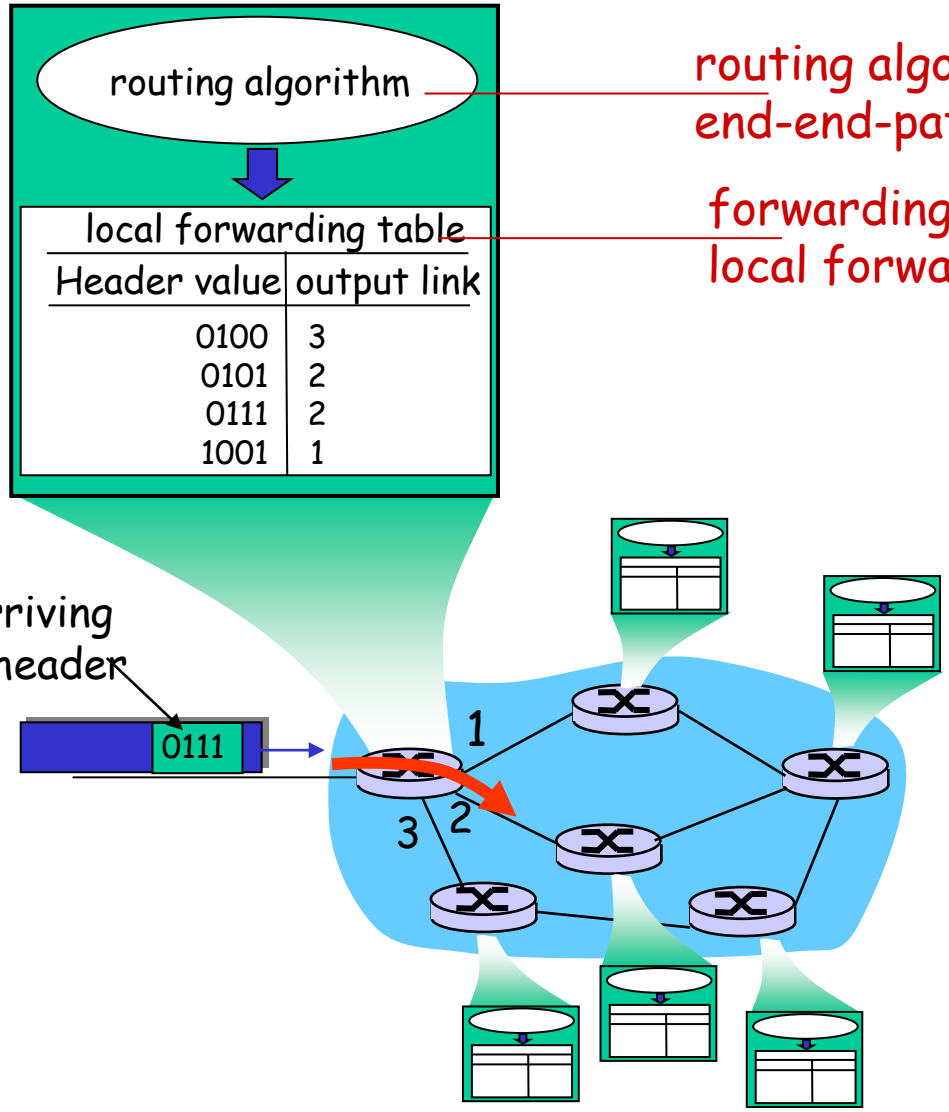
Type	Code	description
0	0	echo reply (ping)
3	0	dest. network unreachable
3	1	dest host unreachable
3	2	dest protocol unreachable
3	3	dest port unreachable
3	6	dest network unknown
3	7	dest host unknown
4	0	source quench (congestion control - not used)
8	0	echo request (ping)
9	0	route advertisement
10	0	router discovery
11	0	TTL expired
12	0	bad IP header

Traceroute: Μια εφαρμογή με βάση το TTL

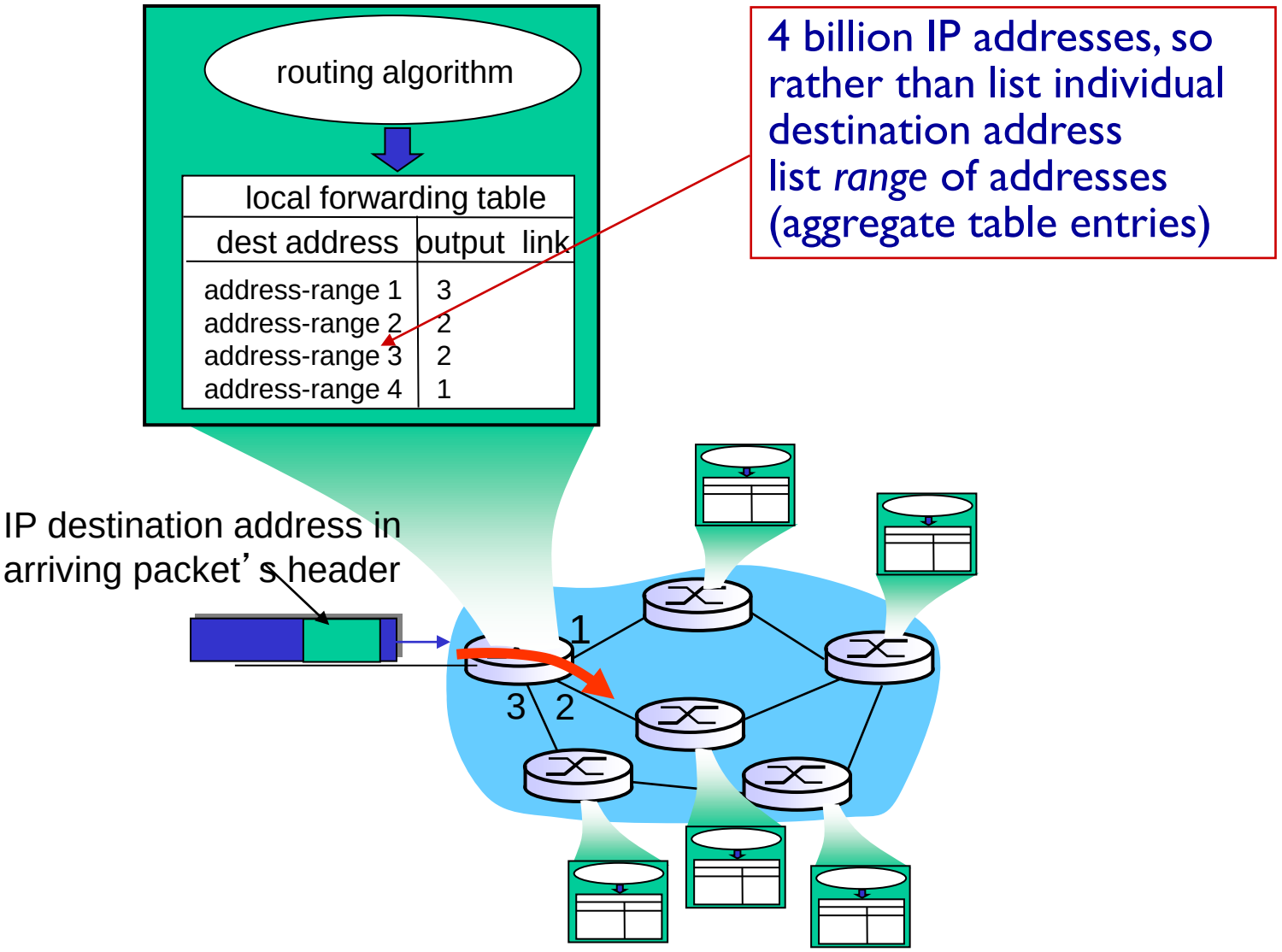
- Το Traceroute μας βοηθάει σε διαδικασίες test ή debugging (διαδικασία εντοπισμού και διόρθωσης λαθών).
- Στέλνει ένα IP πακέτο με TTL=1 σε ένα κόμβο παραλήπτη.
- Ο 1ος Router στην διαδρομή μεταξύ αποστολέα-παραλήπτη θα μειώσει το TTL στην τιμή 0, και επομένως θα στείλει ένα μήνυμα «ICMP Time Exceed» στον αποστολέα.
- Ακολούθως, ο αποστολέας θα στείλει πάλι το ίδιο πακέτο στον παραλήπτη με TTL=2, οπότε ο 2ος Router στην διαδρομή μεταξύ αποστολέα-παραλήπτη θα είναι αυτός που θα μηδενίσει το TTL και θα στείλει το μήνυμα «ICMP Time Exceed» στον αποστολέα.
- Η διαδικασία αυτή επαναλαμβάνεται από τον αποστολέα μέχρι το πακέτο να φθάσει στον παραλήπτη (αυξάνοντας κάθε φορά το TTL κατά 1).
- Έτσι, ο αποστολέας μπορεί να μάθει από ποιους Routers πέρασε το datagram.
- Για καλύτερη εκτίμηση των συνολικών χρόνων RTT (Round Trip Time), κάθε φορά το datagram στέλνεται εις τριπλούν.

Interplay between routing and forwarding

Να θυμάσαι!!!



Ειδικώτερα: Datagram forwarding table

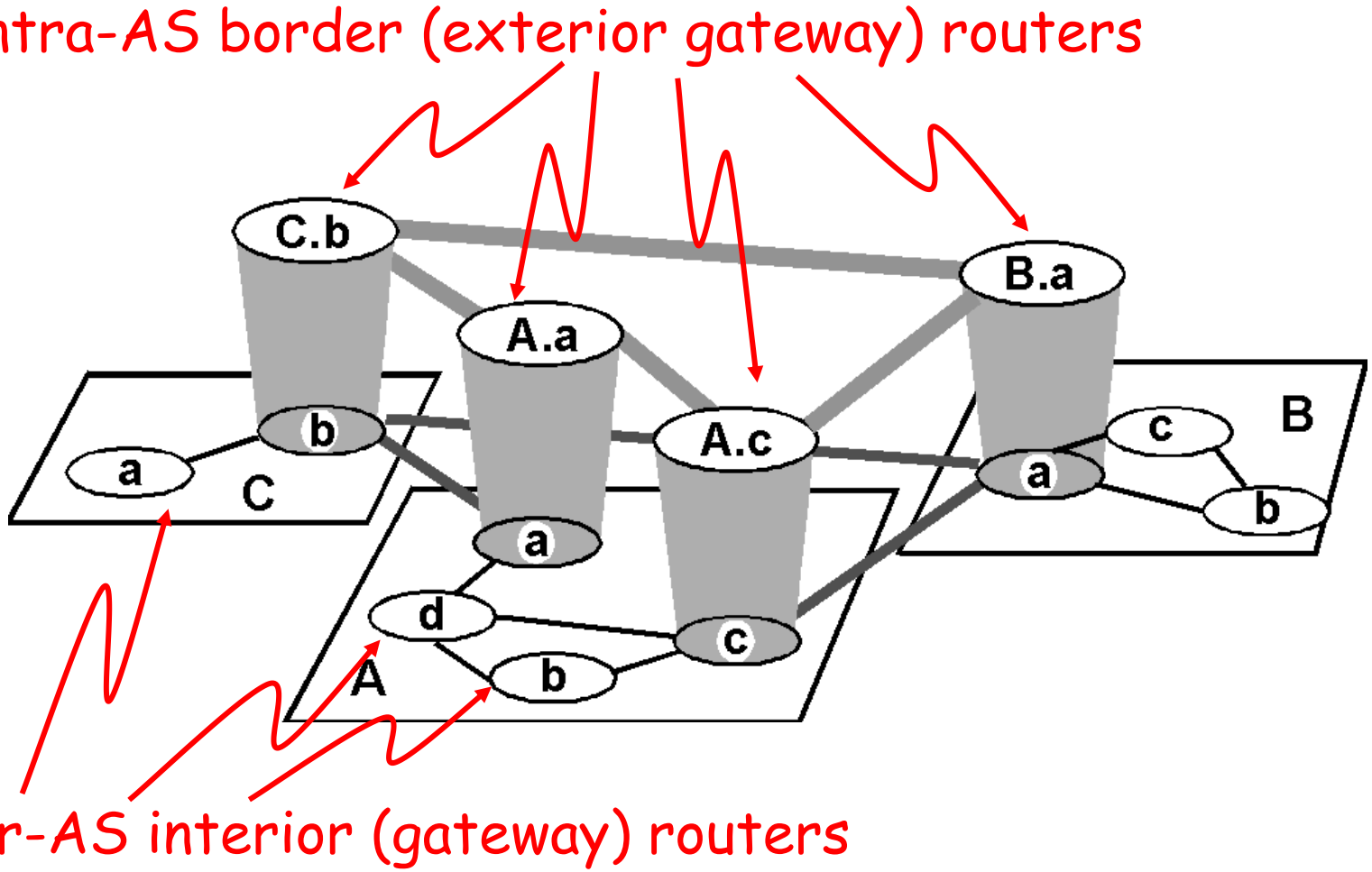


Routing in the Internet

- 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 Routing

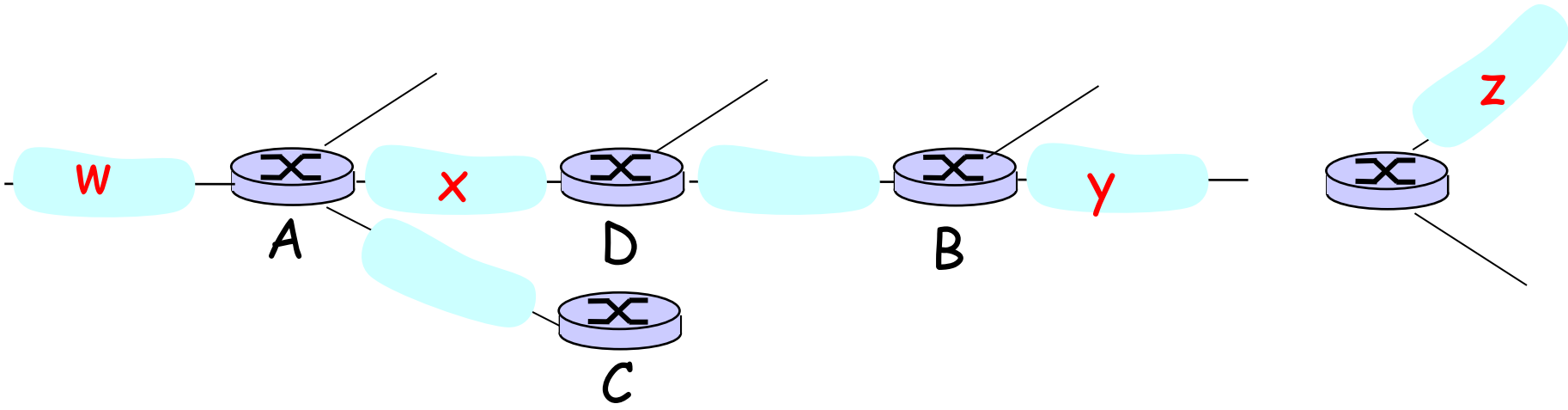
- Also known as Interior Gateway Protocols (IGP)
- Most common IGPs:
  - RIP: Routing Information Protocol
  - OSPF: Open Shortest Path First
  - IGRP: Interior Gateway Routing Protocol (Cisco propr.)

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

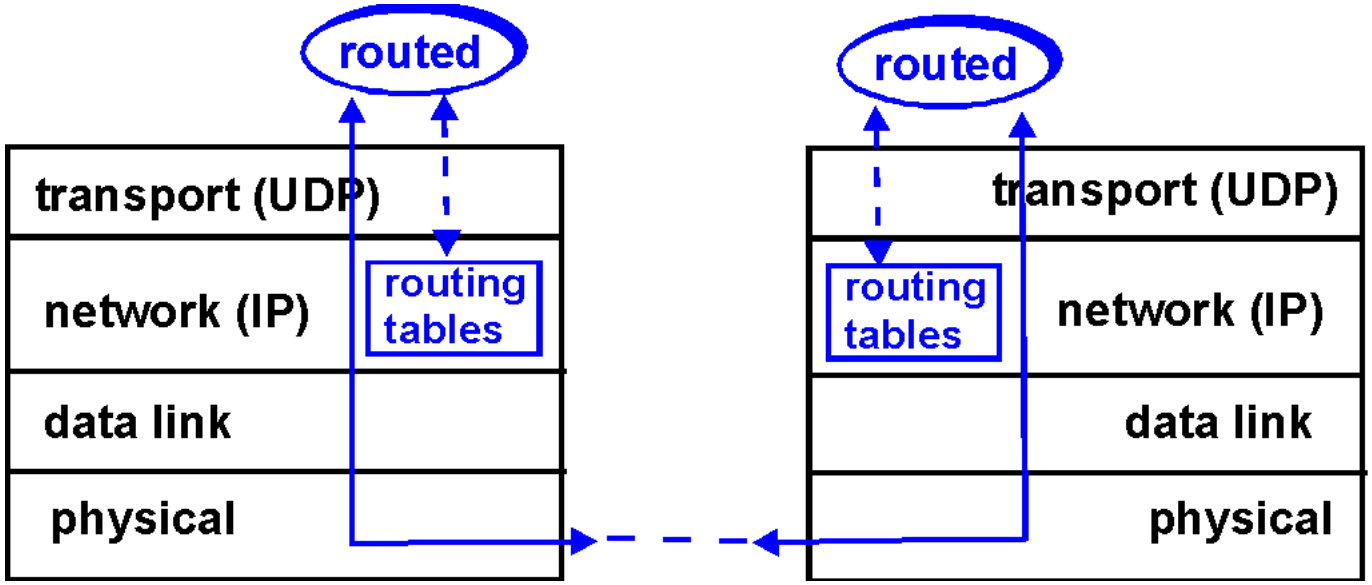
RIP: Link Failure and Recovery

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 (αθώα ψέματα) used to prevent ping-pong loops (**infinite distance = 16 hops**)

RIP Table processing

- RIP routing tables managed by a**pplication-level** process called **route-d (daemon)**
- advertisements sent in **UDP** packets, periodically repeated (30 sec)



RIP Table example (continued)

Router: *giroflee.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 (class D)
- Loopback interface (for debugging) – 127.0.0.1

OSPF (Open Shortest Path First)

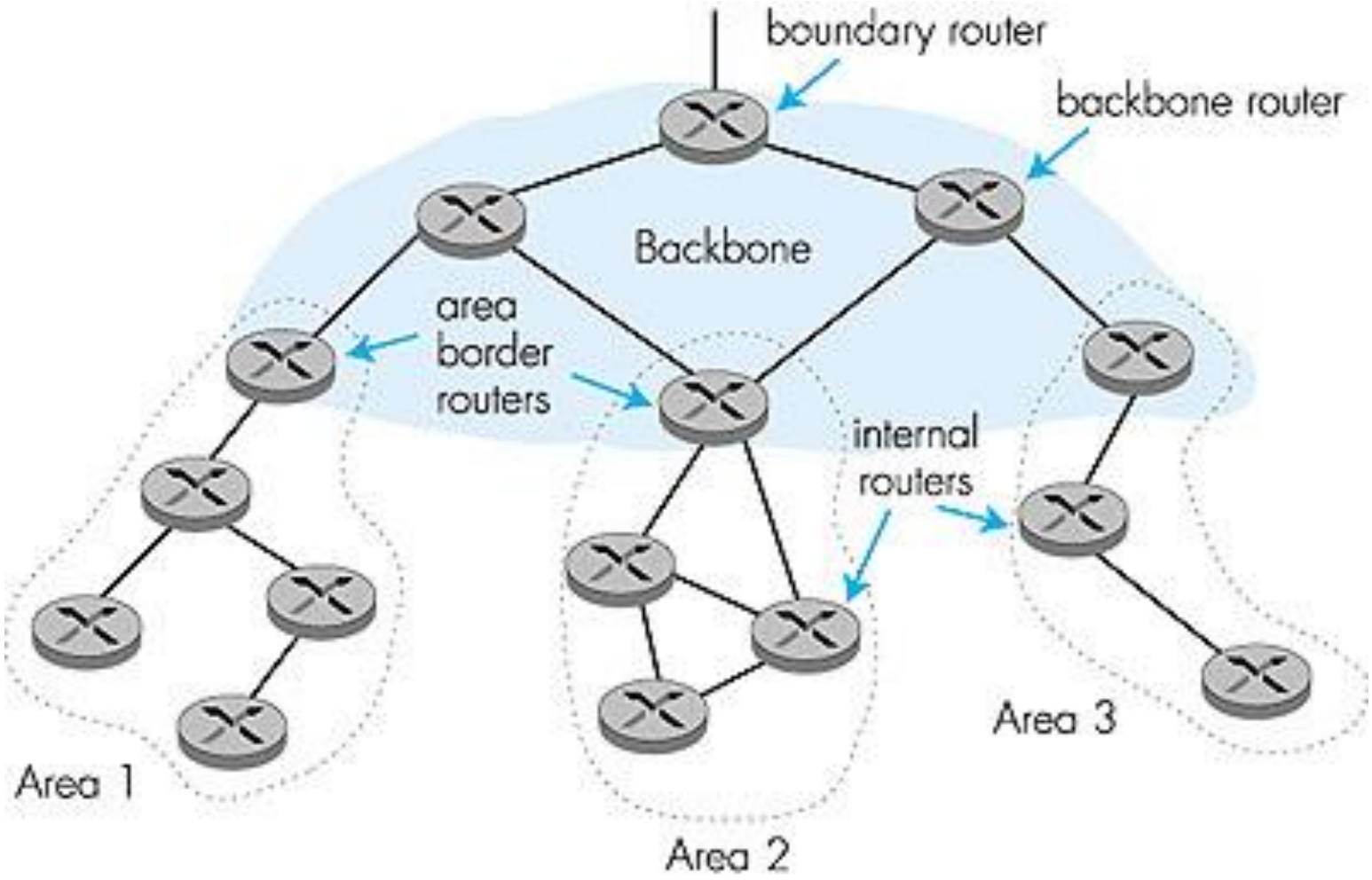
- "open": publicly available
- Uses Link State algorithm
  - LS packet dissemination
  - Topology map at each node
  - Route computation using Dijkstra's algorithm

- 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



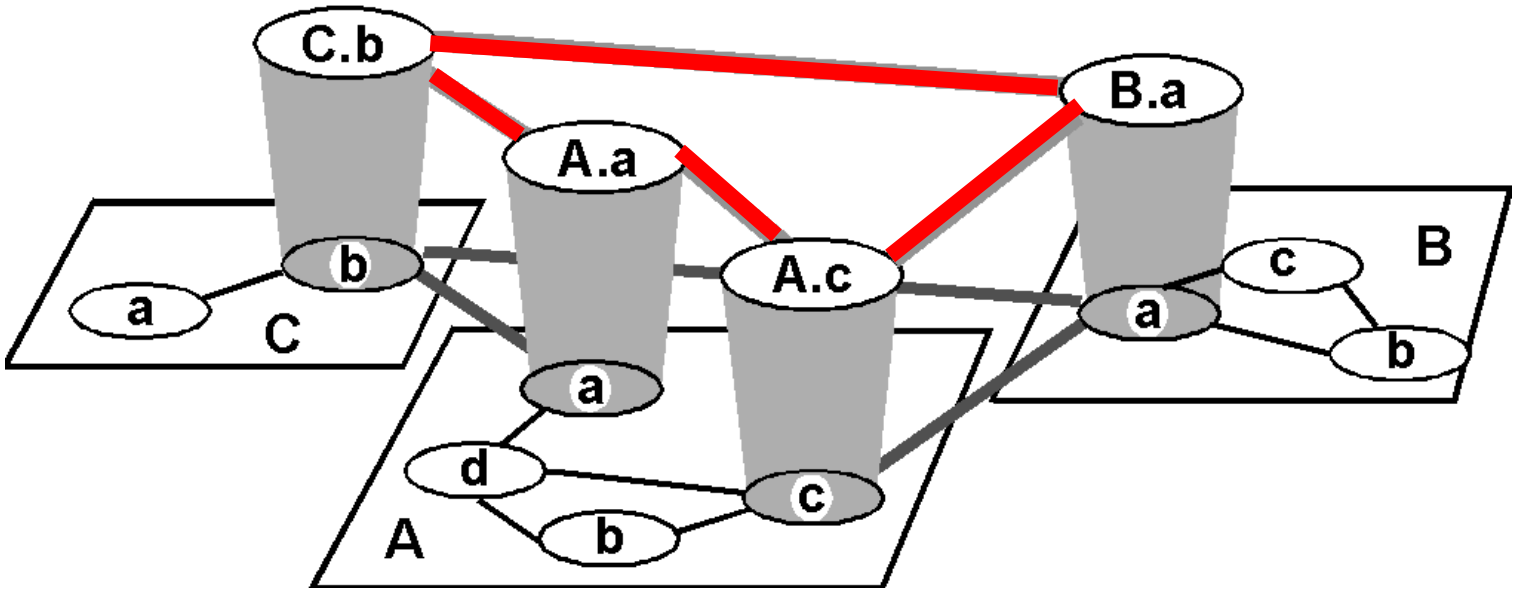
Hierarchical OSPF

- 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*

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:

Path (X,Z) = X,Y1,Y2,Y3,…,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:

  Path (W,Z) = w, Path (X,Z)
- Note: X can control incoming traffic by controling it 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

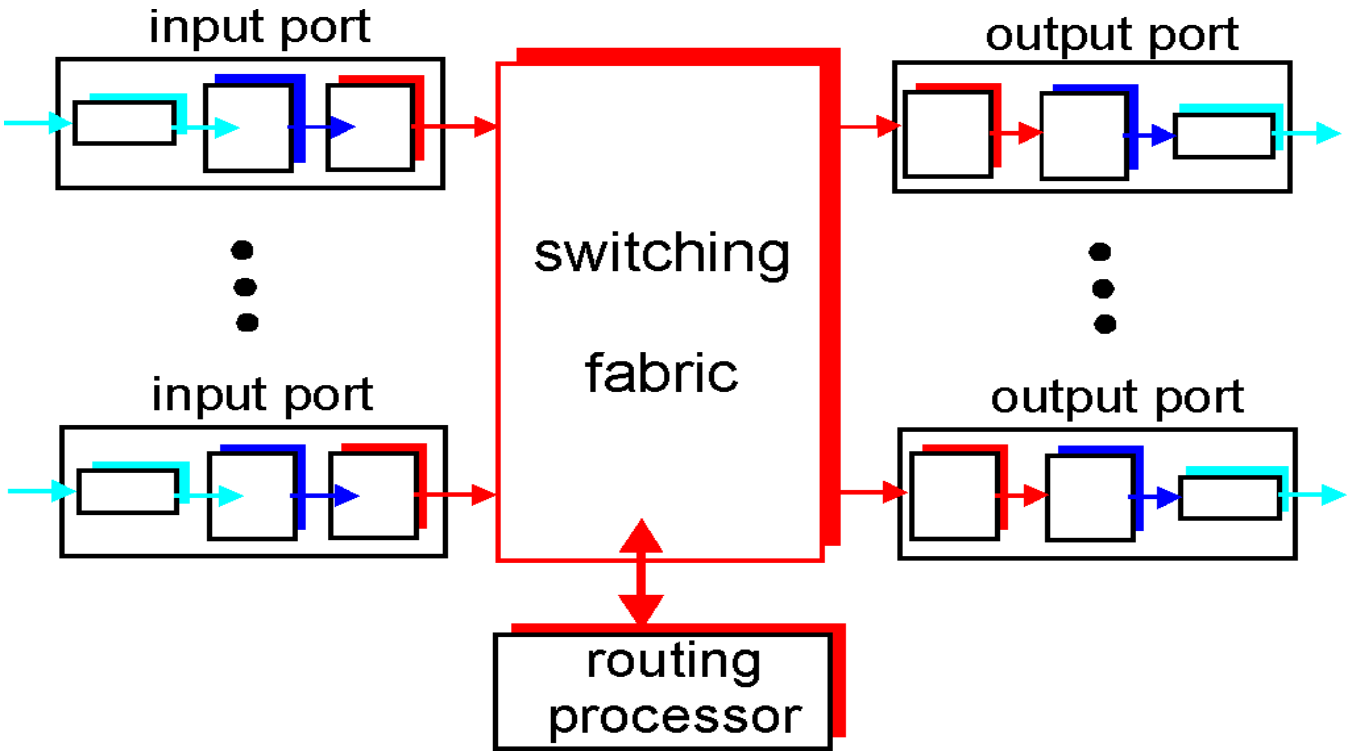
Performance:

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

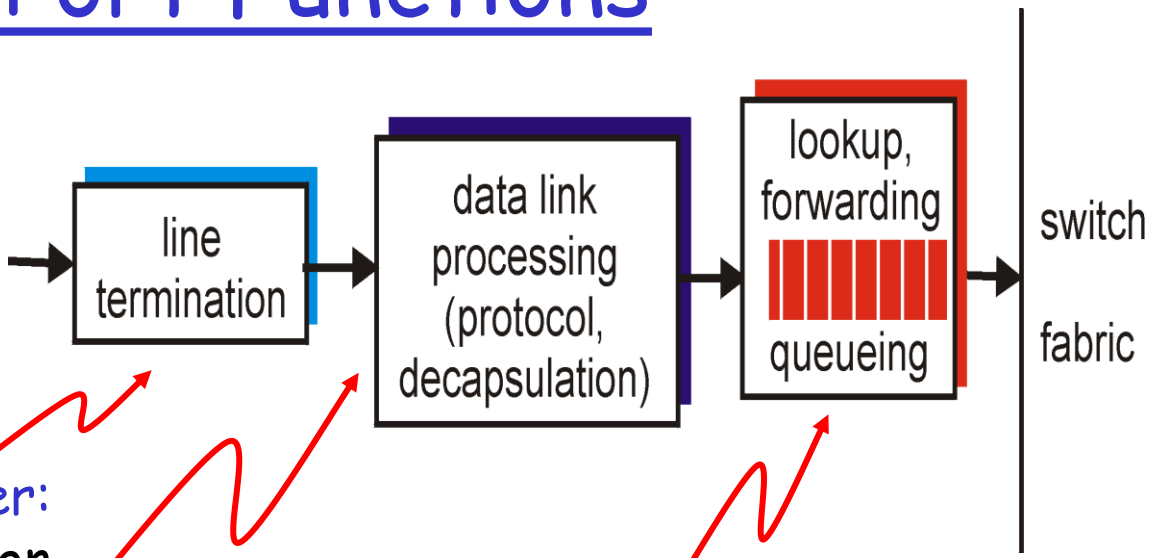
Router Architecture Overview

Two key router functions:

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



Input Port Functions



Physical layer:
bit-level reception

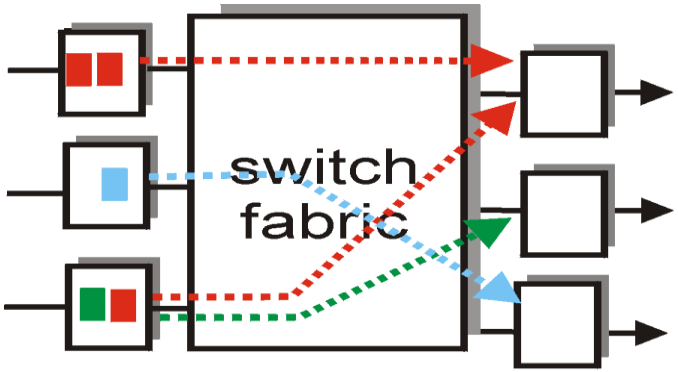
Data link layer:
e.g., Ethernet
see chapter 5

Decentralized switching:

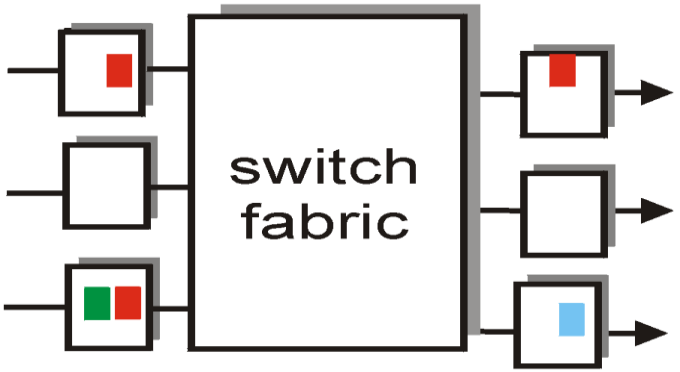
- given datagram dest., lookup output port using routing table in input port memory
- goal: complete input port processing at 'line speed'
- queuing: if datagrams arrive faster than forwarding rate into switch fabric

Input Port Queuing

- Fabric slower that 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

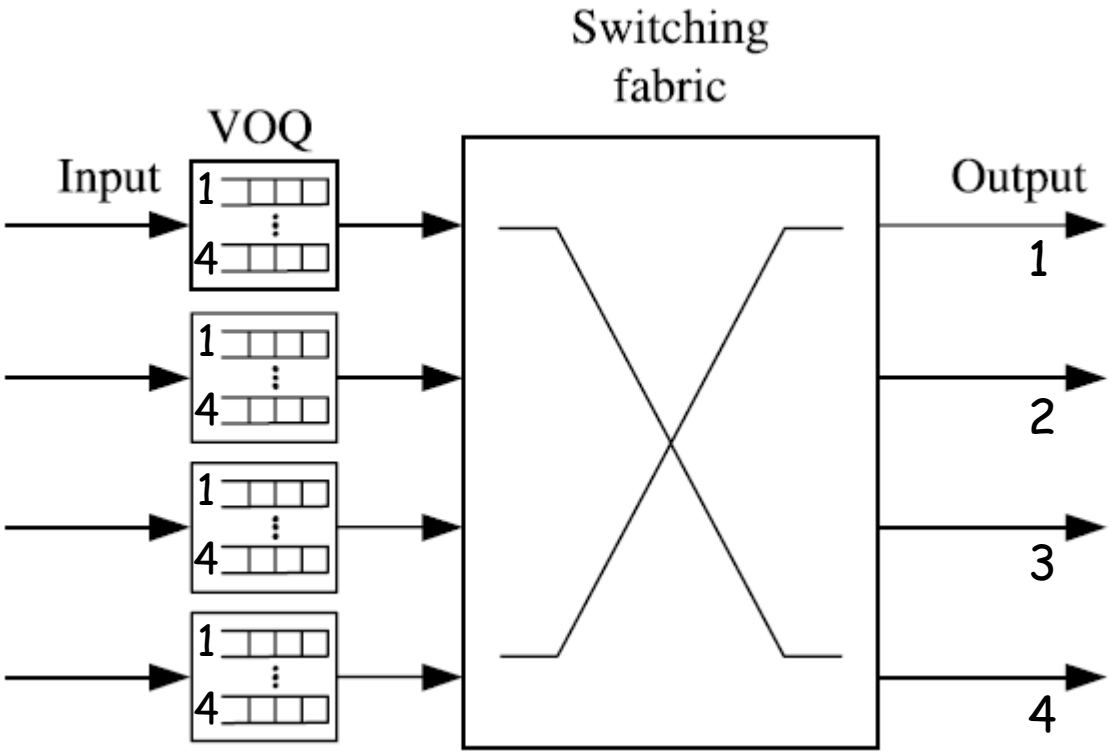


green packet experiences HOL blocking

Virtual Output Queues

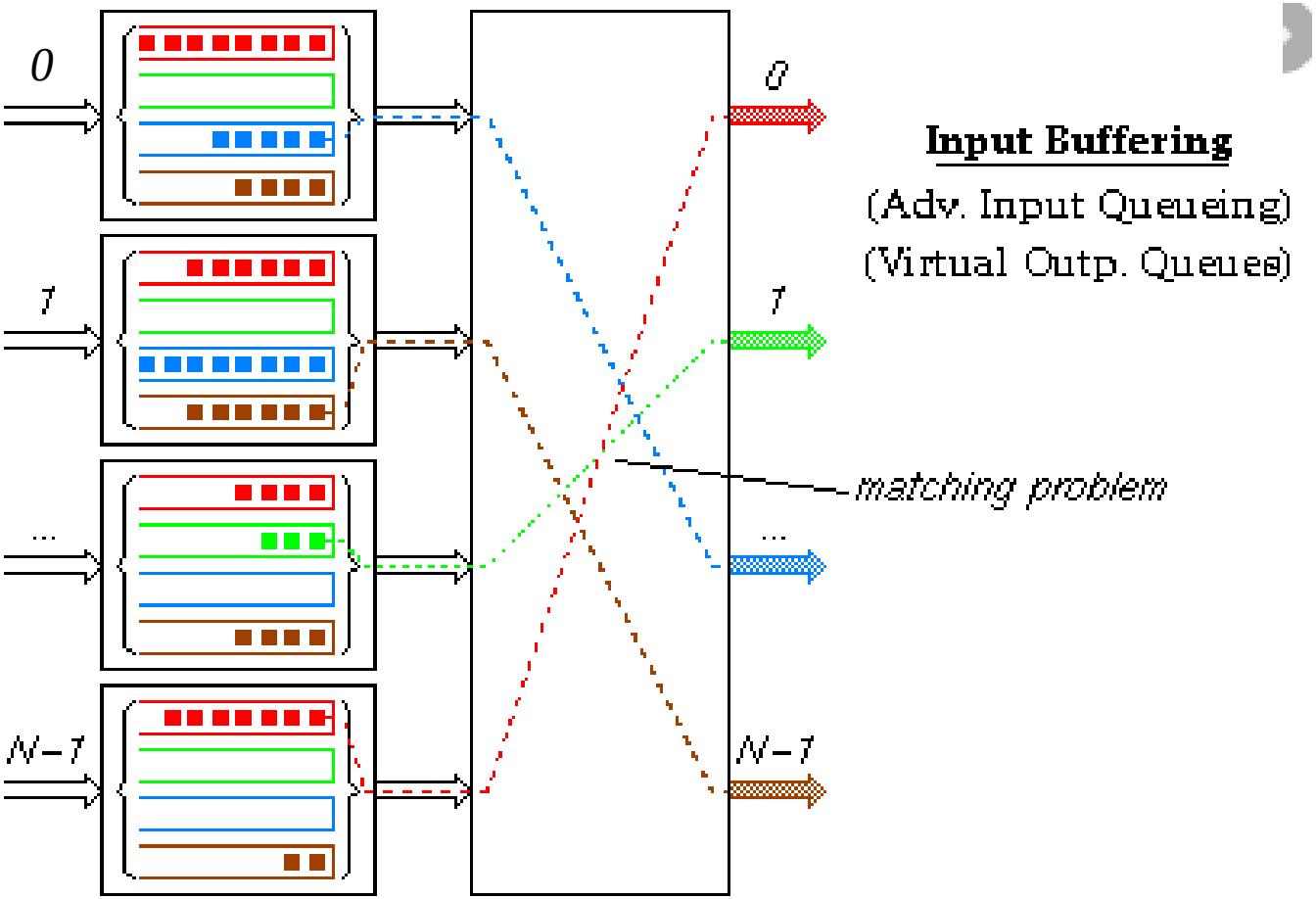
VOQ: Λύση στο πρόβλημα του HOL

Ανάλογο: χώρος στα αυτοκίνητα για να στρίβουν δεξιά

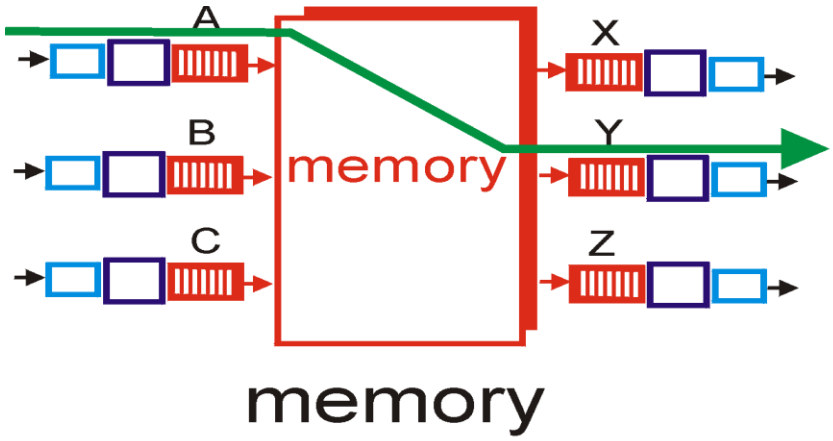


Virtual Output Queues

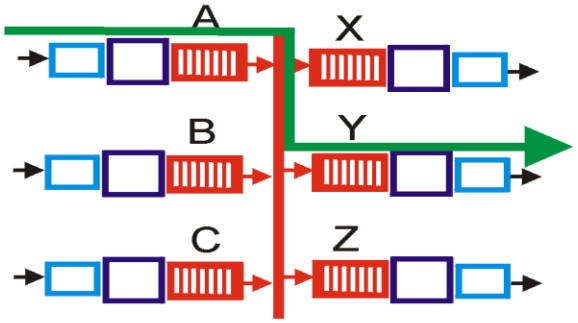
Η λύση στο πρόβλημα HOL blocking



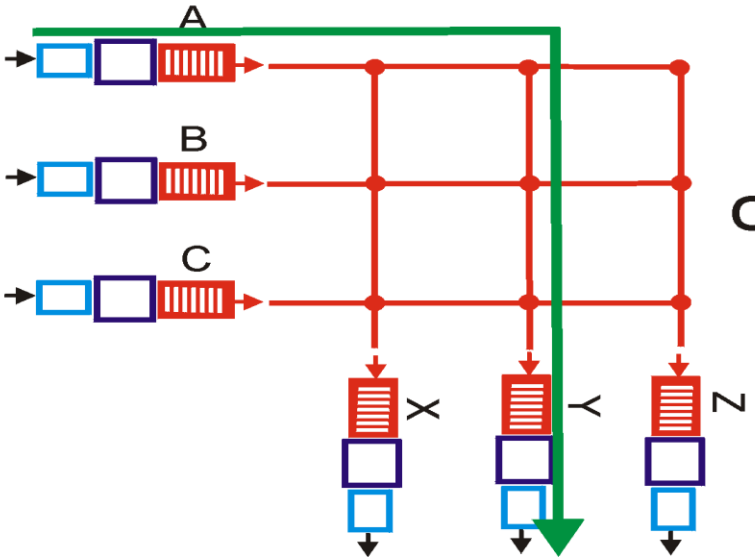
Three types of switching fabrics



memory



bus

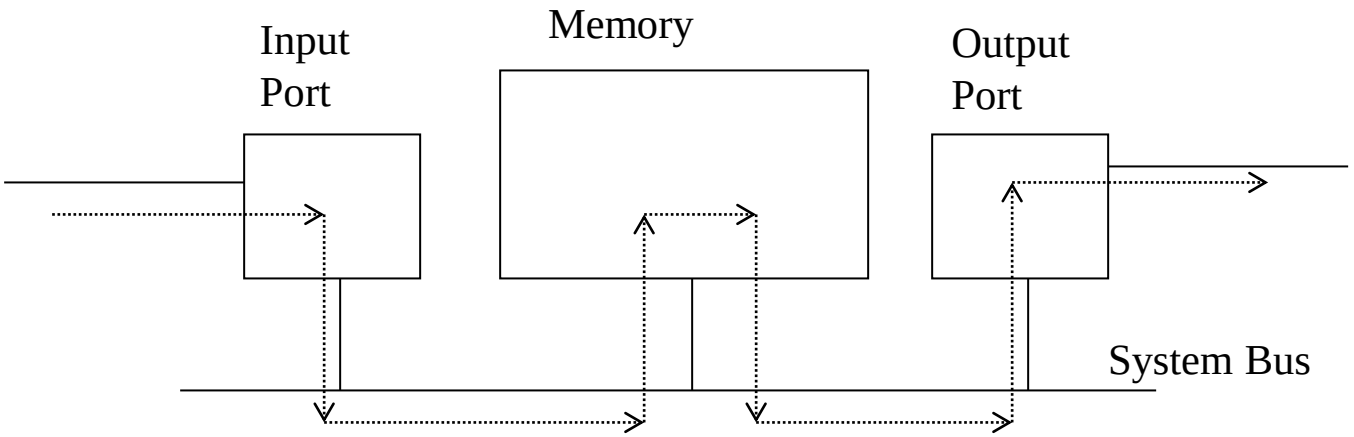


crossbar

Switching Via Memory

First generation routers:

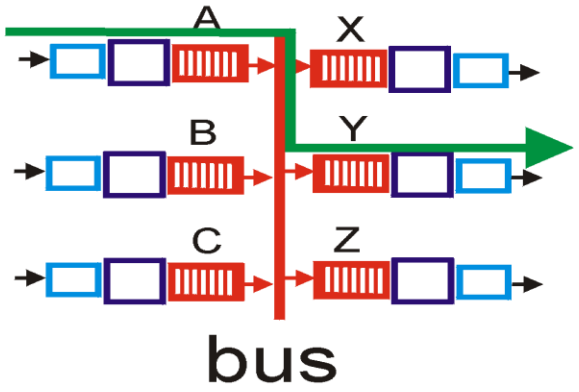
- packet copied by system's (single) CPU
- speed limited by memory bandwidth (2 bus crossings per datagram)



Modern routers:

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

Switching Via Bus

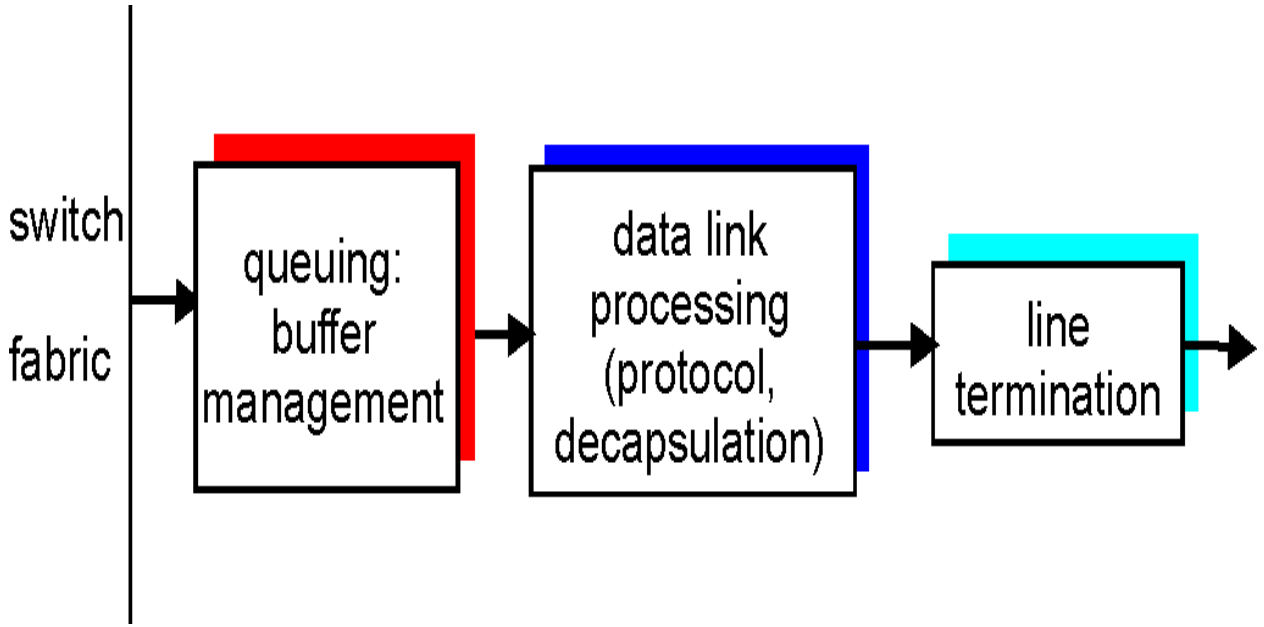


- datagram from input port memory to output port memory via a shared bus
- bus contention: switching speed limited by bus bandwidth
- 1 Gbps bus, Cisco 1900: sufficient speed for access and enterprise routers (not regional or backbone)

Switching Via An Interconnection Network

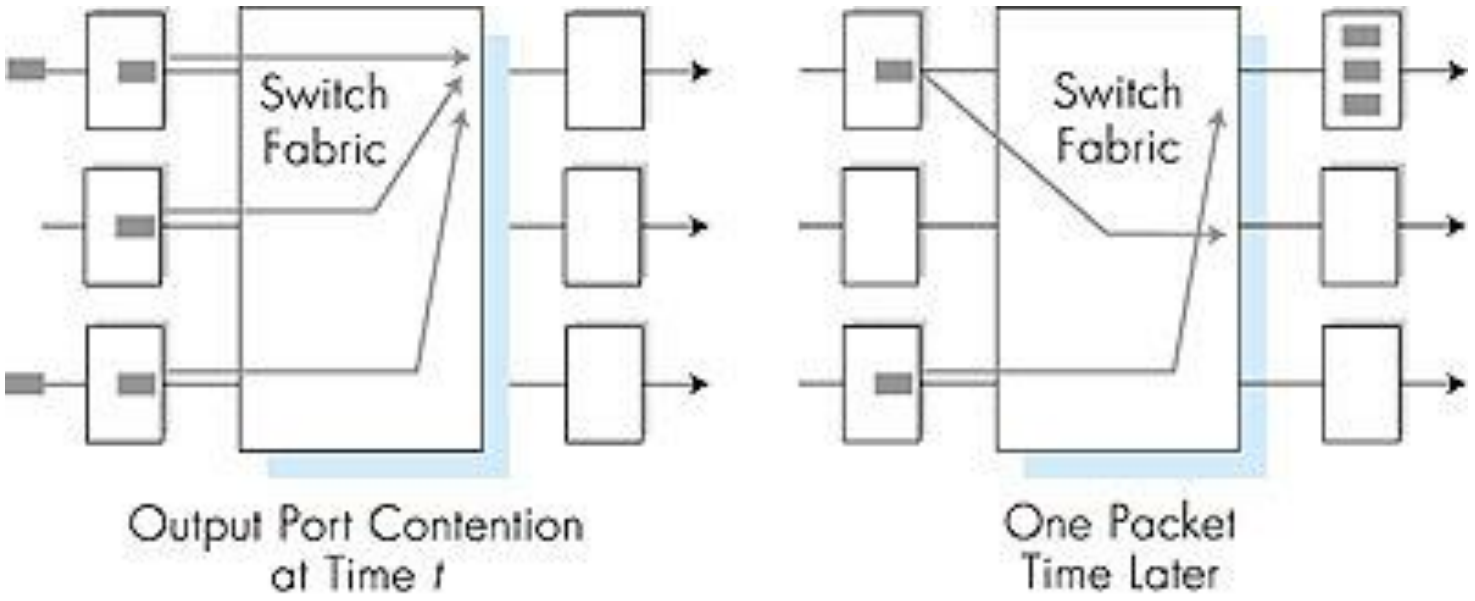
- overcome bus bandwidth limitations
- Banyan networks, other interconnection nets initially developed to connect processors in multiprocessor
- Advanced design: fragmenting datagram into fixed length cells, switch cells through the fabric.
- Cisco 12000: switches Gbps through the interconnection network

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



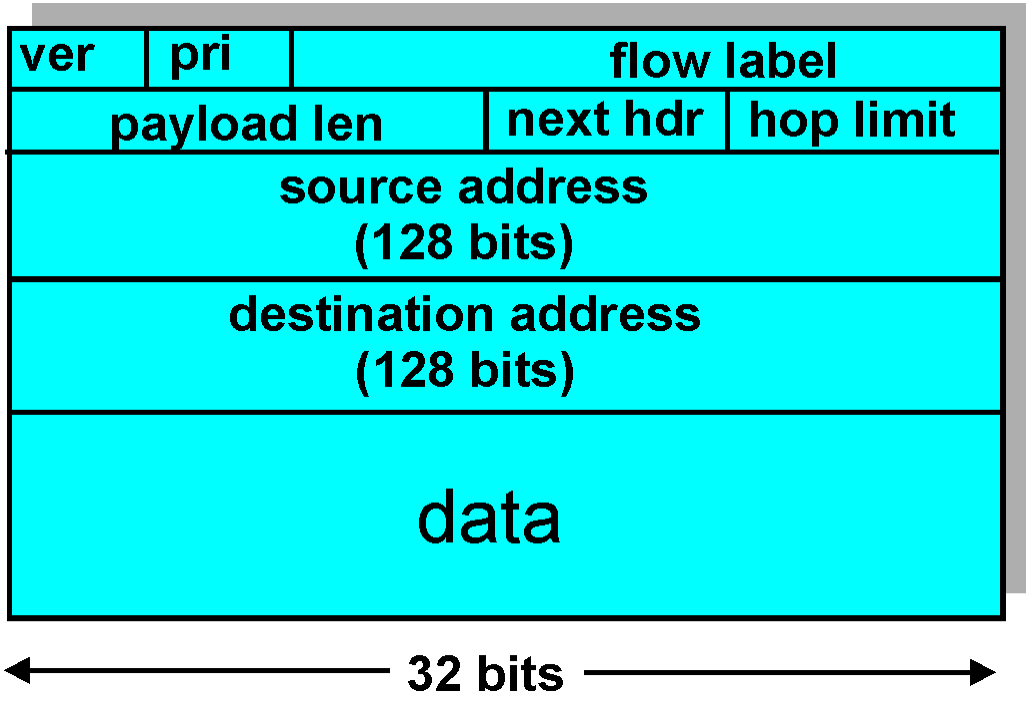
- buffering when arrival rate via switch exceeeds ouput line speed
- *queueing (delay) and loss due to output port buffer overflow!*

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



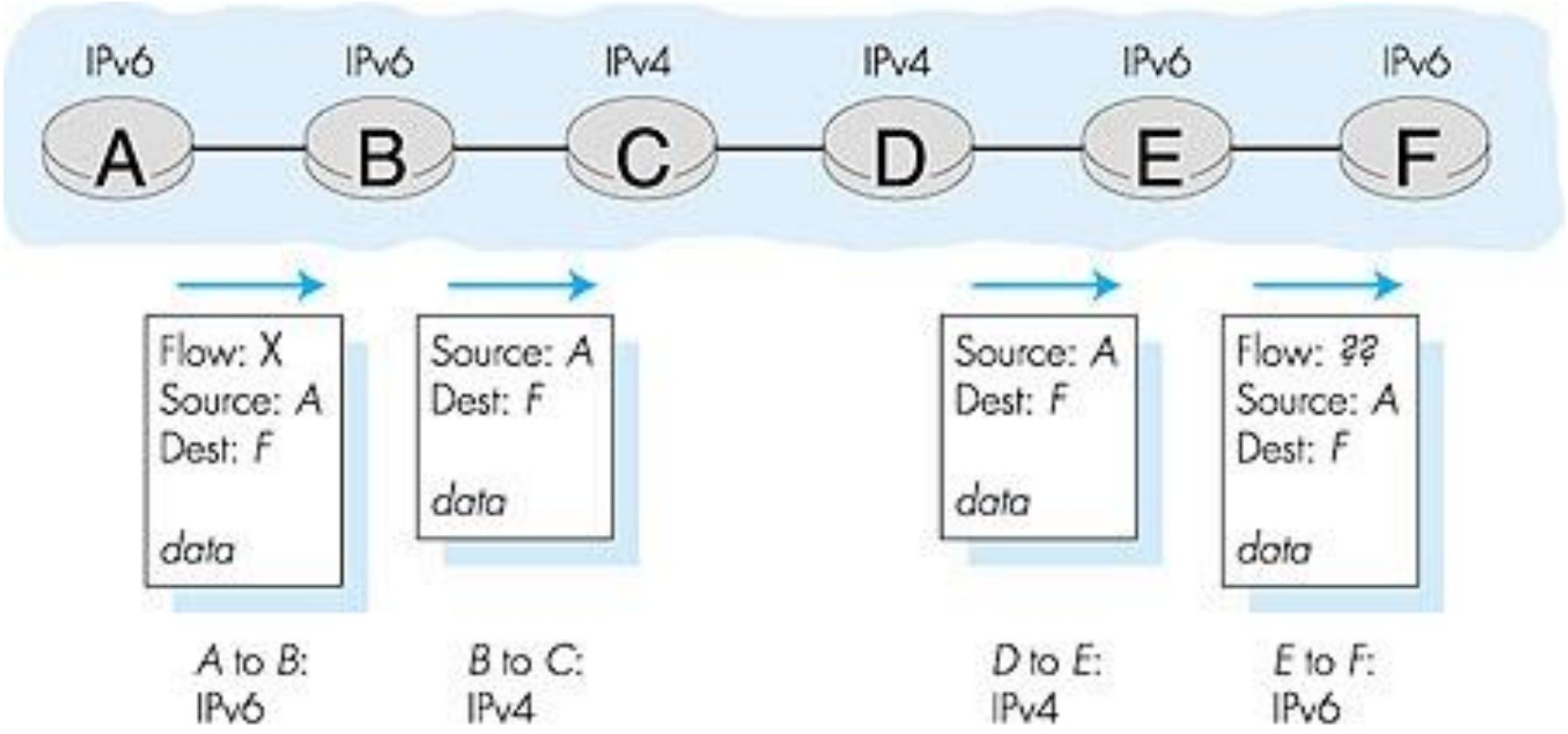
Other Changes from IPv4

- *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

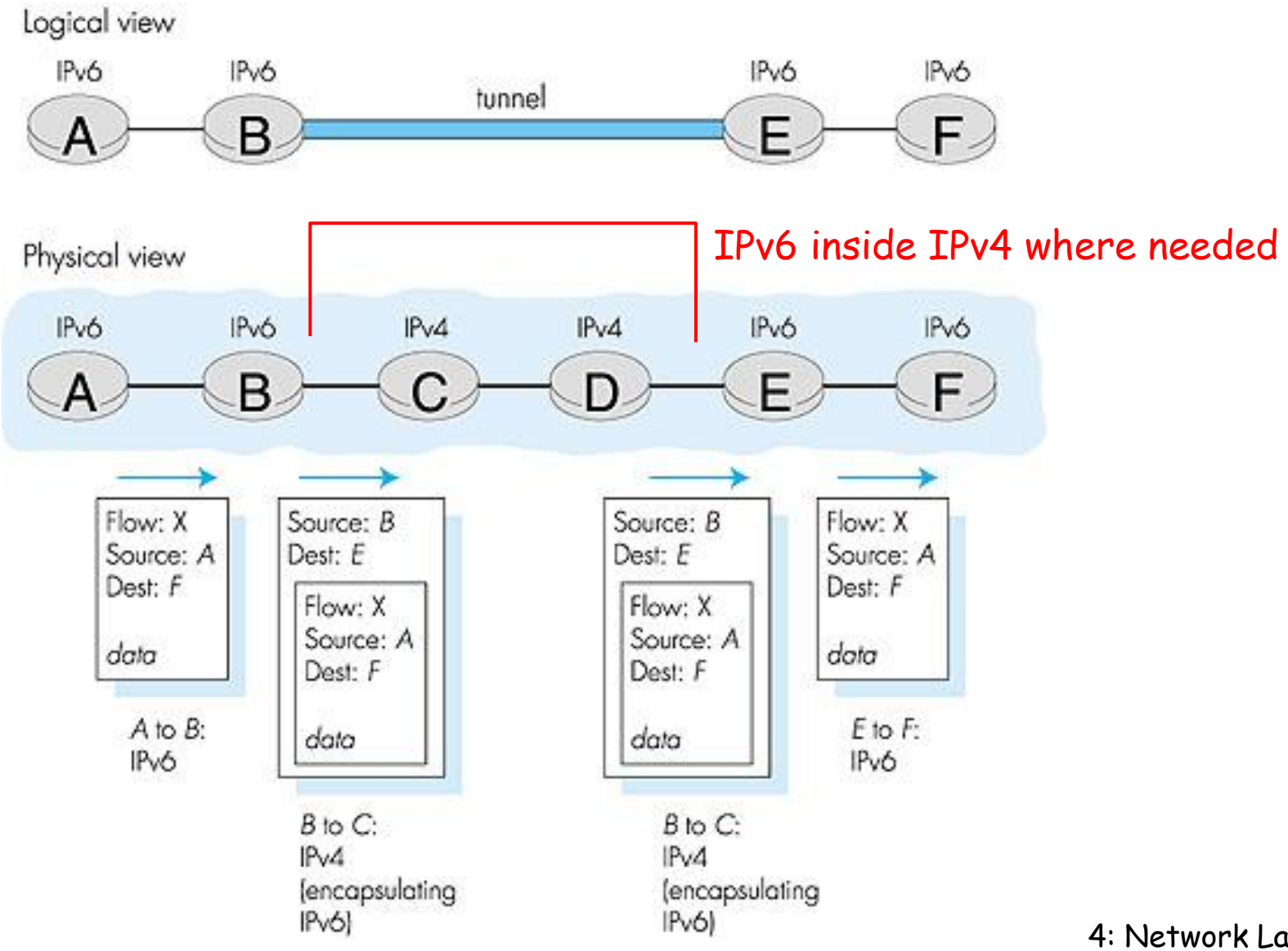
Transition From IPv4 To IPv6

- Not all routers can be upgraded simultaneous
  - no "flag days"
  - How will the network operatewith 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 n IPv4 datagram among IPv4 routers

Dual Stack Approach



Tunneling


Logical view
IPv6
IPv6
tunnel
IPv6
IPv6
A
B
E
F
Physical view
IPv6 inside IPv4 where needed
IPv6
IPv6
IPv4
IPv4
IPv6
IPv6
A
B
C
D
E
F
Flow: X
Source: A
Dest: F
data
A to B:
IPv6
Source: B
Dest: E
Flow: X
Source: A
Dest: F
data
B to C:
IPv4
(encapsulating
IPv6)
Source: B
Dest: E
Flow: X
Source: A
Dest: F
data
B to C:
IPv4
(encapsulating
IPv6)
Flow: X
Source: A
Dest: F
data
E to F:
IPv6