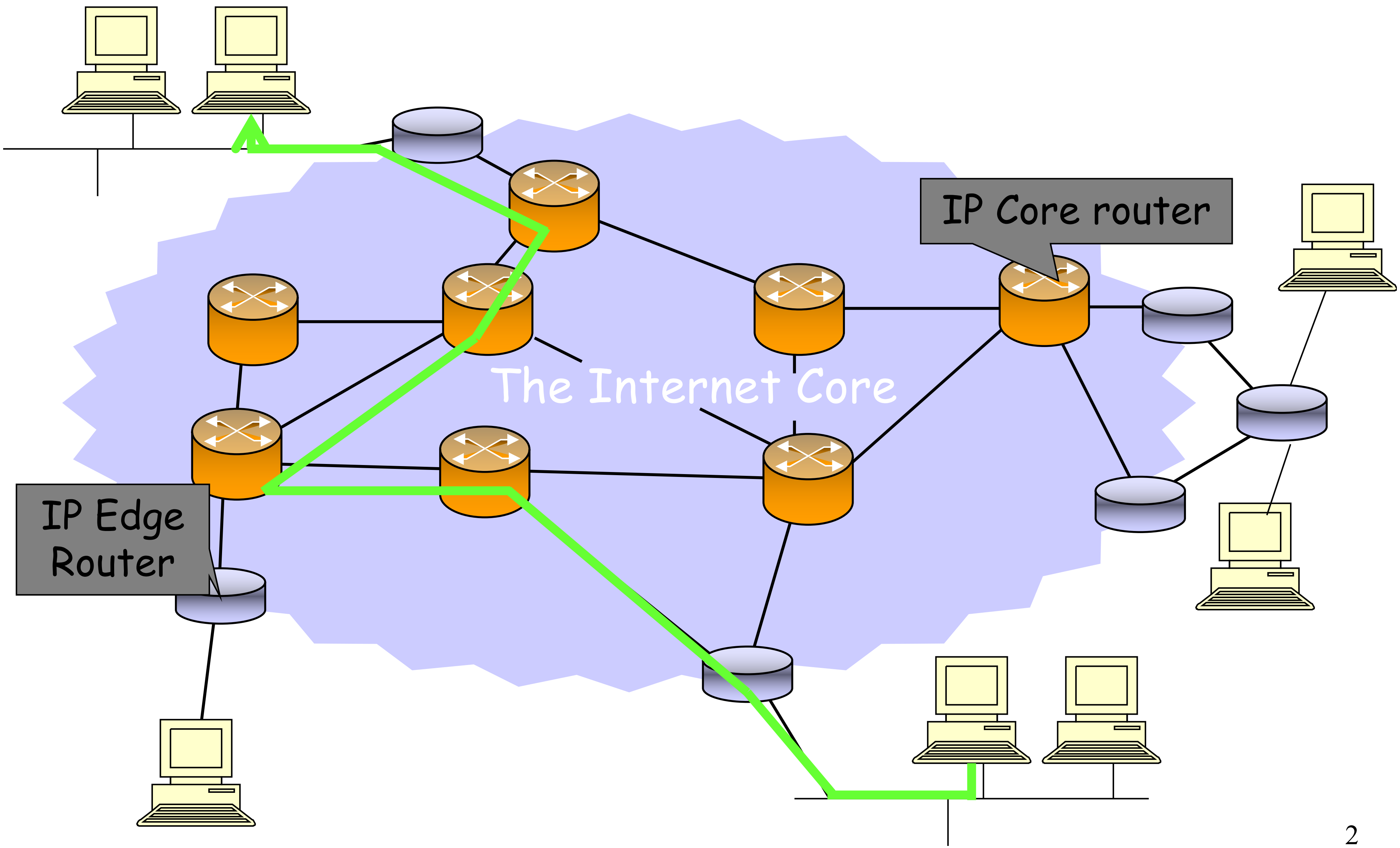
Routers Technologies & Evolution for High-Speed Networks

C. Pham
Université de Pau et des Pays de l'Adour
http://www.univ-pau.fr/~cpham
Congduc.Pham@univ-pau.fr

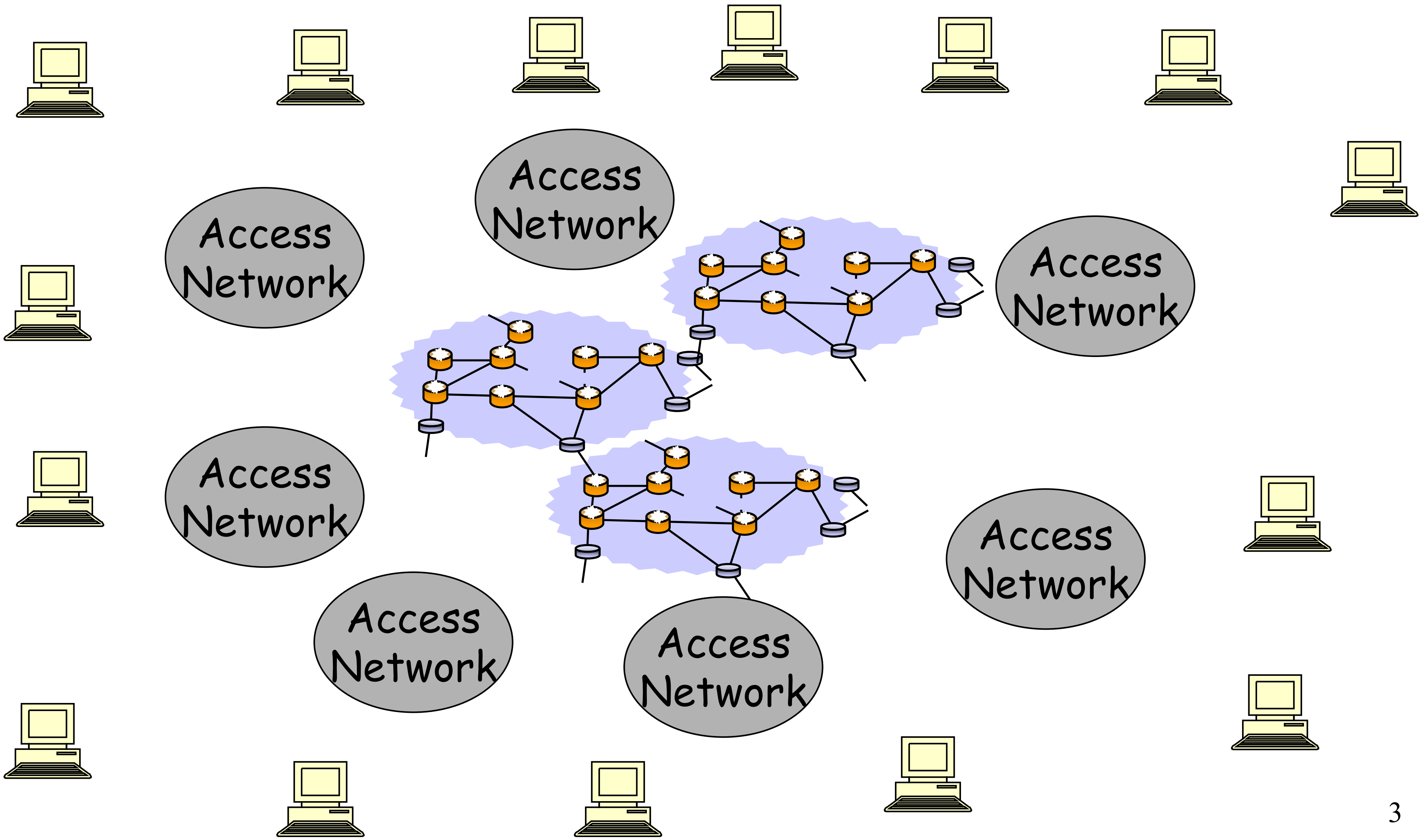
Router Evolution slides from
Nick McKeown, Pankaj Gupta

nickm@stanford.edu
www.stanford.edu/~nickm

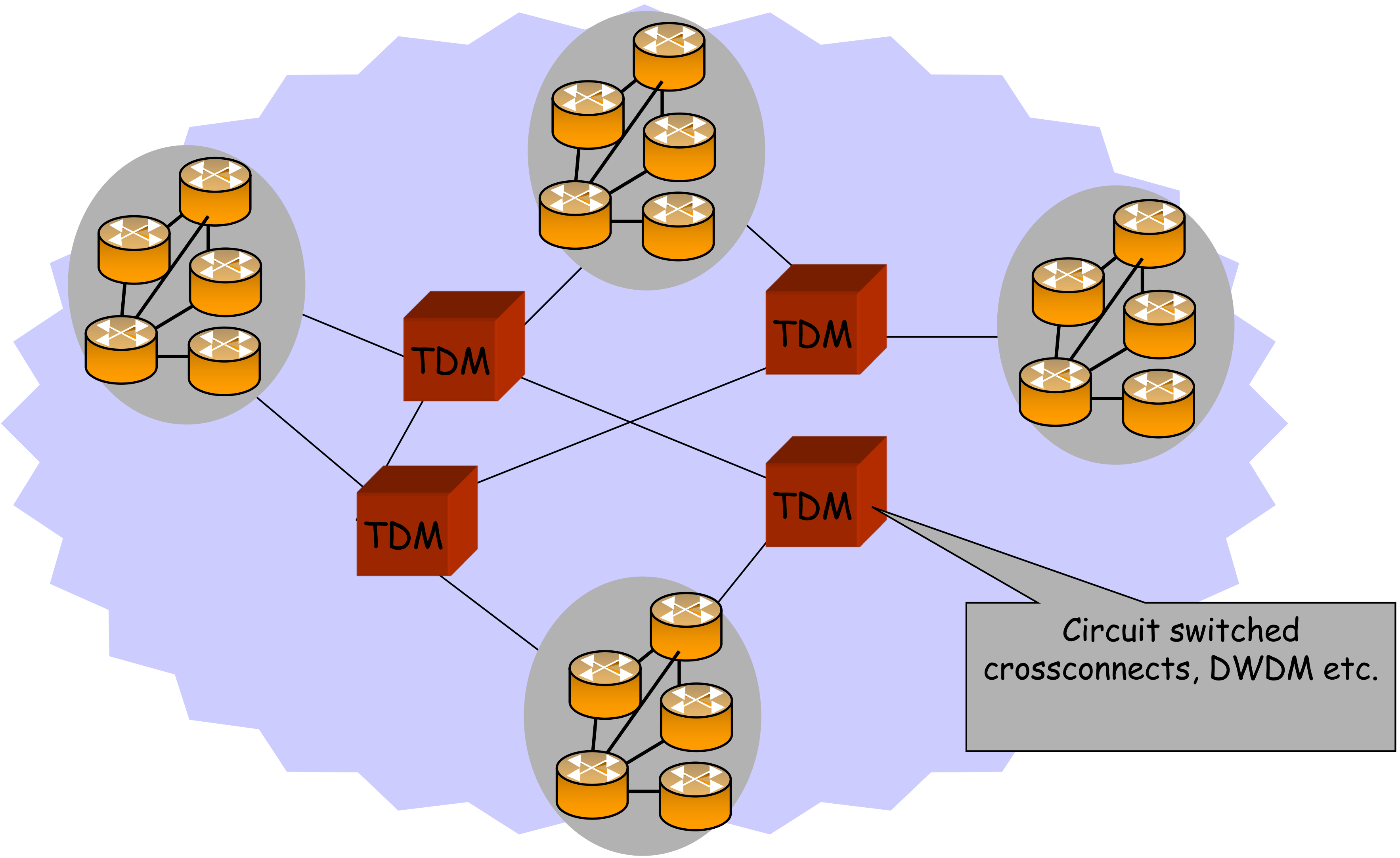
"The Internet is a mesh of routers"



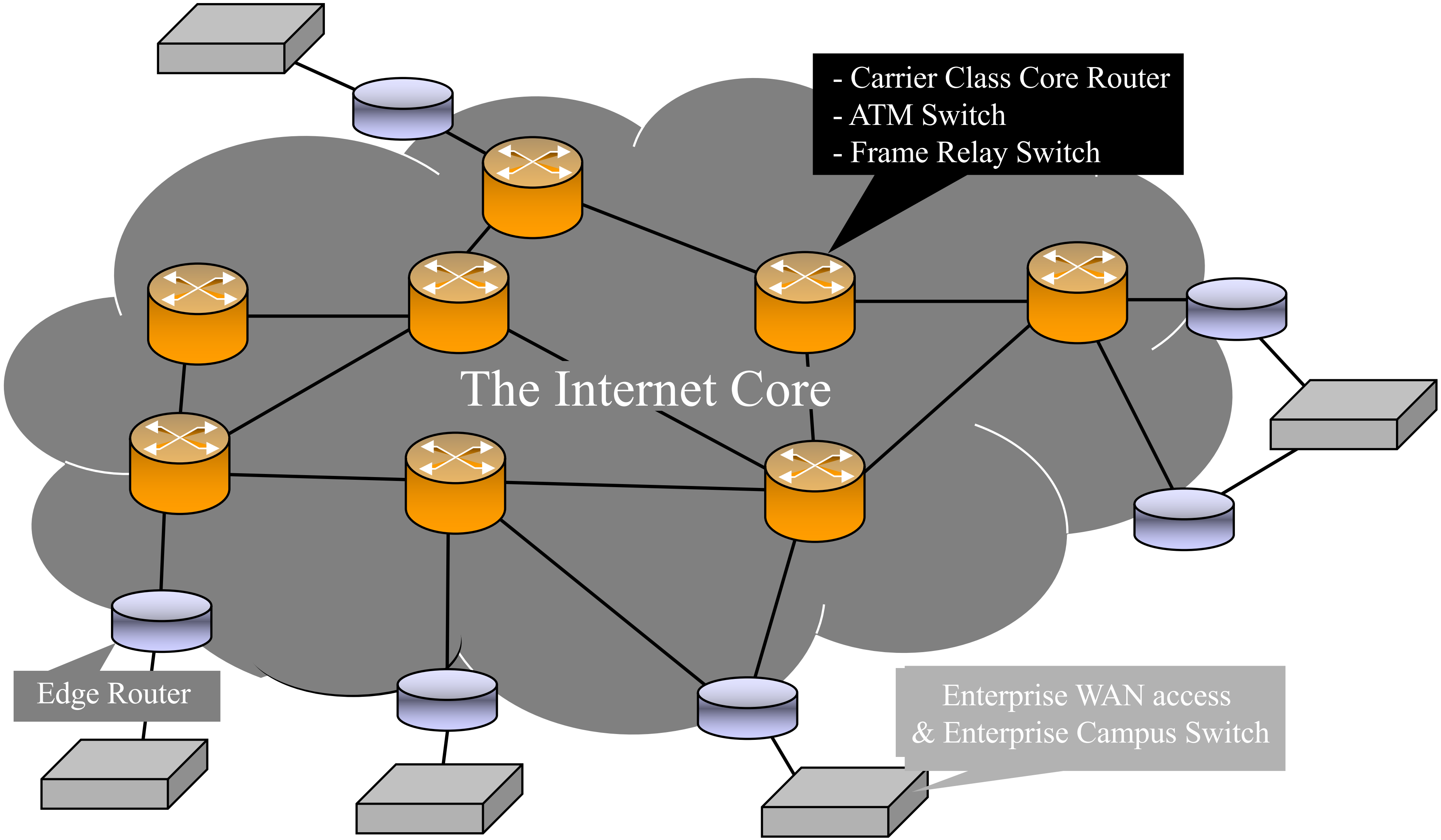
The Internet was a mesh of IP routers, ATM switches, frame relay, TDM, …



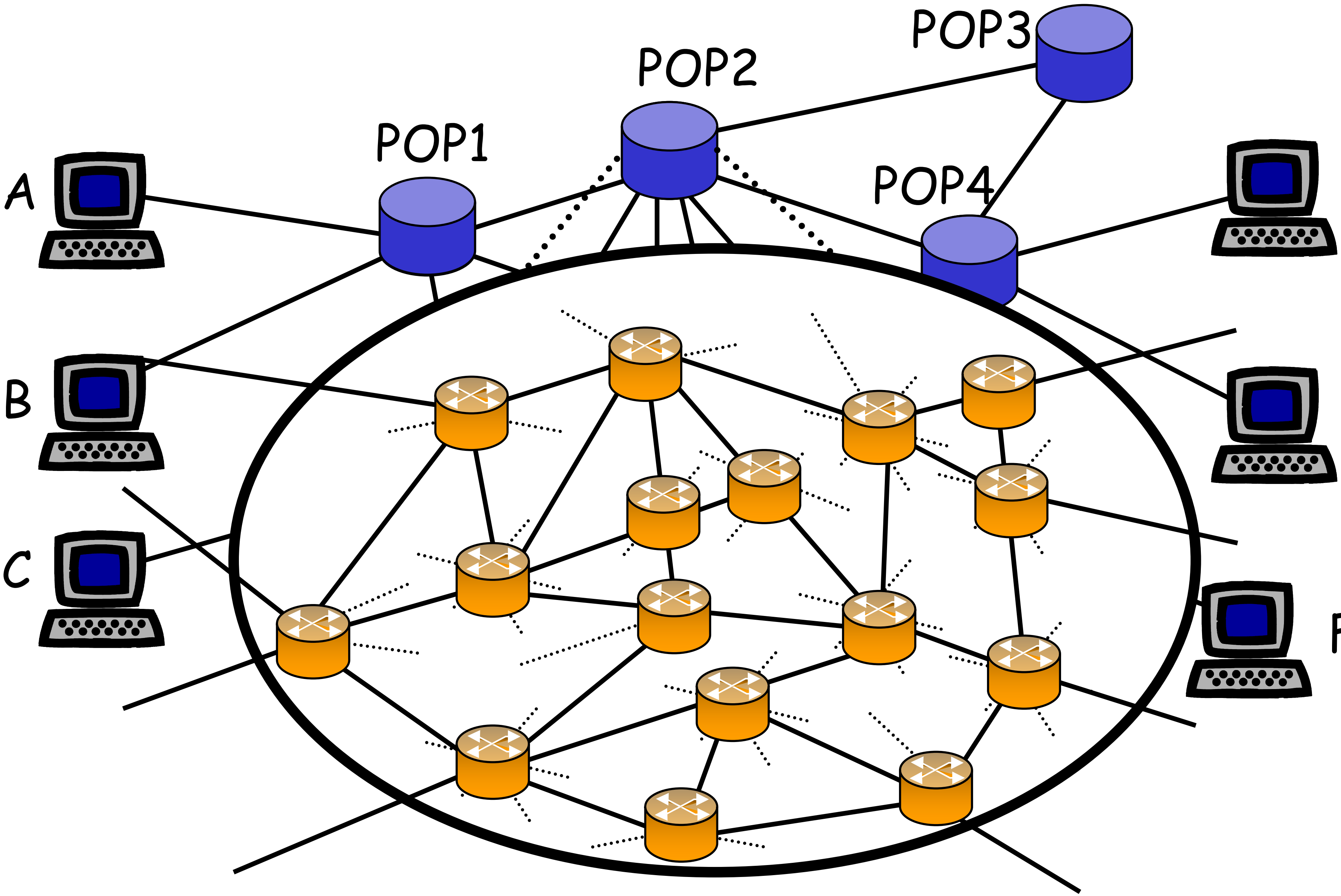
Now, the Internet is a mesh of routers mostly interconnected by SONET/SDH



Where high performance packet switches are used

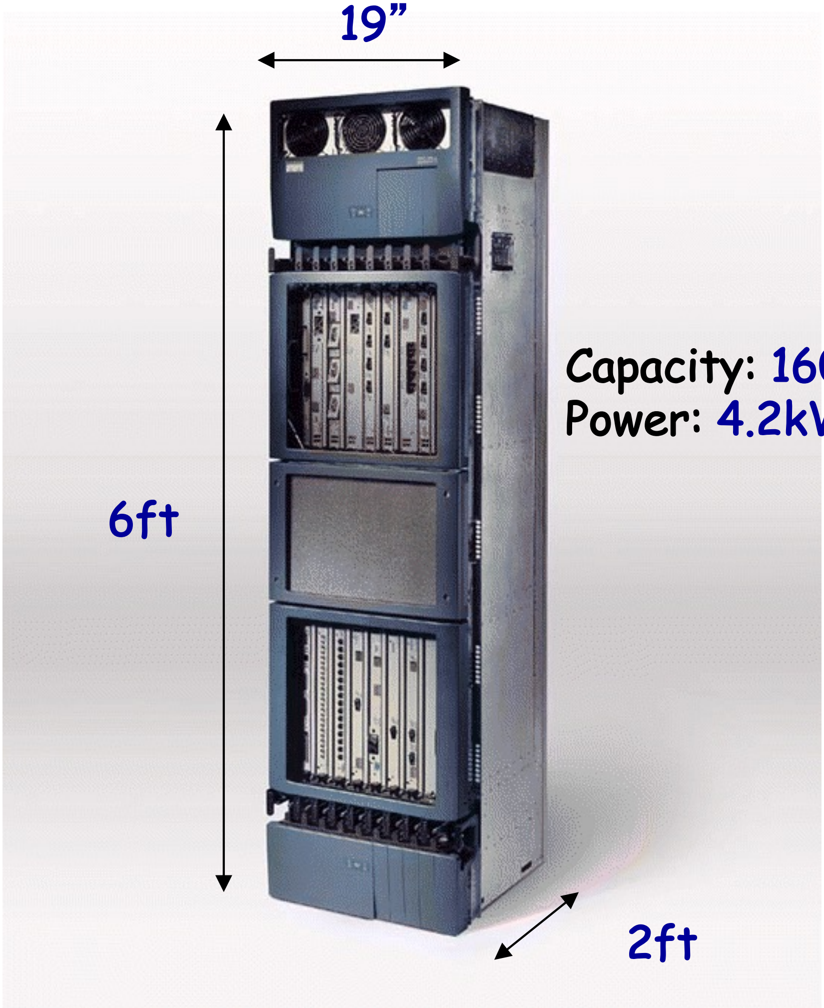


Ex: Points of Presence (POPs)



What a Router Looks Like

Cisco GSR 12416

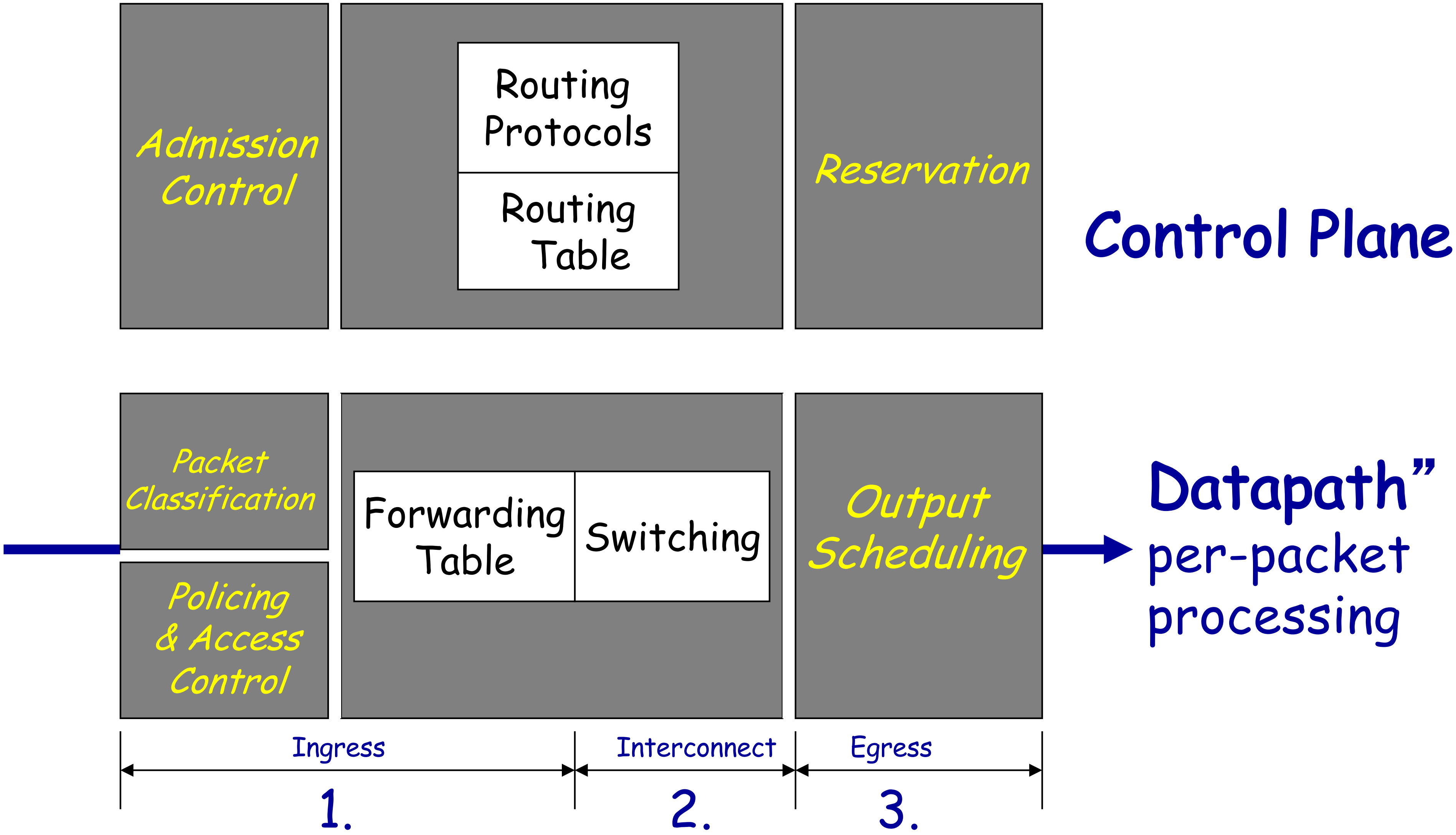


Capacity: 160Gb/s
Power: 4.2kW

Juniper M160

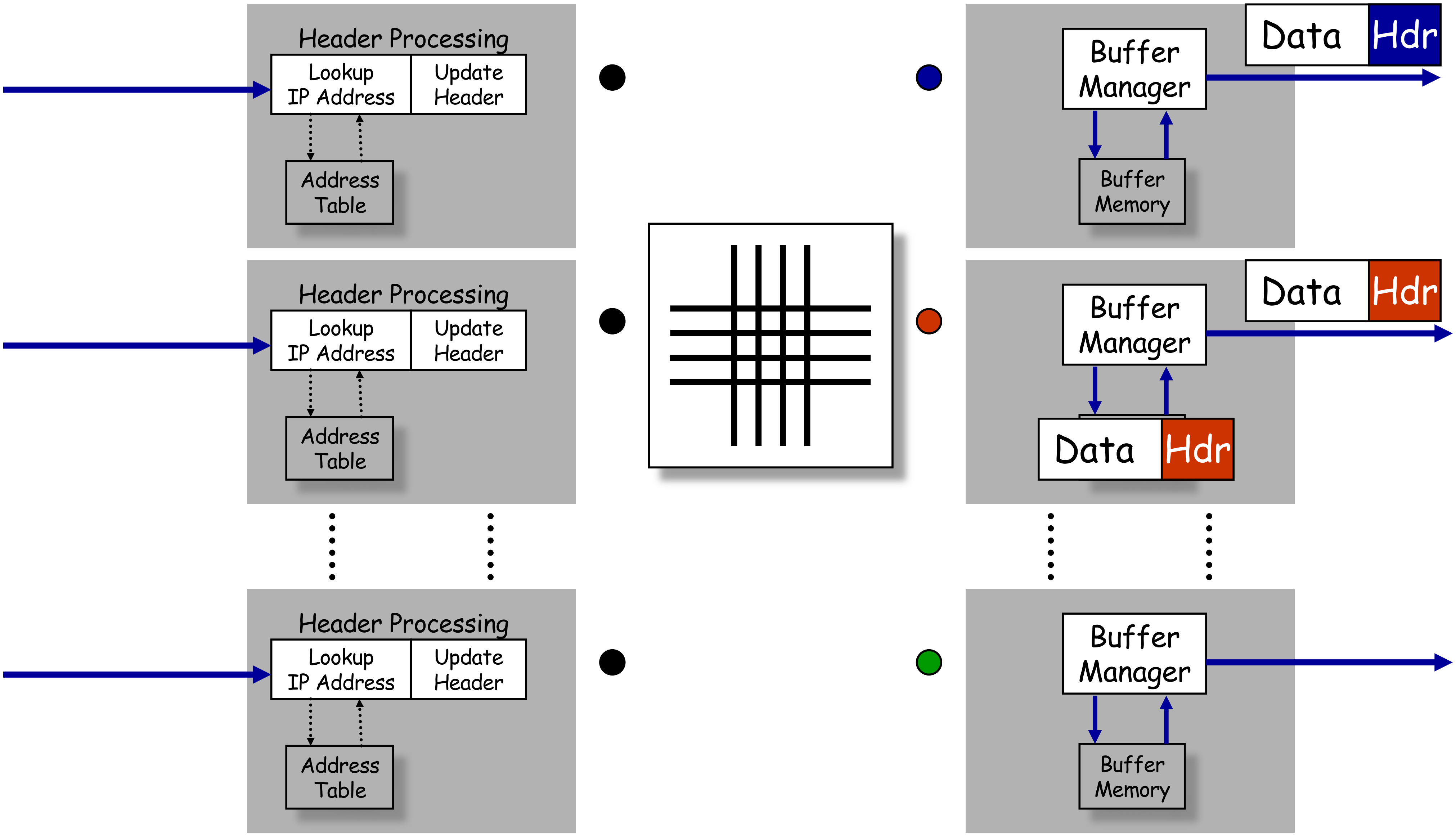
Capacity: 80Gb/s
Power: 2.6kW

Basic Architectural Components



Basic Architectural Components

Datapath: per-packet processing



Routing constraints

Year	Throughput (Gbps)	40B (Mpps)	84B (Mpps)	354B (Mpps)
1997-98	0.155	0.48	0.23	0.054
1998-99	0.622	1.94	0.92	0.22
1999-00	2.5	7.81	3.72	0.88
2000-01	10.0	31.25	14.88	3.53
2002-03	40.0	125	59.52	14.12
2010	200	625	297.6	70.6
2016	1000	3125	1488	353
GEthernet	1.0	3.13	1.49	0.35

Added by C. Pham

Flow-aware vs Flow-unaware Routers

- Flow-aware router: keeps track of flows and perform similar processing on packets in a flow
- Flow-unaware router (packet-by-packet router): treats each incoming packet individually

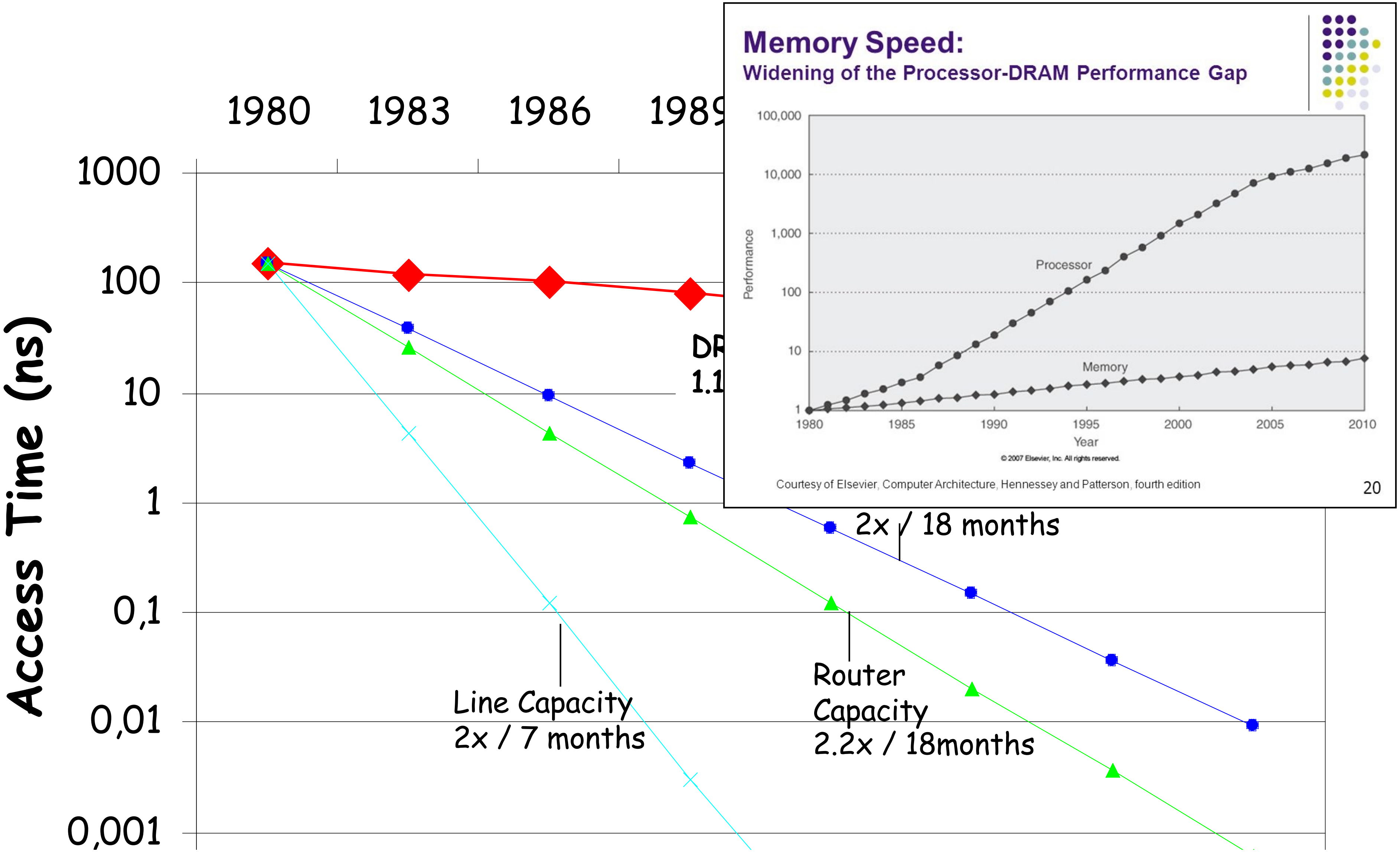
Special Processing Requires Identification of Flows

- All packets of a flow obey a pre-defined rule and are processed similarly by the router
- E.g. a flow = (src-IP-address, dst-IP-address), or a flow = (dst-IP-prefix, protocol) etc.
- Router needs to identify the flow of every incoming packet and then perform appropriate special processing

Examples of special processing

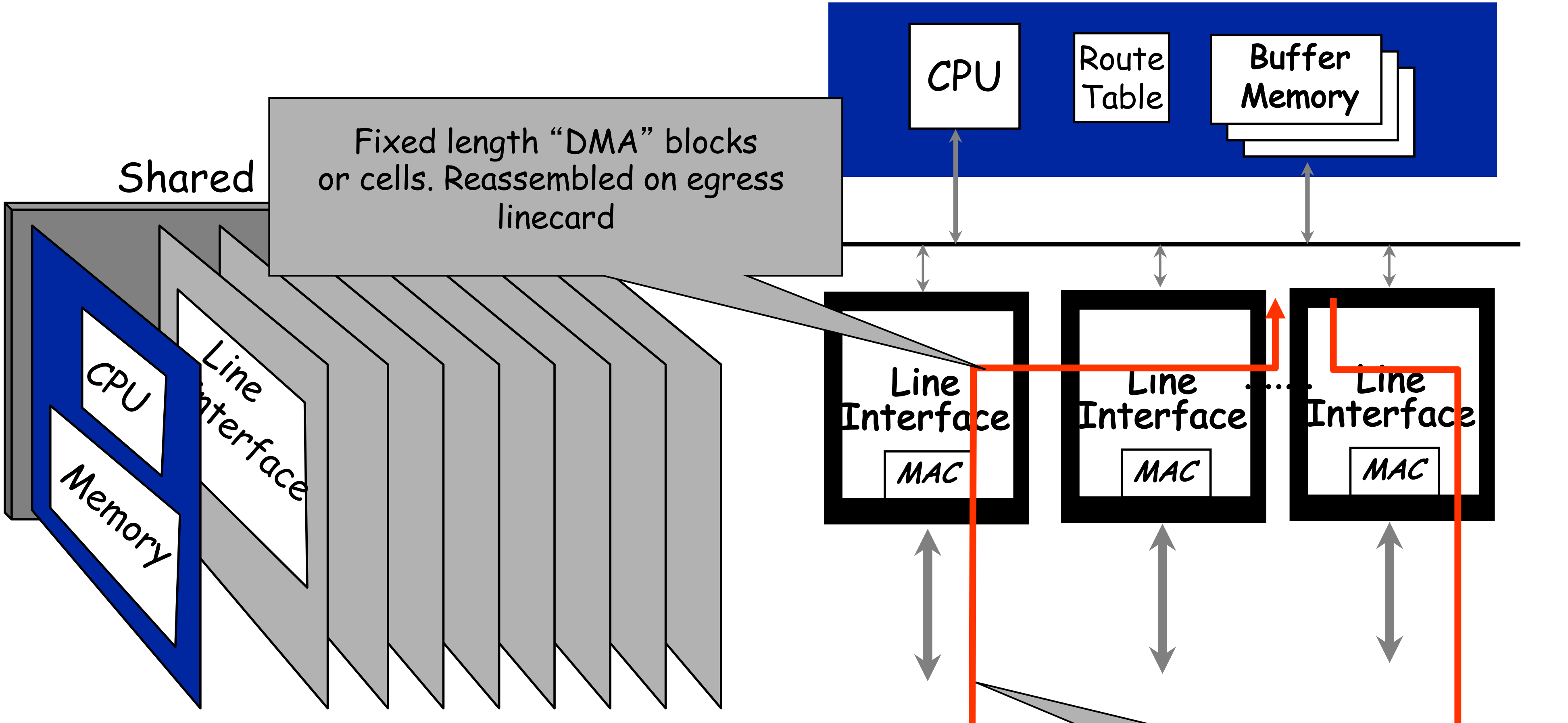
- Filtering packets for security reasons
- Delivering packets according to a pre-agreed delay guarantee
- Treating high priority packets preferentially
- Maintaining statistics on the number of packets sent by various routers

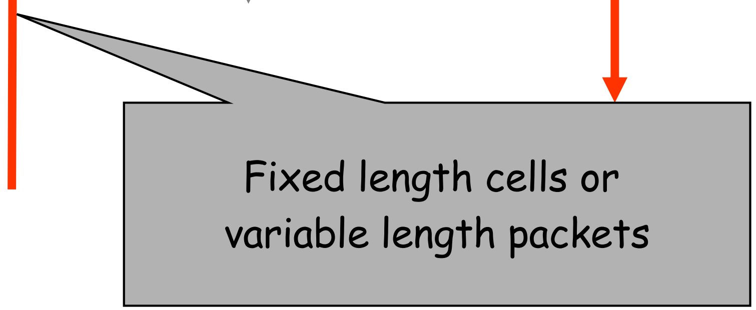
Memory limitation



Added by C. Pham

First Generation Routers



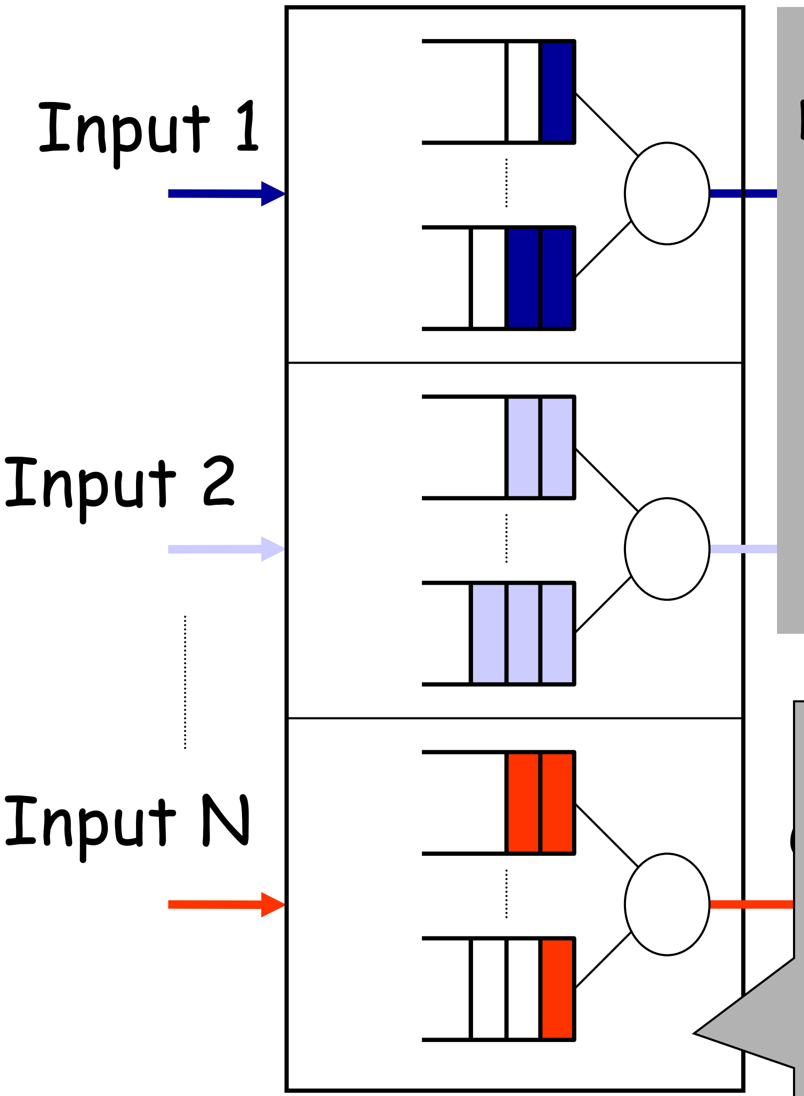


Most Ethernet switches and cheap packet routers

Bottleneck can be CPU, host-adaptor or I/O bus

First Generation Routers

Queueing Structure: Shared Memory



Numerous work has proven and made possible:

- Fairness
- Delay Guarantees
- Delay Variation Control
- Loss Guarantees
- Statistical Guarantees

Large, single dynamically allocated memory buffer:
N writes per "cell" time
N reads per "cell" time.
Limited by memory bandwidth.

Limitations (1)

- First generation router built with 133 MHz Pentium
  - Instruction time is 7.51ns
  - Mean packet size 500 bytes
  - Interrupt is 10 μs, memory access take 50 ns
  - Per-packet processing time is 200 instructions = 1.504 μs
- Copy loop

```
register <- memory[read_ptr]
memory [write_ptr] <- register
read_ptr <- read_ptr + 4
write_ptr <- write_ptr + 4
counter <- counter -1
if (counter not 0) branch to top of loop
```

- 4 instructions + 2 memory accesses = 130.08 ns
- Copying packet takes 500/4 *130.08 = 16.26 μs; interrupt 10 μs
- Total time = 27.764 μs => speed is 144.1 Mbps
- Amortized interrupt cost balanced by routing protocol cost

Limitations (2)

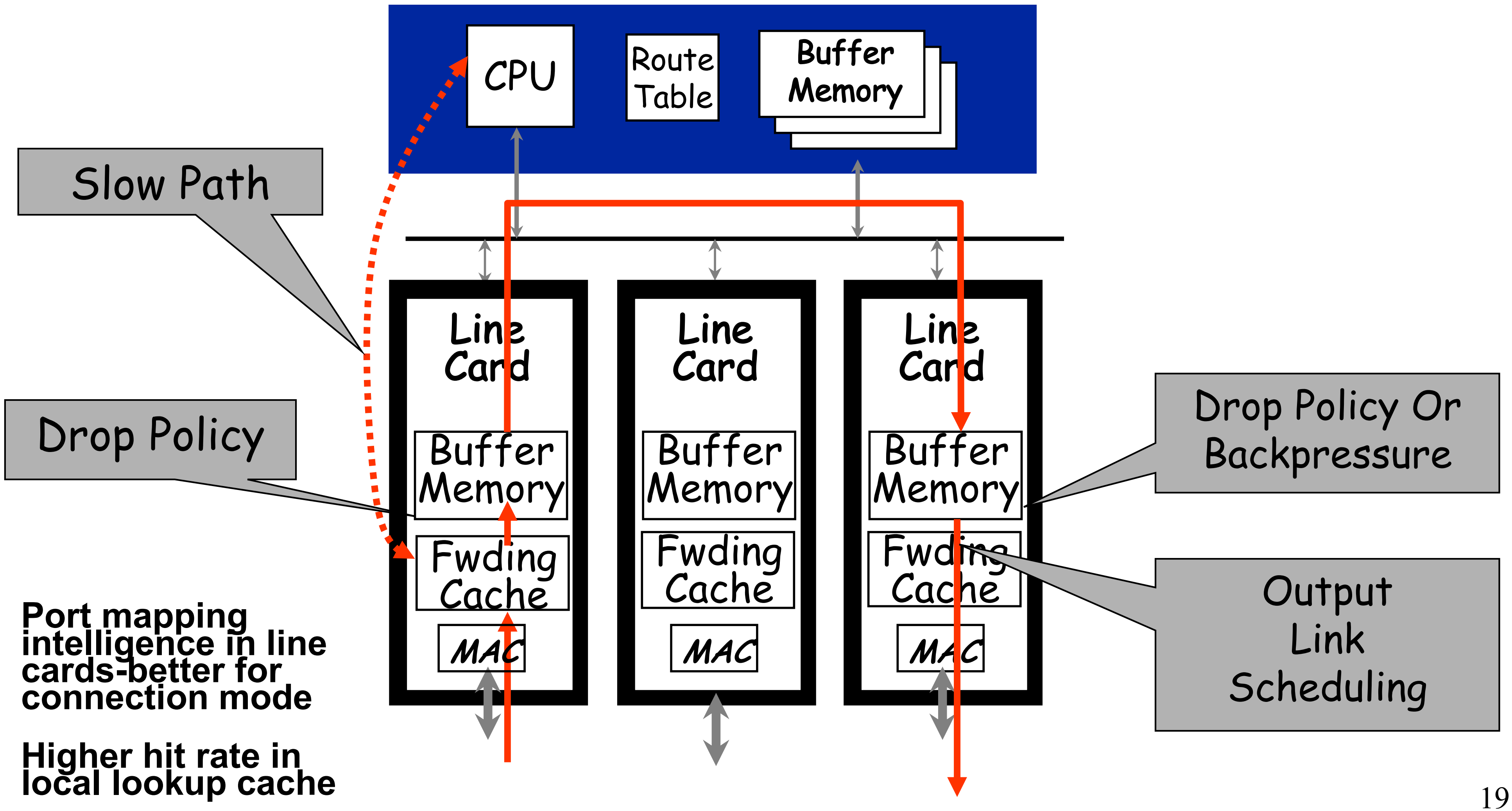
- First generation router built with 4 GHz i7
  - Instruction time is 0.25 ns
  - Mean packet size 500 bytes
  - Negligible interrupt~0, memory access take 5 ns
  - Per-packet processing time is 200 instructions = 50 ns
- Copy loop

```
register <- memory[read_ptr]
memory [write_ptr] <- register
read_ptr <- read_ptr + 8
write_ptr <- write_ptr + 8
counter <- counter -1
if (counter not 0) branch to top of loop
```

- 4 instructions + 2 memory accesses = 11 ns
- Copying packet takes 500/8 *11 = 687.5 ns
- Total time = 687.5 ns => speed is 5.8 Gbps

Added by C. Pham

Second Generation Routers

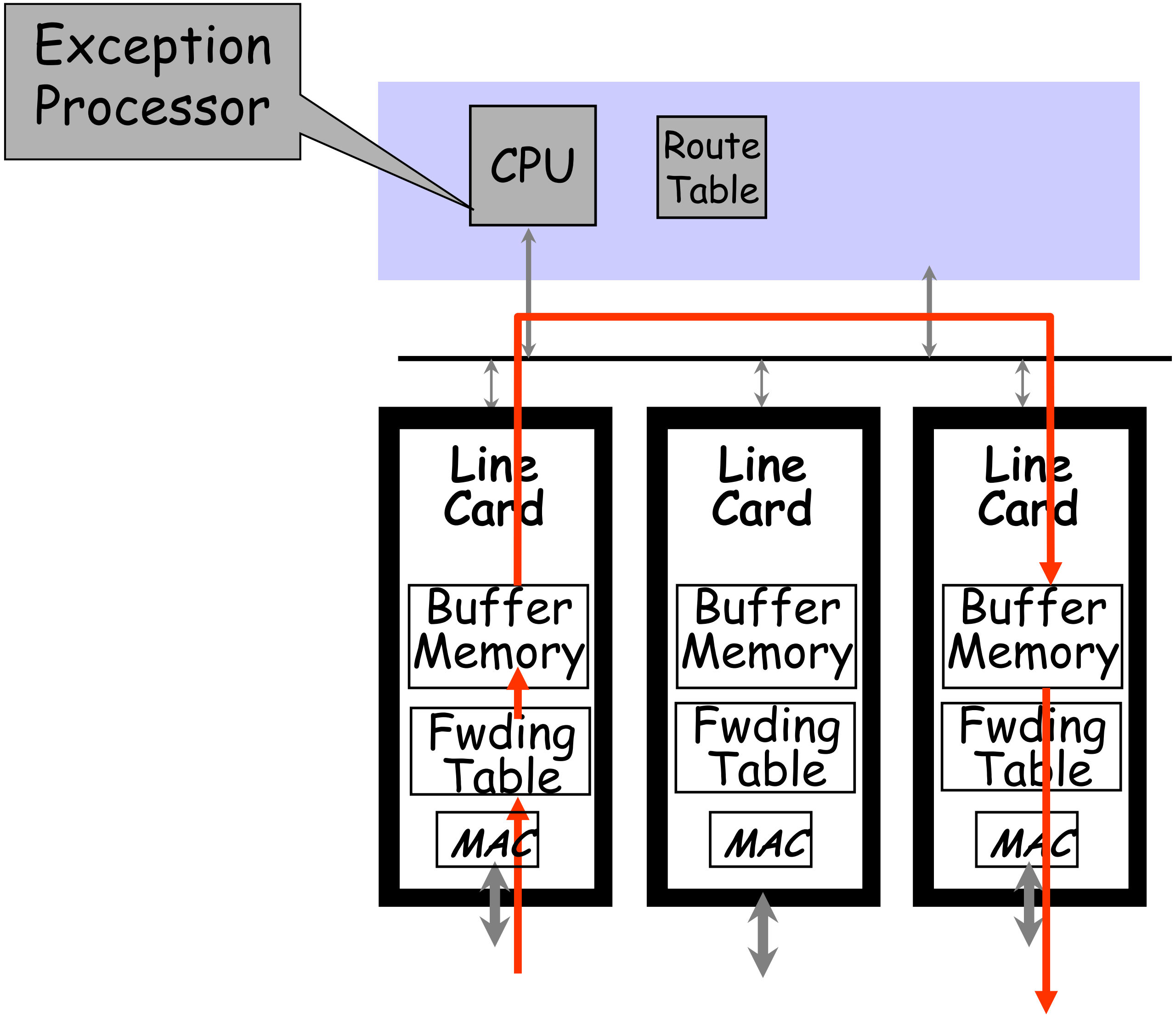


Port mapping intelligence in line cards-better for connection mode

Higher hit rate in local lookup cache

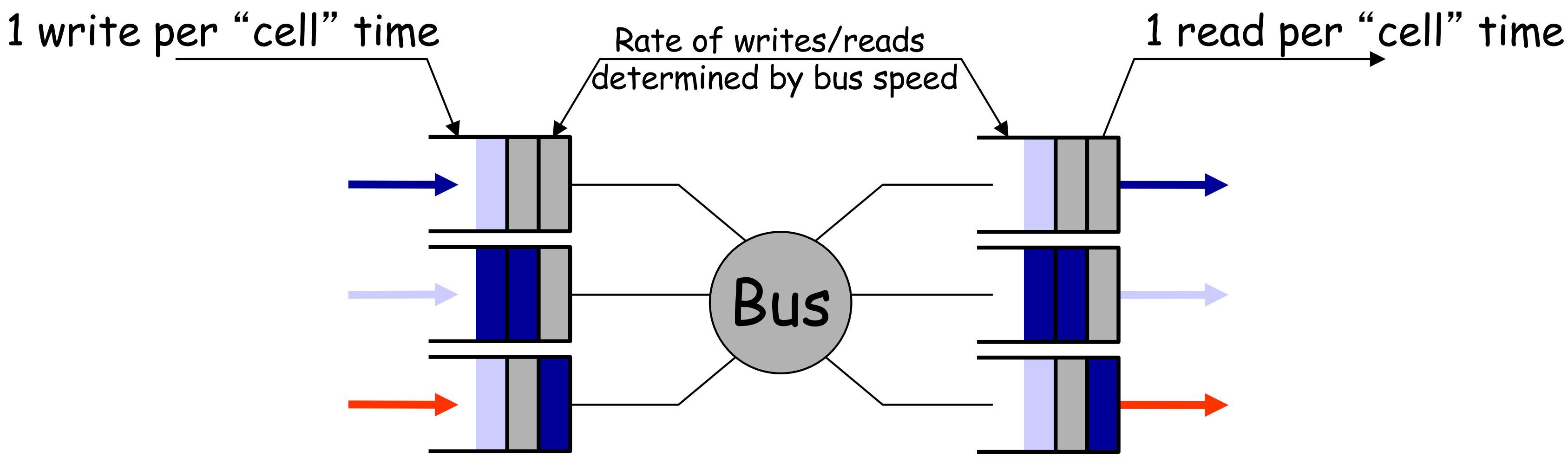
Second Generation Routers

As caching became ineffective



Second Generation Routers

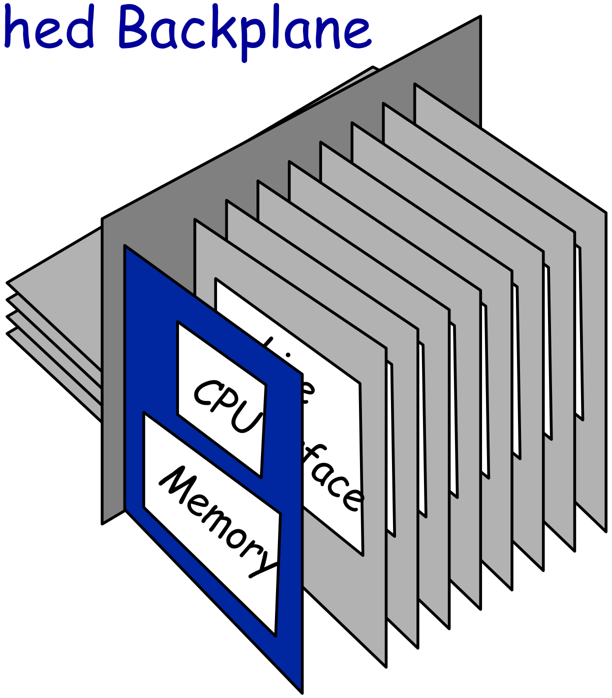
Queueing Structure: Combined Input and Output Queueing

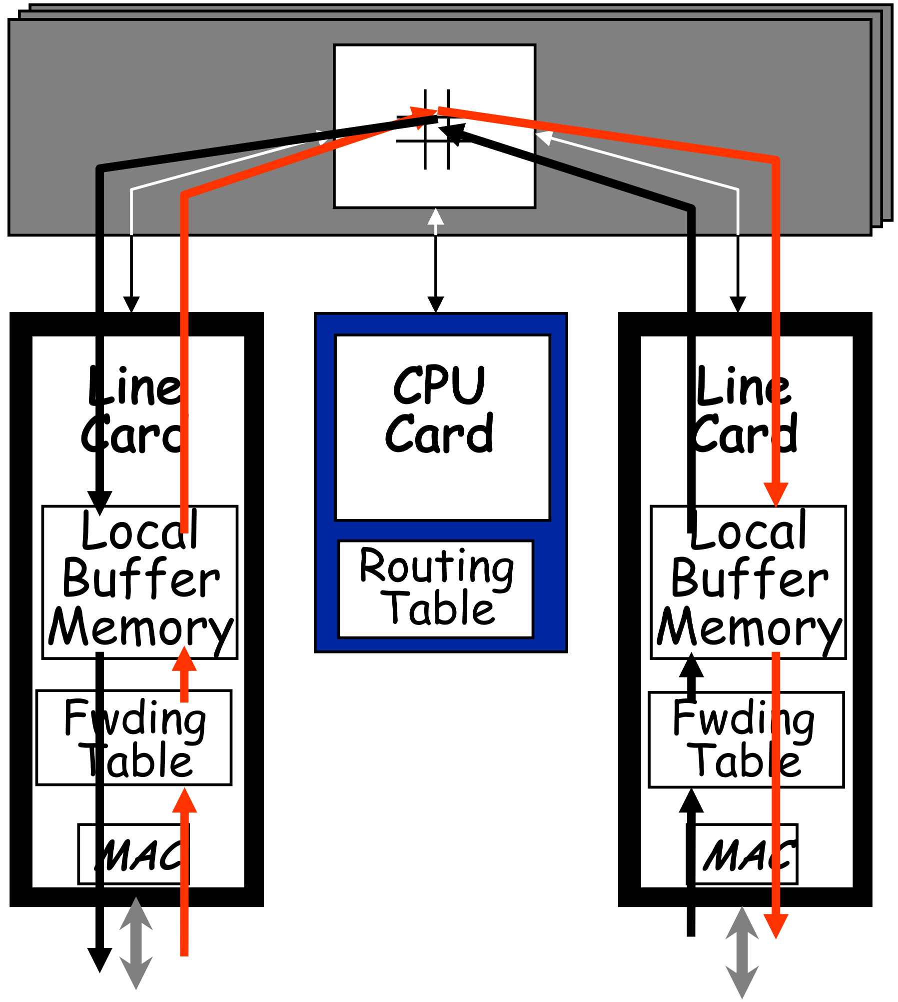


Third Generation Routers

- Third generation switch provides parallel paths (fabric)

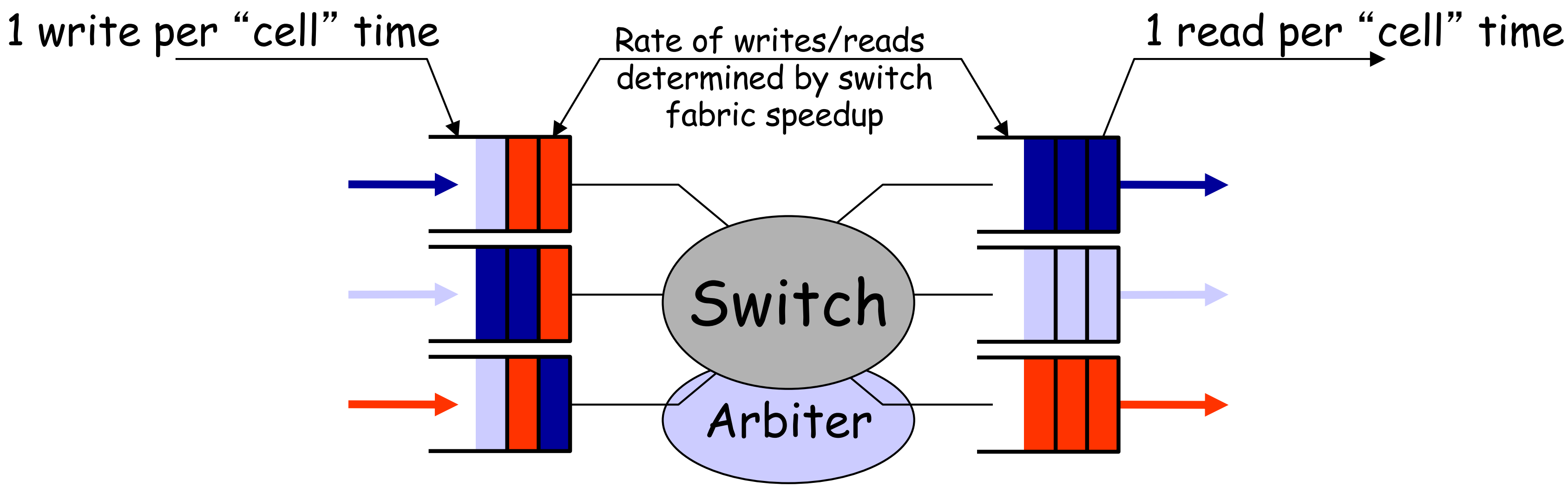
Switched Backplane





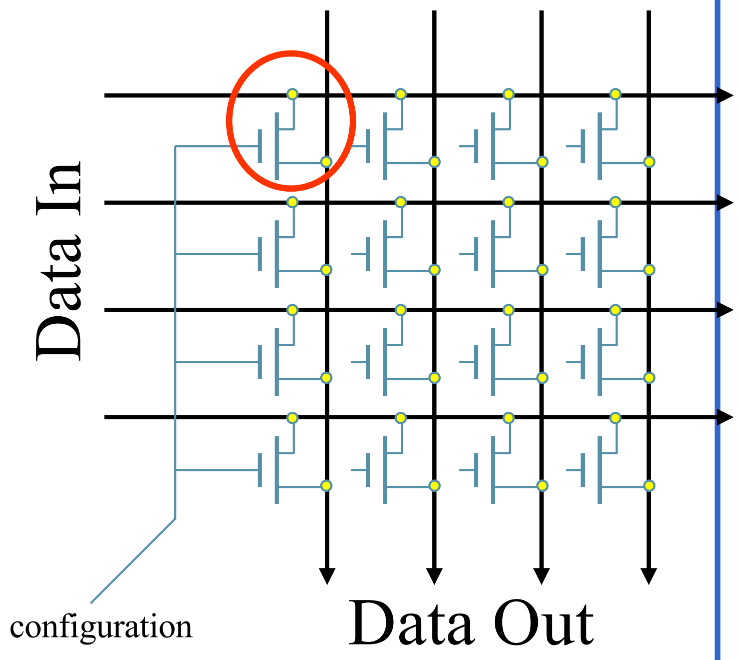
Third Generation Routers

Queueing Structure



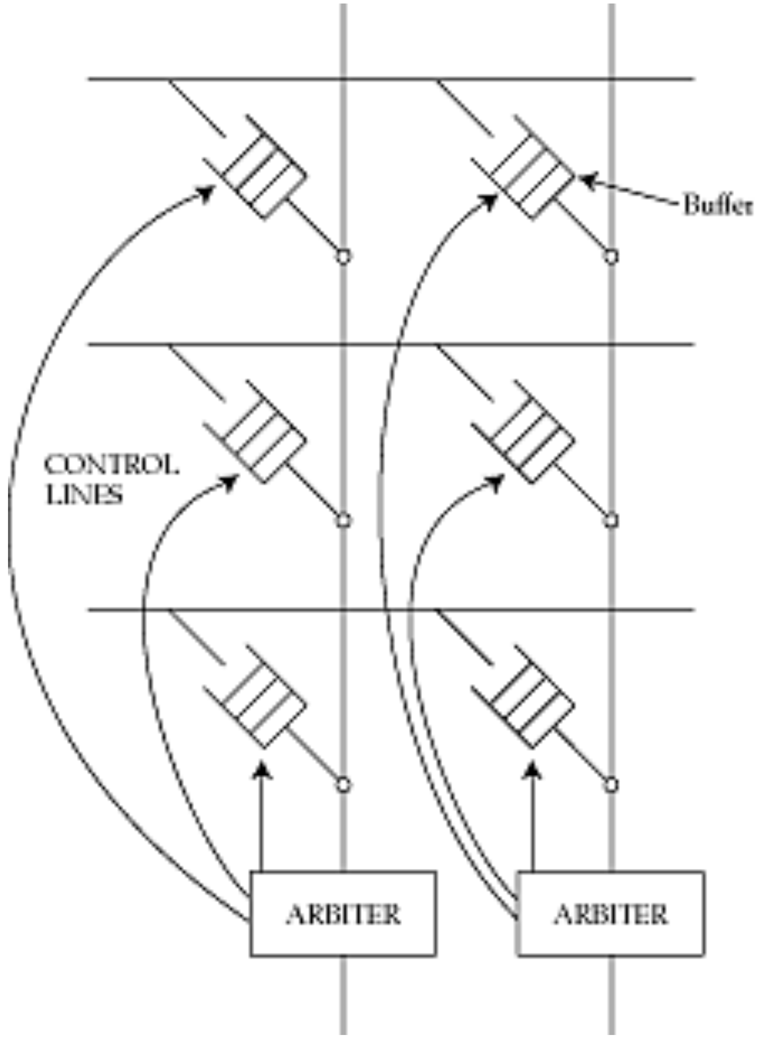
Review: crossbar, general design

- **Simplest possible space-division switch**
- **Crosspoints can be turned on or off, long enough to transfer a packet from an input to an output**
- **Expensive**
  - need N2 crosspoints
  - time to set each crosspoint grows quadratically



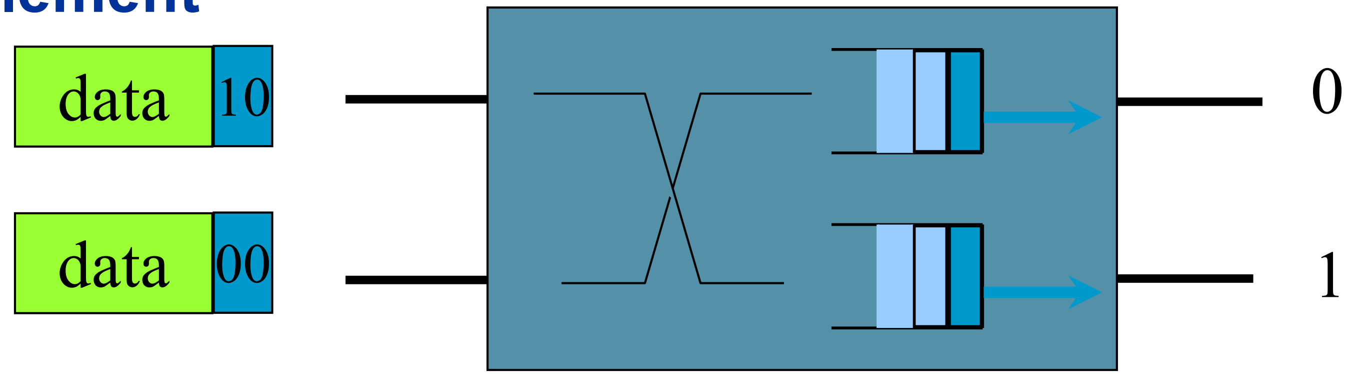
Switch Fabrics: Buffered crossbar (packets)

- What happens if packets at two inputs both want to go to same output?
- Can defer one at an input buffer
- Or, buffer cross-points: complex arbiter



Switch fabric element

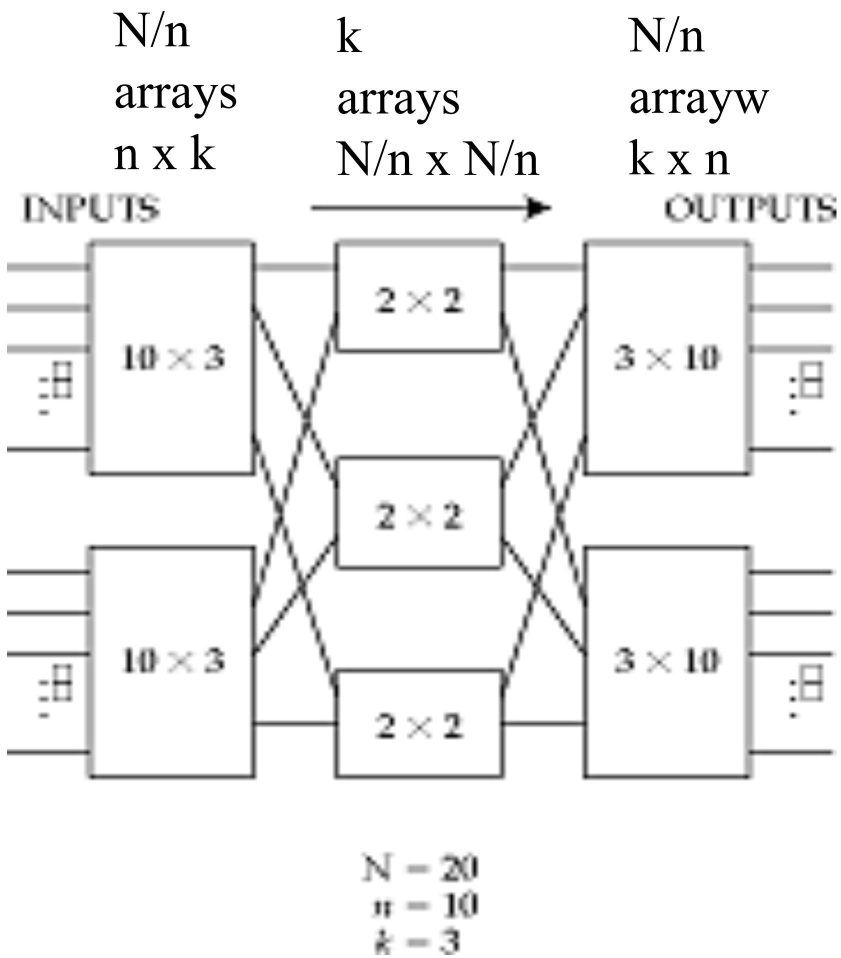
- Goal: towards building "self-routing" fabrics
- Can build complicated fabrics from a simple element



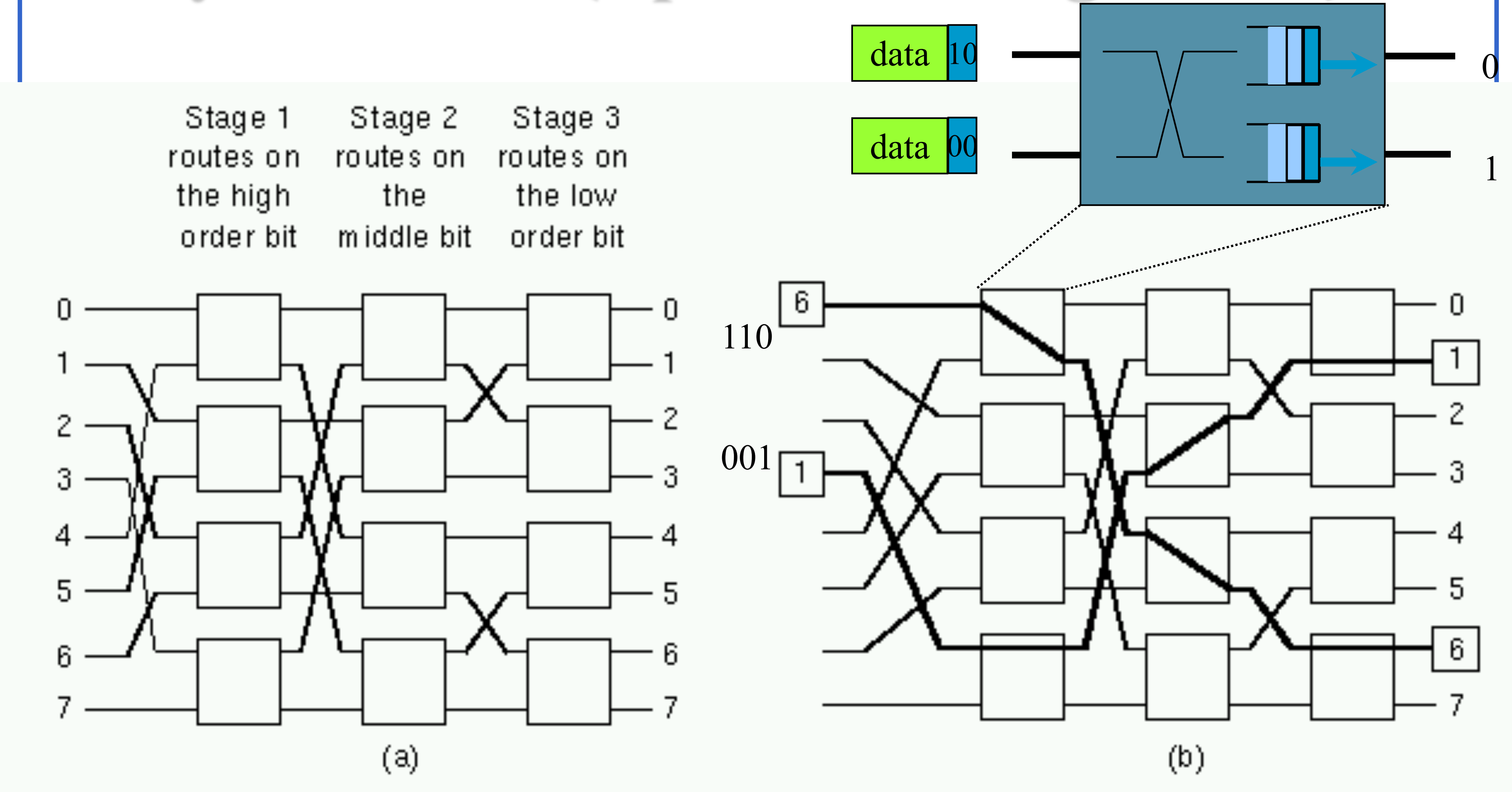
- Routing rule: if 0, send packet to upper output, else to lower output
  - If both packets to same output, buffer or drop

Multistage crossbar

- In a crossbar during each switching time only one cross-point per row or column is active
- Can save crosspoints if a cross-point can attach to more than one input line
- This is done in a multistage crossbar

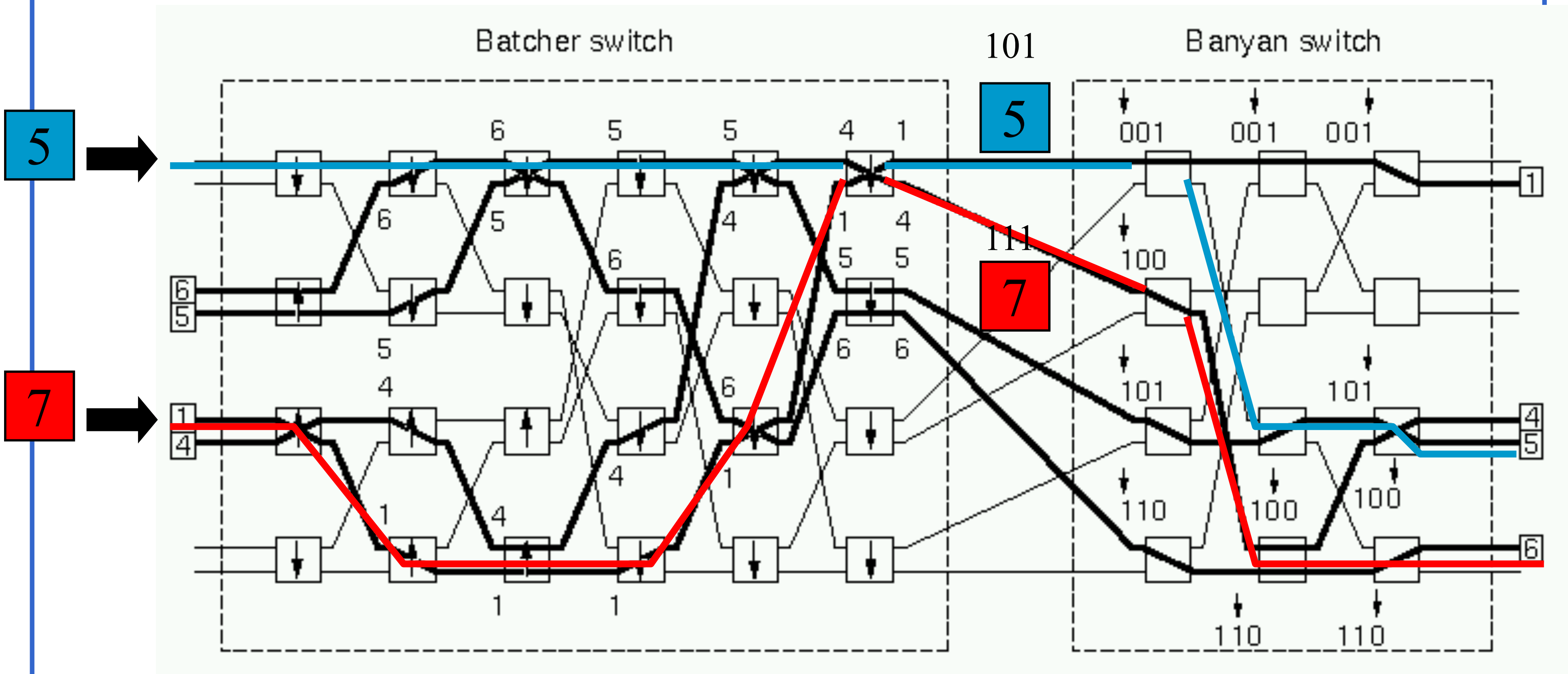


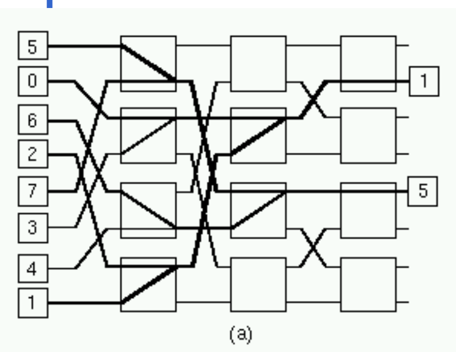
Banyan element (1 possible configuration)



ATM has boosted research on high-performance switches

Batcher-Banyan switch

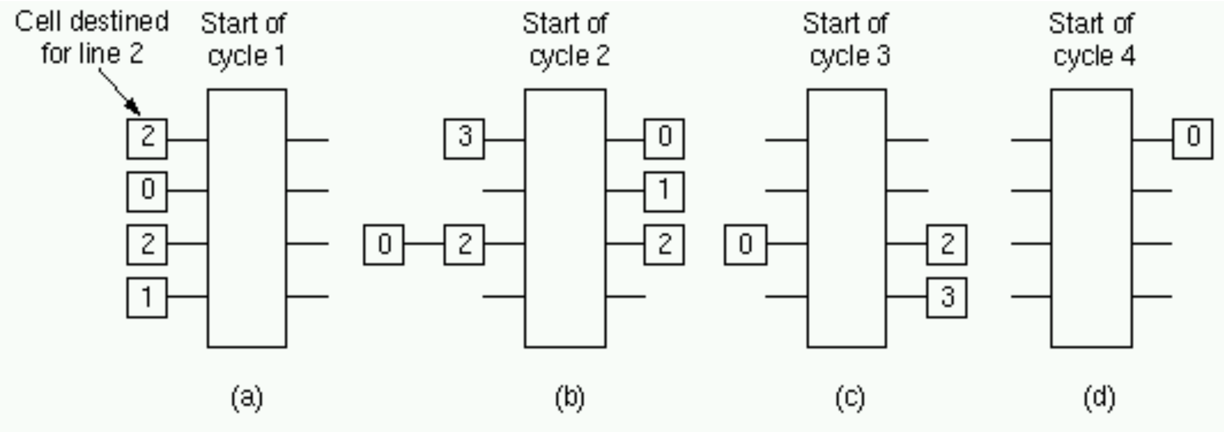




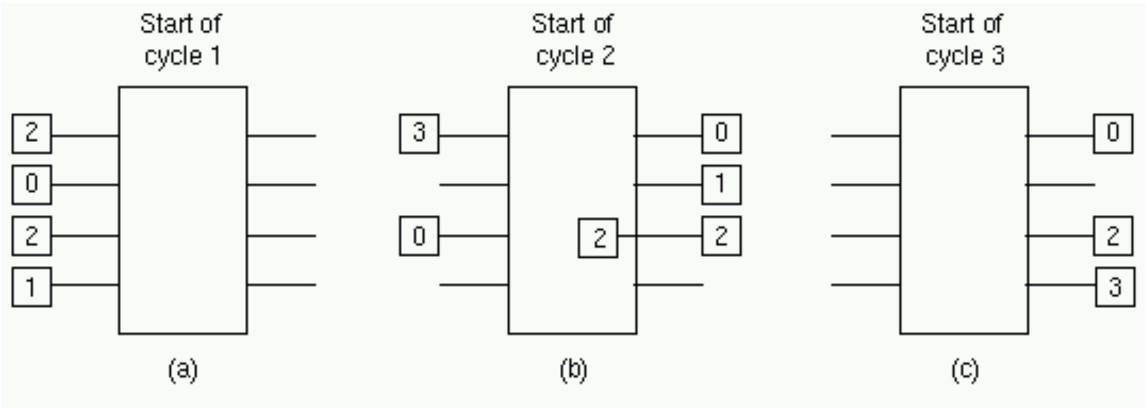
a same direction than arrow if *a* > *b*,
a opposite direction if a is alone

Buffer management

- Input buffers

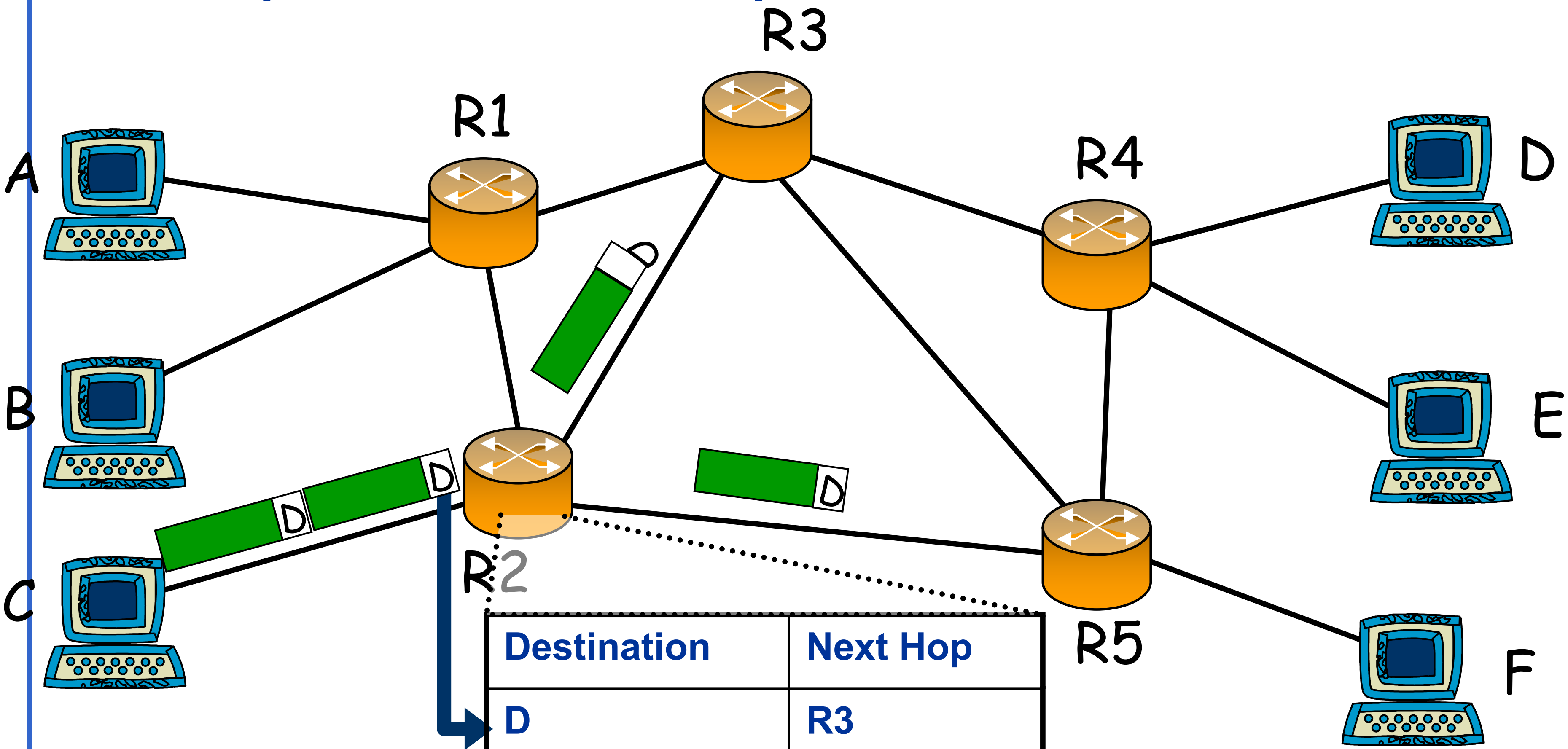


- Output buffer



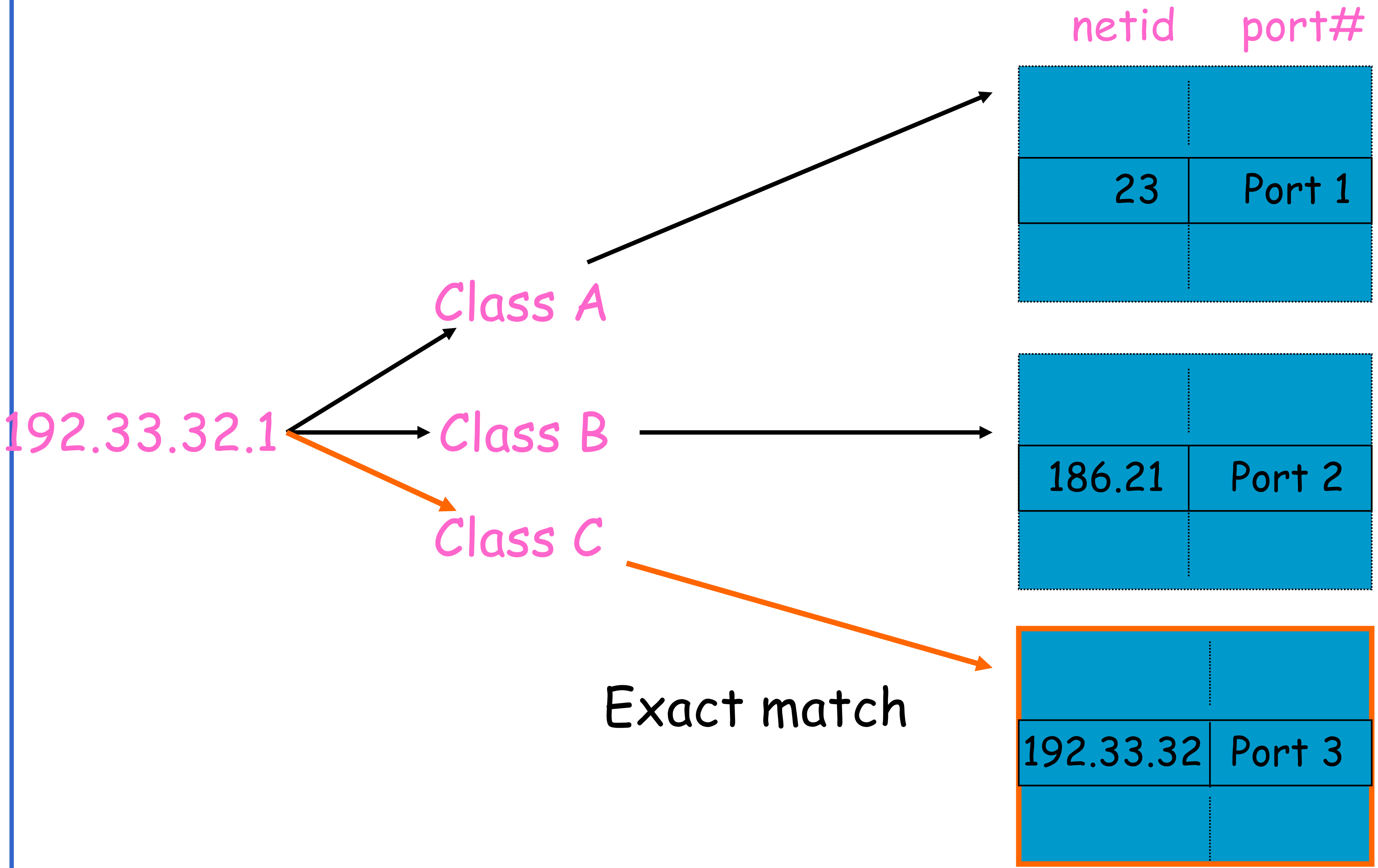
Still, cost of datagram packet switching

- With IP datagram mode, packet lookup is performed for each packet

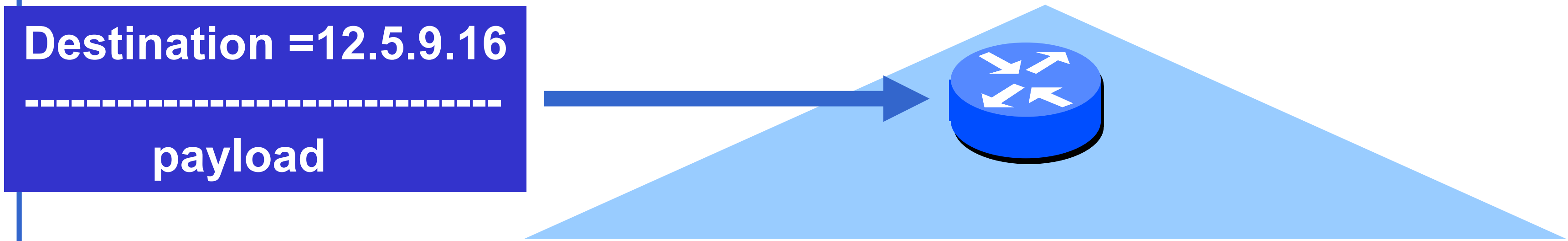


Destination	Next Hop
D	R3
E	R3
F	R5

Class-based lookups



Using longest prefix (CIDR: classless routing)



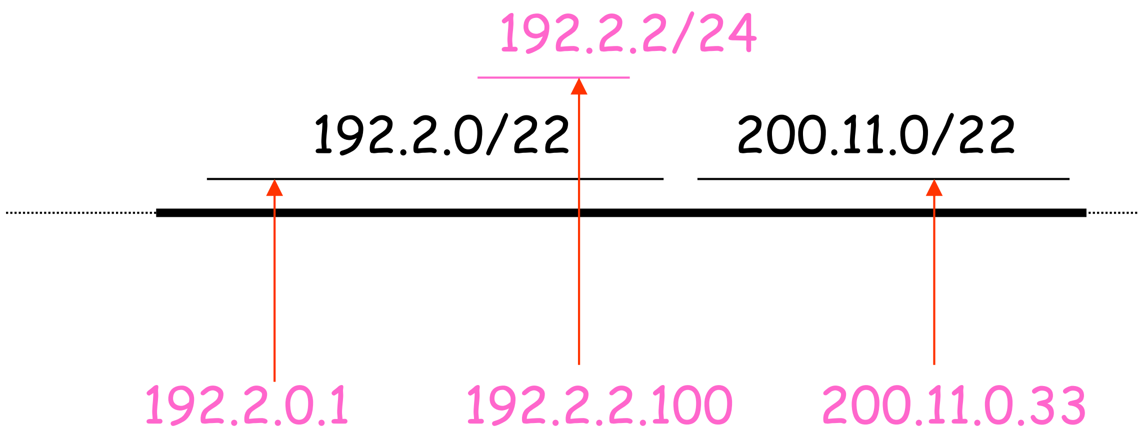
	Prefix	Next Hop	Interface
OK	0.0.0.0/0	10.14.11.33	ATM 5/0/9
better	12.0.0.0/8	10.14.22.19	ATM 5/0/8
even better	12.4.0.0/15	10.1.3.77	Ethernet 0/1/3
best!	12.5.8.0/23	attached	Serial 1/0/7

IP Forwarding Table

CIDR/VLSM lookup

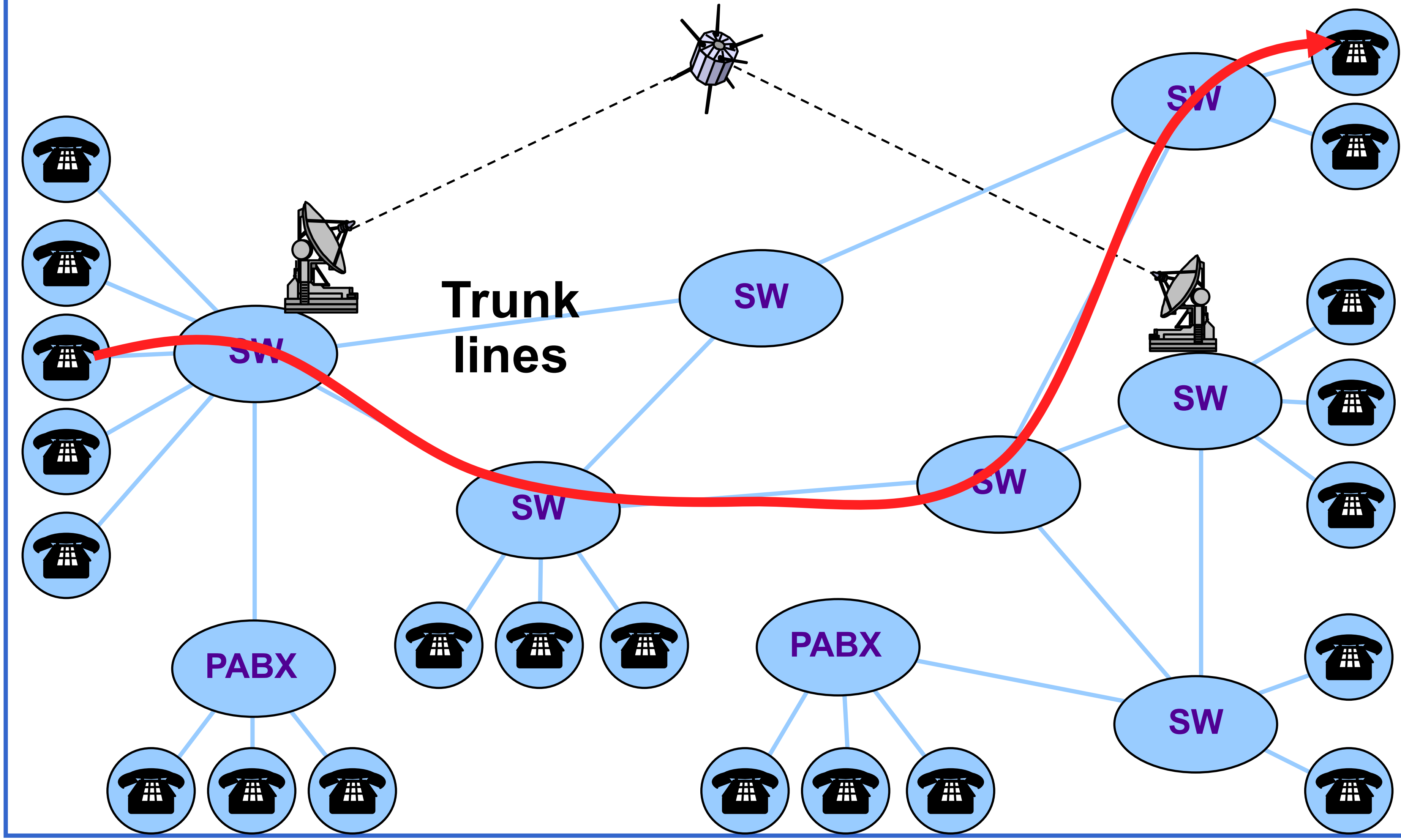
Find the most specific route, or the longest matching prefix among all the prefixes matching the destination address of an incoming packet

192.2.0/22, R2
192.2.2/24, R3
200.11.0/22, R4



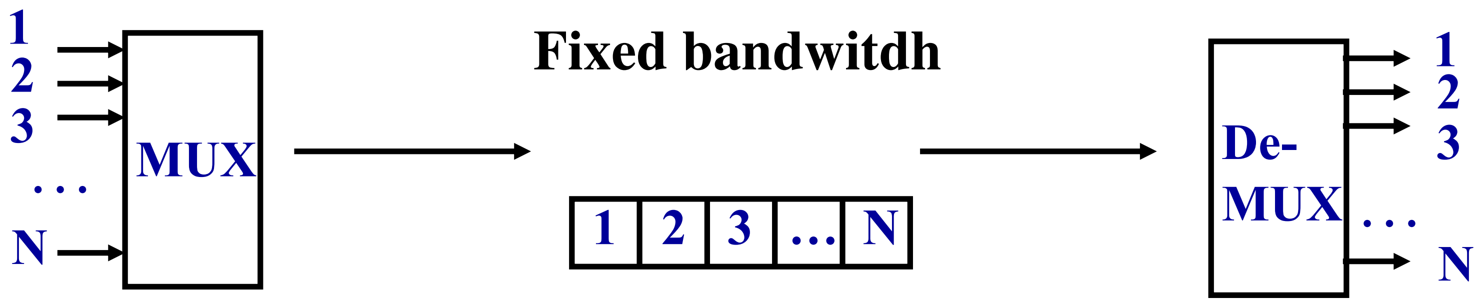
Cost of packet lookup is further increased!!!

Reliability of circuit switching



Traditional circuit in telephony

Simple, efficient, but low flexibility and wastes resources

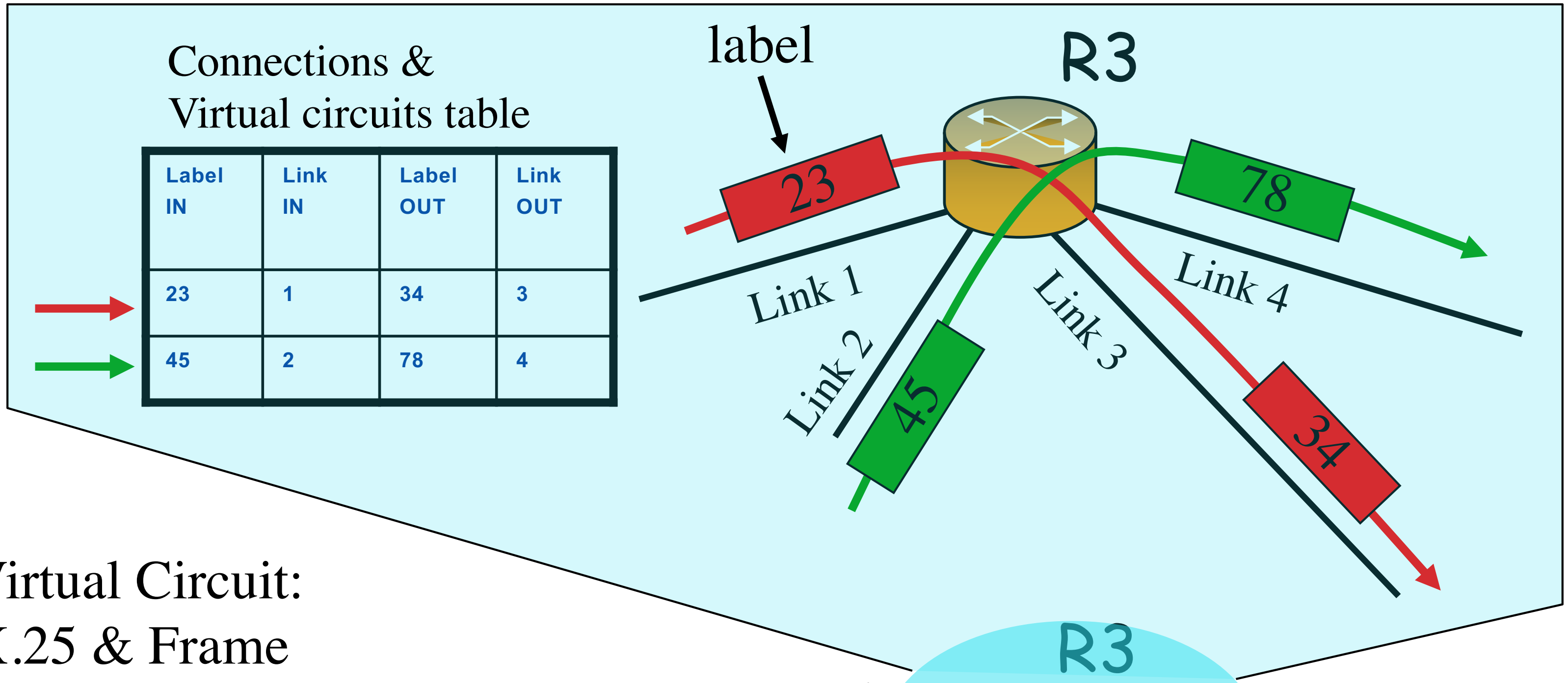


1 sample every 125us gives a 64Kbits/s channel

Packet-switching with <u>virtual circuit:</u>

take advantages of both worlds

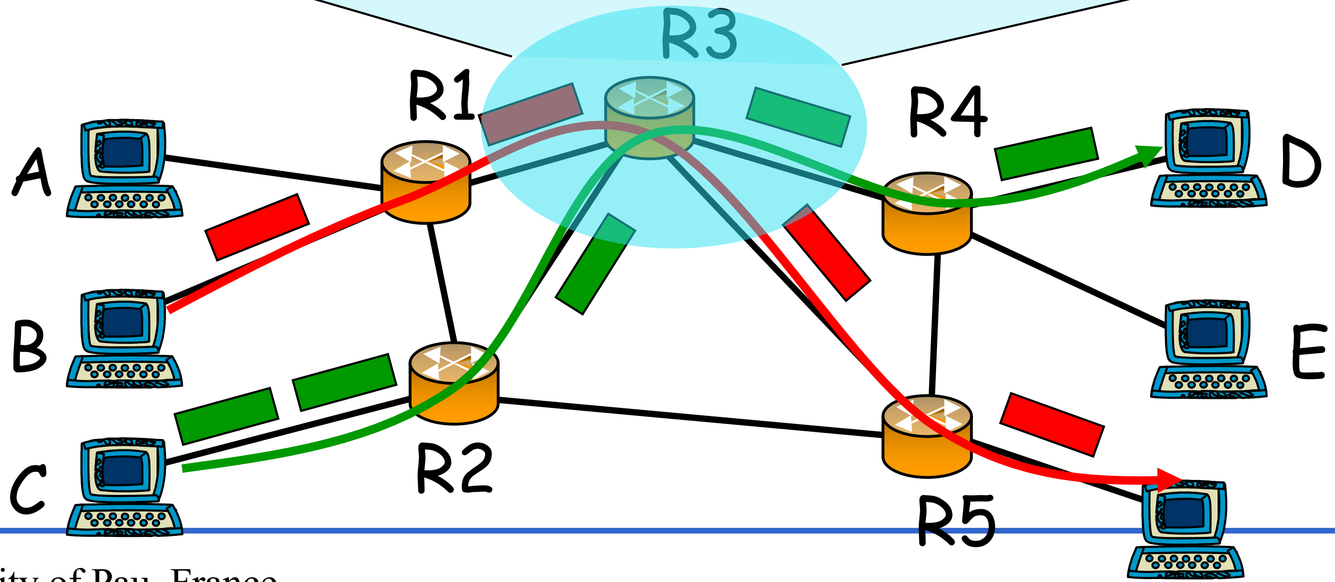
Virtual Circuit



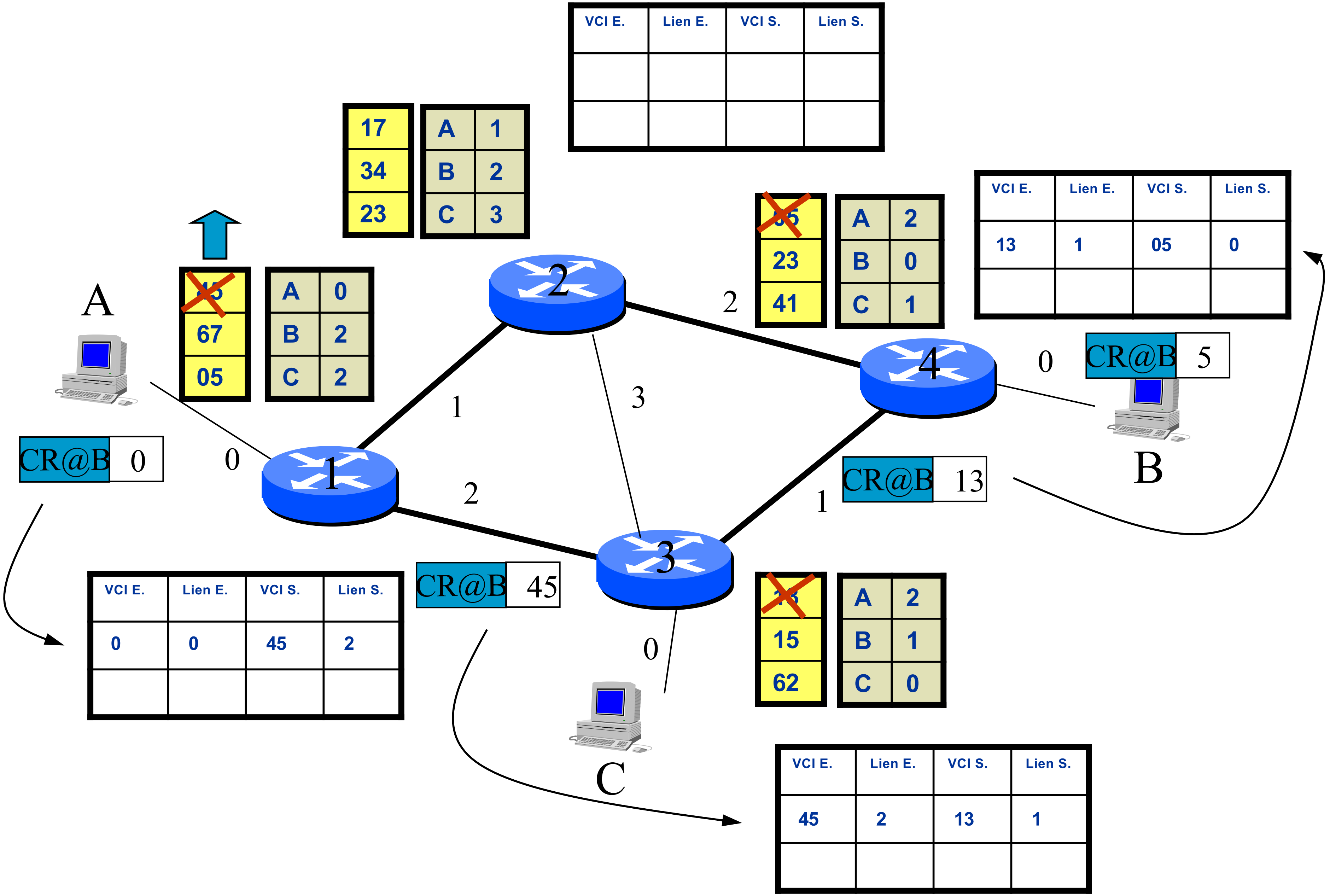
Label IN	Link IN	Label OUT	Link OUT
23	1	34	3
45	2	78	4

Virtual Circuit: X.25 & Frame Relay.

ATM: same principle but much smaller packet size



Setting up a virtual circuit (1)



~~45~~	
67	
05	

A	0
B	2
C	2

17	
34	
23	

A	1
B	2
C	3

VCI E.	Lien E.	VCI S.	Lien S.

~~05~~	
23	
41	

A	2
B	0
C	1

VCI E.	Lien E.	VCI S.	Lien S.
13	1	05	0

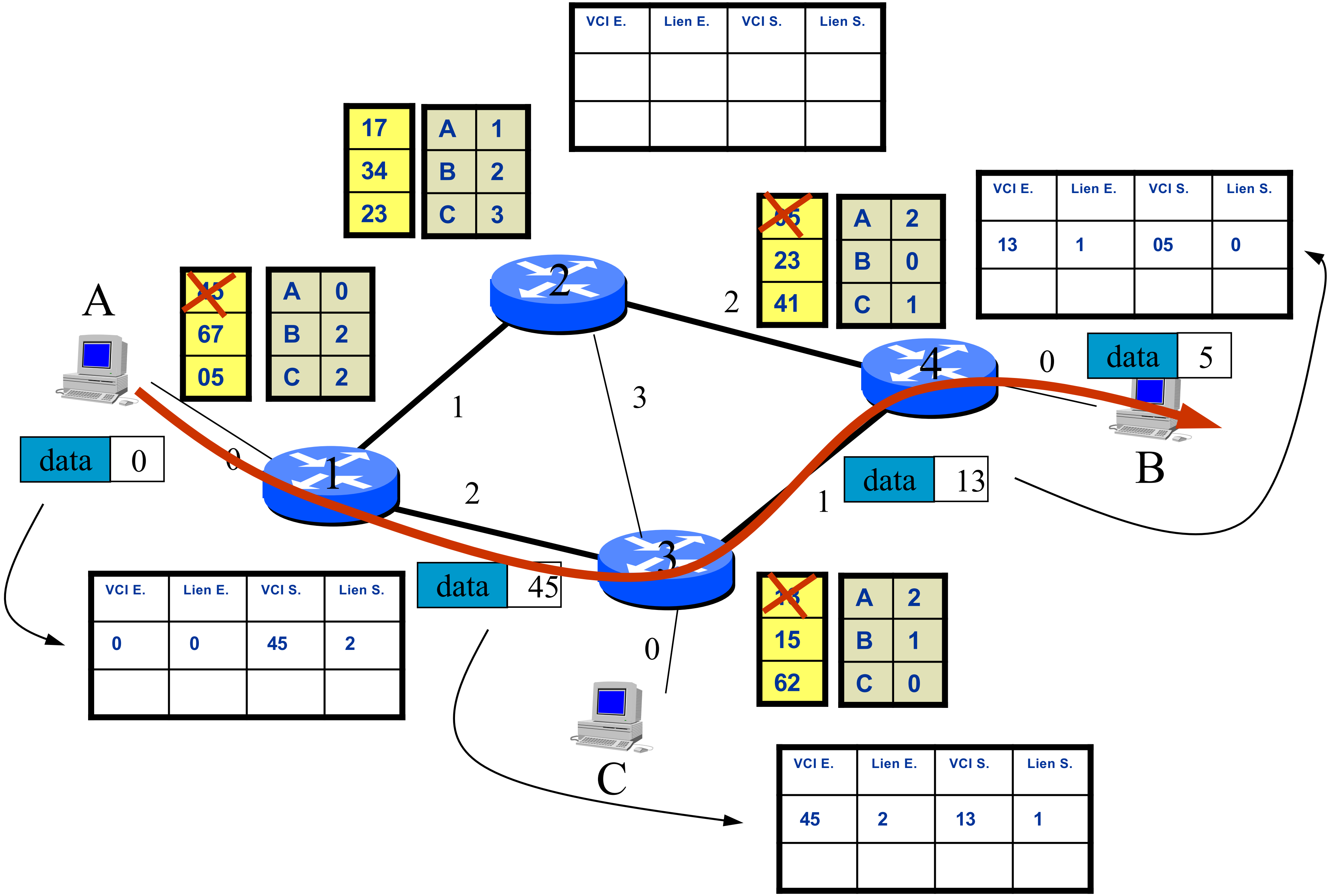
VCI E.	Lien E.	VCI S.	Lien S.
0	0	45	2

~~13~~	
15	
62	

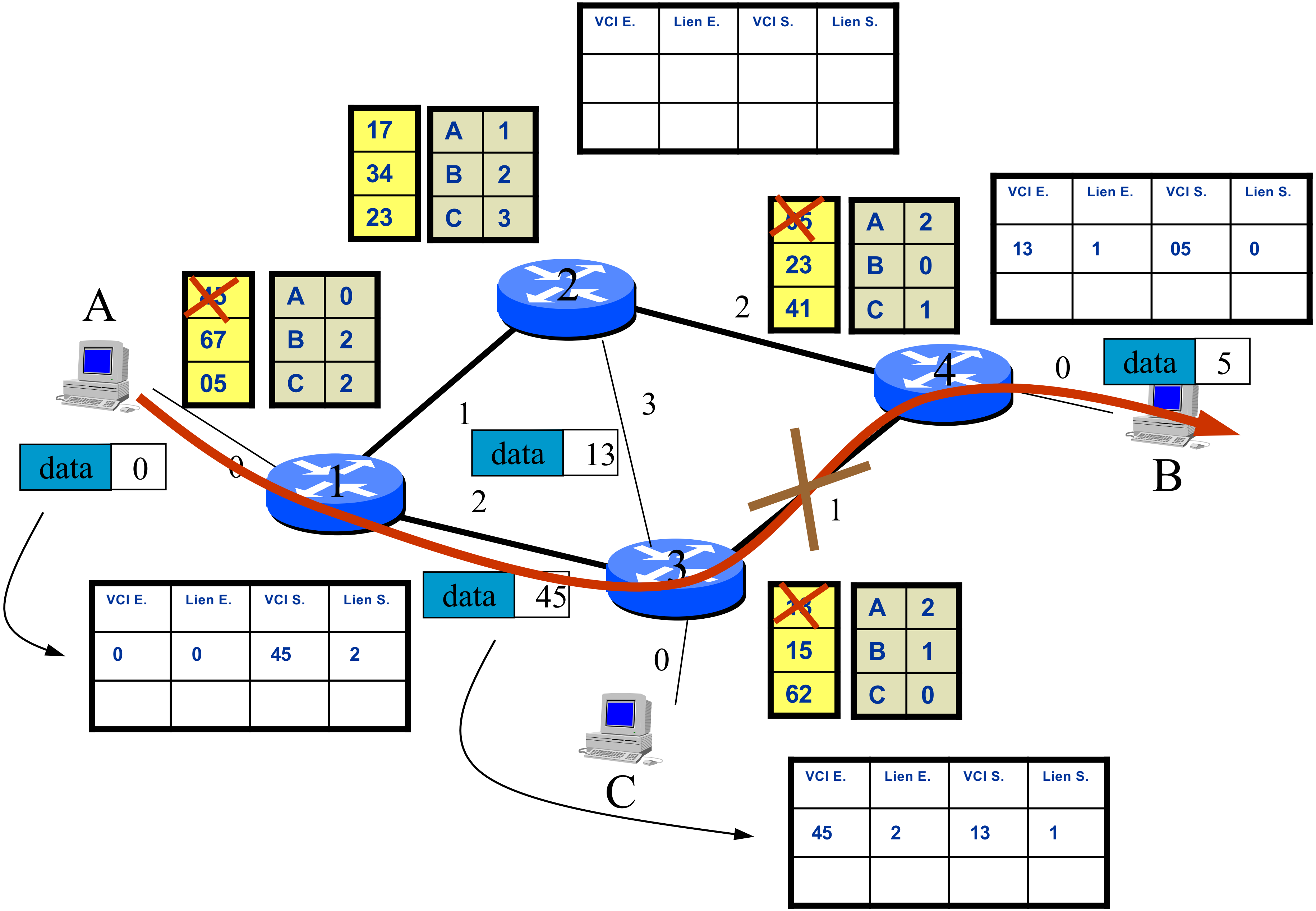
A	2
B	1
C	0

VCI E.	Lien E.	VCI S.	Lien S.
45	2	13	1

Setting up a virtual circuit (2)

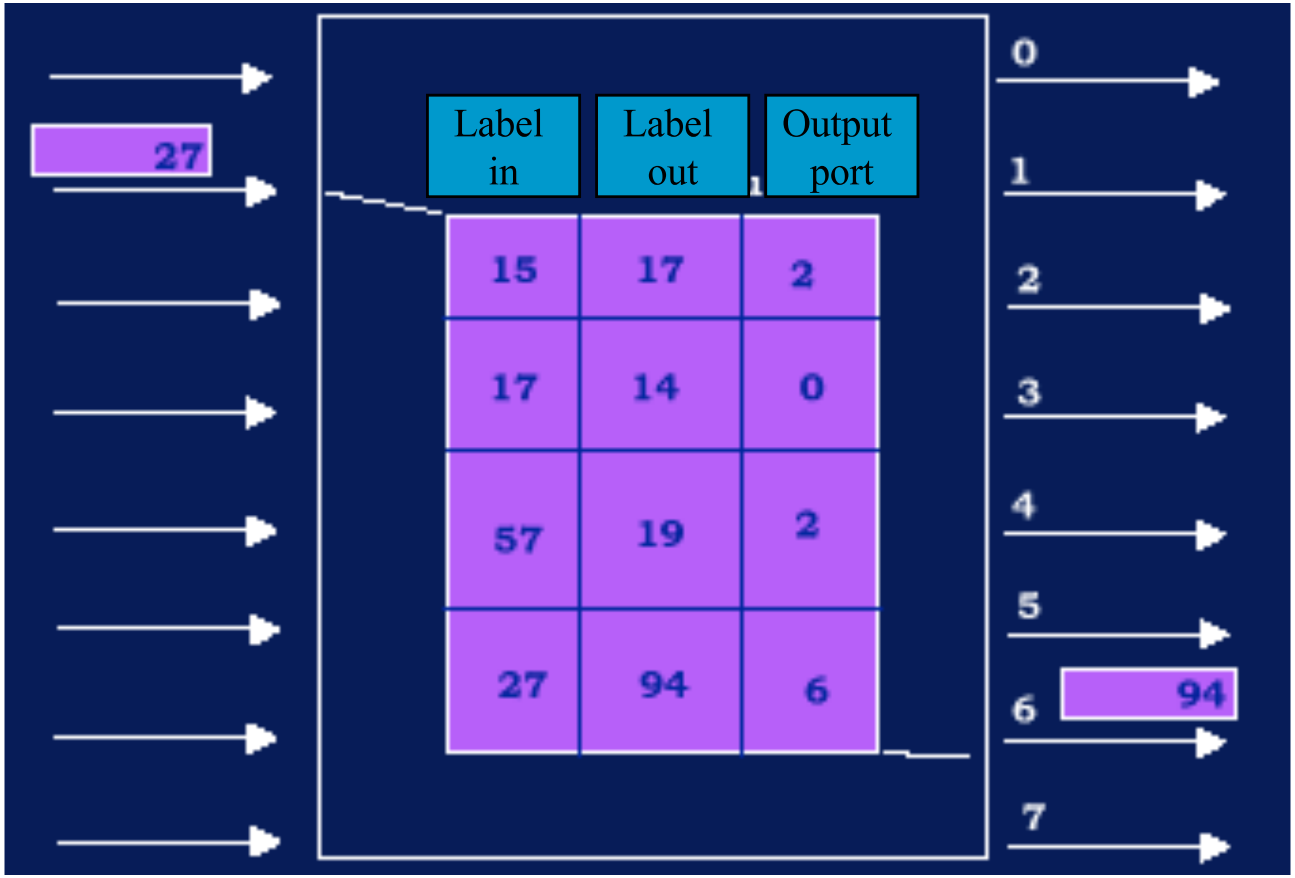


Link failure with virual circuit

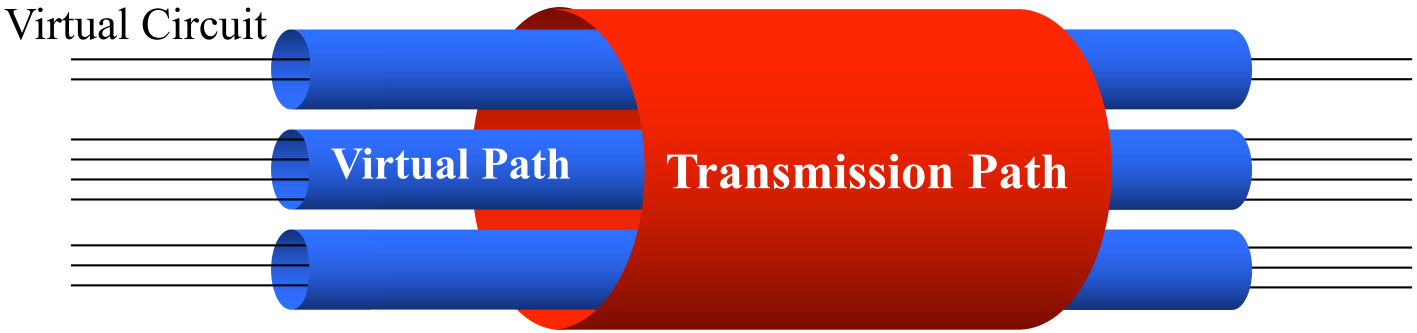


Using virtual circuit to decrease lookup cost

- Introduced by X.25, Frame Relay, ATM
- Use labels to forward packets/cells

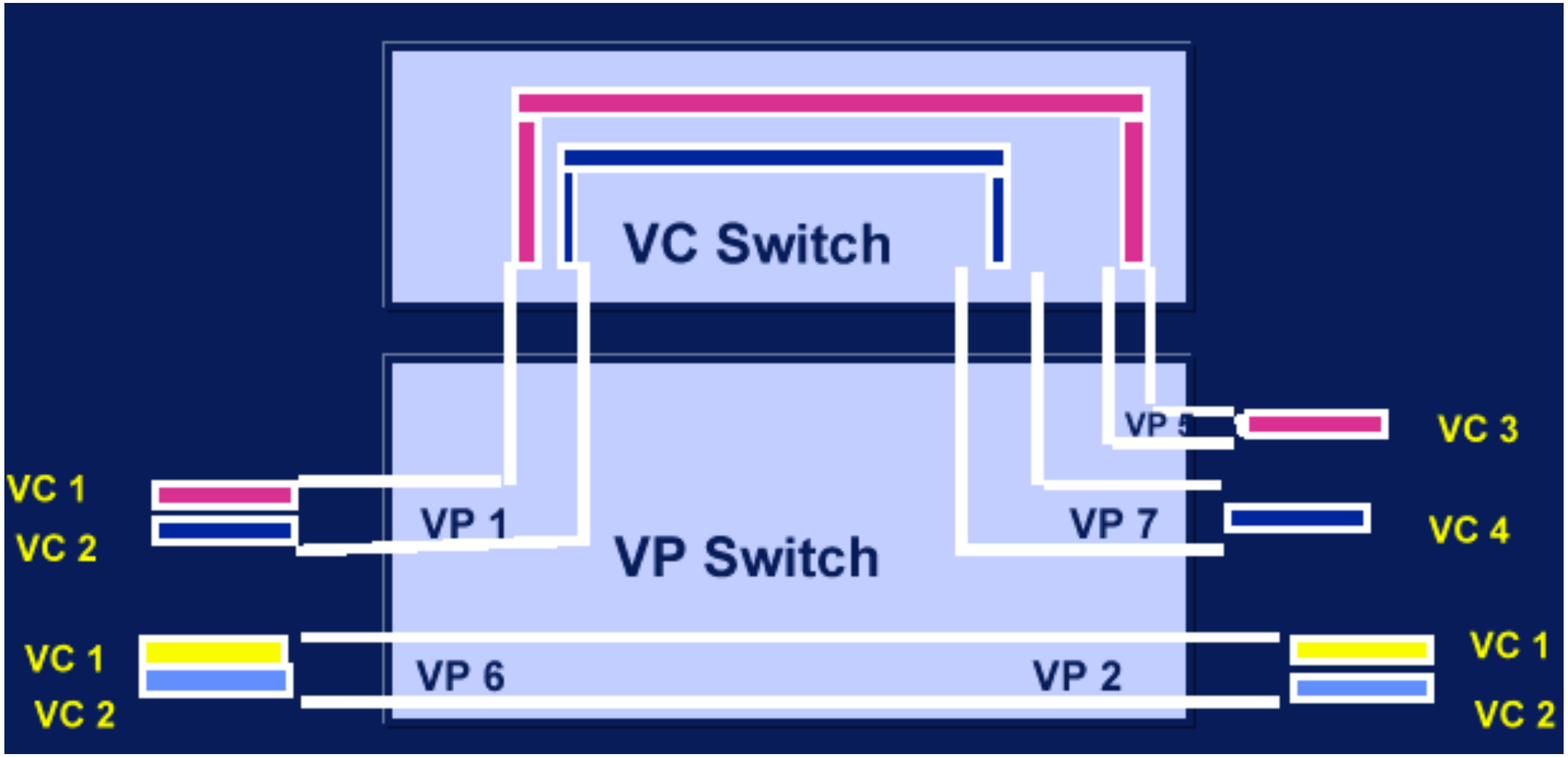


VC & VP: introducing hierarchy

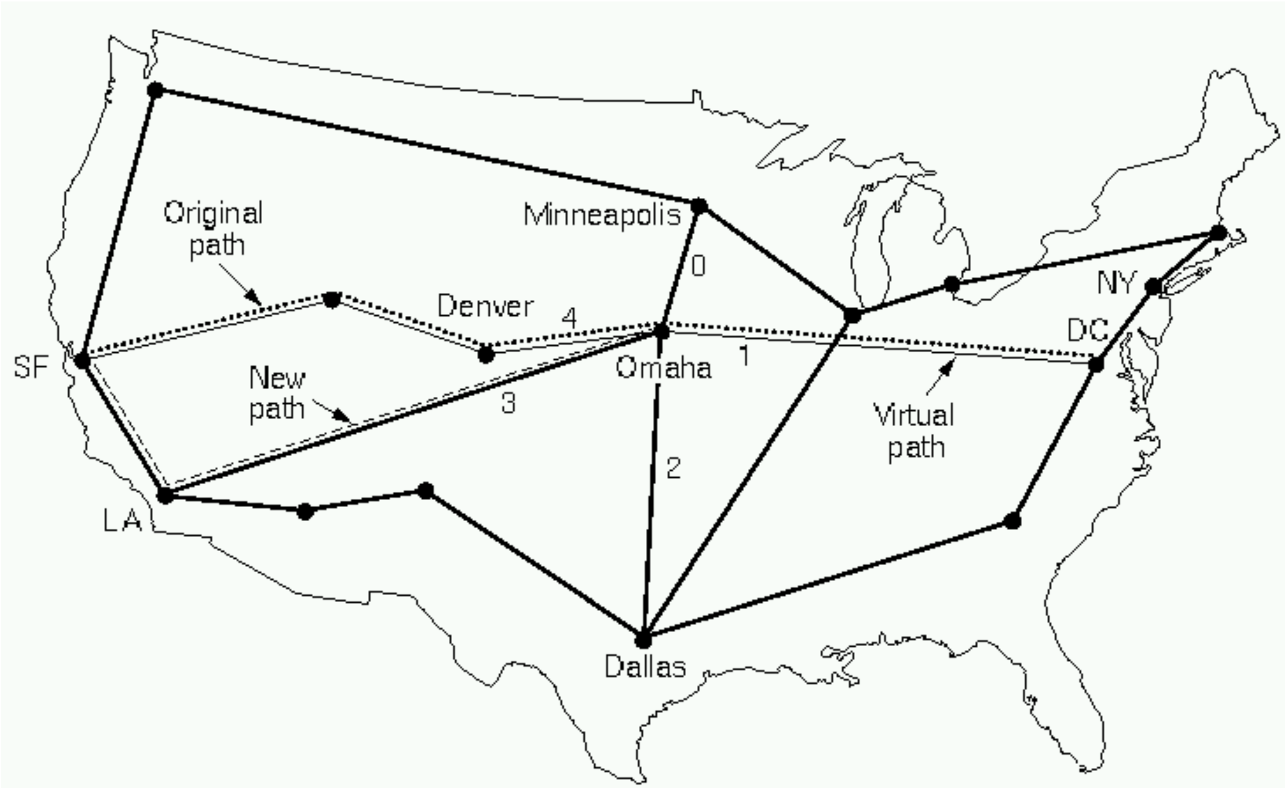


- A VPC = 1 VP or a concatenation of several VPs.
- A VCC = 1 VC or a concatenation of several VCs.
- A VP contains several VCs
- Avantages
  - Simple connection setup for most used paths
  - Easy definition of Virtual Private Networks (VPN),
  - Simplier traffic management: traffics with different constraints can be transported in different VPs for isolation.

2 level switching

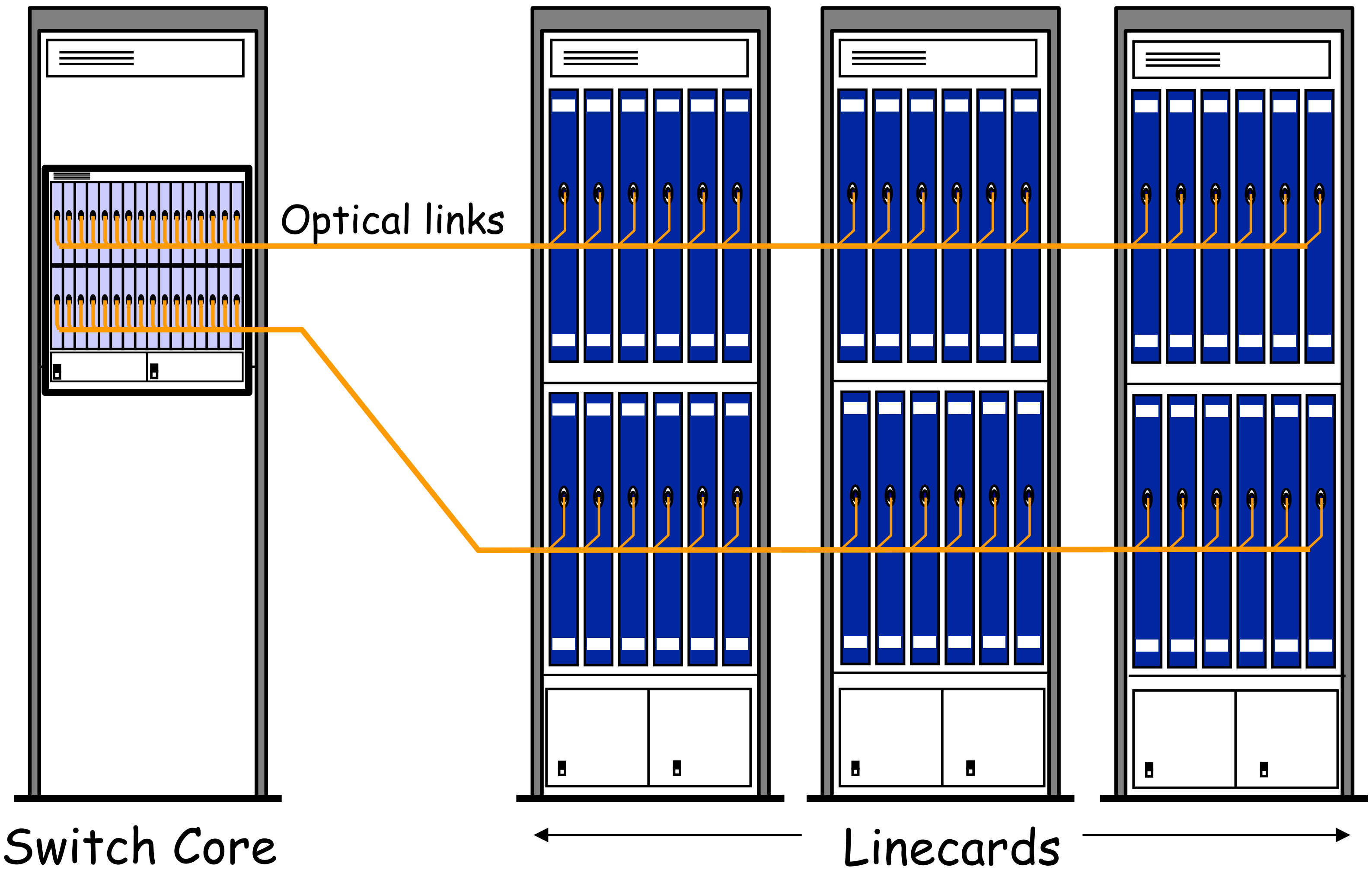


Advantages of VP and VC hierarchy



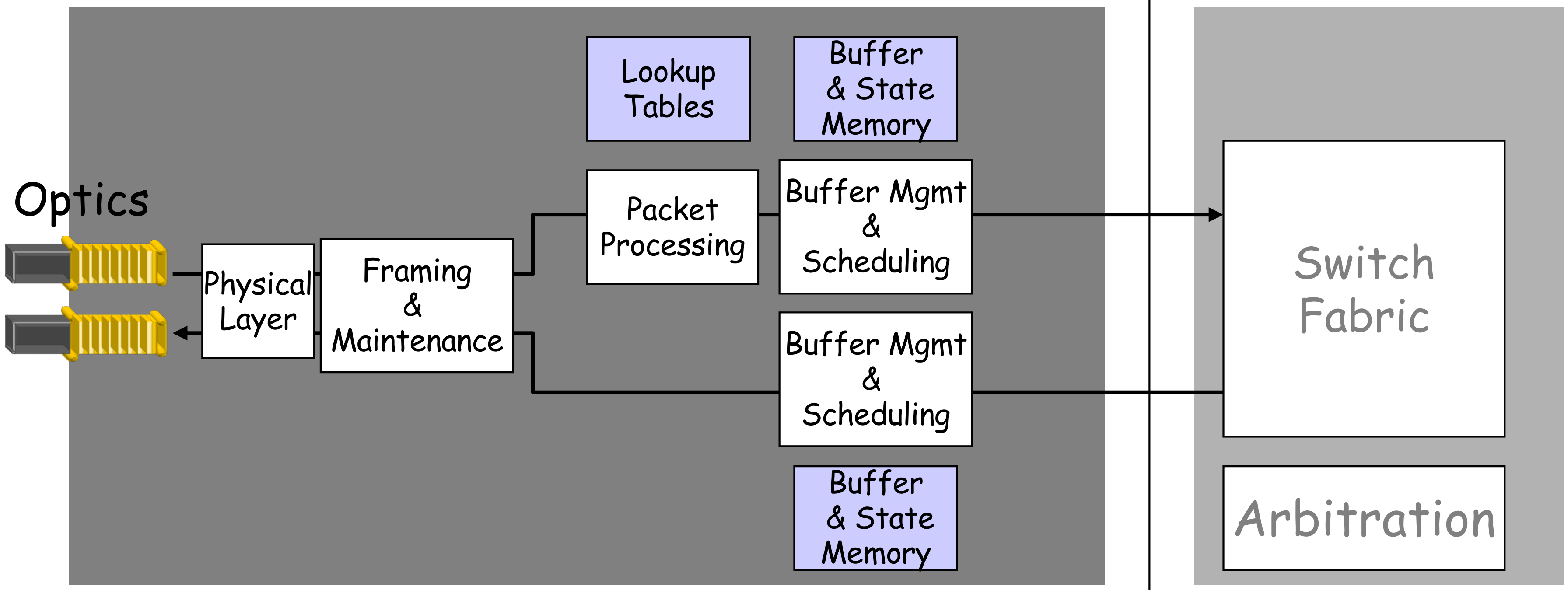
- Re-routing a VP automatically re-routes all VCs of the VP
- Towards Traffic Engineering!!

Optics in routers



Complex linecards

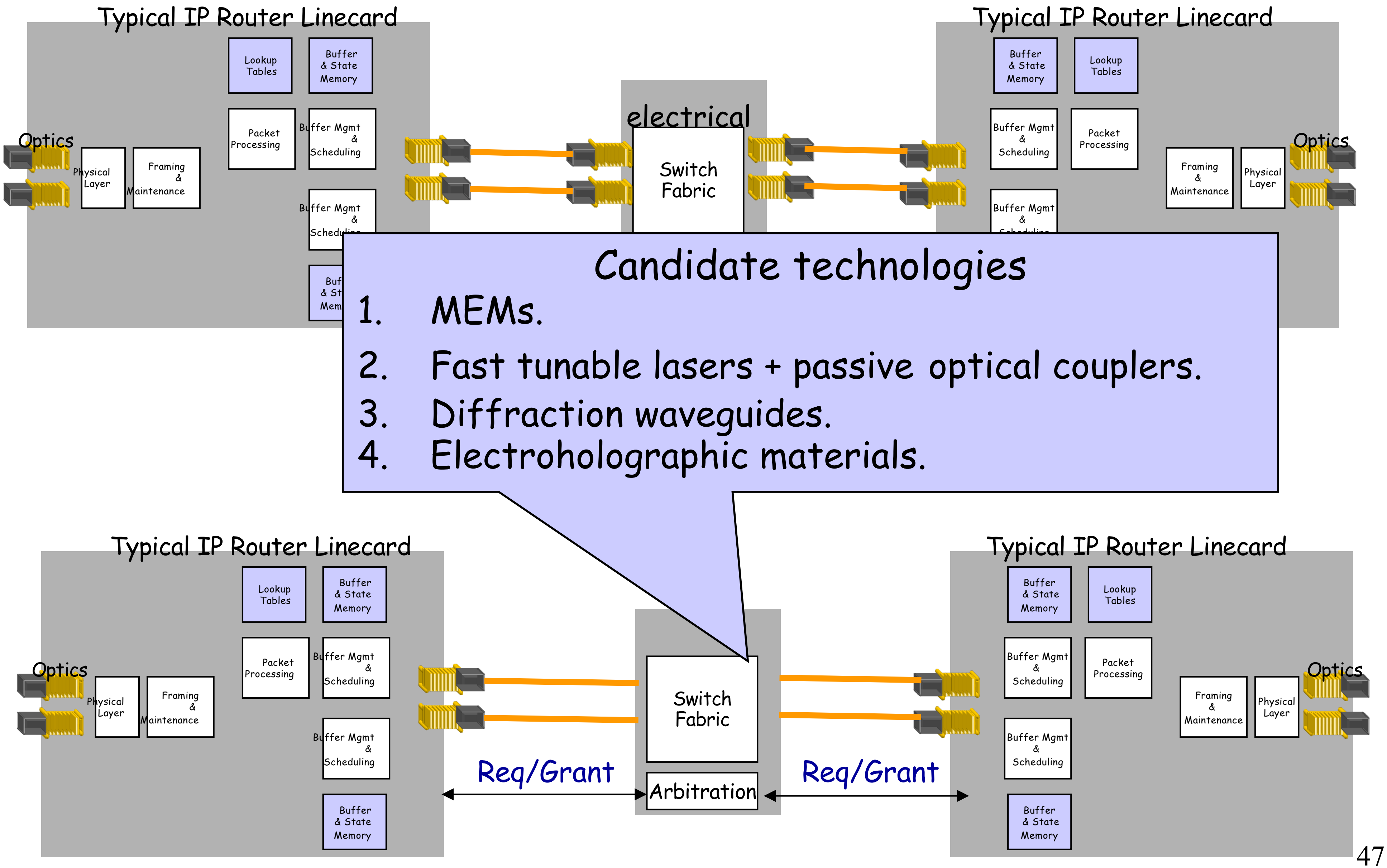
Typical IP Router Linecard



10Gb/s linecard:

- Number of gates: 30M
- Amount of memory: 2Gbits
- Cost: >$20k
- Power: 300W

Replacing the switch fabric with optics



Candidate technologies

1. MEMs.
2. Fast tunable lasers + passive optical couplers.
3. Diffraction waveguides.
4. Electroholographic materials.