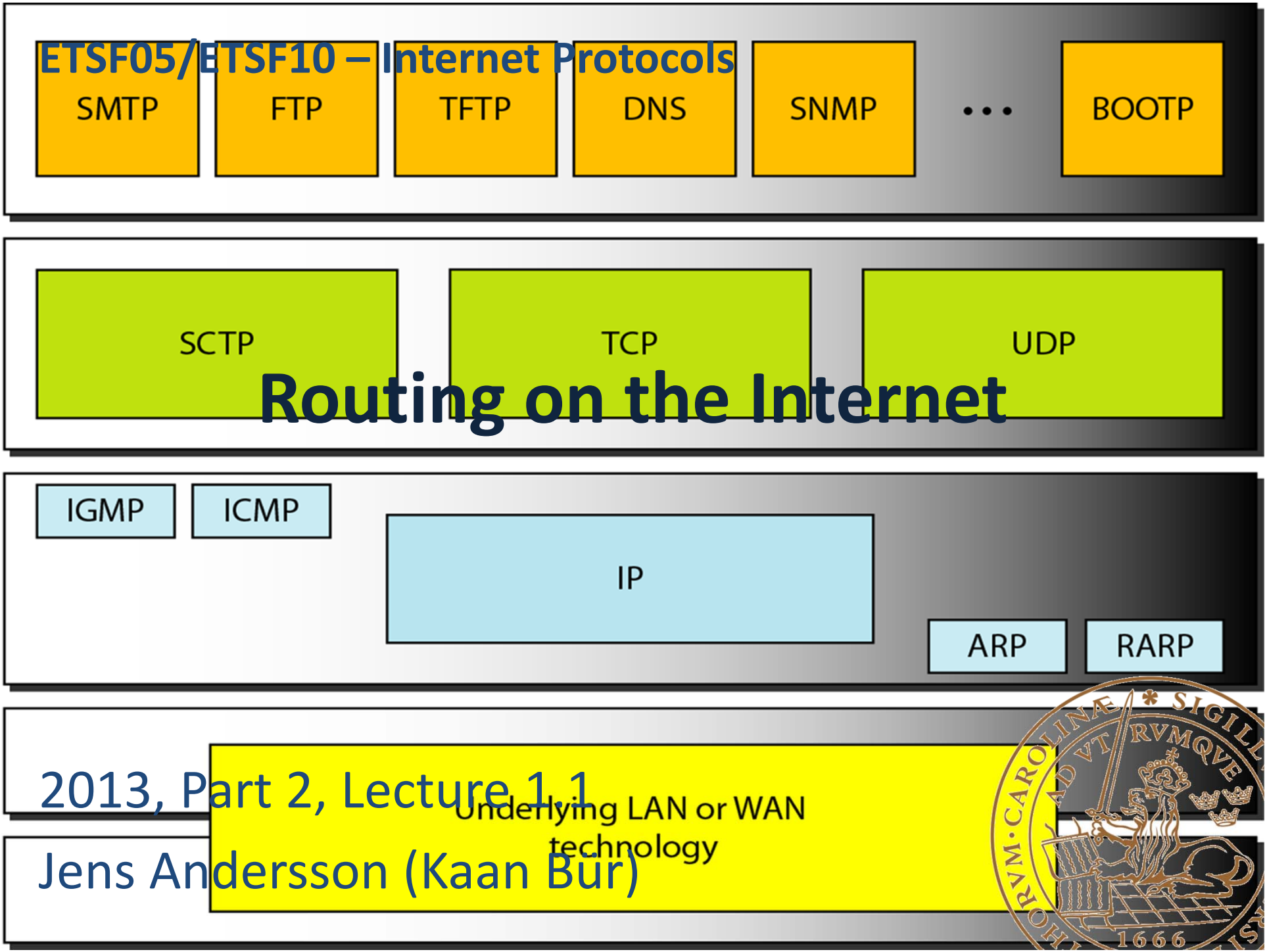
ETSF05/ETSF10 – Internet Protocols



Routing on the Internet

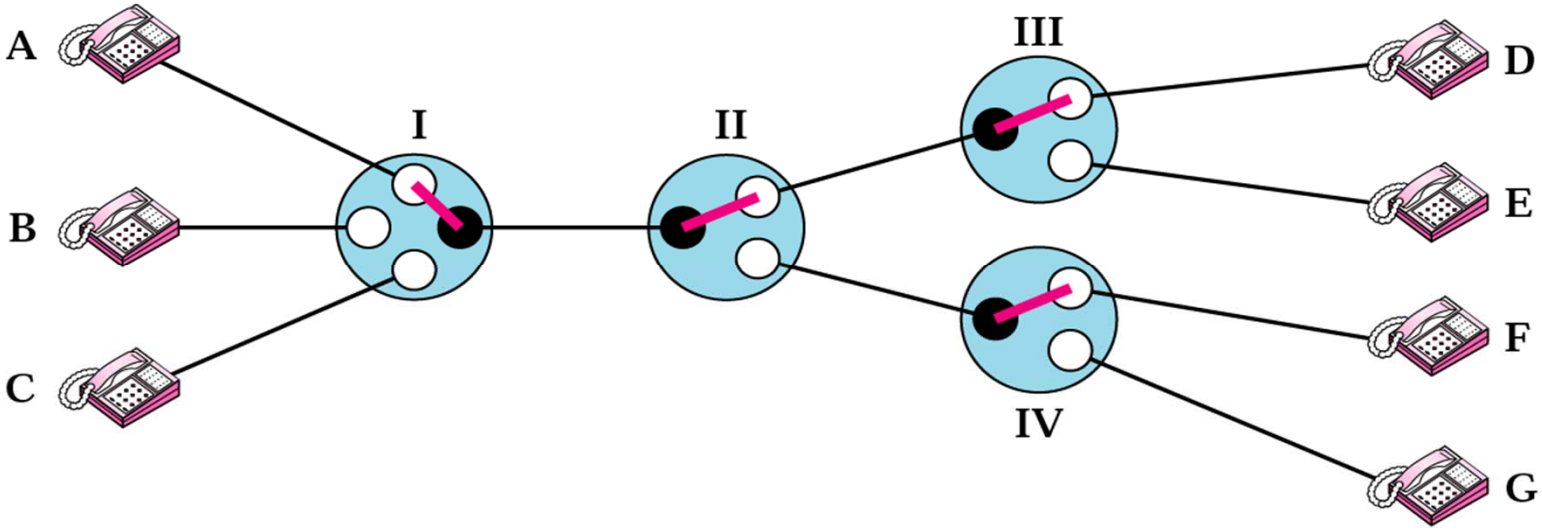
2013, Part 2, Lecture 1.1

Jens Andersson (Kaan Bür)

Routing on the Internet

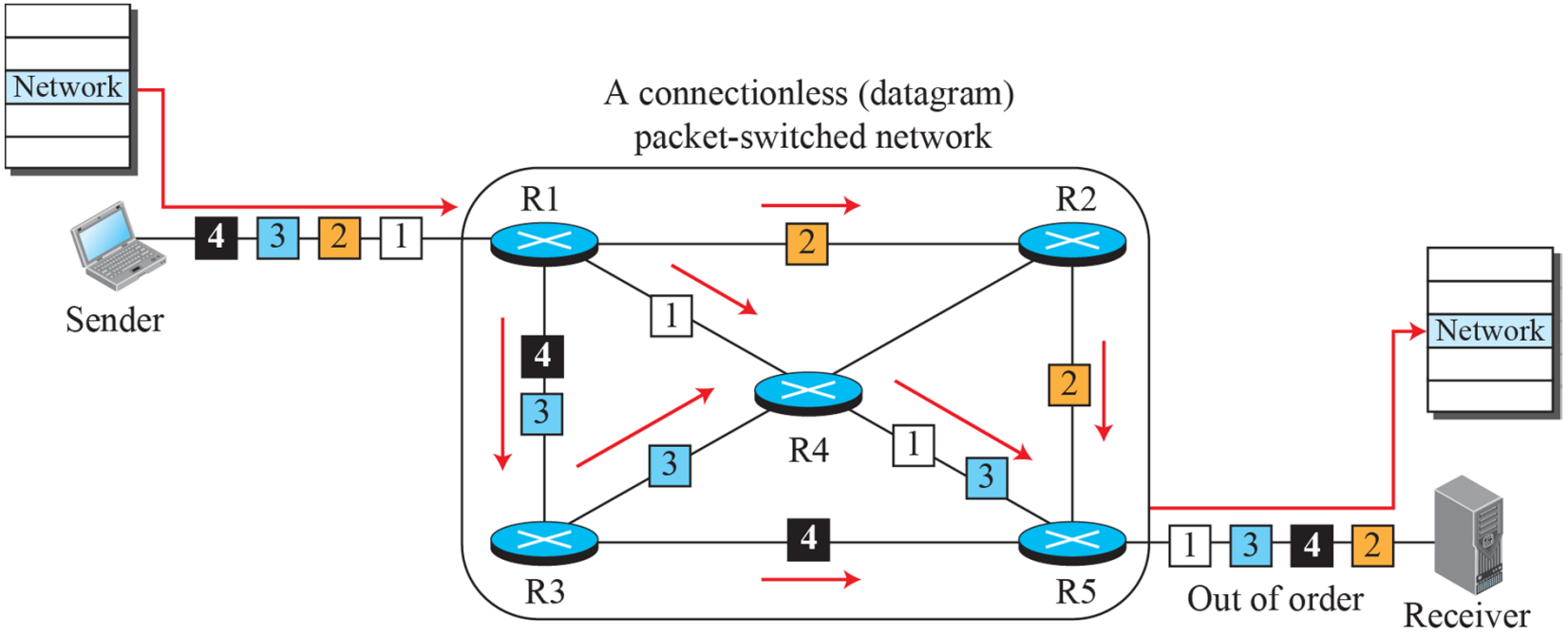
- Router architecture
  [ed.5 ch.8.4.2]
- Routing concepts and algorithms
  [ed.5 ch.20.1-2]
- Unicast routing protocols (part 1)
  [ed.5 ch.20.3]

Circuit switched routing



Packet-switched Routing

- Choosing an optimal path
  - According to a cost metric
  - Decentralised: each router has full information

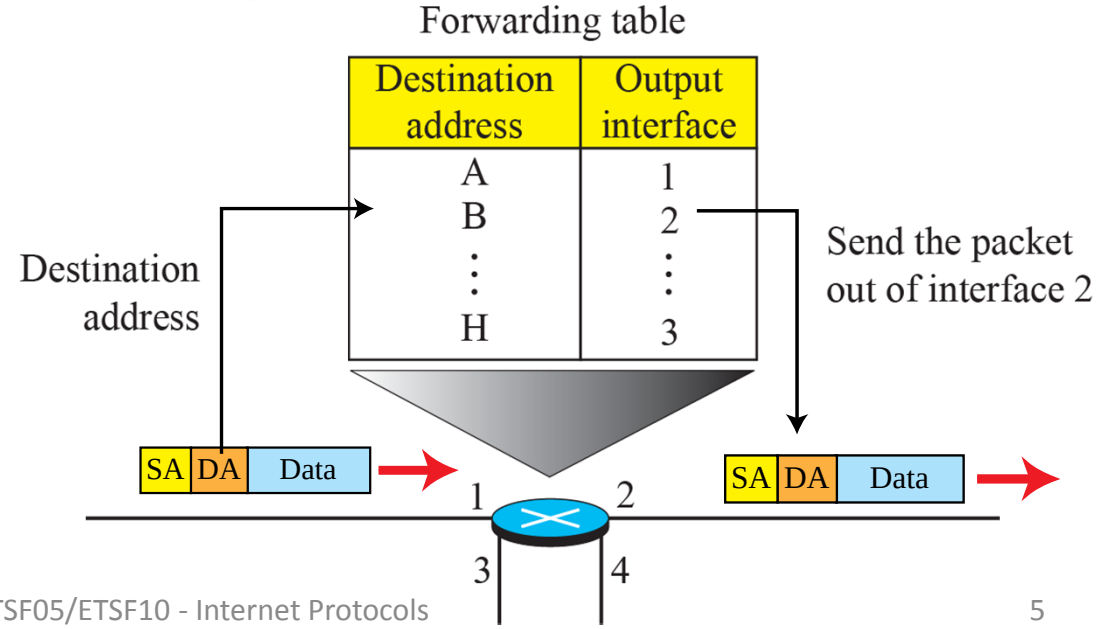


Router

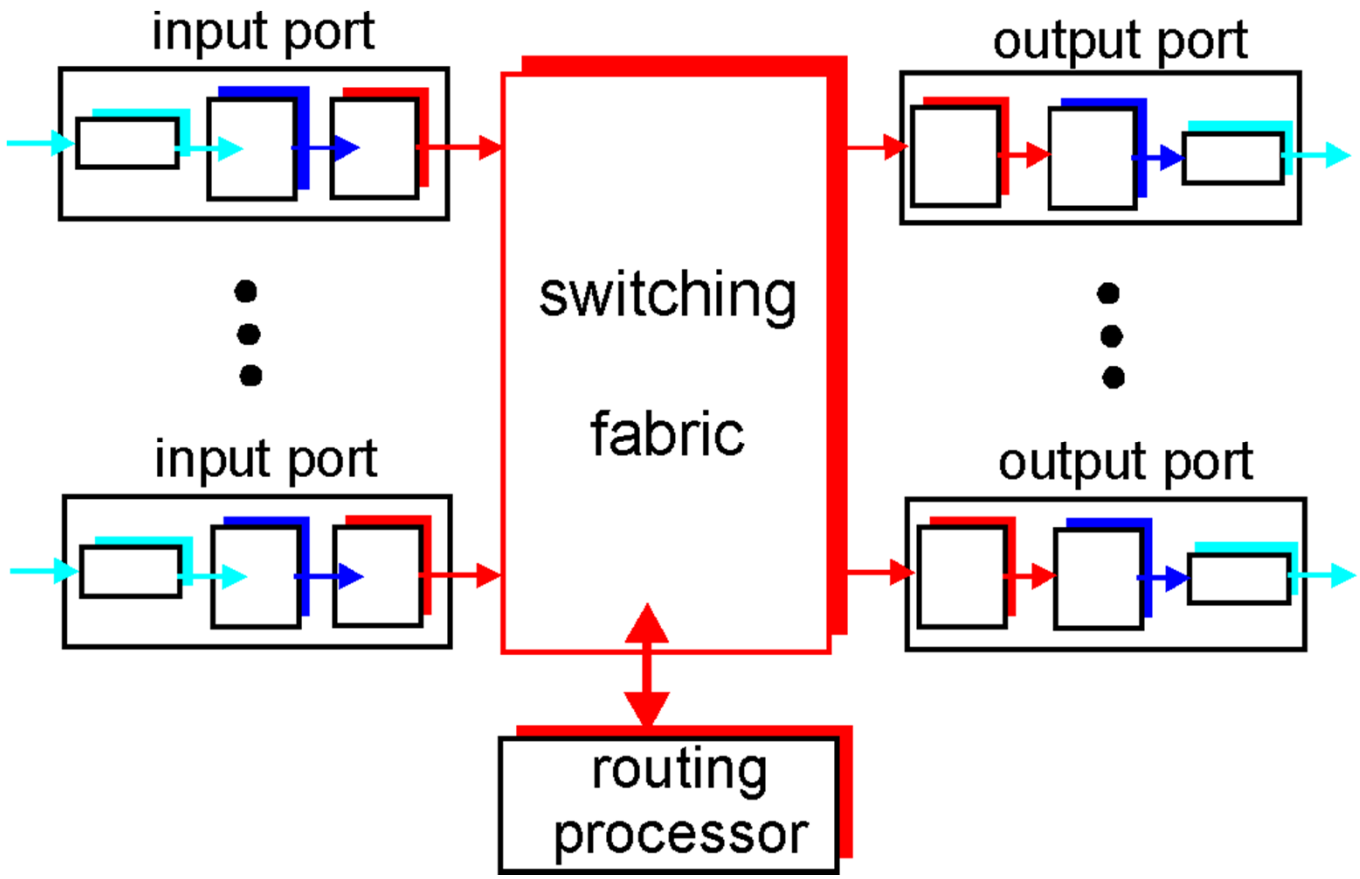
- Internetworking device
  - Passes data packets between networks
  - Checks ***Network Layer*** addresses
  - Uses Routing/forwarding tables

Two functions:

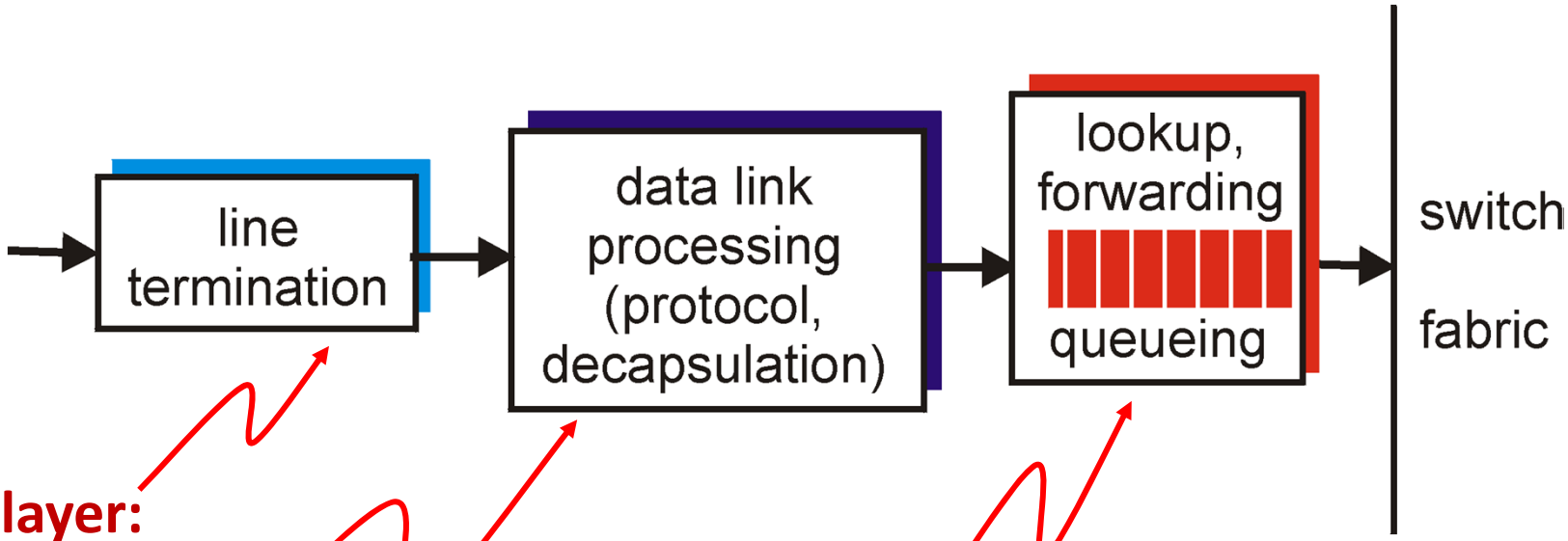
1. Routing
2. Forwarding



Router Architecture Overview



Input Port



Physical layer:
bit-level reception

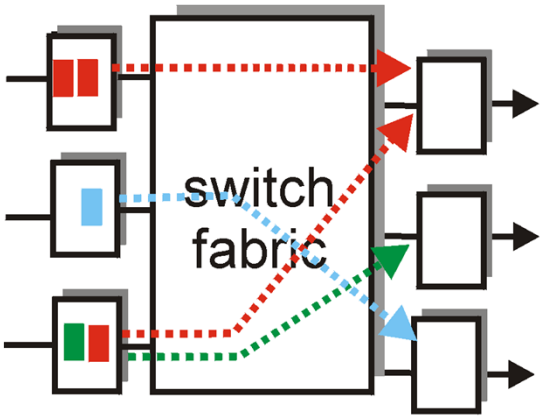
Data link layer:
e.g., Ethernet

Decentralized switching:

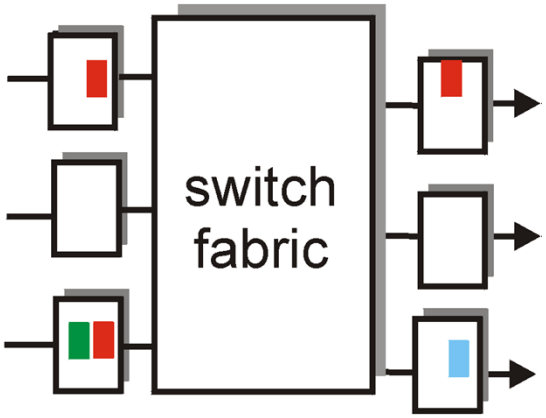
- Given destination, lookup output port using routing table in input port memory
- Goal: complete input port processing at 'line speed'

Input Port Queuing

- Fabric slower that sum of input ports → queuing
- Head-of-the-Line (HOL) blocking: Datagram at front of queue prevents others in queue from proceeding
- Delay and loss due to input buffer overflow

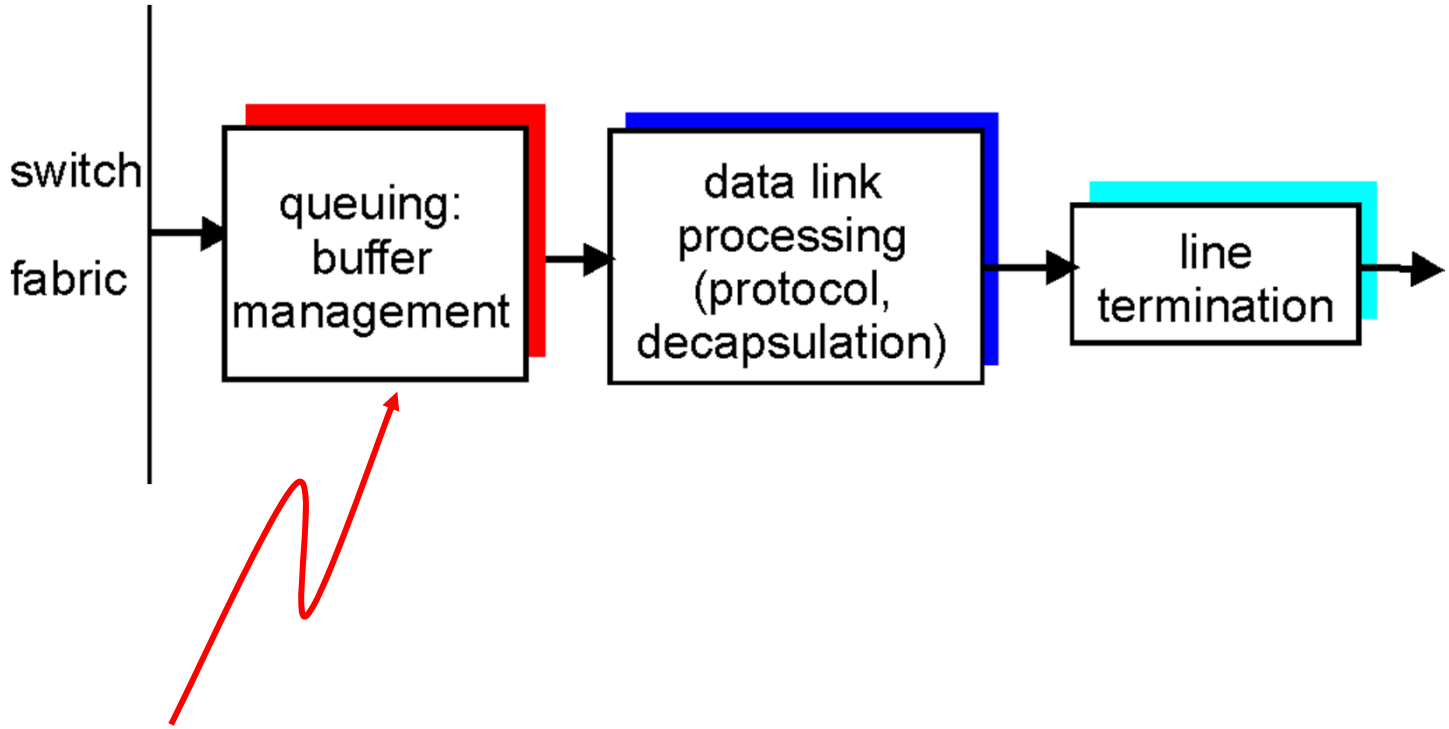


output port contention at time t - only one red packet can be transferred



green packet experiences HOL blocking

Output Port

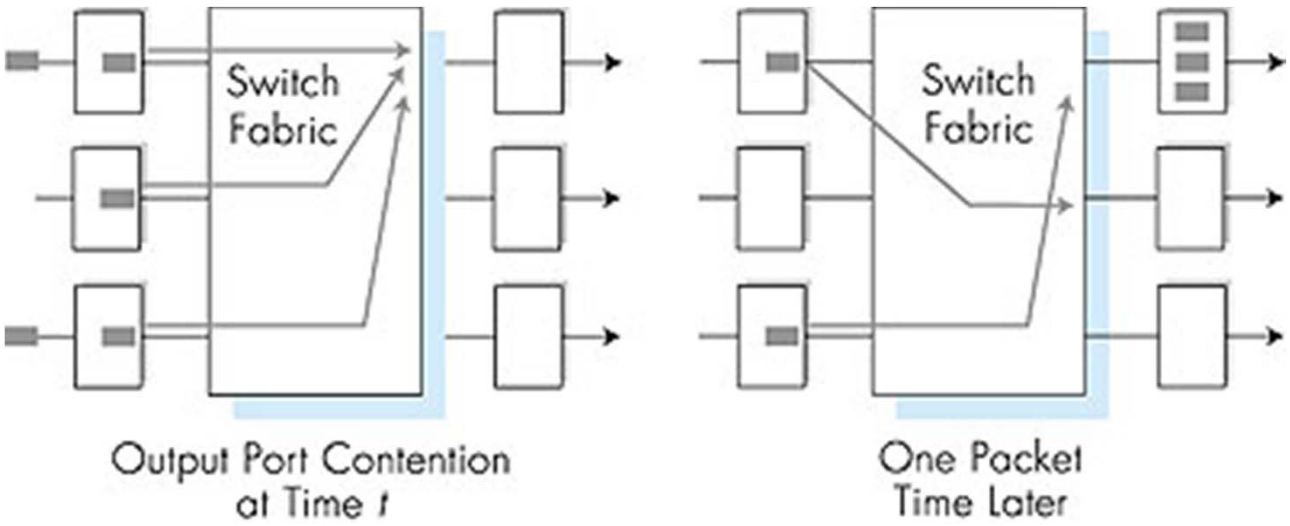


Priority Scheduling:

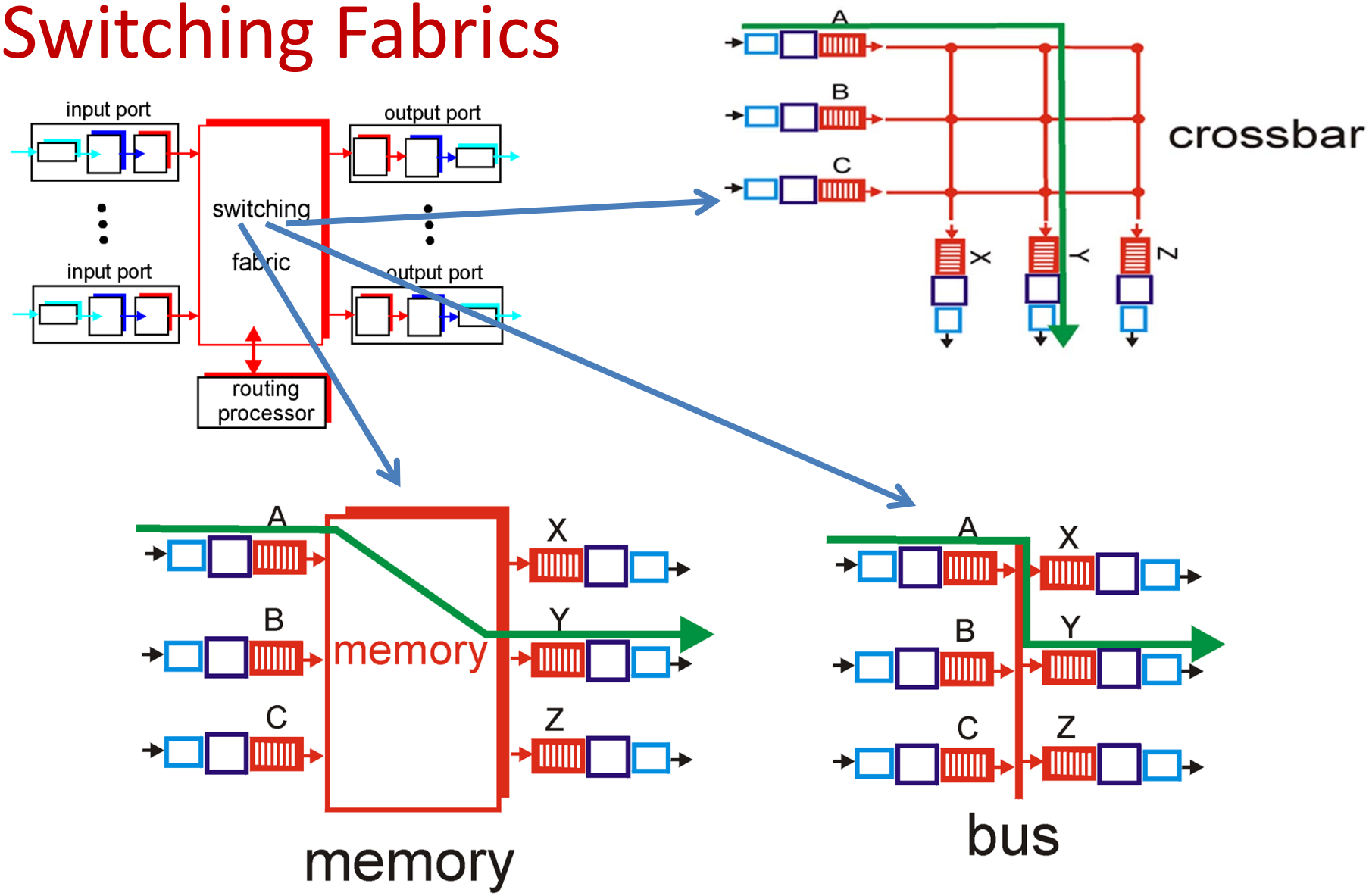
- Scheduling discipline may choose among queued datagrams for transmission

Output Port Queuing

- Datagrams' arrival rate through the switch exceeds the transmission rate of the output line → buffering
- Delay and loss due to output port buffer overflow

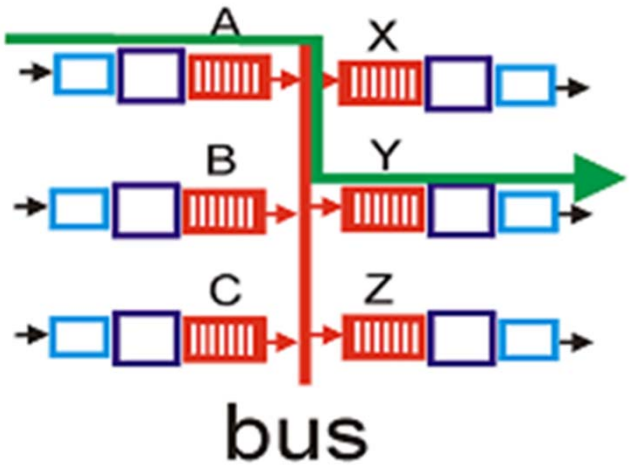


Switching Fabrics



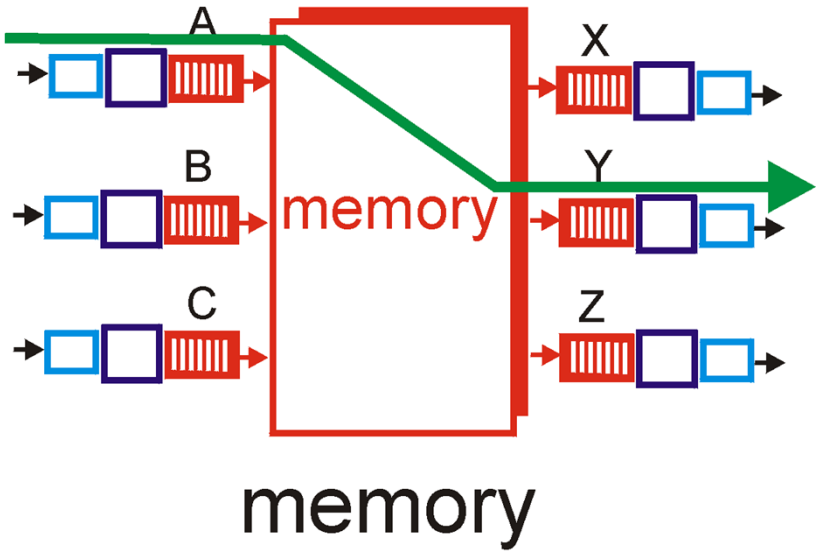
Switching via Bus

- Datagram from input port buffer to output port buffer via shared bus
- Bus contention: Switching speed limited by bus bandwidth



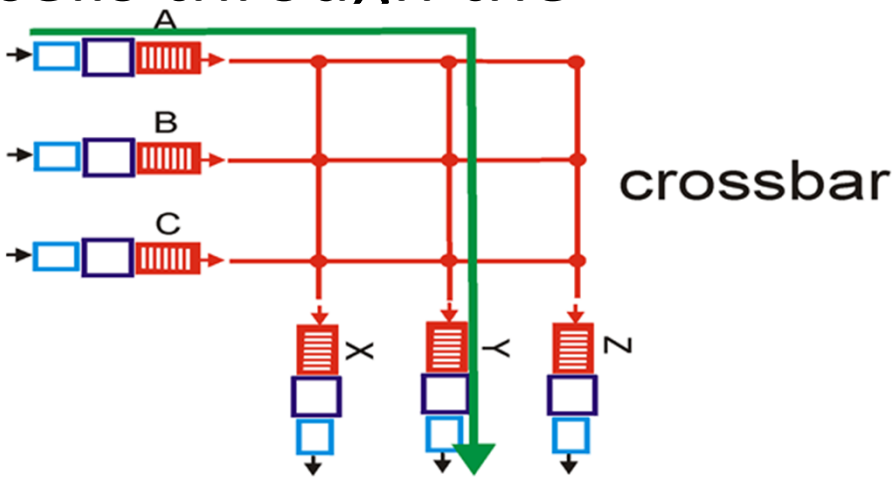
Switching via Memory

- First generation:
  - Packet copied by system CPU
  - Speed limited by memory bandwidth
- Next generation:
  - Input port processor performs lookup and copying into memory
- Today:
  - Specialised mechanisms



Switching via crossbar

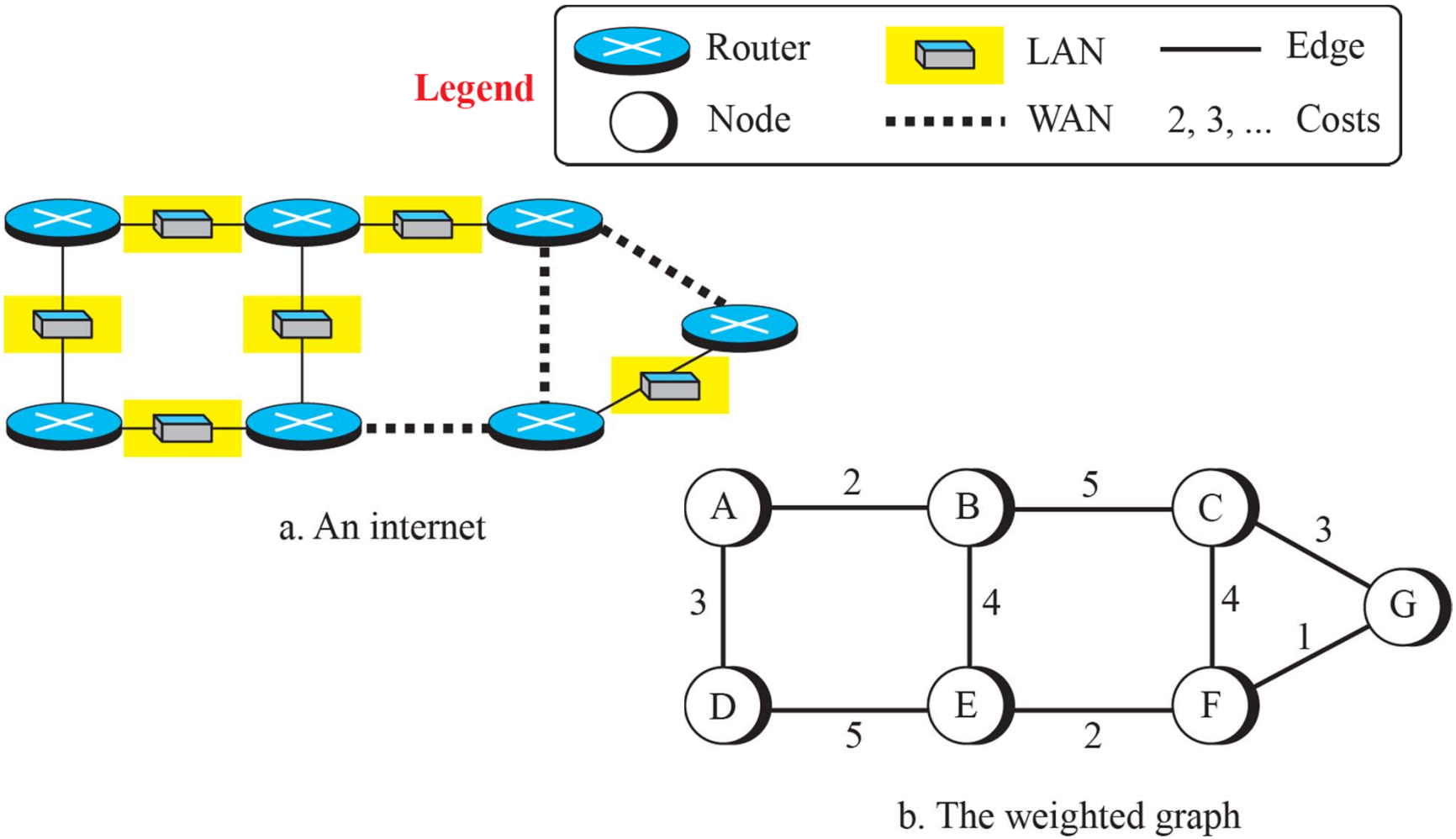
- overcome bus bandwidth limitations
- interconnection nets initially developed to connect processors in multiprocessor
- Advanced design: fragmenting datagram into fixed length cells, switch cells through the fabric.



Common Cost Metrics

- Alternatives at the link level
  - Hop count
  - Inverse of the link bandwidth
  - Delay
  - Dynamically calculated
  - Administratively assigned
  - Combination
- Traffic monitored → metrics adjusted

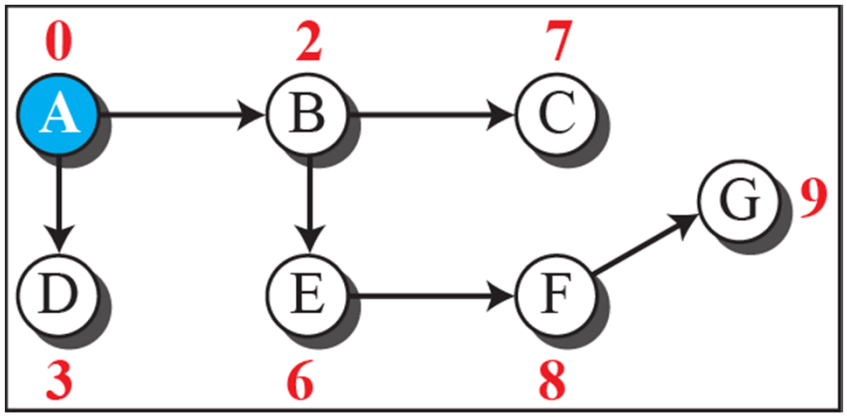
Graphical representation of a net

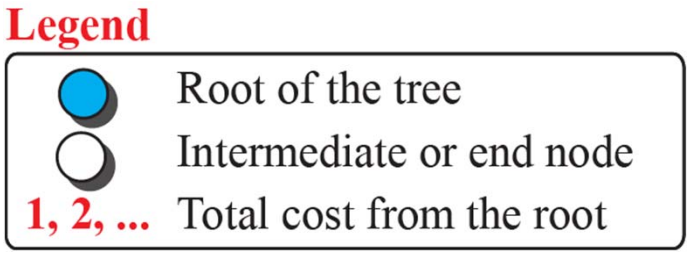


a. An internet

b. The weighted graph

What is an end node?

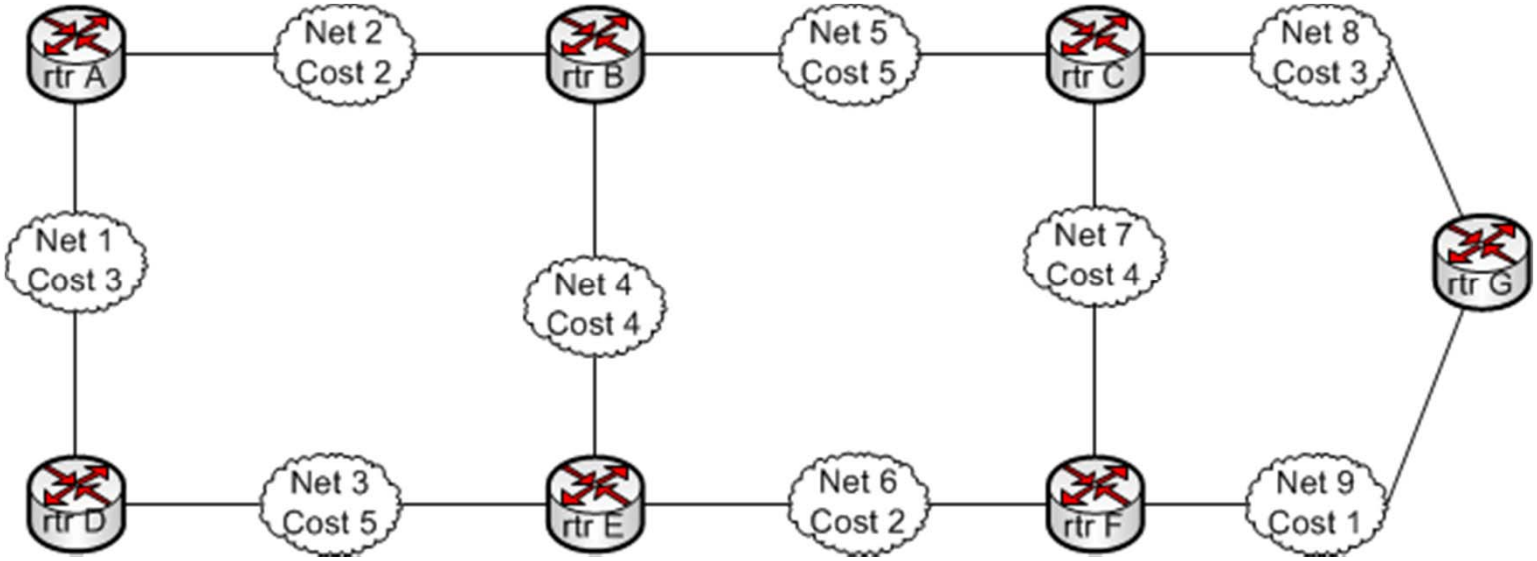




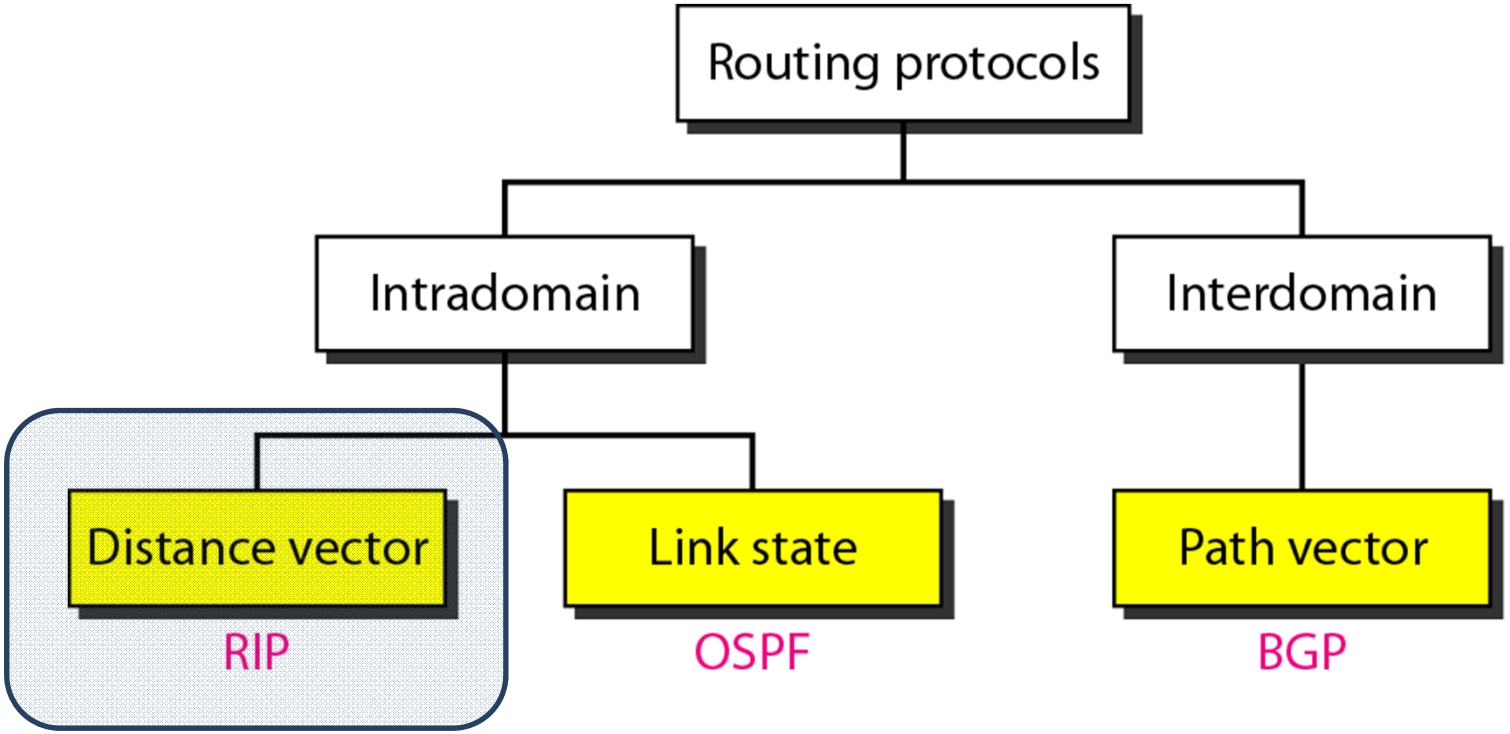
Problem: The LANs are our destinations/end nodes, not the routers

A more realistic representation

- Solution: Nets and routers are all nodes in the tree.
- Routers hold tables how to reach nets and what is the *next hop* for to get there



Routing Algorithms and Protocols



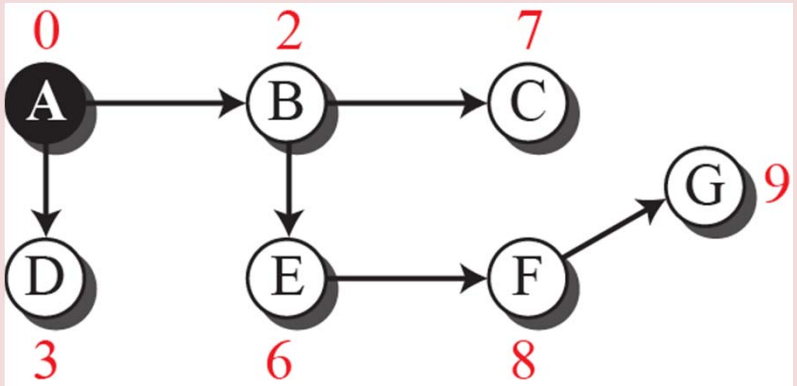
RIP (Routing Information Protocol)

- Included in BSD-UNIX Distribution in 1982
- Distance metric:
  - **# of hops** (max 15) to destination network
- Distance vectors:
  - exchanged among neighbours every 30” via Response Message (advertisement)
- Implementation:
  - Application layer protocol, uses UDP/IP

Distance Vector Routing

- Best path info **shared** locally
  - Periodically
  - Upon any change
- Routing tables **updated** for
  - New entries
  - Cost changes
- Metric
  - Not necessarily hop count!

Tree and Distance Vector

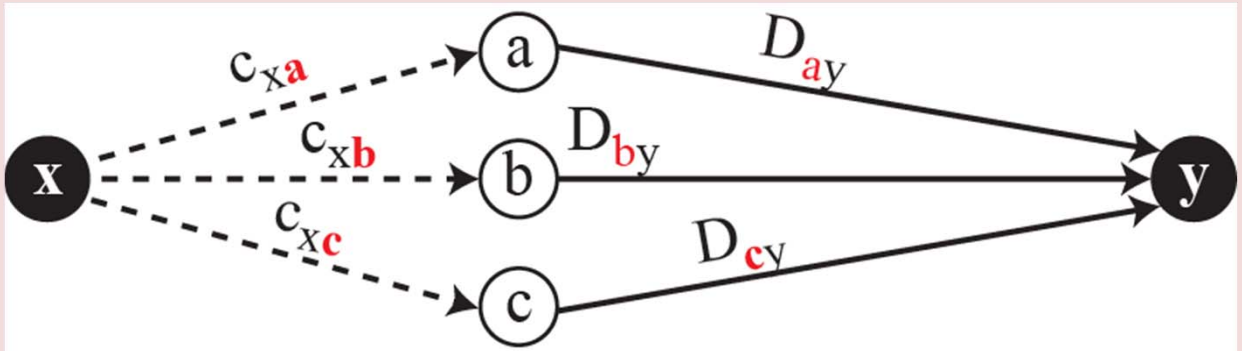


a. Tree for node A

	A
A	0
B	2
C	7
D	3
E	6
F	8
G	9

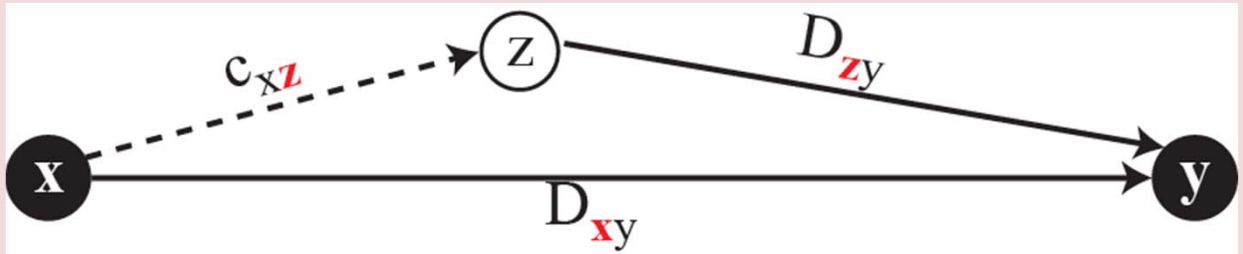
b. Distance vector for node A

Bellman-Ford



a. General case with three intermediate nodes

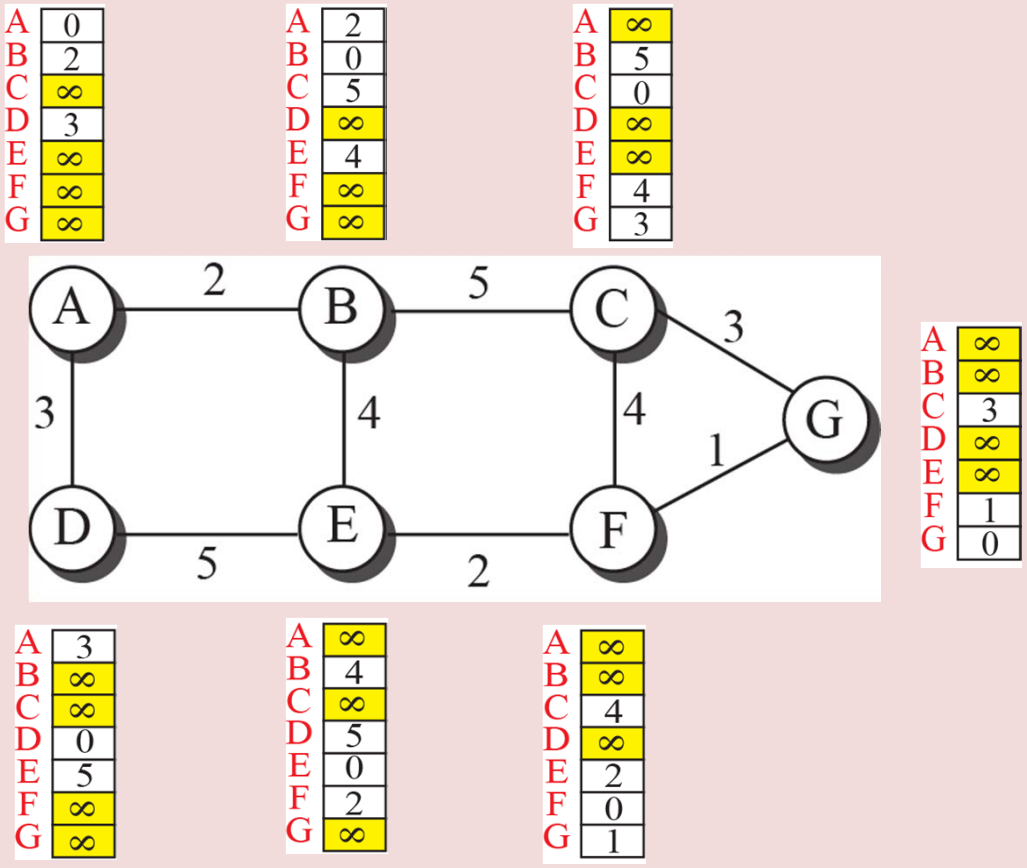
$$D_{xy} = \min\{(c_{xa} + D_{ay}), (c_{xb} + D_{by}), (c_{xc} + D_{cy}) \dots\}$$



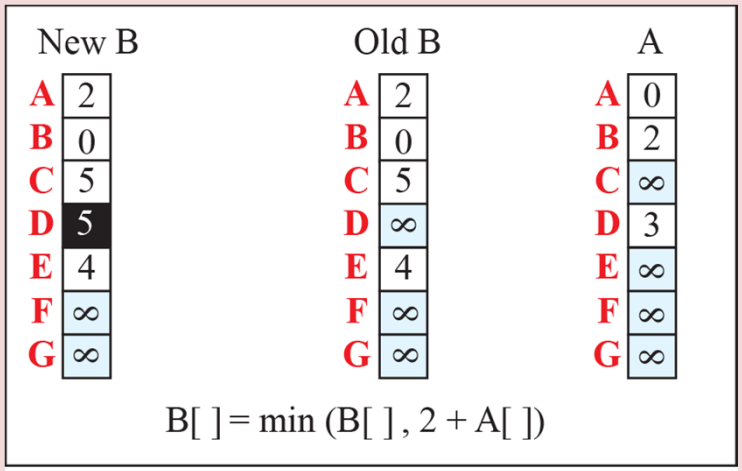
b. Updating a path with a new route

$$D_{xy} = min\{D_{xy}, (c_{xz} + D_{zy})\}$$

Initial Distance Vectors

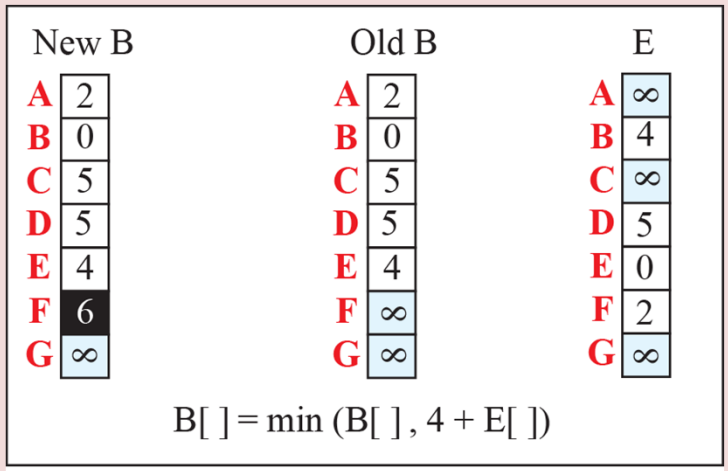


Updating Distance Vectors



a. First event: B receives a copy of A's vector.

Note:
X[ ]: the whole vector



b. Second event: B receives a copy of E's vector.

A RIP Forwarding/Routing Table

Destination=net	Cost	Next hop=router
123	3	A
32	5	D
16	3	A
7	2	-

RIP update message

- Contains the whole forwarding table
- Add 1 to cost in received message
- Change next hop to sending router
- Apply RIP updating algorithm

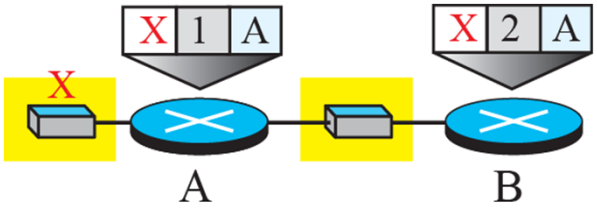
RIP Updating Algorithm (Bellman-Ford)

```
if (advertised destination not in table)
    {
    add new entry // rule #1
    }
else if (adv. next hop = next hop in table)
    {
    update cost // rule #2
    }
else if (adv. cost < cost in table)
    {
    replace old entry // rule #3
    }
```

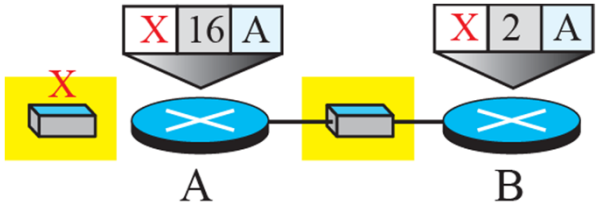
RIP Example

- From textbook Figure 20.18

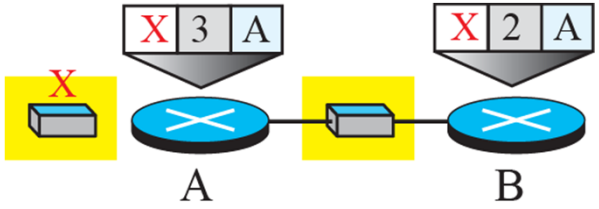
Two node instability/Count to inifinity



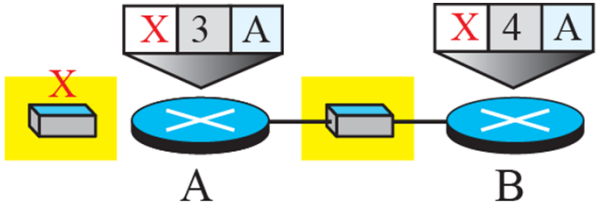
a. Before failure



b. After link failure

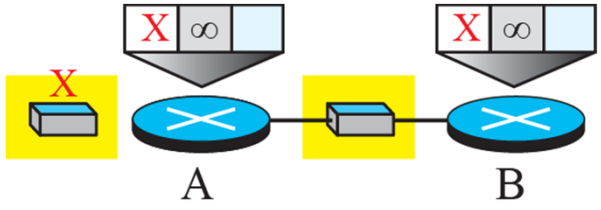


c. After A is updated by B



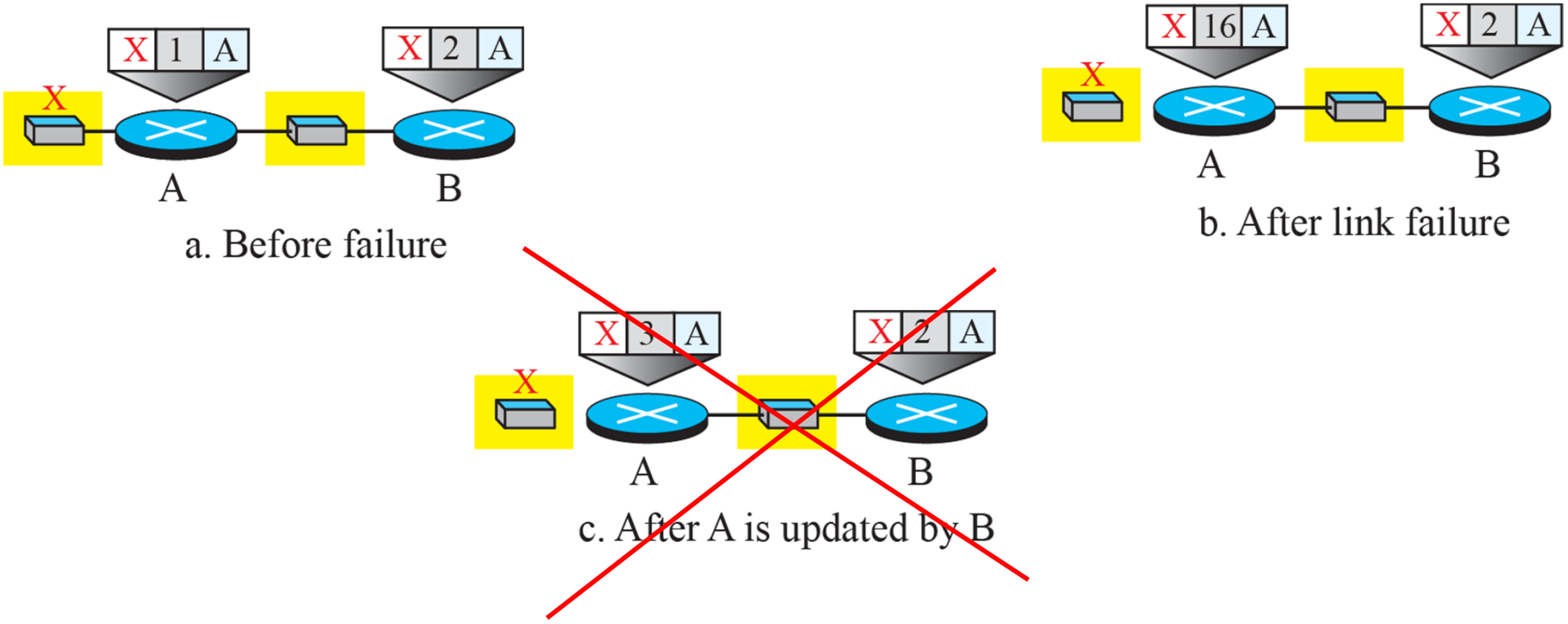
d. After B is updated by A

• • •



e. Finally

Split Horizon breaks Count to inifinity

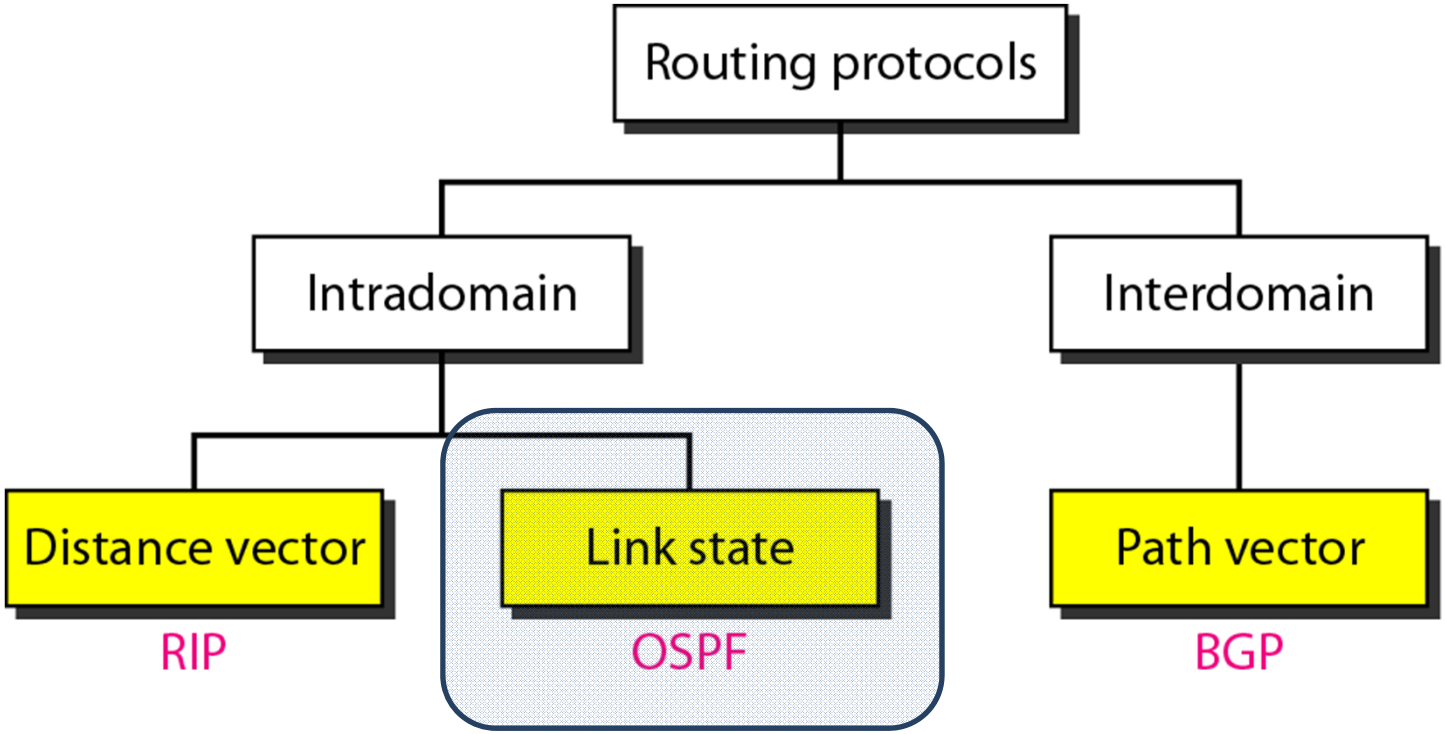


I have a route to X, but I got it from A so I won't tell A about it!

RIP: Link Failure and Recovery

- If no advertisement heard after 180”
  - Neighbour/link declared dead
  - Routes via neighbour invalidated (infinite distance = 16 hops)
  - New advertisements sent to neighbours (triggering a chain reaction if tables changed)
  - “Poison reverse” used to prevent count to infinity loops
  - “Good news travel fast, bad news travel slow”

Routing Algorithms and Protocols



OSPF (Open Shortest Path First)

- Divides domain into areas
  - Limits flooding for efficiency
  - One ”backbone” area connects all
- Distance metric:
  - Cost to destination network

Link State Routing

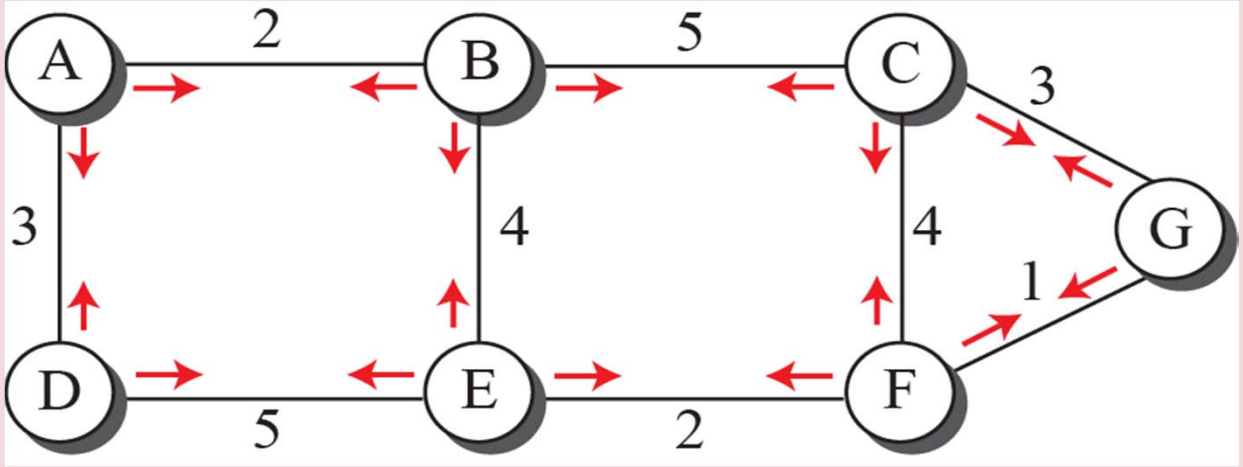
- Local topology info **flooded globally**
  - Periodically (very seldom …)
  - Upon any change
- Routing tables **updated** for
  - Link state changes
  - Cost changes

Link State information (LSP)

Node	Cost
B	2
D	3

Node	Cost
A	2
C	5
E	4

Node	Cost
B	5
F	4
G	3



Node	Cost
C	3
F	1

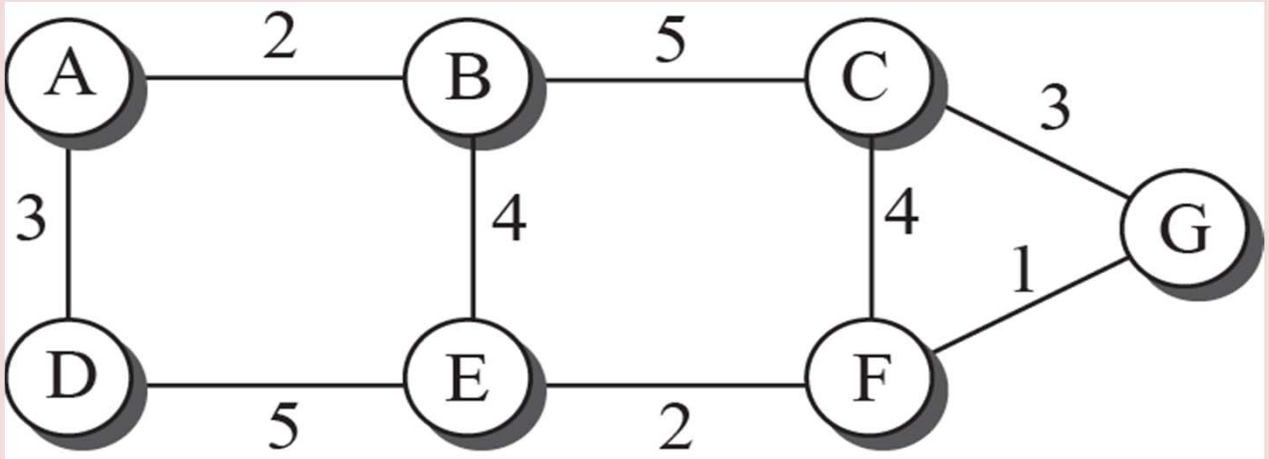
Node	Cost
A	3
E	5

Node	Cost
B	4
D	5
E	2

Node	Cost
C	4
E	2
G	1

Updates on changes!

Link State DataBase



a. The weighted graph

	A	B	C	D	E	F	G
A	0	2	∞	3	∞	∞	∞
B	2	0	5	∞	4	∞	∞
C	∞	5	0	∞	∞	4	3
D	3	∞	∞	0	5	∞	∞
E	∞	4	∞	5	0	2	∞
F	∞	∞	4	∞	2	0	1
G	∞	∞	3	∞	∞	1	0

b. Link state database

Same DB in all nodes!

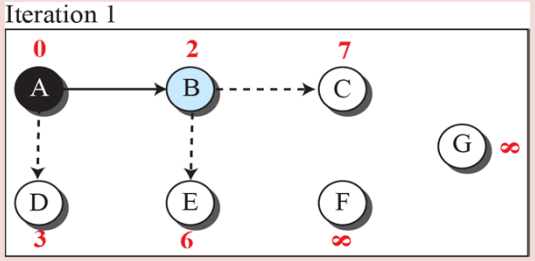
Tree Generation Algorithm (Dijkstra)

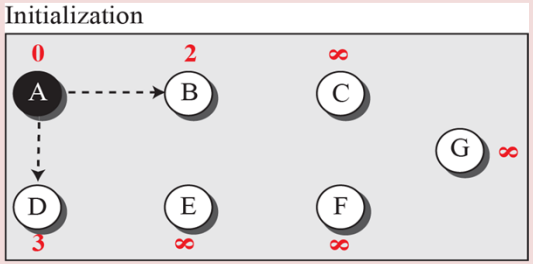
```
put yourself to tentative list
while tentative list not empty
   {
   pick node which can be reached
             with least cumulative cost
   add it to your tree*
   put its neighbours to tentative list**
             with cumulative costs to reach them
   }
```

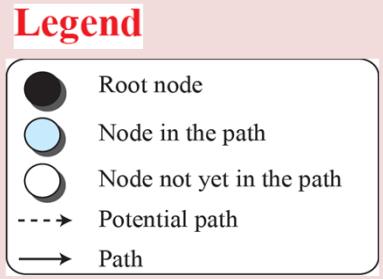
*(a.k.a. permanent list)

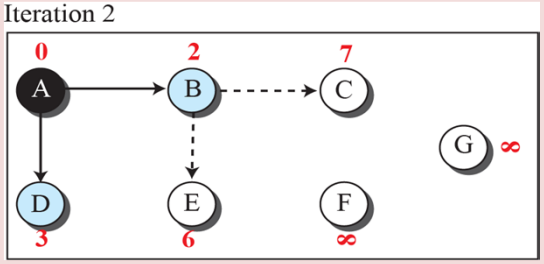
**(if not already there)

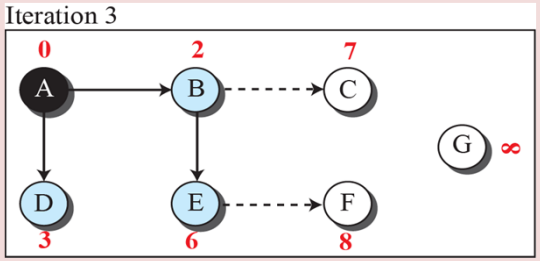
Building Least Cost Trees

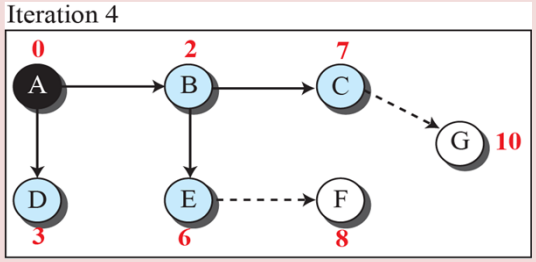


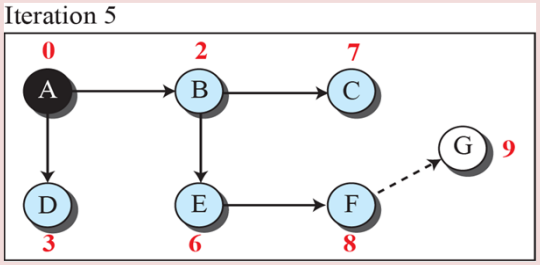


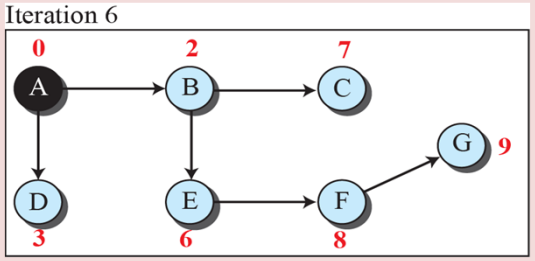




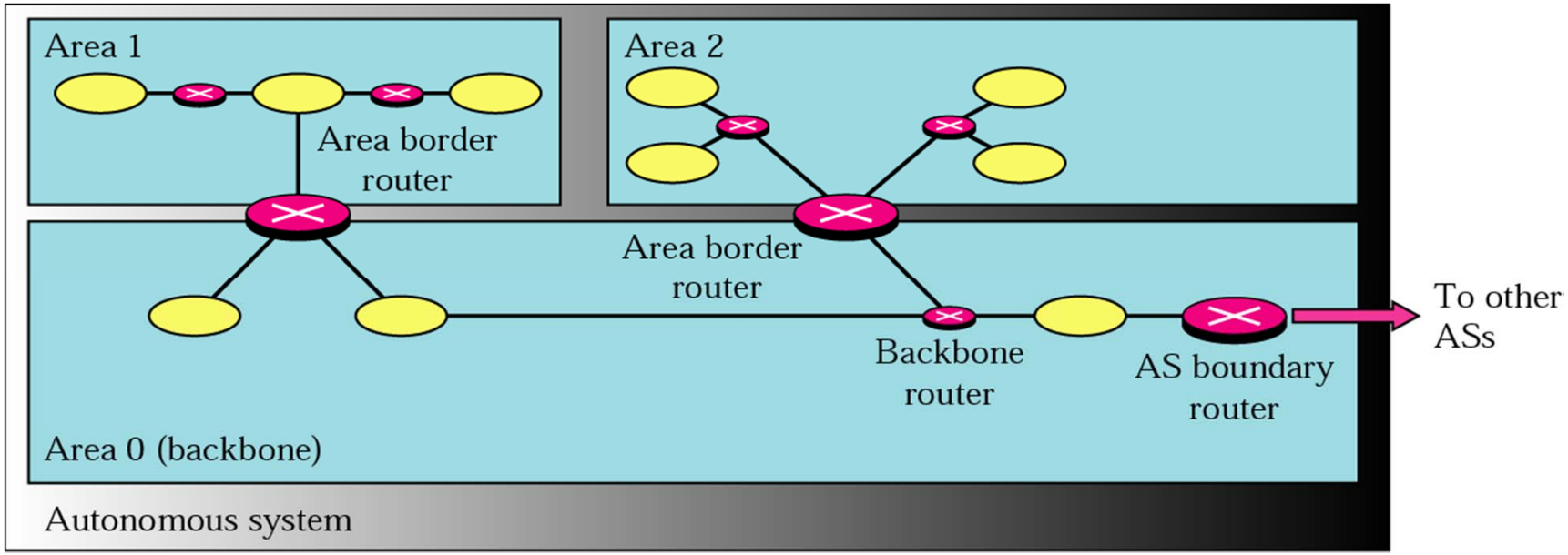






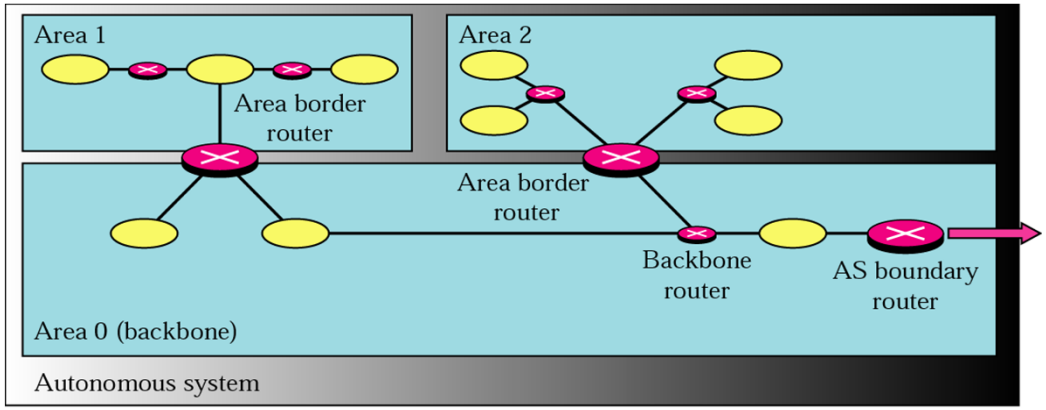


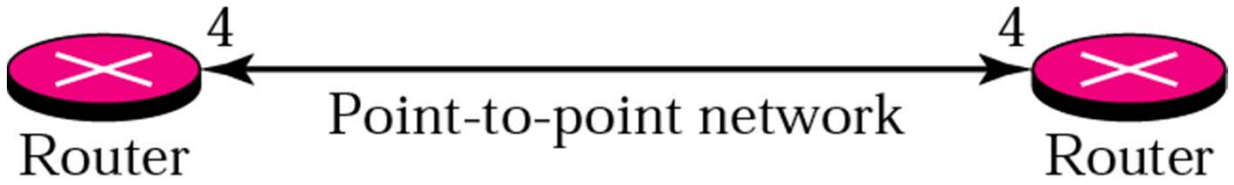
Areas, Router and Link Types



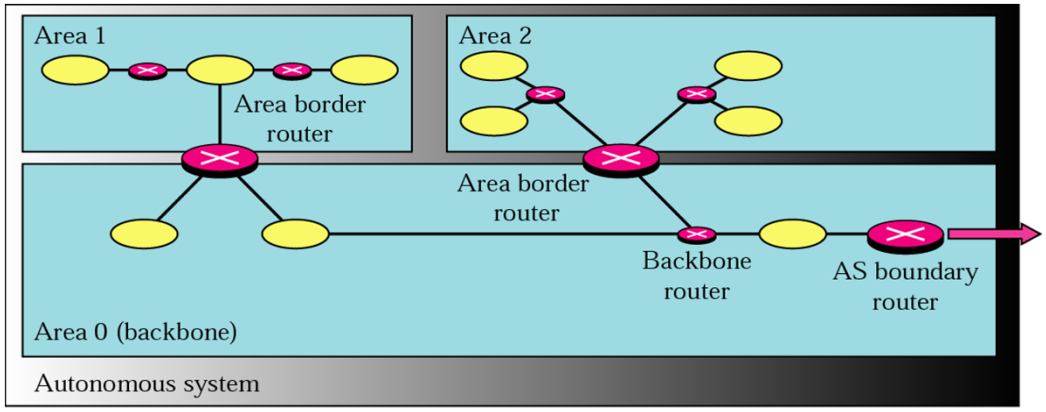
Point-to-Point Link

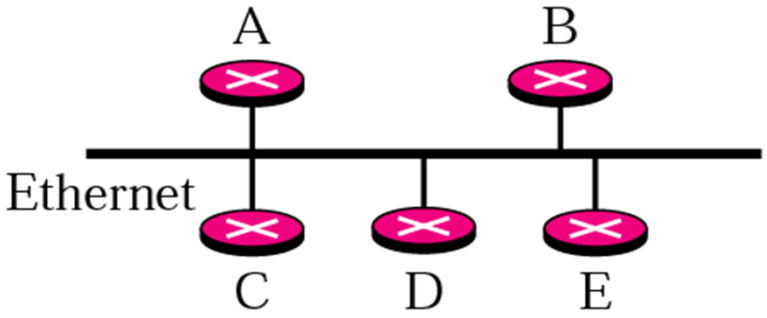
- Connects two routers
- No need for addresses



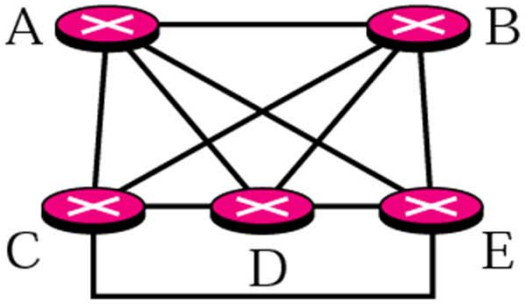


Transient Link

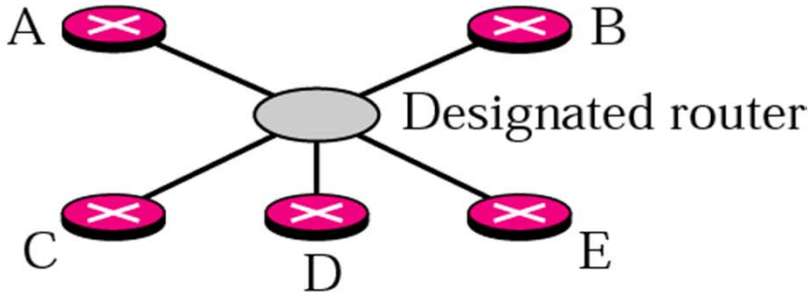




a. Transient network

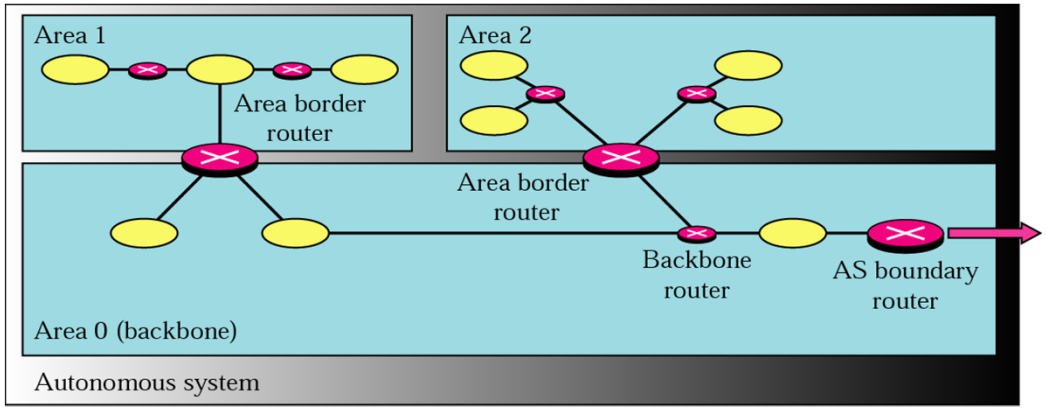


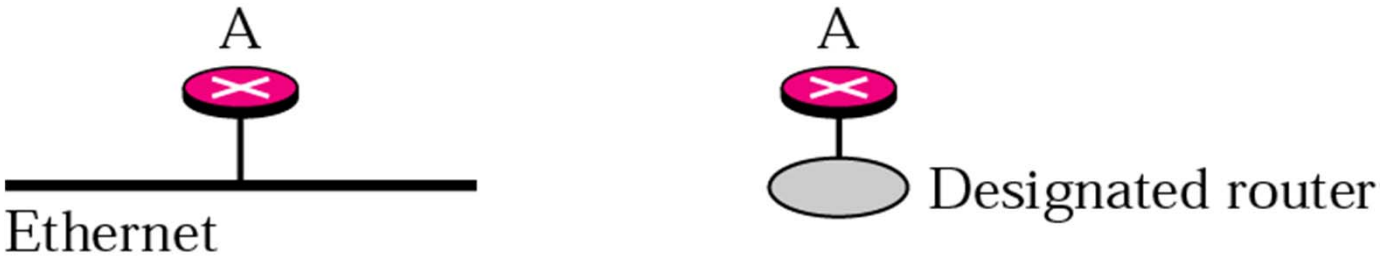
b. Unrealistic representation



c. Realistic representation

Stub Link

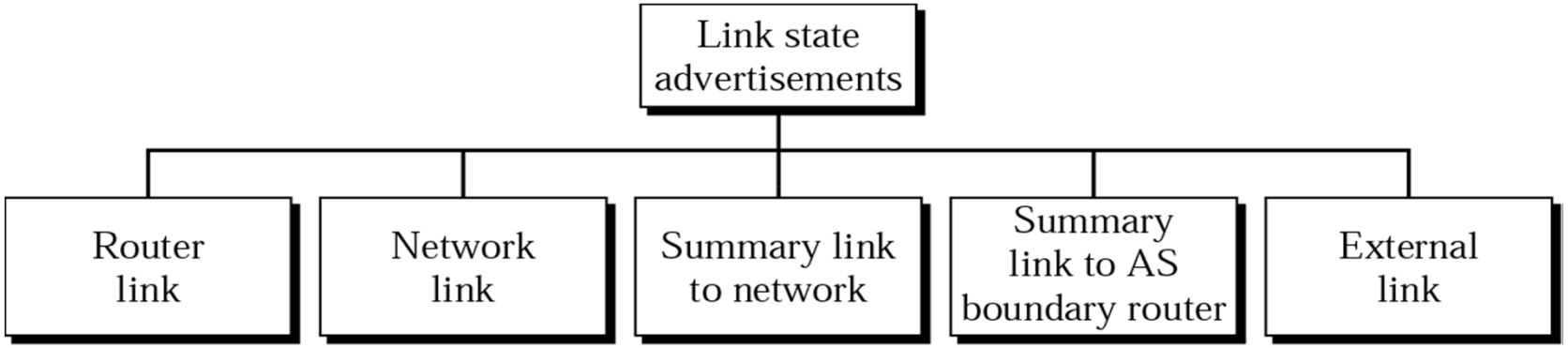




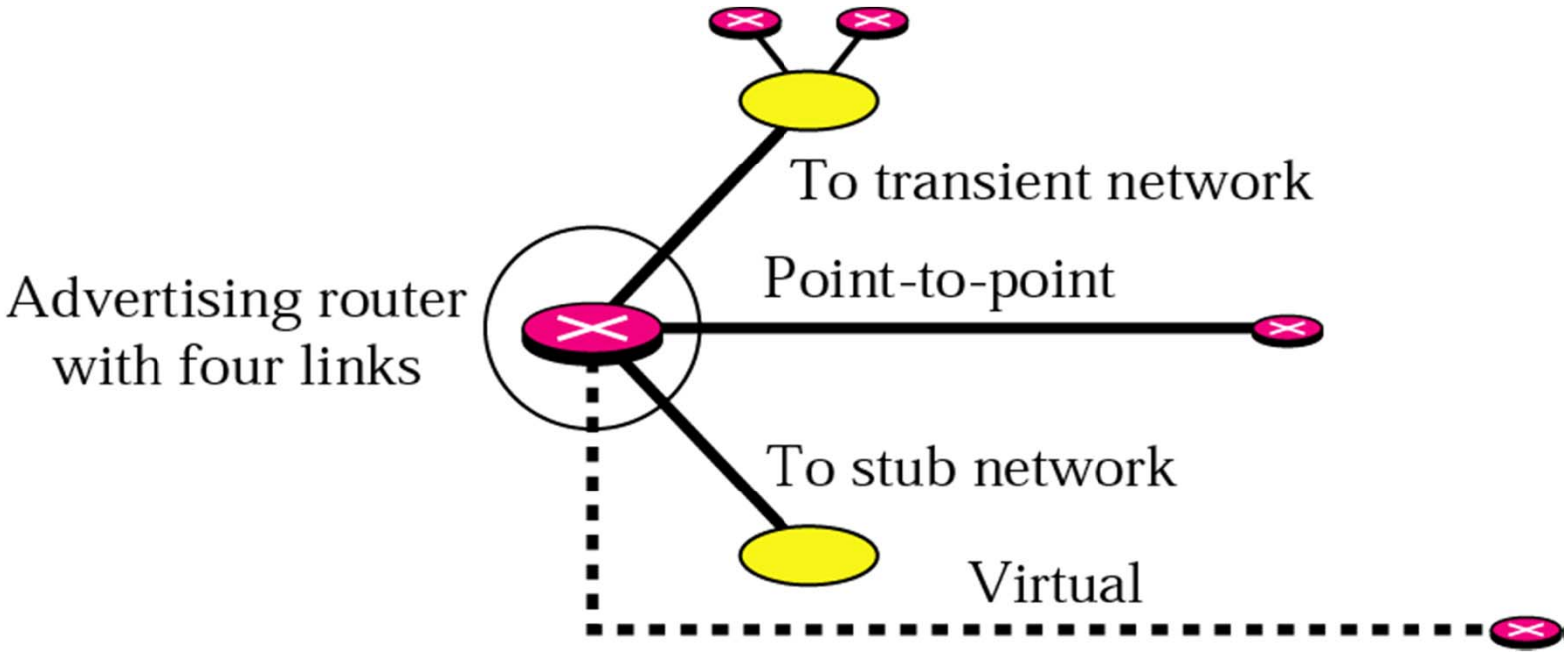
a. Stub network b. Representation

Link State Advertisements

- What to advertise?
  - Different entities as nodes
  - Different link types as connections
  - Different types of cost

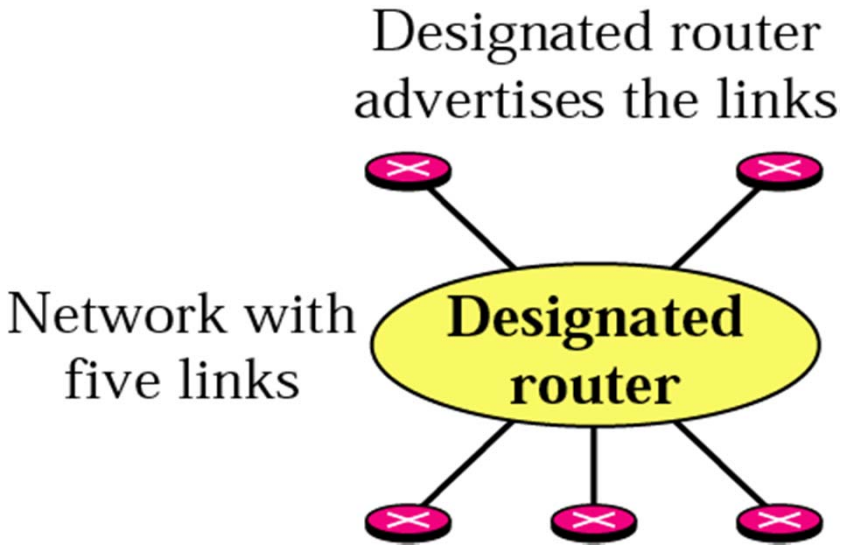


Router Link Advertisement



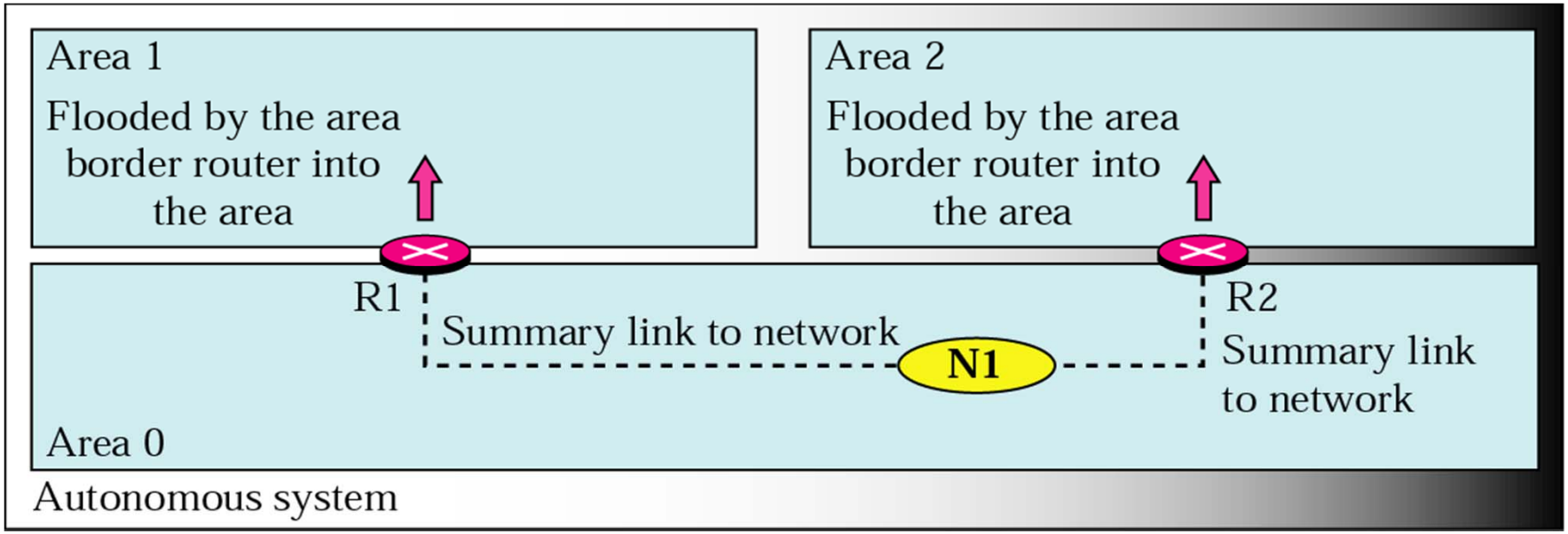
Network Link Advertisement

- Network is a passive entity
  - It cannot advertise itself



Summary Link to Network

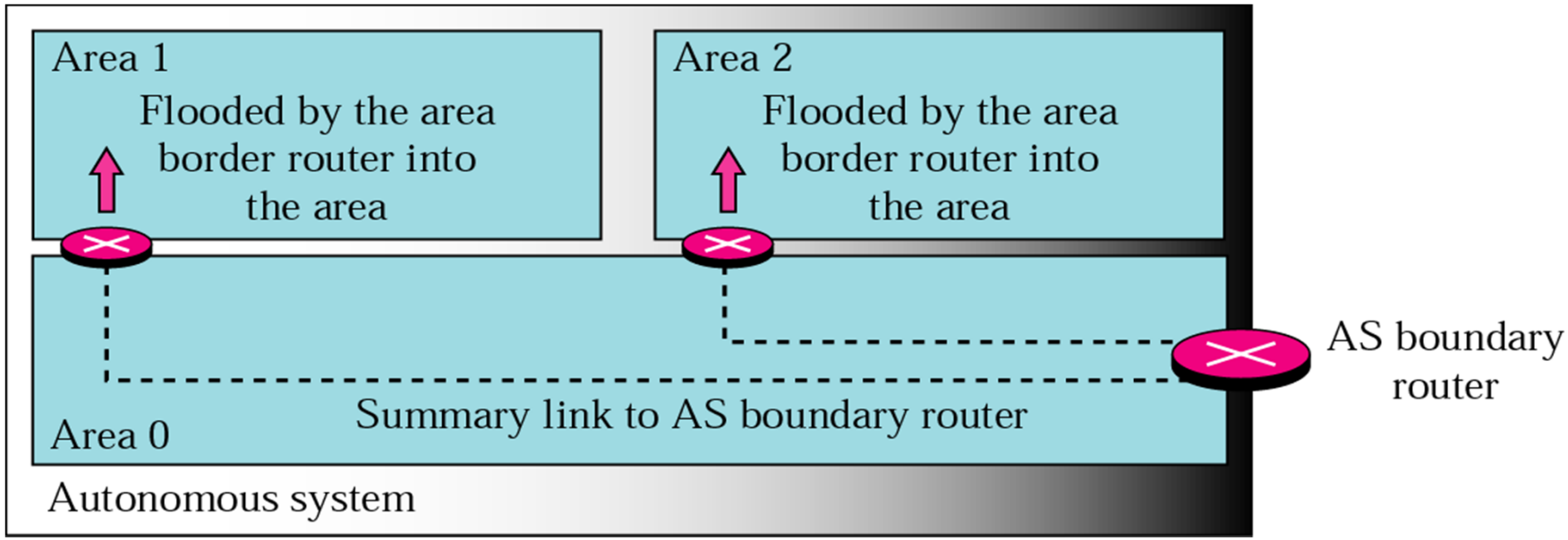
- Done by area border routers
  - Goes through the backbone



Summary Link to AS Boundary Router

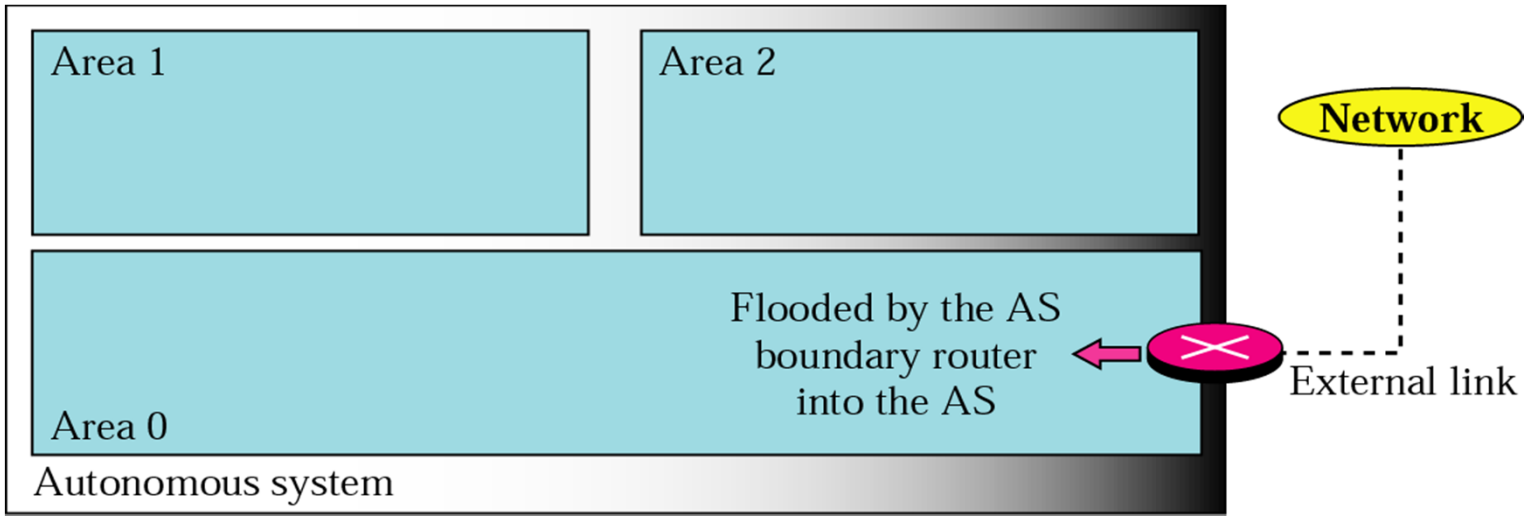
- Links to other domains

  ”autonomous systems”



External Link Advertisement

- Link to a single network outside the domain



Hello message

- Find neighbours
- Keep contact with neighbours: I am still alive!
- Sent out periodically (typically every 10th second)
- If no hellos received during holdtime (typically 30 seconds), neighbour declared dead.

- Compare RIP update messages