Communication Networks:

Technology & Protocols

Stavros Tripakis (stavros@eecs)

Lectures 9, 10, 11

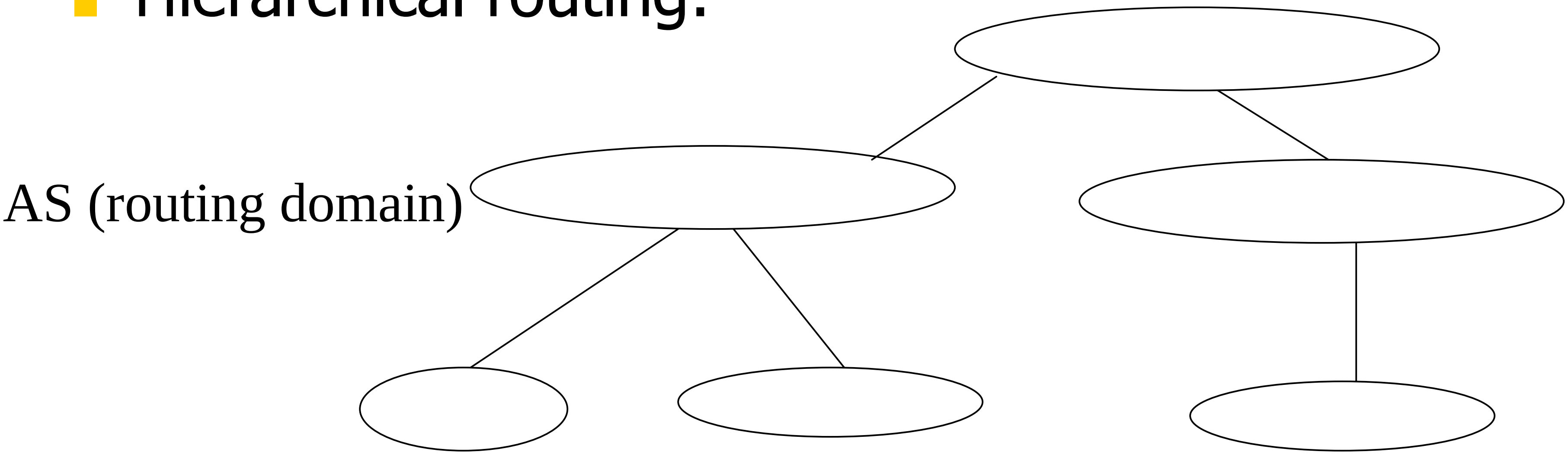
September 13, 15, 17

Logistics

- Web site:
  - www.cs.berkeley.edu/~amc/eecs122
- **Homework 3** (due Friday 9/17) is available on web-site.
  - Homework 2 due today.
- Book typo: Subnetting, pg. 59 & fig. 3.10, **replace: D $\otimes$ M by: D $\otimes$ N**.

Internet routing: summary of previous lecture

- Goal of routing: interconnect a large number of heterogeneous networks.
- Major concerns: scalability, robustness.
- Hierarchical routing:



Internet routing: summary of previous lecture

- *Intradomain routing* : routing inside an AS.
  - Packet forwarding in LANs.
  - Distance-vector routing (Bellman-Ford's shortest-path algorithm, RIP protocol).
  - Link-state routing (Dijkstra's shortest-path algorithm, OSPF protocol).
- *Interdomain routing* : routing across many ASs.
  - EGP and BGP.

Packet-forwarding in LANs

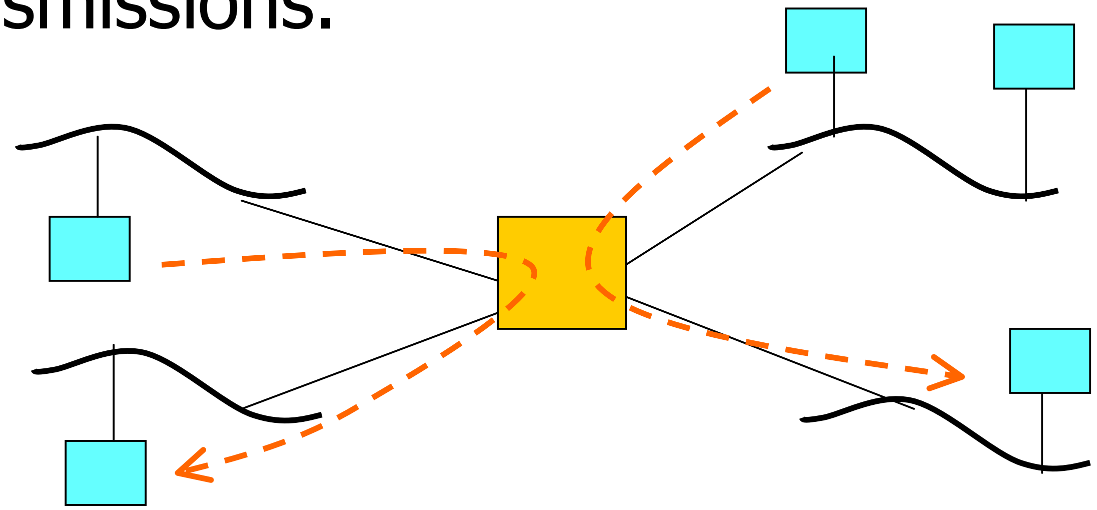
- ARP table (or cache):

IP address	LAN address
C	x, Ethernet
D	y, Ethernet
E	z, FDDI
…	

- Basic operations:
  - If A doesn't have an entry for B, it **broadcasts** message "B, are you on my LAN? If yes, give me your interface address".
  - If B is in the LAN, it replies, and A adds an entry in its ARP cache.

Switched Ethernets

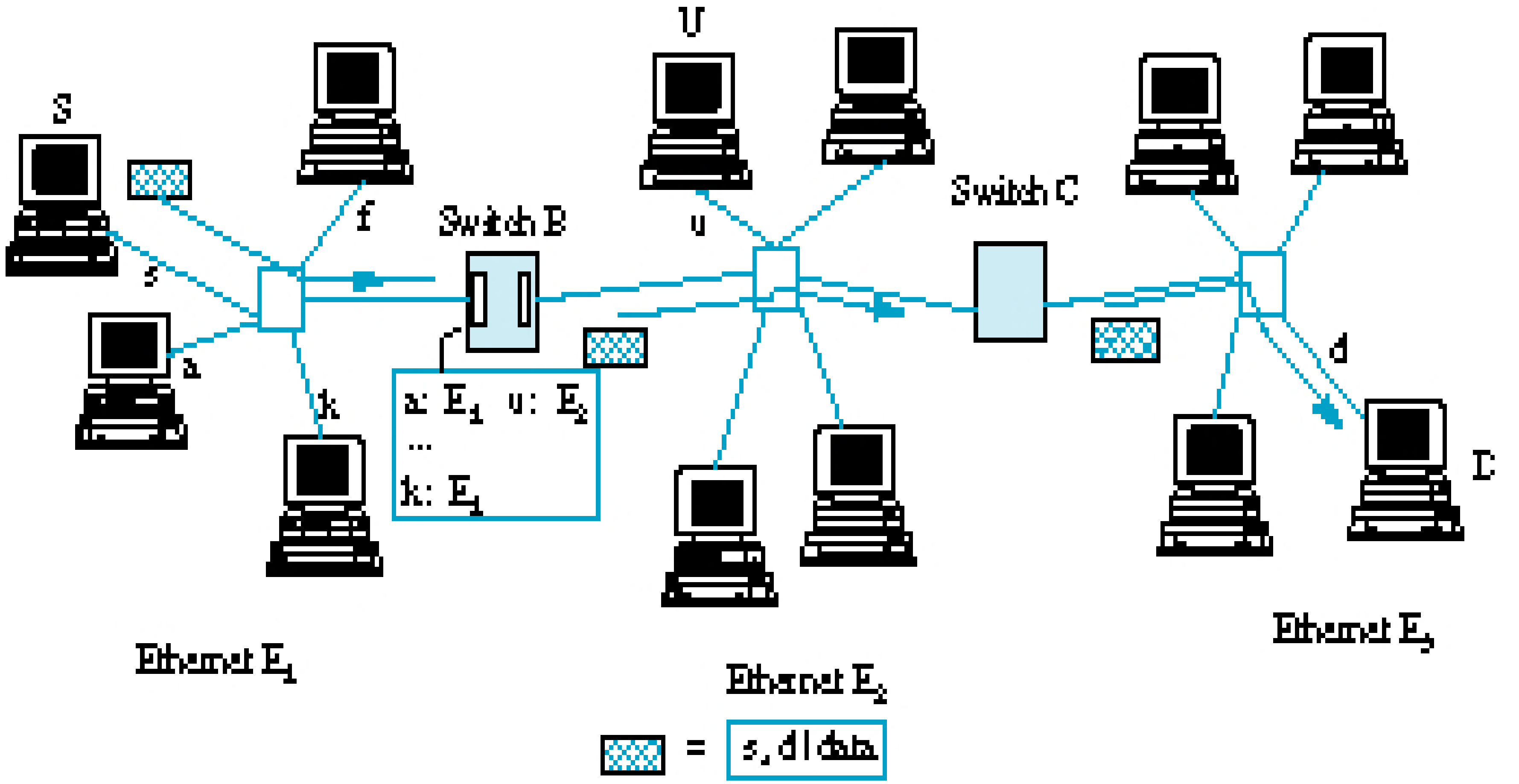
- In a single Ethernet two hosts cannot transmit simultaneously (collisions).
- A **switch** can break-up an Ethernet into many Ethernets, allowing a number of simultaneous transmissions.



Packet forwarding in switched Ethernets

- Switch has table with entries: (Eth.addr., Port)
- When receiving packet (s, d | data) at input port Eth1, the switch looks-up its table for d.
  - If (d, Eth1) is there, do nothing.
  - If (d, Eth2) is there, forward packet to port 2.
  - If d is not there, forward packet to all ports but 1.
- How is the table updated?
  - Upon receiving (s, d | data) on Eth1, add (s, Eth1).
  - Remove old entries (timeout).

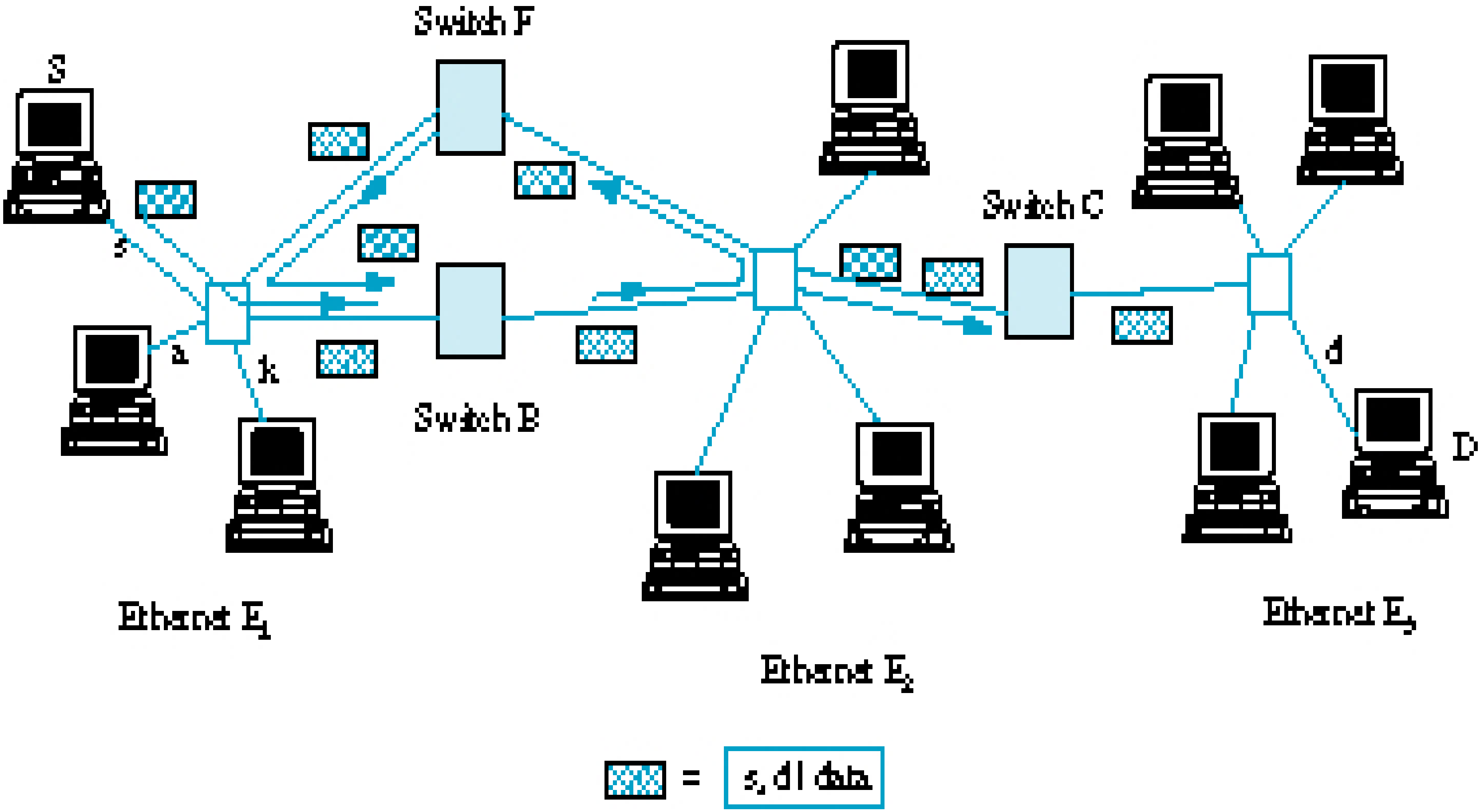
Packet forwarding in switched Ethernets



Packet forwarding in switched Ethernets

- Why not completely replace the hub of an Ethernet by a switch ?
  - Cost.
  - Number of input/output ports.

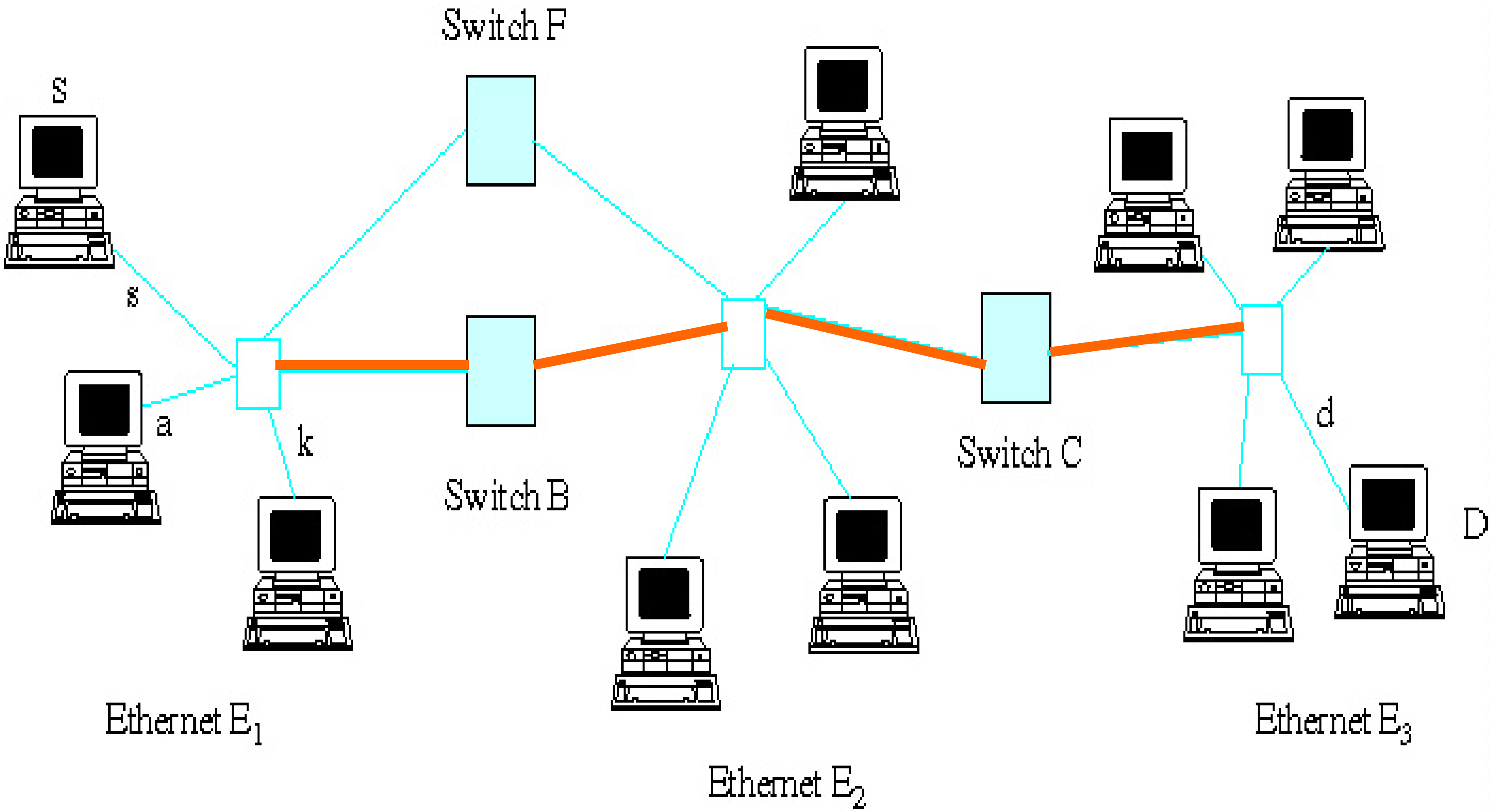
Packet forwarding in switched Ethernets: loops



Packet forwarding in switched Ethernets: loops

- Networks with loops desirable for reliability.
- However, loops should be avoided in forwarding:
  - Record all forwarded packets, do not re-forward already forwarded packet. Problem: too much bookkeeping.
  - Temporarily disable some of the links to break the loop: form **spanning-tree** (network without loops, where all hosts are connected).

Packet forwarding in switched Ethernets: loops

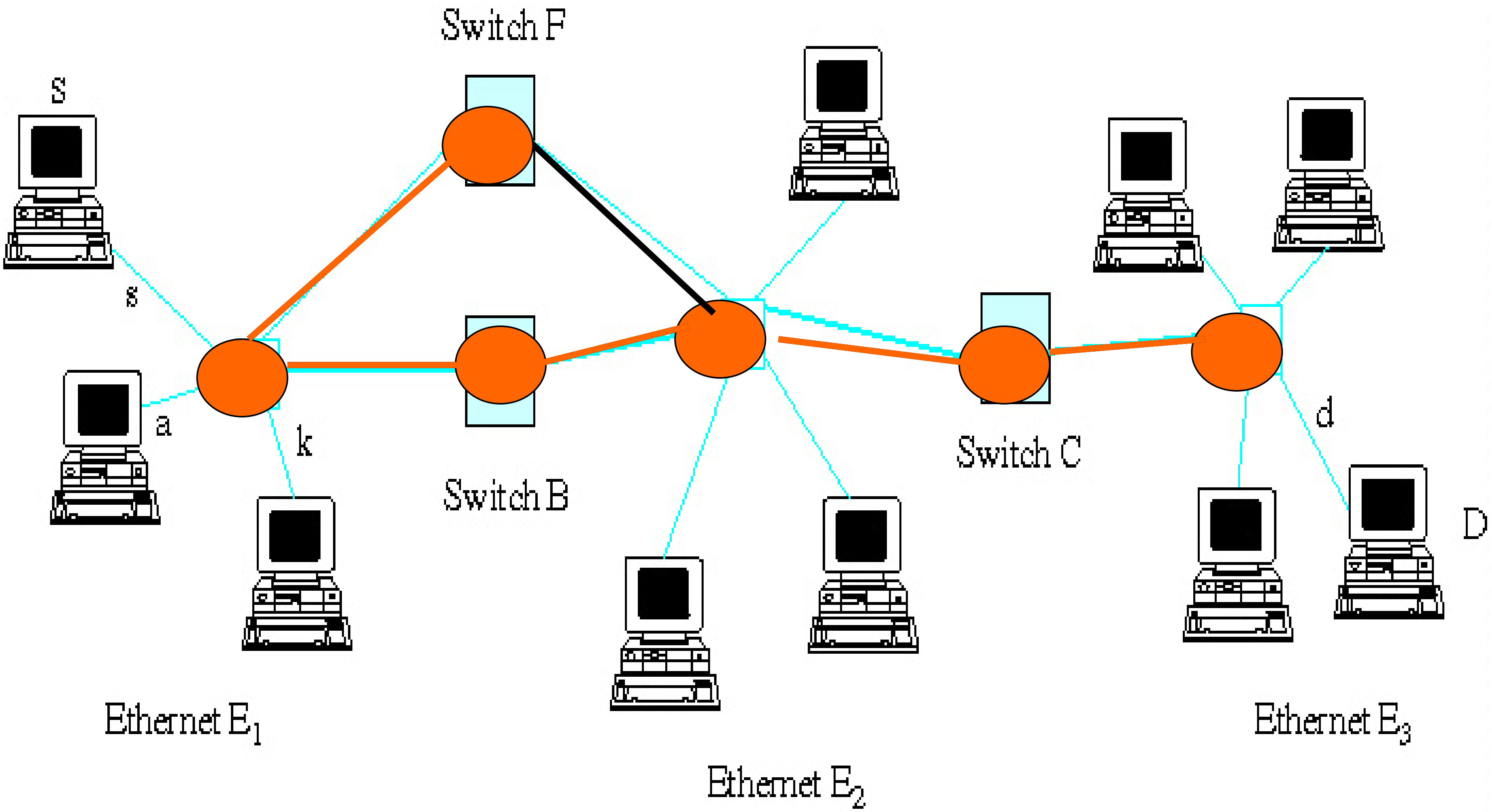


Packet forwarding in switched Ethernets: loops

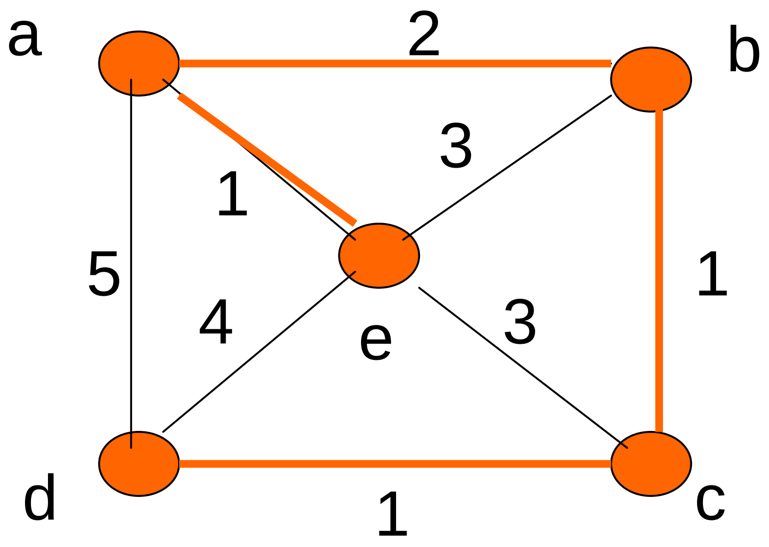
Minimum spanning-tree algorithm:

- Network represented as a **graph**:
  - Nodes of the graph are Ethernets or switches.
  - Edge means a switch is connected to an Ethernet.
- Spanning tree: a sub-graph connecting all nodes (actually all “ethernet” nodes).
- Minimum spanning tree: a spanning tree with minimum number of edges.

Switched Ethernet as a graph for spanning tree algorithm



Spanning-tree algorithm: example



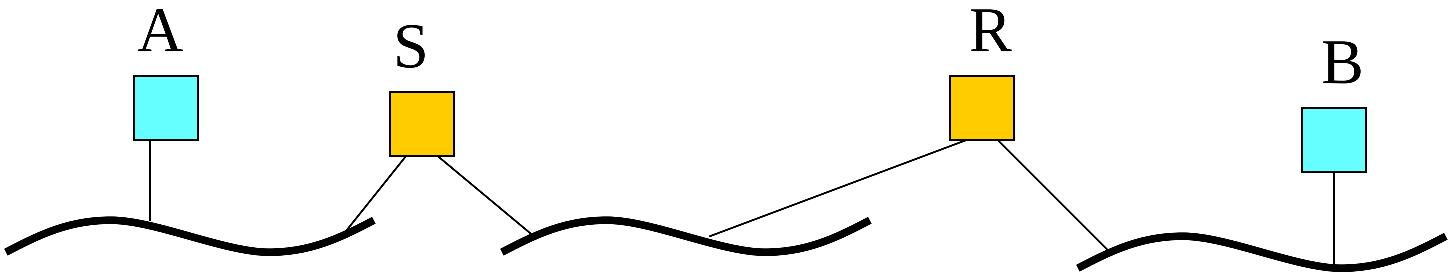
Nodes covered	Cost of tree
a	0
a, e	1
a, e, b	3
a, e, b, c	4
a, e, b, c, d	5

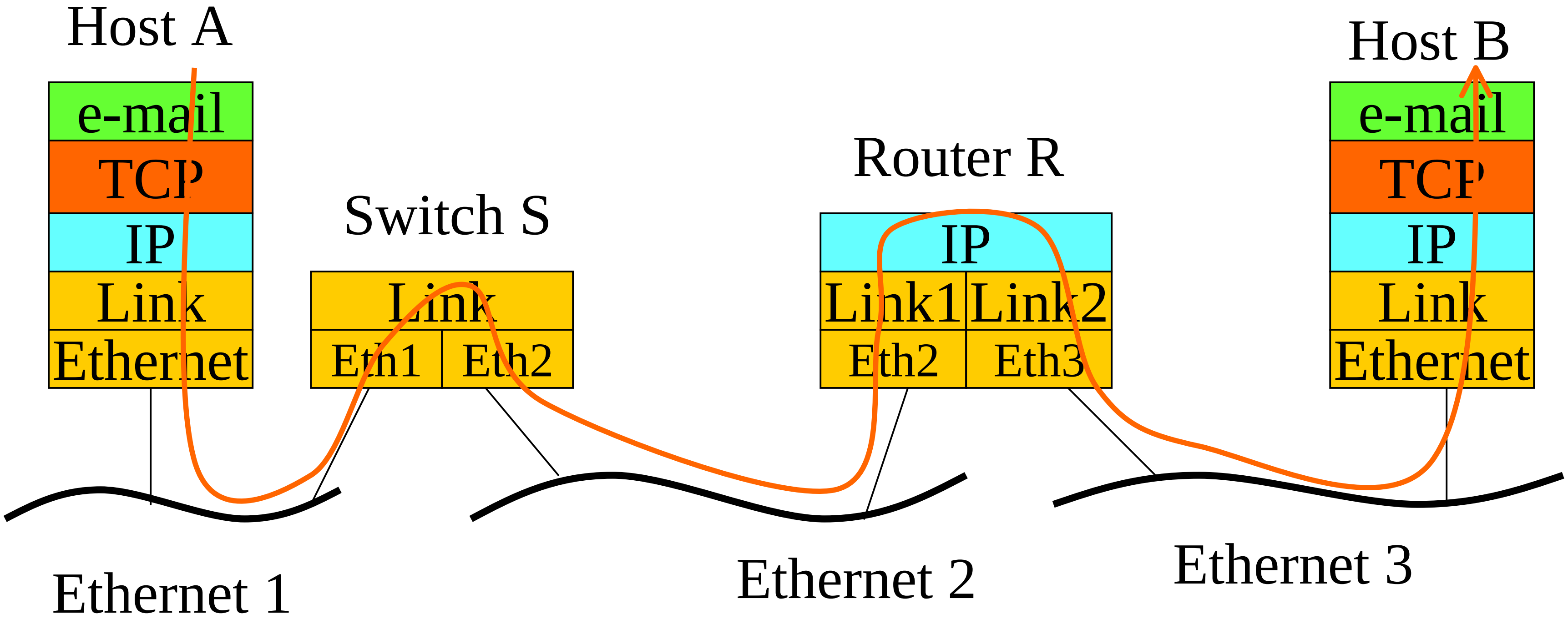
- Choose an initial node as a **root**.
- Repeat until all nodes are added to the tree:
    - add the node which least increases the cost of the tree.

Spanning-tree algorithm

- Centralized (Prim's): O(m log(m)) complexity, where m is number of edges in the graph.
- **Distributed** algorithm: non-trivial (IEEE 802.1 standard).
  - Has to be implemented among switches.
  - Switches are "blind": they only communicate by messages.
  - Steps in the algorithm to be "agreed upon" by all switches, e.g. election of root switch.
  - Final distributed knowledge has to be consistent.

Switches and routers




Host A

e-mail
TCP
IP
Link
Ethernet

Switch S

Link	
Eth1	Eth2

Router R

IP	
Link1	Link2
Eth2	Eth3

Host B

e-mail
TCP
IP
Link
Ethernet

Ethernet 1

Ethernet 2

Ethernet 3

Intradomain routing

- Network (routing domain) viewed as a **weighted graph**, where:
  - nodes are routers;
  - an edge (R1, R2) means routers R1 and R2 are connected physically (e.g., by point-to-point link, or on the same LAN);
  - the weight of an edge corresponds to a **metric** (latency, capacity, loss probability).

Distance-vector routing (Bellman-Ford's shortest-path)

- Used in (old) RIP (routing information protocol), BSD public distribution of TCP/IP.
- Bellman-Ford algorithm: given a weighted graph and a destination node D, find the shortest path from each node in the graph to D.
- Routers exchange **distance-vectors** to **neighbor** routers, e.g., (R1:5, R2:3, R3:7).
- Update routing table based on received distance vectors.

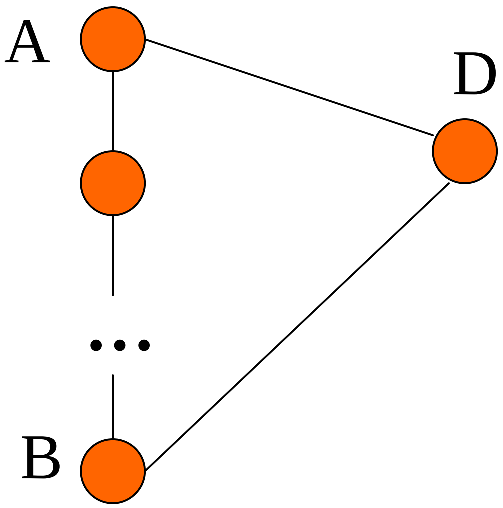
Distance-vector routing: problems

- Convergence is slow.
- Loops can be formed, due to routing table inconsistency: packets being forwarded from router to router and never reach the destination.
- Loops might last for a long time:
  - until convergence, or
  - count to infinity problem.

Link-state routing (Dijkstra's shortest-path)

- Used in (current) OSPF (open shortest path first) protocol, by IETF.
- Dijkstra's algorithm: given a weighted graph and a source node A, find the shortest path from A to each other node in the graph.
- Routers send **link-state** packets (R1,R2,7), to **all other** routers in the same routing domain (**flooding**).
- Each router learns the current state of the whole network and runs Dijkstra’s algorithm to build its own routing tables.

Link-state routing: loops



- Links A-D and B-D fail simultaneously.
- A updates its route to D through B and sends LSP to B, saying A-D is down.
- B updates its route to D through A and sends LSP to A, saying B-D is down.
- Until the LSPs are received, there is a loop between A and B for packets to D.
- Loop is **transient**: disappears when one of the LSPs is received.

Link-state vs. distance-vector

- Experience has shown OSPF to be better than RIP in **stability** (robustness to network changes):
  - distance-vector converges very slowly, loops can last for long periods of time.
  - link-state converges very quickly, loops are transient.
- RIP is distributed, whereas OSPF is centralized (flooding).
  - OSPF creates more routing traffic (flooding).