15-441 Computer Networks

Ethernet and Switch Design

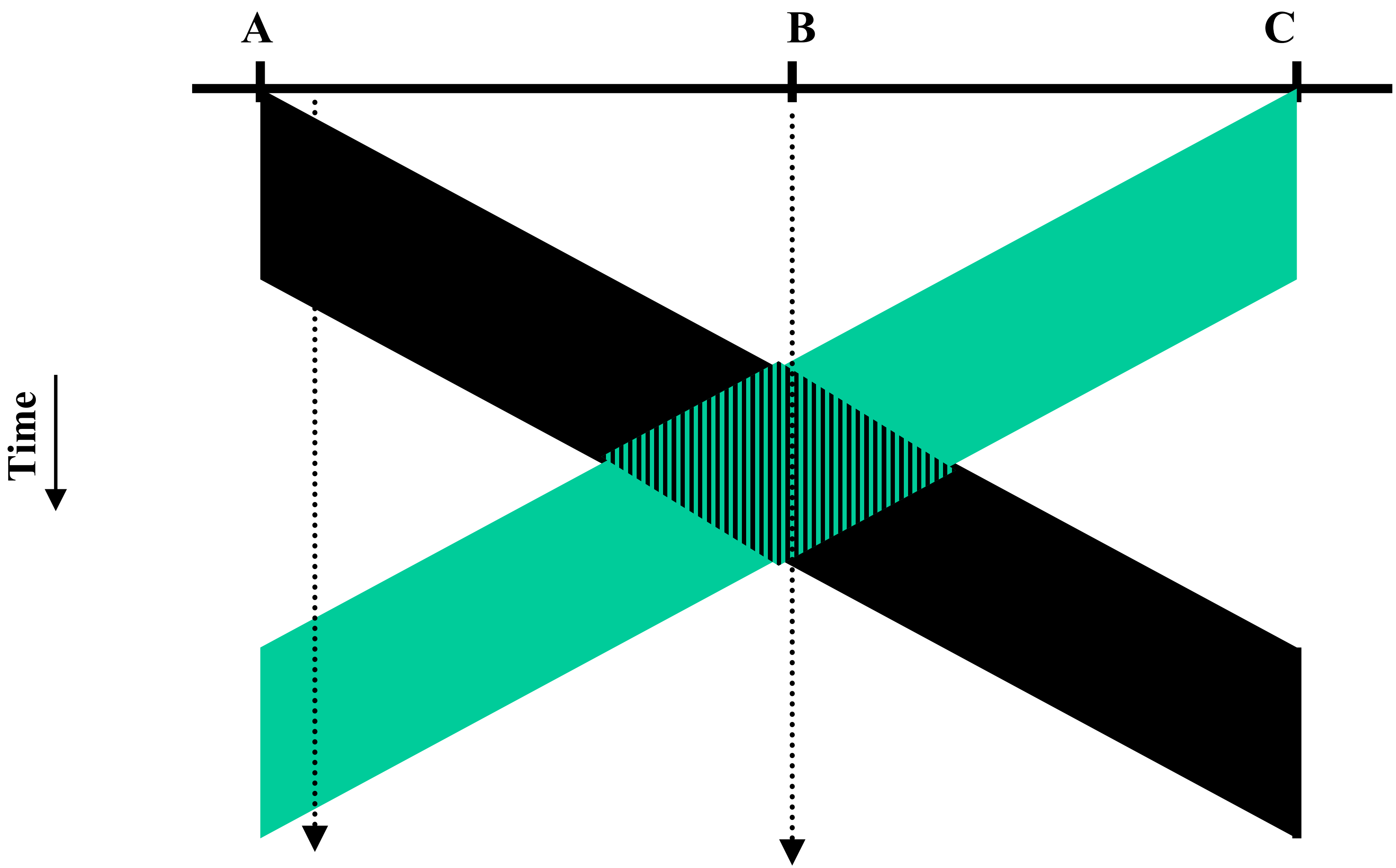
Professor Hui Zhang

hzhang@cs.cmu.edu

Ethernet Packet Size/Cable Length

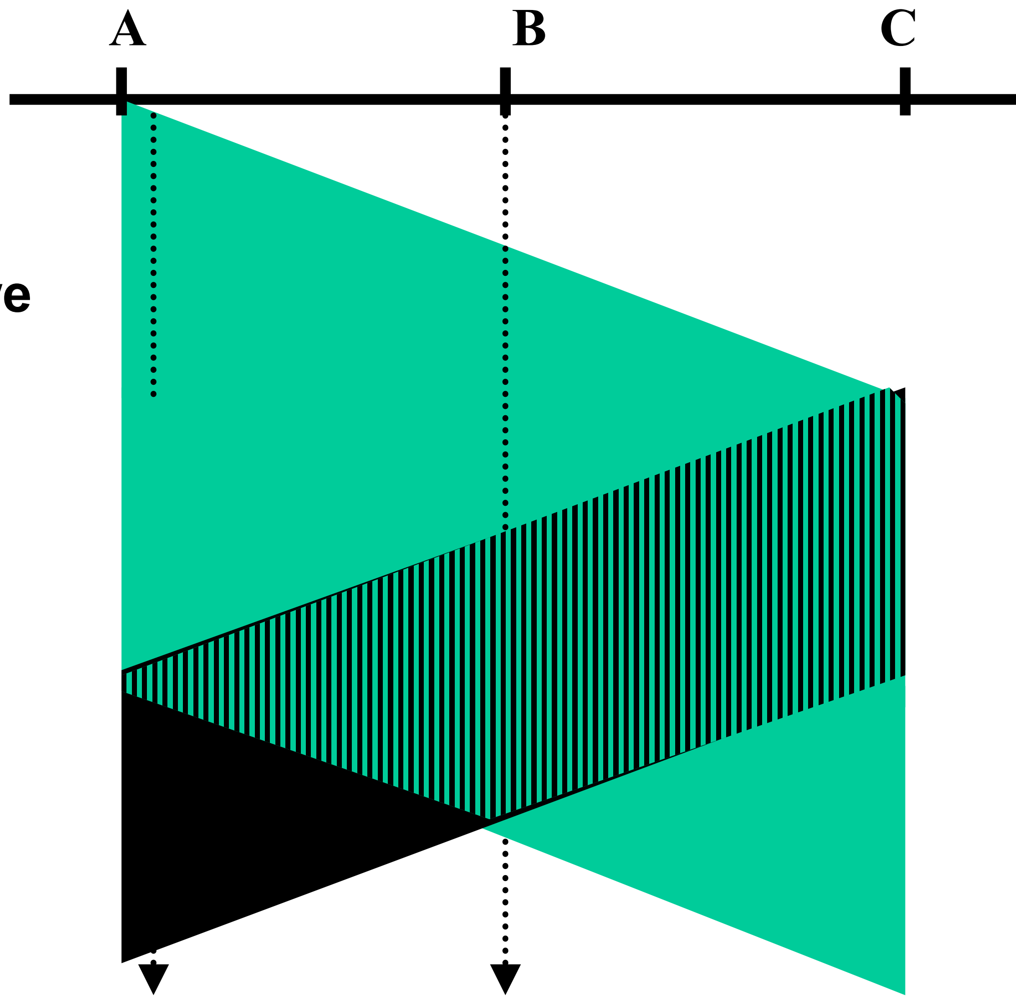
- Ethernet specifies a maximum packet size, why?
- Ethernet specifies a minimum packet size, why?
- Ethernet specifies a maximum cable length, why?

Collision Detection



Collision Detection: Implications

- **All nodes must be able to detect the collision.**
  - Any node can be sender
- **The implication is that either we must have a short wires, or long packets.**
  - Or a combination of both

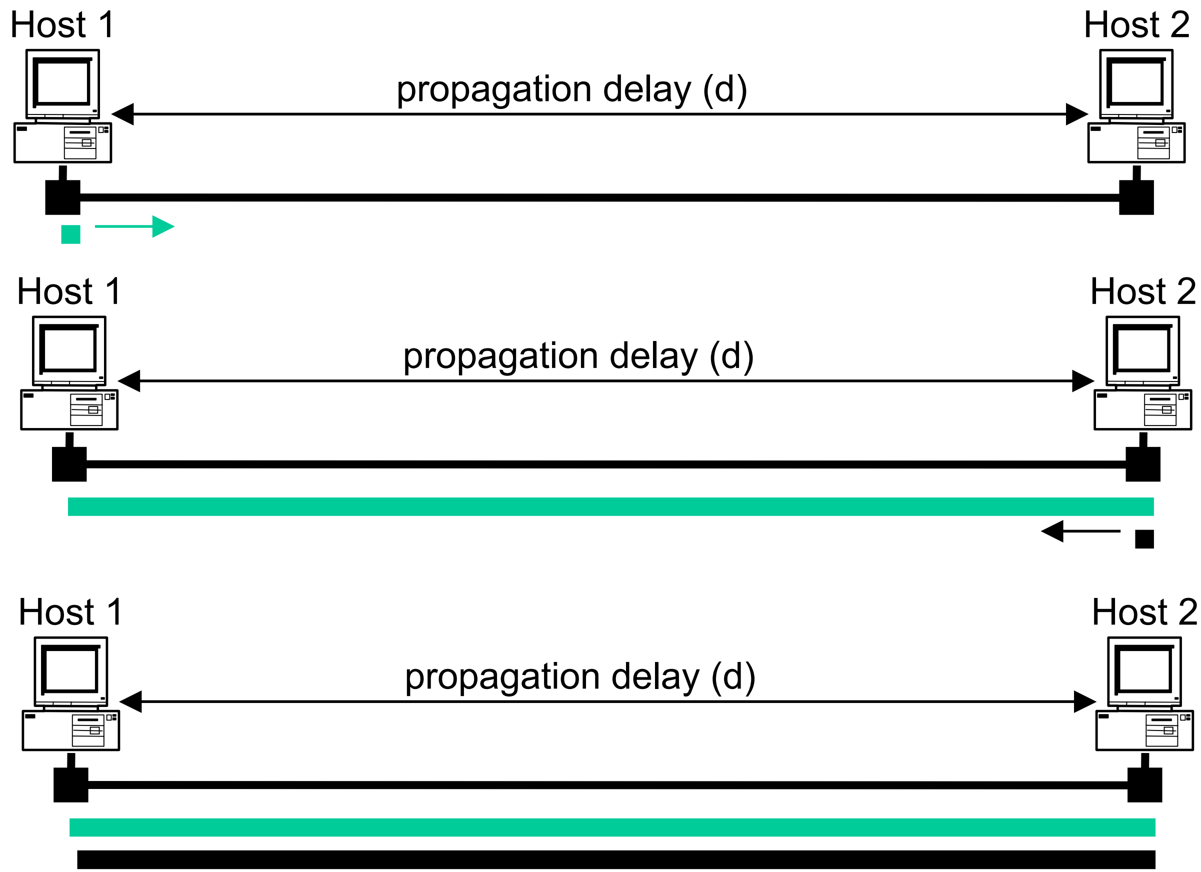


Minimum Packet Size

a) Time = t; Host 1 starts to send frame

b) Time = t + d; Host 2 starts to send a frame just before it hears from host 1’s frame

c) Time = t + 2*d; Host 1 hears Host 2’s frame → detects collision



LAN length = (min_frame_size)*(light_speed)/(2*link_rate) =

= (8*64b)*(2*10^8mps)/(2*10^7 bps) = 5.12 km

Summary: Minimum Packet Size

- **Why put a minimum packet size?**
- **Give a host enough time to detect collisions**
  - A host should be able to detect collision before it finishes the transmission of a packet
- **In Ethernet, minimum packet size = 64 bytes (two 6-byte addresses, 2-byte type, 4-byte CRC, and 46 bytes of data)**
- **If host has less than 46 bytes to send, the adaptor pads (adds) bytes to make it 46 bytes**

Higher Speed Ethernet

- What need to be changed in the Ethernet protocol to make it run 100Mbps?

Hui Zhang

802.3u Fast Ethernet

- **Apply original CSMA/CD medium access protocol at 100Mbps**
- **Must change either minimum frame or maximum diameter: change diameter**
- **Requires**
  - 2 UTP5 pairs
  - 4 UTP3 pairs
  - 1 fiber pair
- **No more “shared wire” connectivity.**
  - Hubs and switches only

802.3z Gigabit Ethernet

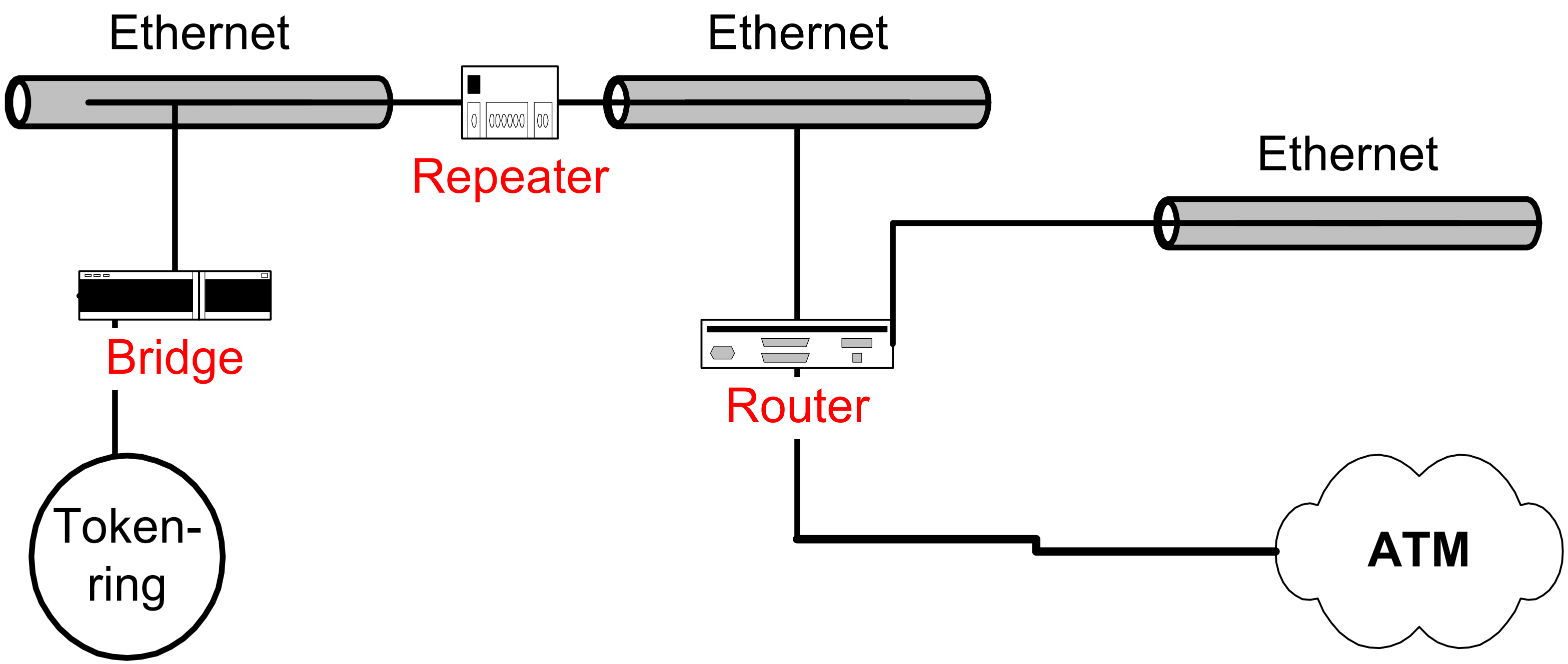
- **Same frame format and size as Ethernet.**
  - This is what makes it Ethernet
- **Full duplex point-to-point links in the backbone are likely the most common use.**
  - Added flow control to deal with congestion
- **Choice of a range of fiber and copper transmission media.**
- **Defining “jumbo frames” for higher efficiency.**

Repeater/Hub/Switch

- **Why do we need the following devices?**
  - Repeaters
  - Hubs
  - Switches
- **What are the differences between them?**

Internetworking

- There are many different devices for interconnecting networks.

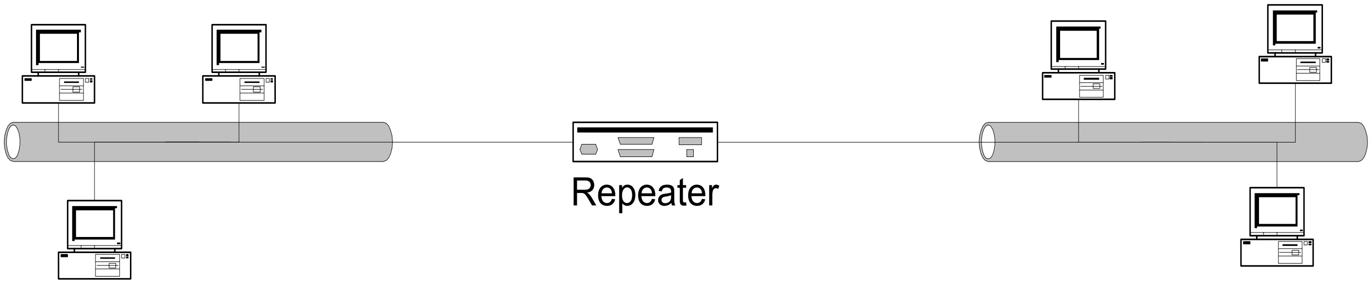


Repeaters

– Used to interconnect multiple Ethernet segments

– Merely extends the baseband cable

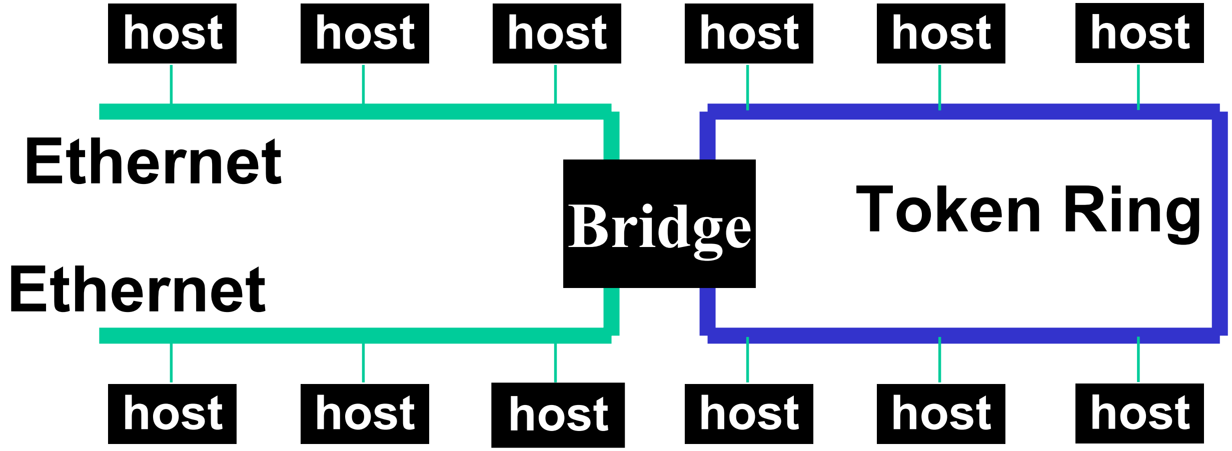
– Amplifies all signals including collisions



Hui Zhang

Building Larger LANs: Bridges

- **Bridges connect multiple IEEE 802 LANs at layer 2.**
  - Only forward packets to the right port
  - Reduce collision domain compared with single LAN
- **In contrast, hubs rebroadcast packets.**
- **Implications:**
  - Performance
  - Distance
  - Type of networks



Ethernet Switches

- **Bridges make it possible to increase LAN capacity.**
  - Packets are no longer broadcasted - they are only forwarded on selected links
- **Ethernet switch is a special case of a bridge: each bridge port is connected to a single host.**
  - Can make the link full duplex (really simple protocol!)
    - Full duplex vs. half duplex vs simplex
    - What are the original CSMA/CD Ethernet?
  - Simplifies the protocol and hardware used (only two stations on the link) – no longer full CSMA/CD
  - Can have different port speeds on the same switch
    - Unlike in a hub, packets can be stored (what is the benefit?)
    - An alternative is to use cut through switching (what is the benefit?)

More Efficient Encoding

- **Ethernet: Manchester**
- **Fast Ethernet: 4B/5B**
  - Borrowed from FDDI
- **Gigabit Ethernet: 8B/10B**
  - Borrowed from HIPPI
- **10Giga Ethernet: 64/66B**
  - Borrowed from SONET
- **Issues**
  - Clock recovery
  - DC balance
  - Clock rate vs. bit rate

Longer Distance Ethernet

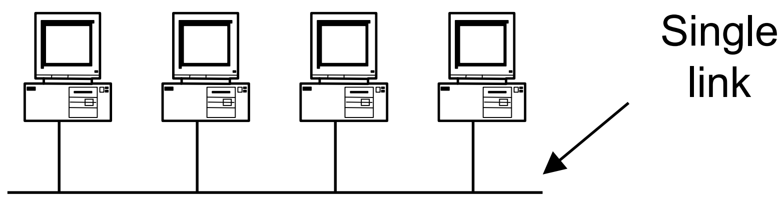
- GbE and 10GbE can run over one of the wavelength of a DWDM link, whose distance can be extended by optical amplification
- Can we have a national Ethernet network?

Wireless Ethernet (802.11)

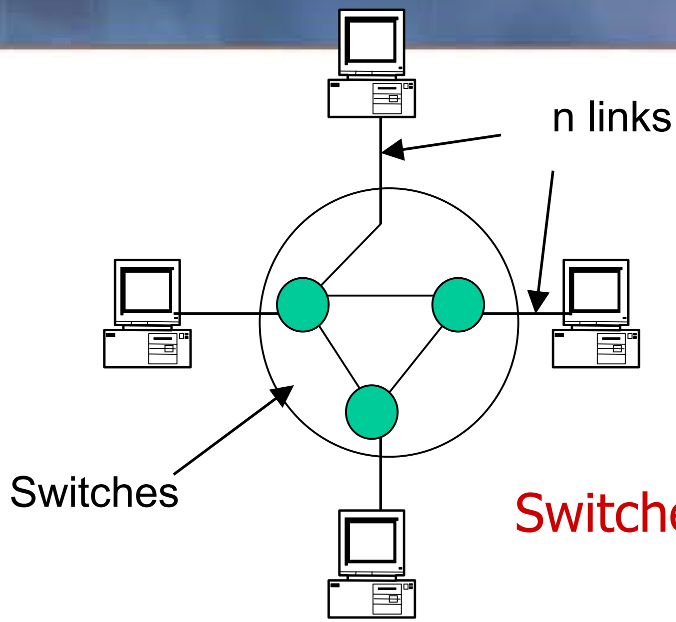
- **Why CSMA/CD is used in Ethernet but not in 802.11?**

Switching: Why?

- **Direct vs. Switched Networks:**



Direct



Switched

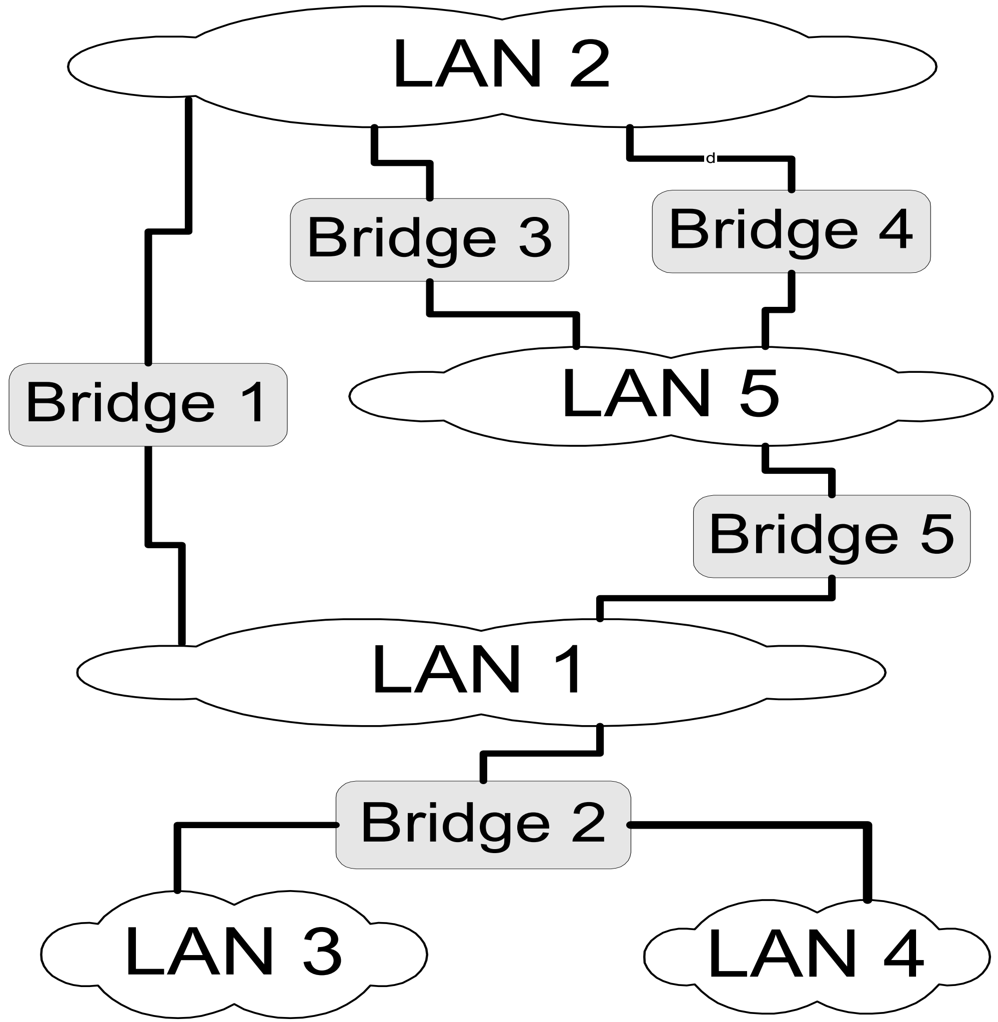
- **Direct Network Limitations:**
  - Distance (coordination delay; propagation limitation)
  - Number of hosts (collisions; shared bandwidth; address tables)
  - Single link technology (cannot mix optical, wireless, …)

- **Internetworking**
  - Distance
  - Performance
  - Multiple types of links and networks

Hui Zhang

Big Ethernet Network

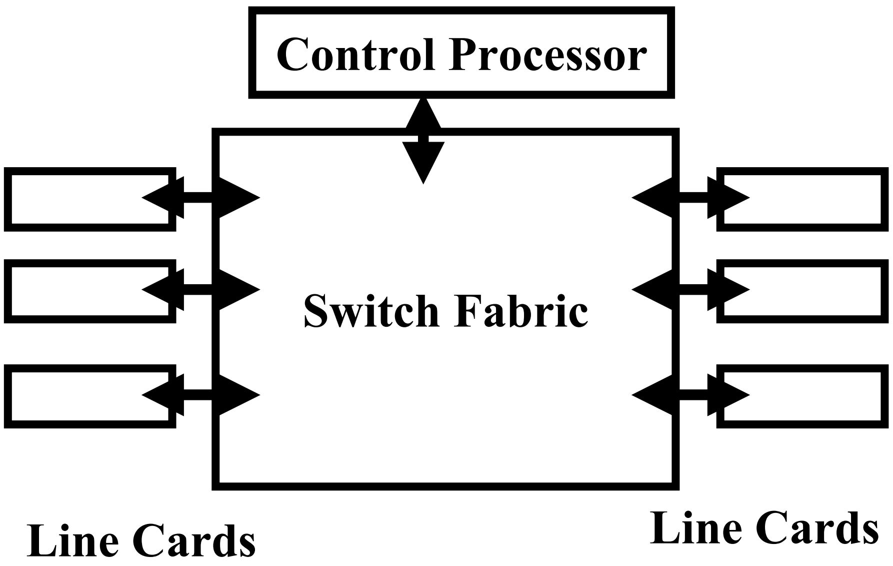
- **With switching and optical Ethernet**
  - Distance is no longer an issue
  - How big Ethernet network can we build?
- **“big” in terms of**
  - Geographical reach
  - Number of hosts



Scalability Issues with Switched Ethernet

- **Broadcast**
  - Still broadcast in switched Ethernet network?
- **Size of forwarding table**
- **Not being able to utilize all links**
  - Only links on spanning tree can forward packets

Structure of A Generic Communication Switch



- **Switches**
  - circuit switch
  - Ethernet switch
  - ATM switch
  - IP router

- **Switch fabric**
  - high capacity interconnect
- **Line card**
  - address lookup in the data path (forwarding)
- **Control Processor**
  - load the forwarding table (routing or signaling)

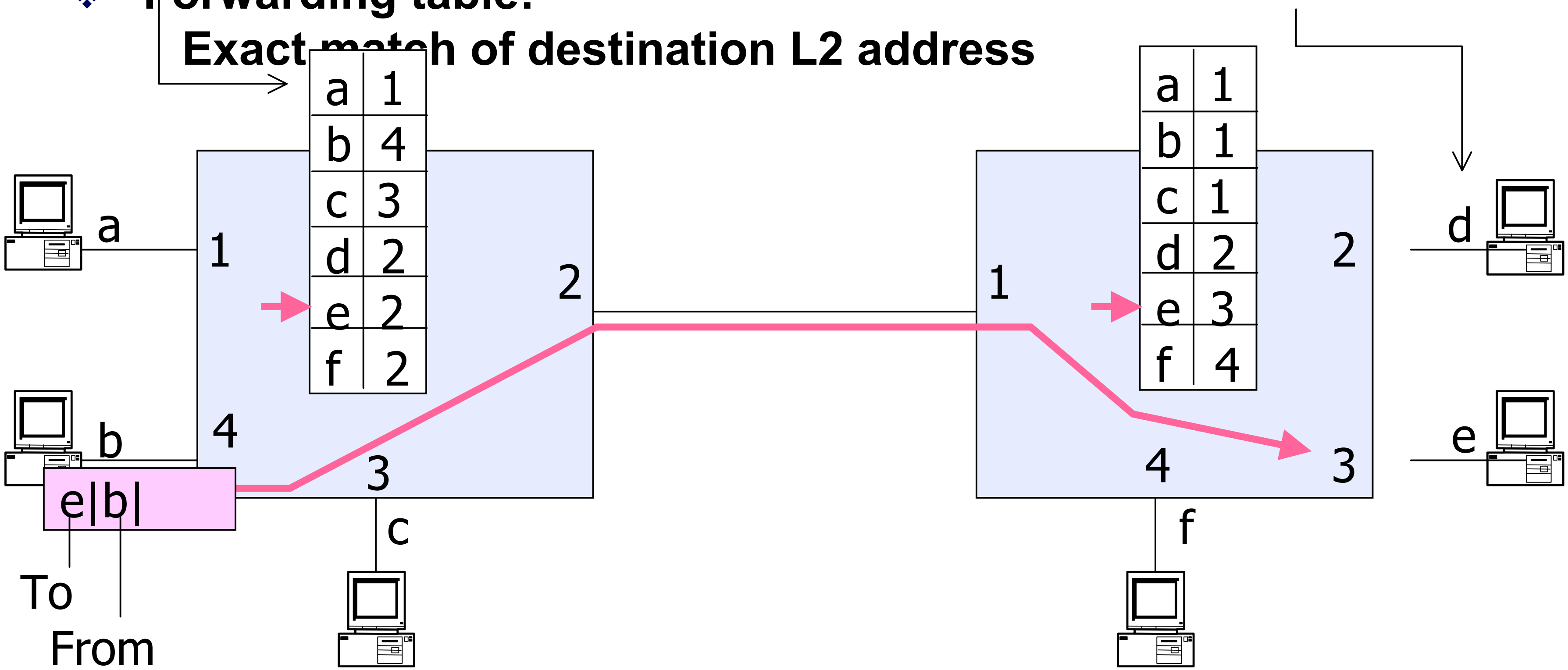
Addressing and Look-up

- **Flat address**
  - Ethernet: 48 bit MAC address
  - ATM: 28 bit VPI/VCI
  - DS-0: timeslot location
- **Limited scalability**
- **High speed lookup**

- **Hierarchical address**
  - IP <network>.<subnet>.<host>
  - Telephone: country.area.home
- **Scalable**
- **Easy lookup if boundary is fixed**
  - telephony
- **Difficult lookup if boundary is flexible**
  - longest prefix match for IP

Datagram: Layer 2 (e.g., Ethernet)

- Flat address space (no structure)
- Forwarding table:
  Exact match of destination L2 address



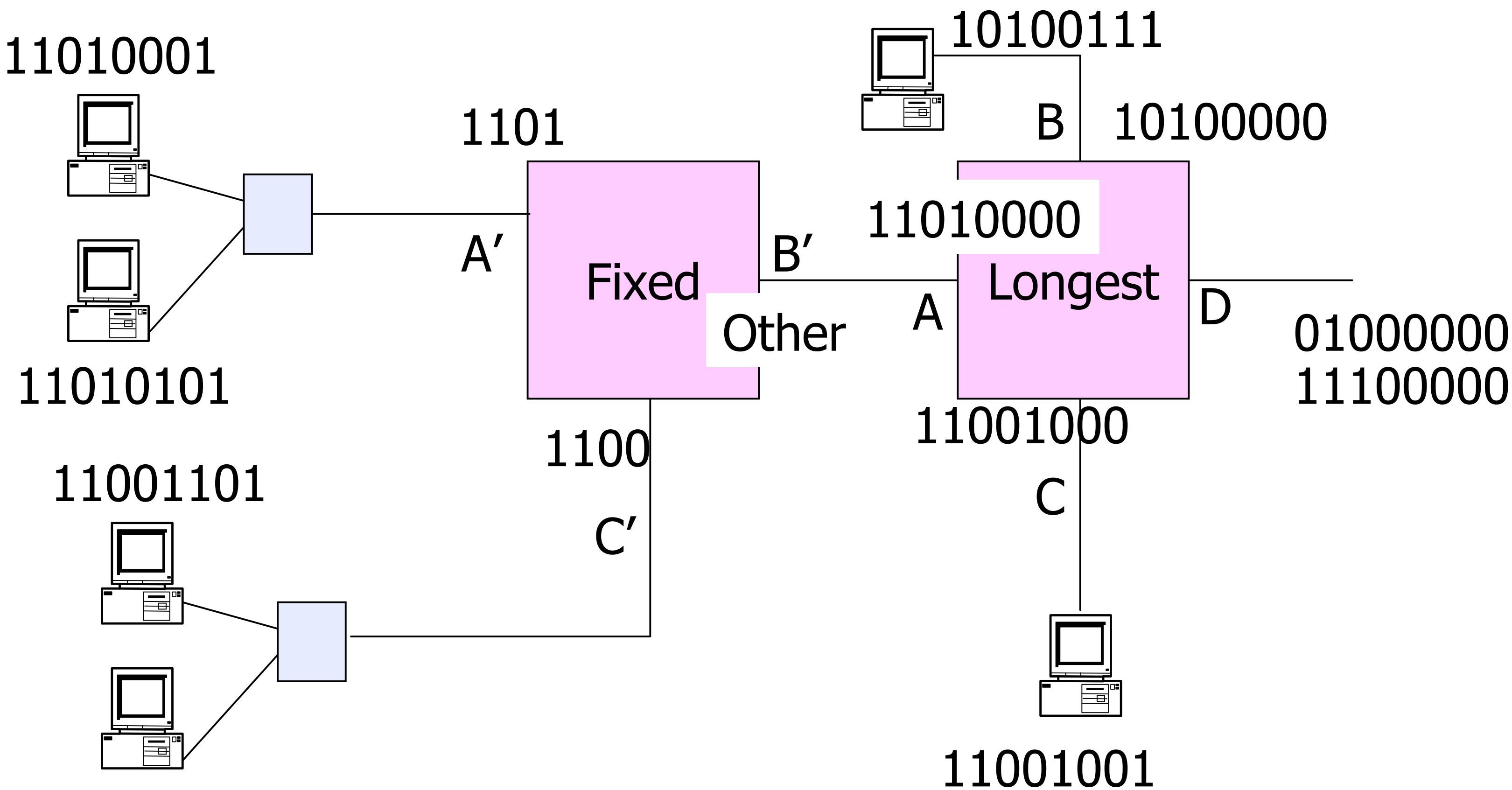
Datagram: Layer 3 (e.g., IP)

E.g. 1

E.g. 2

E.g. 3

- **L3-network (e.g., IP)**
  - Topological structure – match prefix
  - Either fixed prefix length or longest match



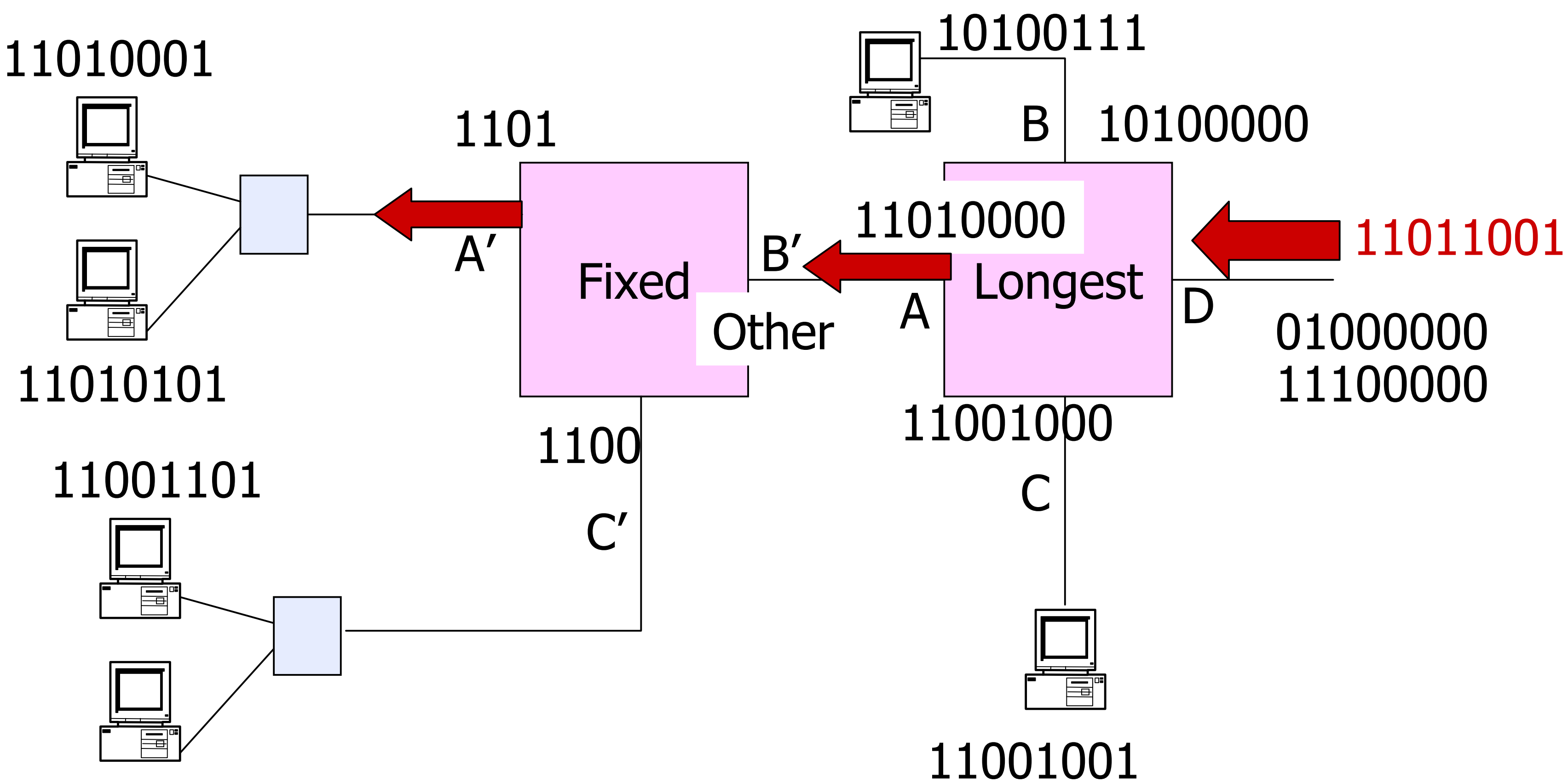
Datagram: Layer 3 (e.g., IP)

E.g. 1

11011001 matches

4 bits at A, 1 at B, 3 at C → A = LPM

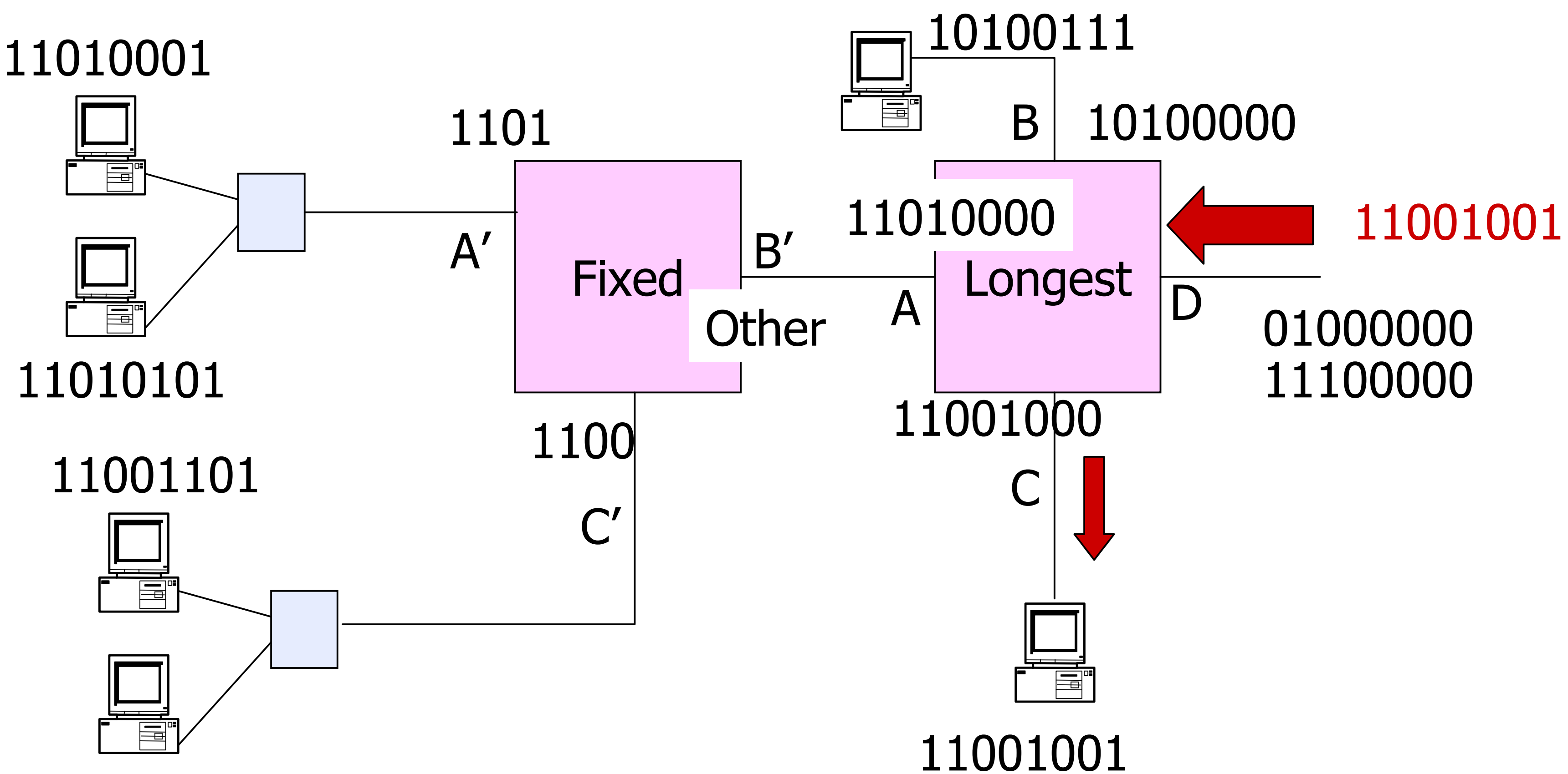
4 bits at A′ → A′ = EM



Datagram: Layer 3 (e.g., IP)

E.g. 2

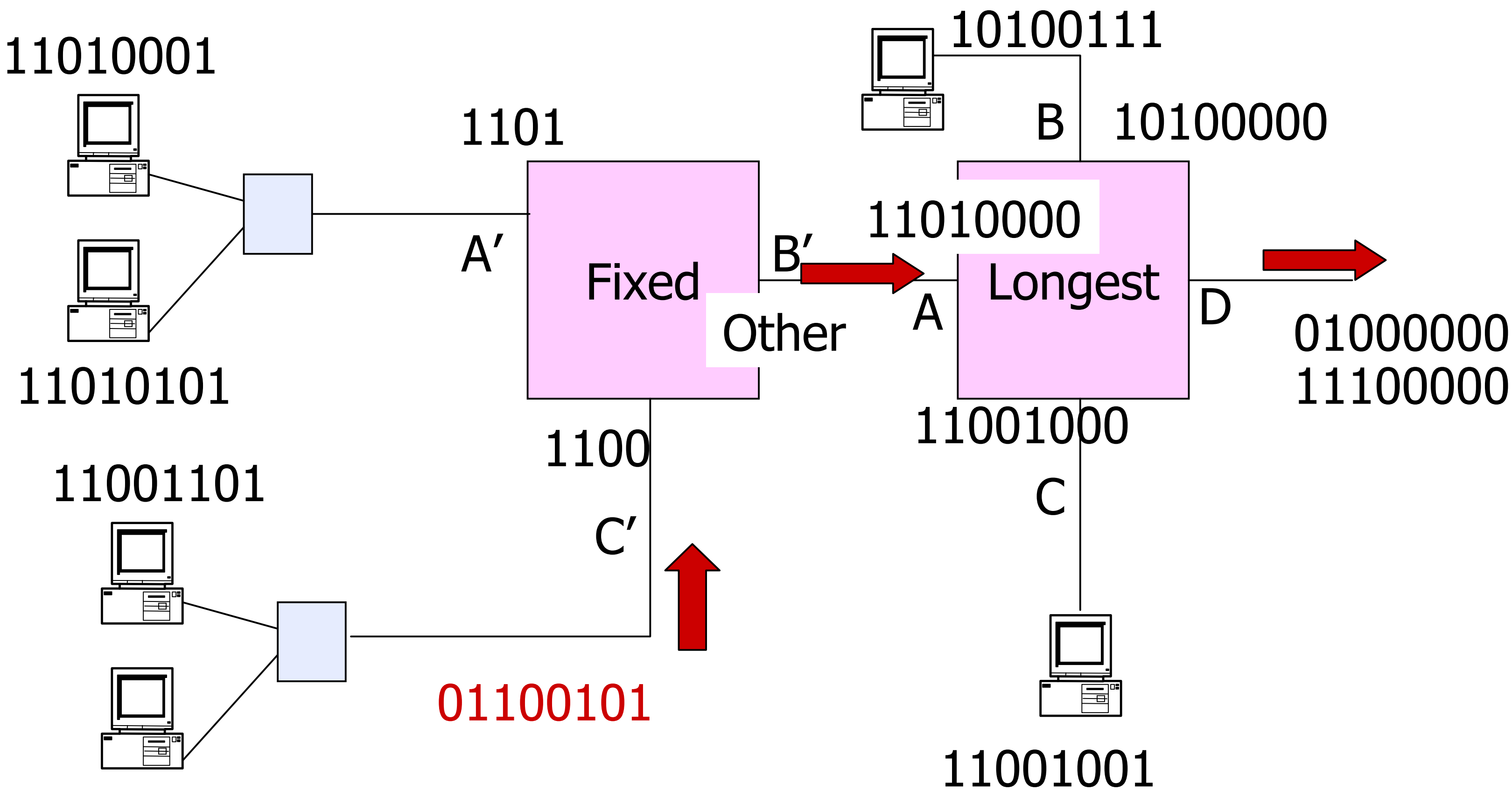
11001001 matches
3 bits at A, 1 at B, 7 at C → C = LPM



Datagram: Layer 3 (e.g., IP)

01100101 matches
- 0 bit at A′ → B′
- 0 bit at B, 0 at C, 2 at D → D = LPM

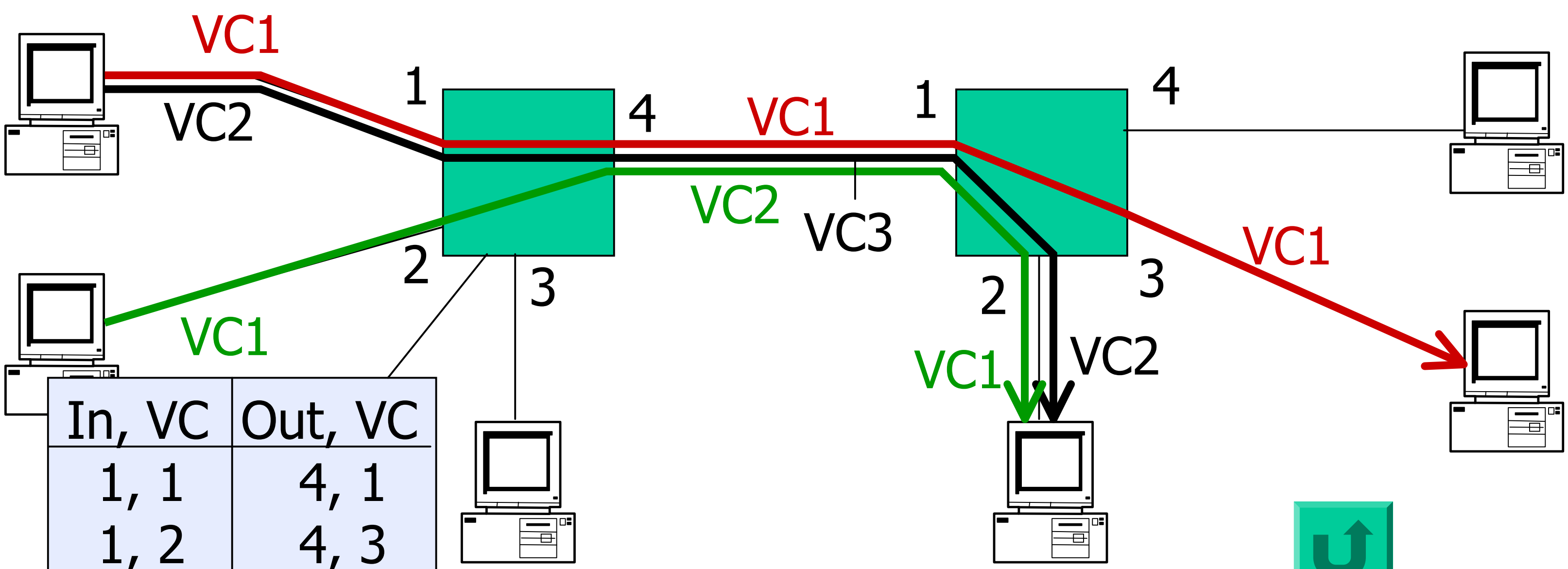
E.g. 3



Virtual Circuit

Connection setup establishes a path through switches

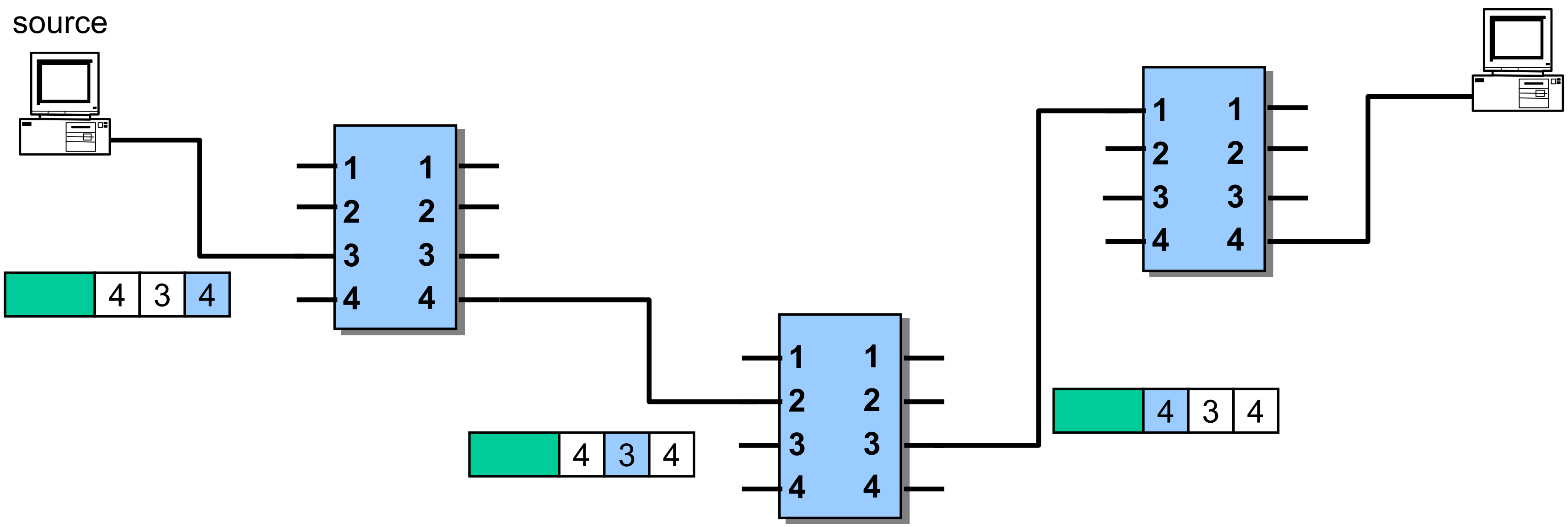
- A virtual circuit ID (VCI) identifies path
- Uses packet switching, with packets containing VCI
- VCIs are often indices into per-switch connection tables; change at each hop



In, VC	Out, VC
1, 1	4, 1
1, 2	4, 3
2, 1	4, 2

Techniques: Source Routing

Each packet specifies the sequence of routers (or of output ports) from source to destination



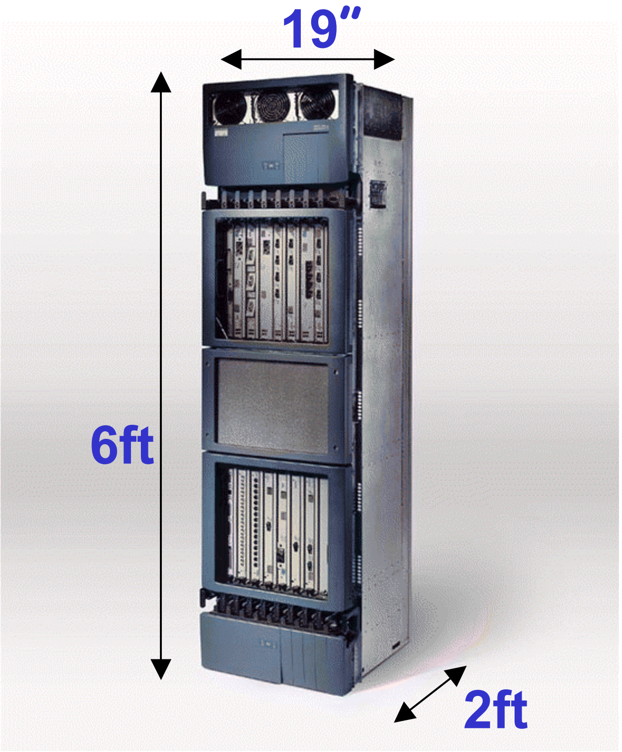
Control Plane

- **Installing entries in forwarding tables at all switches**
  - Forwarding tables need to be consistent so that packets go to the right destination
- **The Brain of a network**
- **Difficulties**
  - Topology keeps on changing
  - Distributed computation
- **Control plane in different networks**
  - Ethernet: address learning and spanning tree
  - Circuit network: routing + signaling
  - IP network: routing

Examples: Cisco GSR - 12416

- WAN Router – Large throughput; SONET links
- Up to 16 line cards at 10 Gbps each
- Crossbar Fabric
- Line Cards:
  1-port OC-192c
  4-port OC48c
  Many others
  (ATM, Ethernet, …)

Cisco GSR 12416



Examples: Juniper M160

- WAN Router – Large throughput; SONET links
- Crossbar Fabric
- Line Cards:
  1-port OC-192c
  4-port OC48c
  Many others
  (ATM, Ethernet, …)

Juniper M160

Capacity:
80Gb/s
Power: 2.6kW

Examples: Cisco 7600

- MAN-WAN Router
- Up to 128 Gbps with Crossbar Fabric
- 10Mbps – 10Gbps LAN Interfaces
- OC-3 to OC-48 SONET Interfaces
- MPLS, WFQ, LLQ, WRED, Traffic Shaping

Examples: Cisco Catalyst 6500

- From LAN to Access
- 48 to 576 10/100 Ethernet Interfaces
- 10 GE, OC-3, OC-12, OC-48, ATM
- QoS, ACL
- Load Balancing; VPN
- Up to 128Gbps (with crossbar)
- L4-7 Switching
- VLAN
- IP Telephony (E1, T1, inline-power Ethernet)
- SNMP, RMON

Examples: Extreme - Summit

- 48 10/100 ports
- 2 GE (SX, LX, or LX-70)
- 17.5Gbps non-blocking
- 10.1 Mpps
- Wire speed L2
- Wire speed L3 static or RIP
- OSPF, DVRMP, PIM, …

Hui Zhang

Examples: Foundry - ServerIron

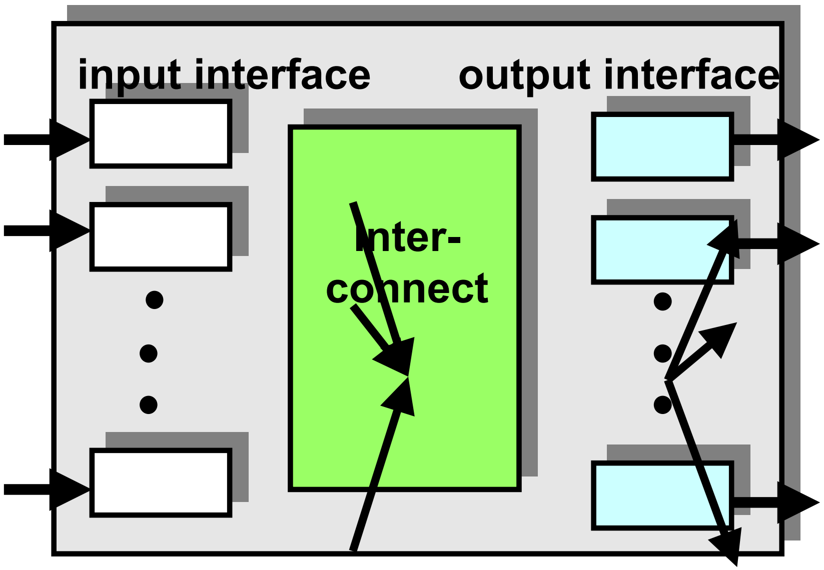
- Server Load Balancing
- Transparent Cache Switching
- Firewall Load Balancing
- Global Server Load Balancing
- Extended Layer 4-7 functionality including URL-, Cookie-, and SSL Session ID-based switching
- Secure Network Address Translation (NAT) and
  Port address

Switching: Characteristics

- **Ports**
  - Fast Ethernet, OC-3, ATM, …
- **Protocols**
  - ST, Link Agg., VLAN, OSPF, RIP, BGP, VPN, Load Balancing, WRED, WFQ
- **Performance**
  - Throughput: slot throughput, box throughput
    - Bits per second, packets per second
  - Latency (where does it come from?)
  - Reliability
  - Power consumption

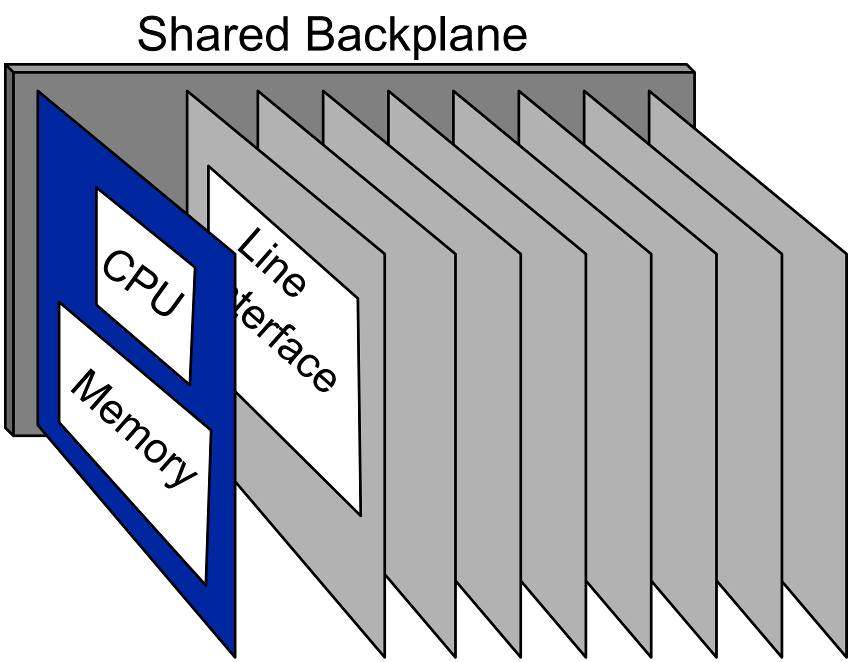
Architectures: Generic

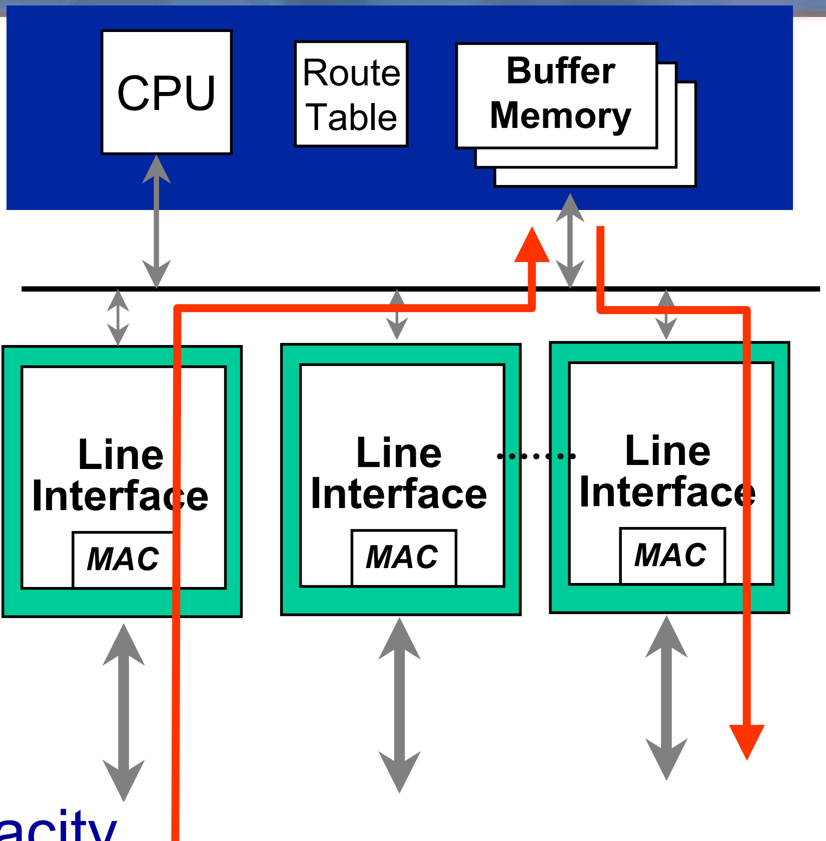
- **Input and output interfaces are connected through an interconnect**
- **A interconnect can be implemented by**
  - Shared memory
    - low capacity routers (e.g., PC-based routers)
  - Shared bus
    - Medium capacity routers
  - Point-to-point (switched) bus
    - High capacity routers



Hui Zhang

Architectures: First Generation



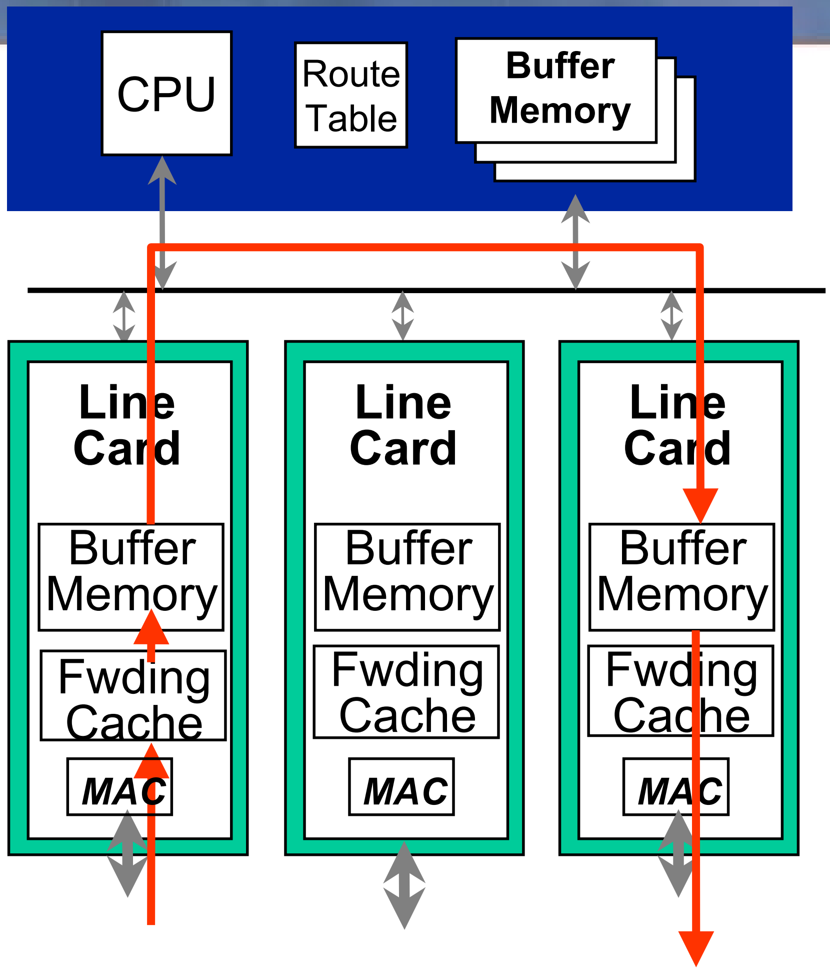


Typically < 0.5Gbps aggregate capacity

Limited by rate of shared memory

Slide by Nick McKeown

Architectures: Second Generation

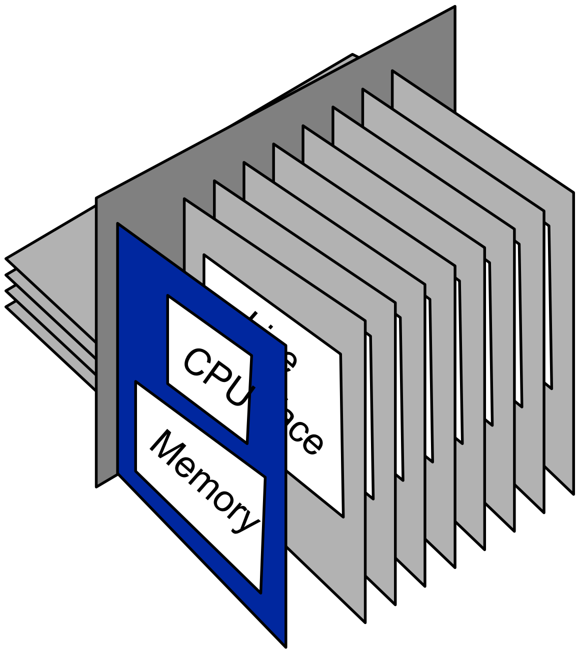


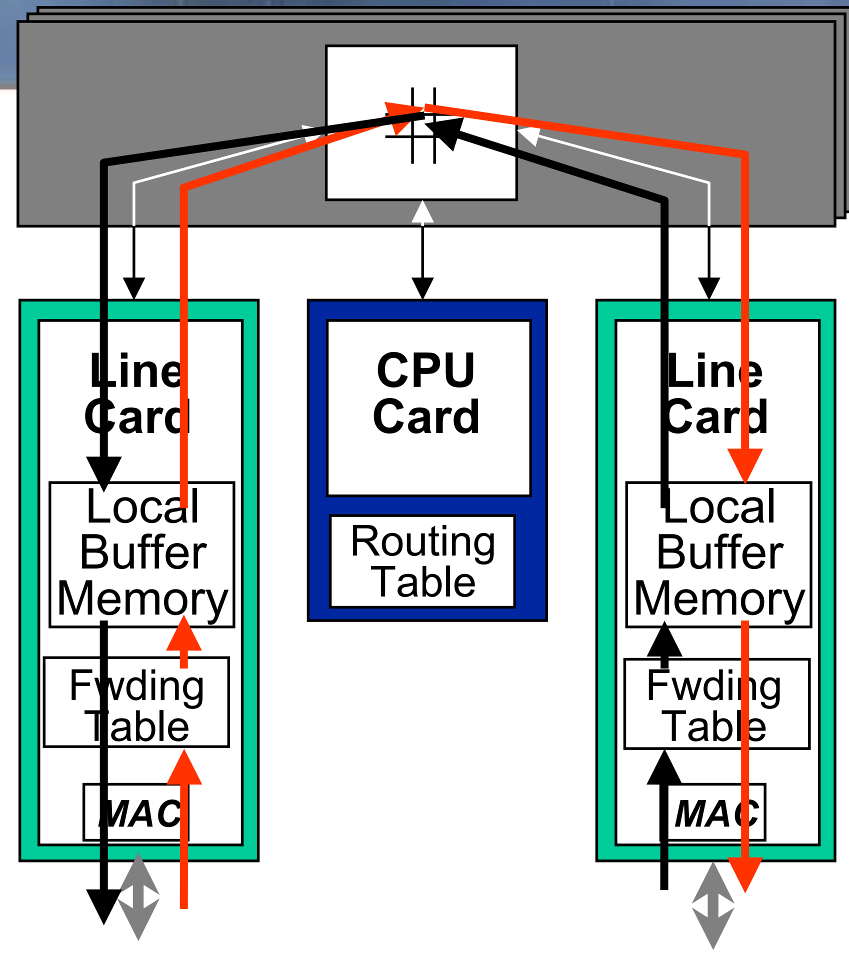
Typically
  < 5Gb/s aggregate capacity
Limited by shared bus

Slide by Nick McKeown

Architectures: Third Generation

Switched Backplane





Typically
< 50Gbps aggregate capacity

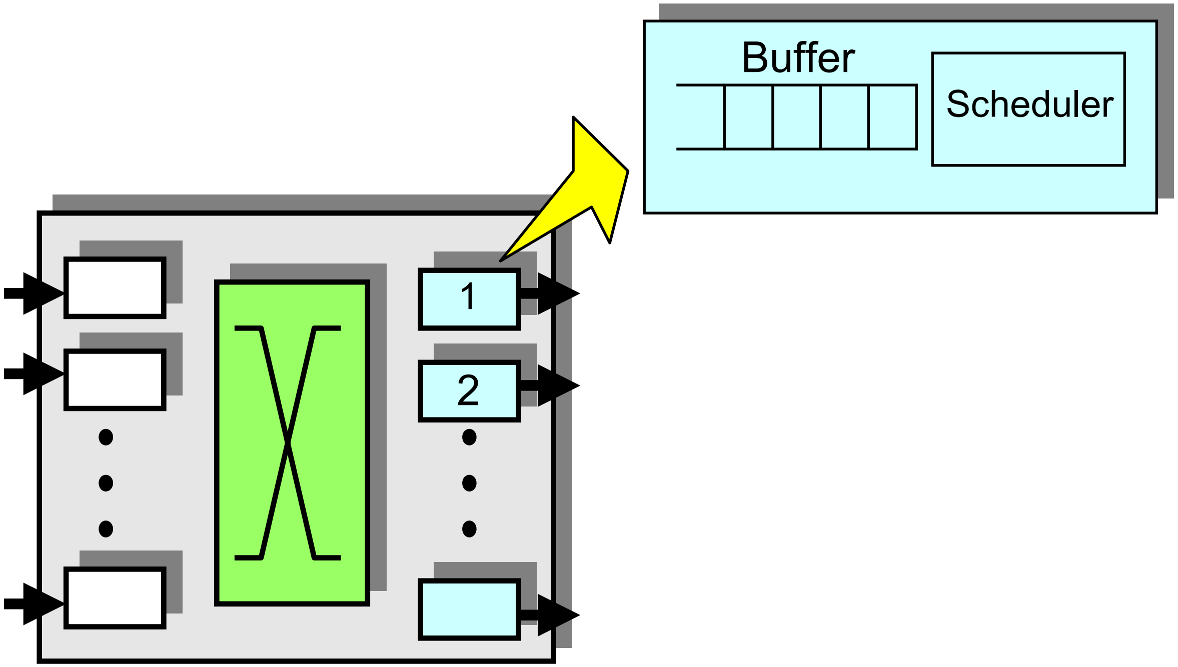
Slide by Nick McKeown

Architectures: Input Functions

- **Packet forwarding: decide to which output interface to forward each packet based on the information in packet header**
  - examine packet header
  - lookup in forwarding table
  - update packet header

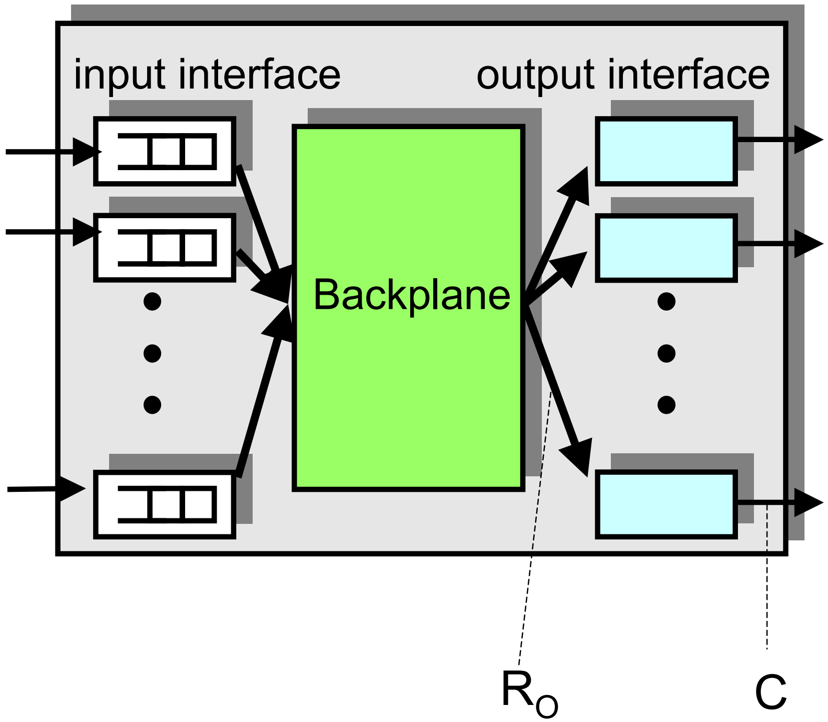
Architectures: Output Functions

- **Buffer management:** decide when and which packet to drop
- **Scheduler:** decide when and which packet to transmit



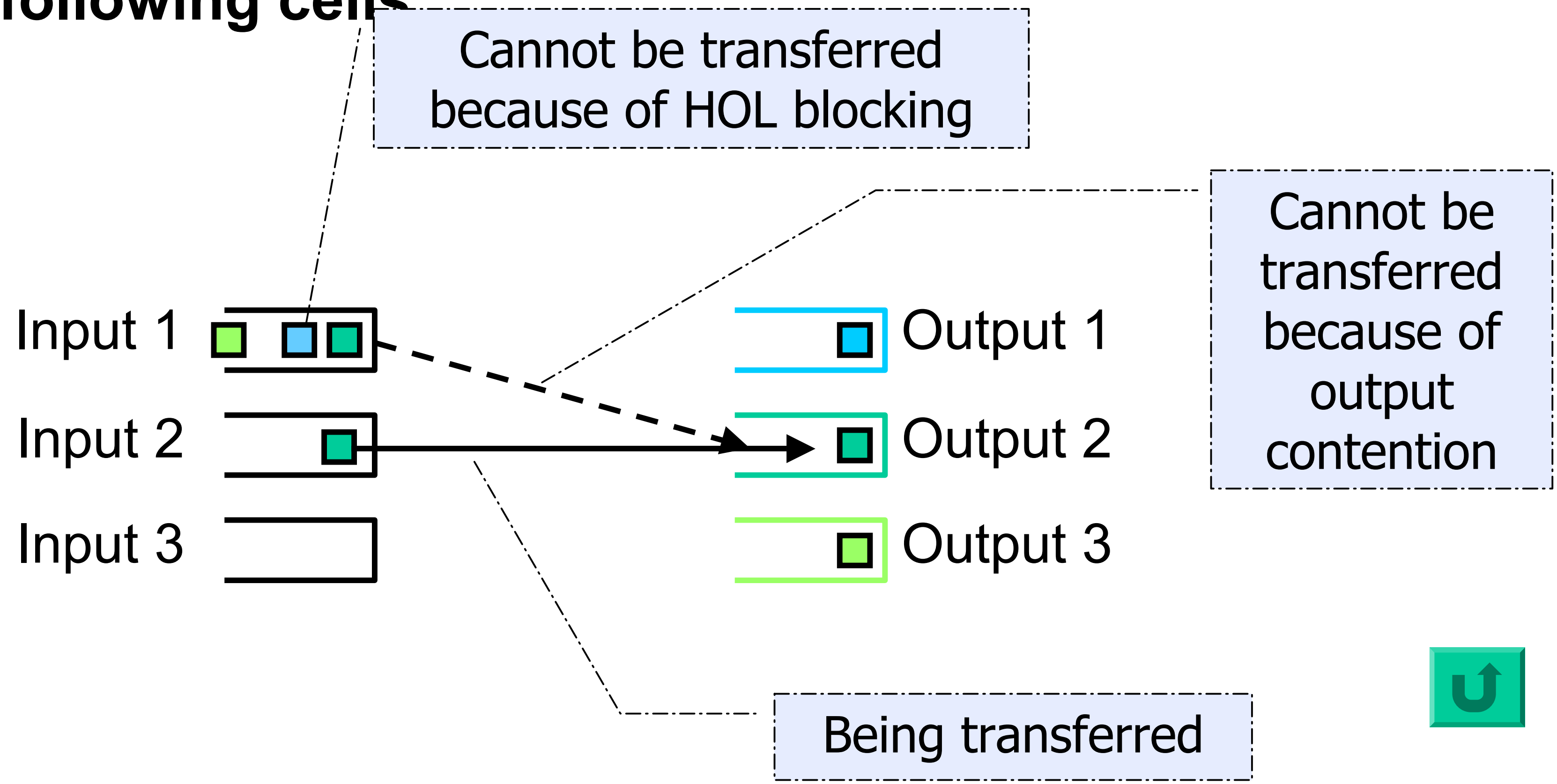
Input Queues

- **Only input interfaces store packets**
- **Advantages**
  - Easy to build
  - Simple algorithms
- **Disadvantages**
  - HOL blocking



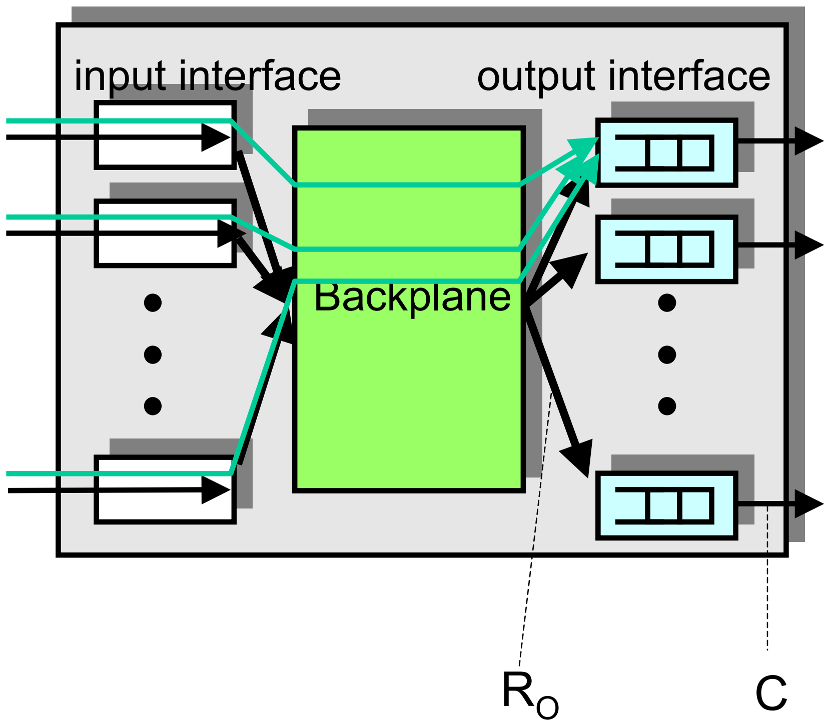
Head-of-line Blocking

- **Buffering at input: packet cell at the head of an input queue cannot be transferred, thus blocking the following cells**



Architectures: Output Queued

- **Only output interfaces store packets**
- **Advantage**
  - Easy to design algorithms: only one congestion point
- **Disadvantage**
  - Requires an output speedup Ro/C = N, where N is the number of interfaces → not feasible for large N



Architectures: Virtual Output Buffers

- **OUT buffers at each input port**
  - Complexity: Matching Problem

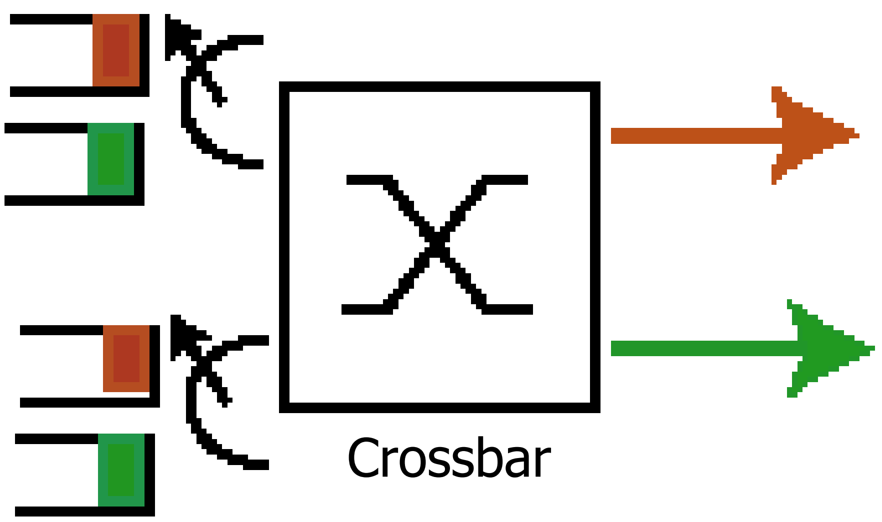
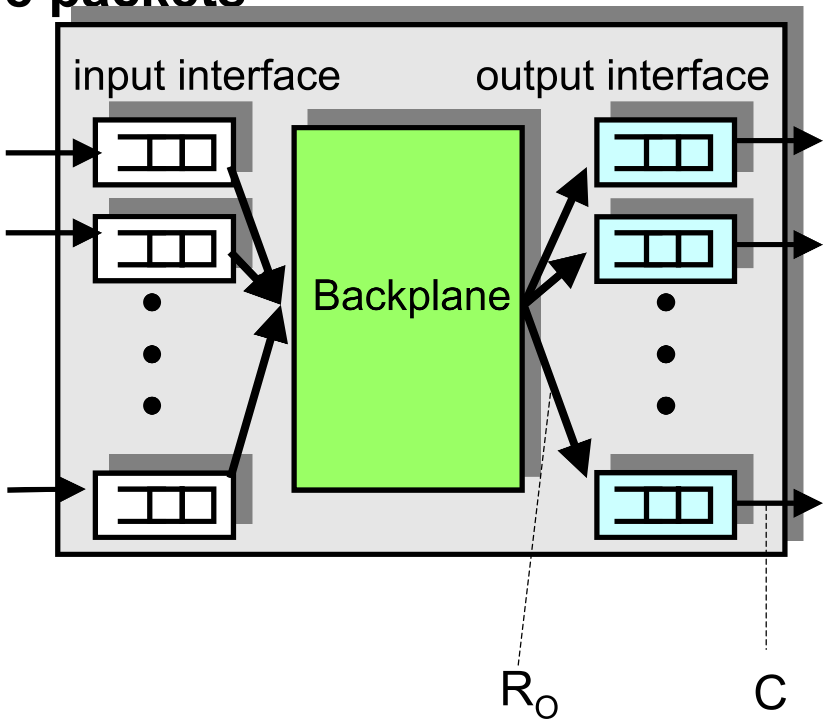


Figure from Prof. Varaiya's notes

Architectures: Combined IN/OUT

- **Both input and output interfaces store packets**
- **Advantages**
  - Easy to built
    - – Utilization 1 can be achieved with limited input/output speedup (<= 2)
- **Disadvantages**
  - Harder to design algorithms
    - – Two congestion points
    - – Need to design flow control



Hui Zhang

Switching: Summary

- **Switching needed for big networks**
- **Internetworking externality**
- **Circuit**
- **Packet – VC: QoS possible**
- **Packet – Datagram**
  - L2: Limited by flat address space
  - L3:
    - Exact Match: Easy lookup – less efficient
    - Longest Prefix Match
- **Switch functions: control and data**
- **Different Architectures:**
  - cost vs. performance