

An Introduction to Packet Switching



Nick McKeown

Assistant Professor of Electrical Engineering
and Computer Science, Stanford University

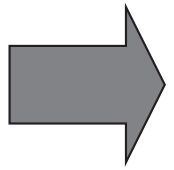
nickm@stanford.edu

<http://www.stanford.edu/~nickm>

Sir William Preece, Chief of the British Postal System, 1876:

"The Americans may have need of the telephone, but we do not. We have plenty of messenger boys."

Outline



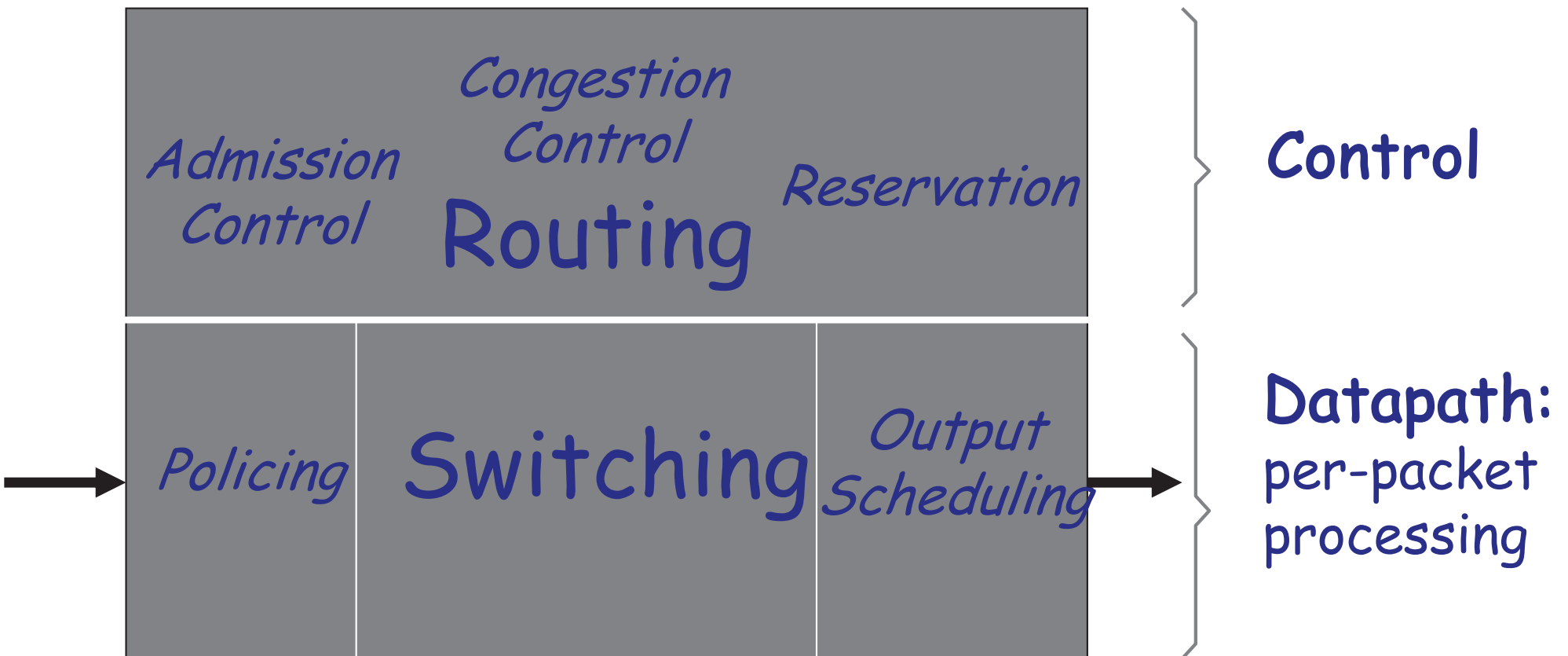
- Introduction
What is a packet-switch?
- The Memory Bandwidth Problem
- Input-Queued Switches
Reducing memory bandwidth requirements
- Combined Input-Output Queued Switches
Making input-queued switches useful
- Parallel Packet Switches
Further reducing memory b/width requirements

Introduction

What is a Packet Switch?

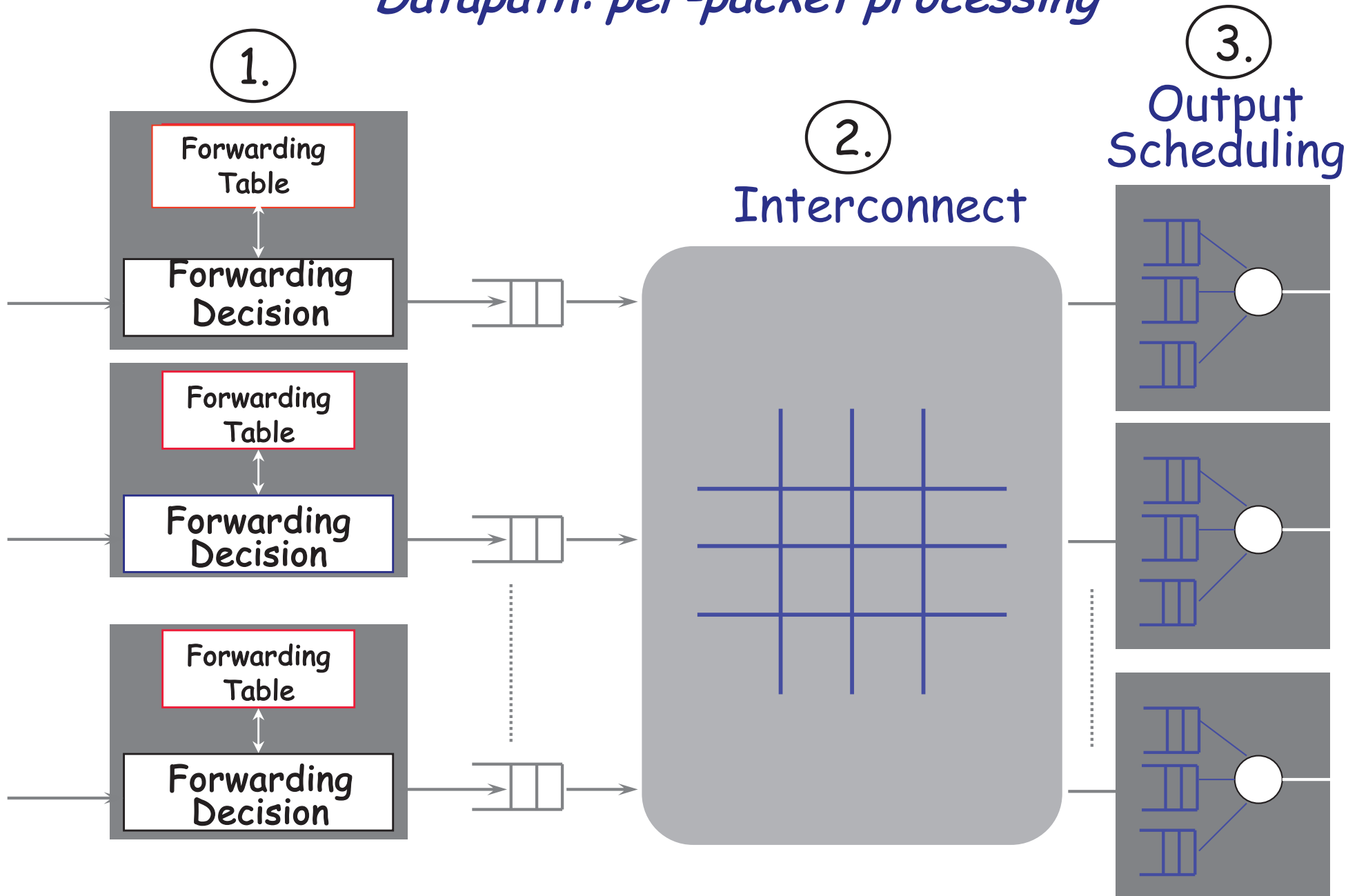
- Introduction
 - What is a packet-switch?*
 - Basic Architectural Components
 - Some Example Packet Switches
 - The Evolution of IP Routers
- The Memory Bandwidth Problem
- Input-Queued Switches
 - Reducing memory bandwidth requirements*
- Combined Input-Output Queued Switches
 - Making input-queued switches useful*
- Parallel Packet Switches
 - Further reducing memory b/width requirements*

Basic Architectural Components

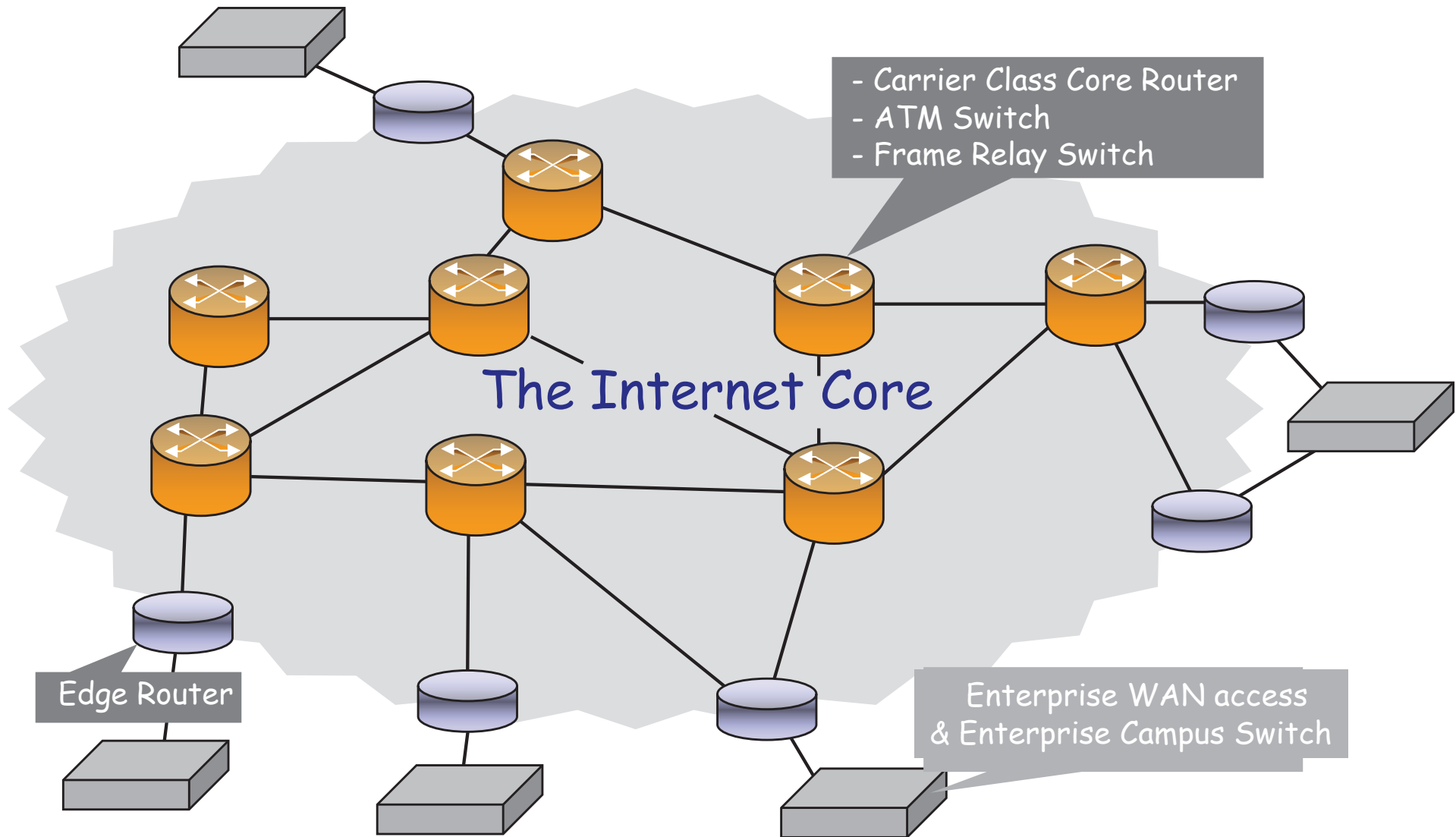


Basic Architectural Components

Datapath: per-packet processing



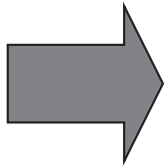
Where high performance packet switches are used



Introduction

What is a Packet Switch?

- Introduction
 - What is a packet-switch?*
 - Basic Architectural Components
 - Some Example Packet Switches
 - The Evolution of IP Routers
- The Memory Bandwidth Problem
- Input-Queued Switches
 - Reducing memory bandwidth requirements*
- Combined Input-Output Queued Switches
 - Making input-queued switches useful*
- Parallel Packet Switches
 - Further reducing memory b/width requirements*



ATM Switch

- Lookup cell VCI/VPI in VC table.
- Replace old VCI/VPI with new.
- Forward cell to outgoing interface.
- Transmit cell onto link.

Ethernet Switch

- Lookup frame DA in forwarding table.
 - If known, forward to correct port.
 - If unknown, broadcast to all ports.
- Learn SA of incoming frame.
- Forward frame to outgoing interface.
- Transmit frame onto link.

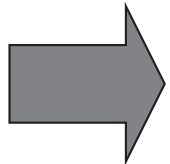
IP Router

- Lookup packet DA in forwarding table.
 - If known, forward to correct port.
 - If unknown, drop packet.
- Decrement TTL, update header Cksum.
- Forward packet to outgoing interface.
- Transmit packet onto link.

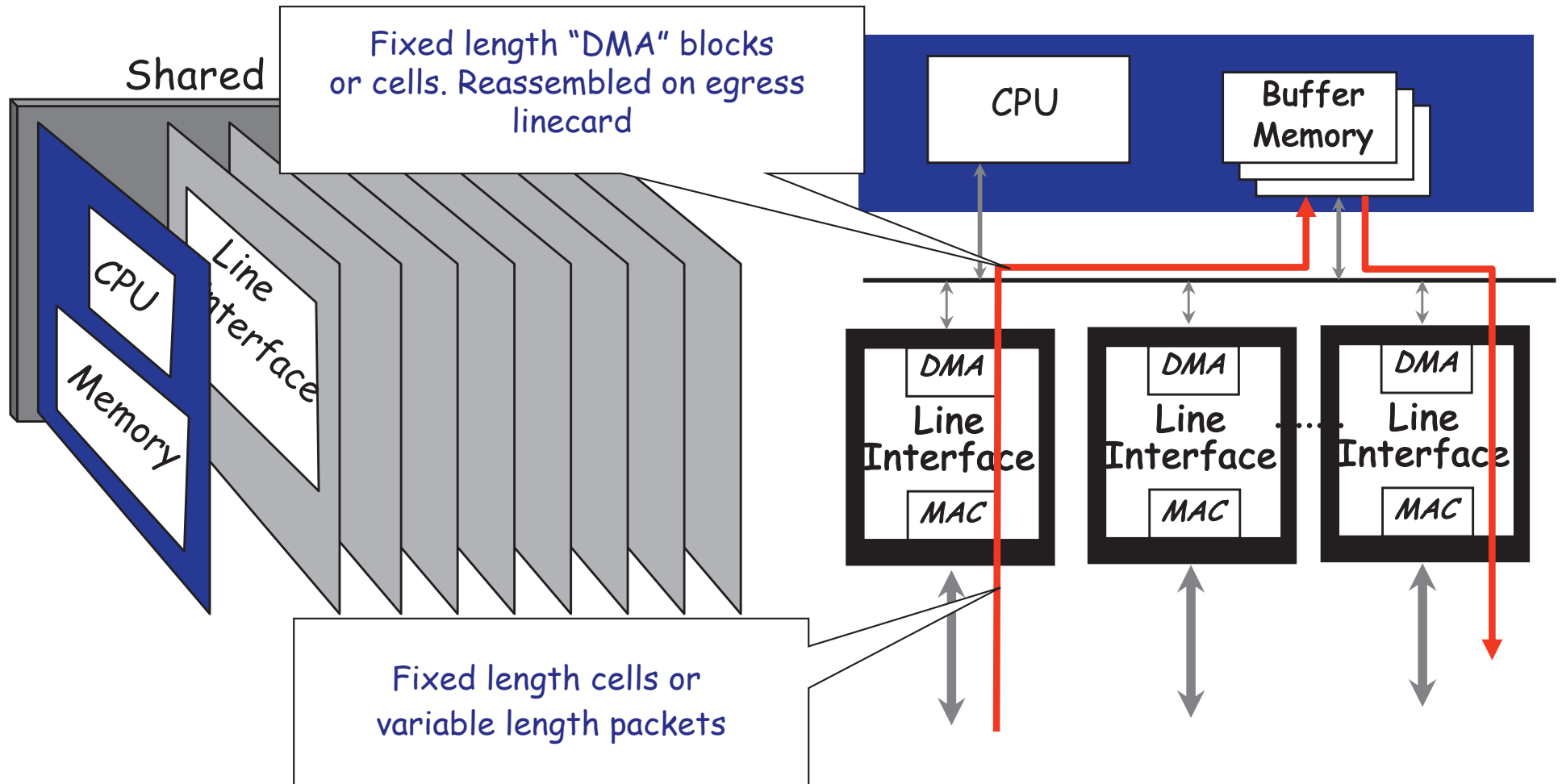
Introduction

What is a Packet Switch?

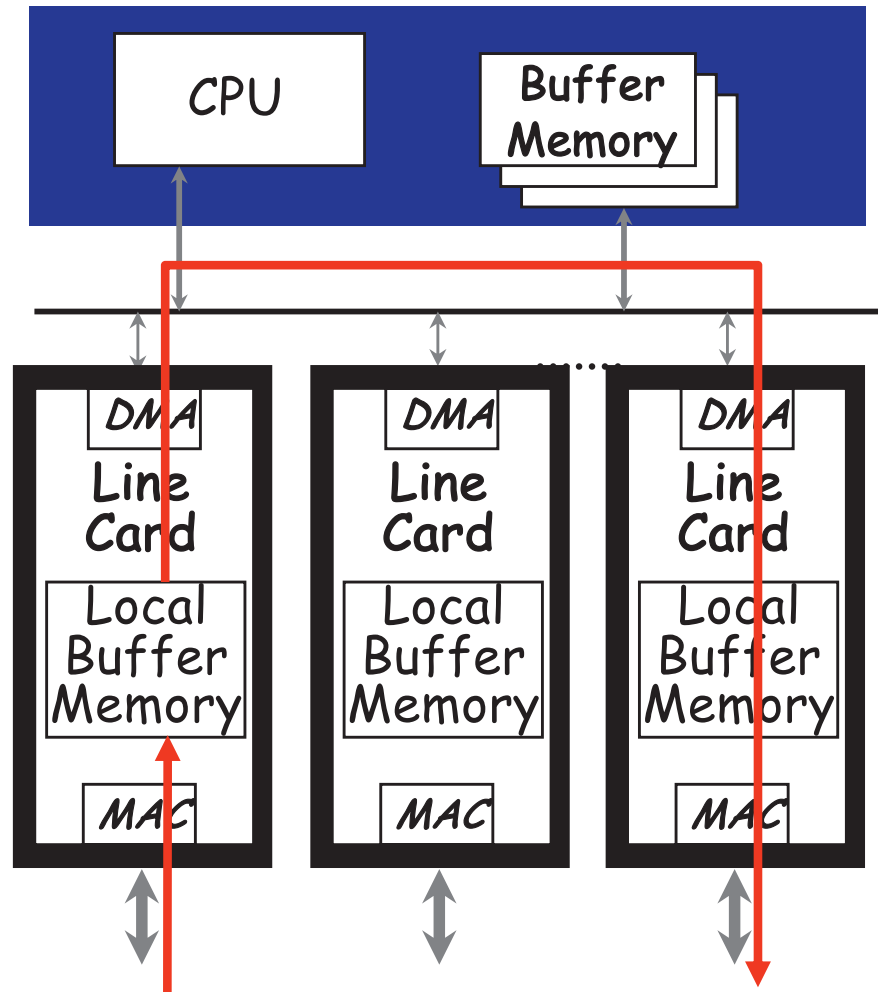
- Introduction
 - What is a packet-switch?*
 - Basic Architectural Components
 - Some Example Packet Switches
 - The Evolution of IP Routers
- The Memory Bandwidth Problem
- Input-Queued Switches
 - Reducing memory bandwidth requirements*
- Combined Input-Output Queued Switches
 - Making input-queued switches useful*
- Parallel Packet Switches
 - Further reducing memory b/width requirements*



First Generation Packet Switches

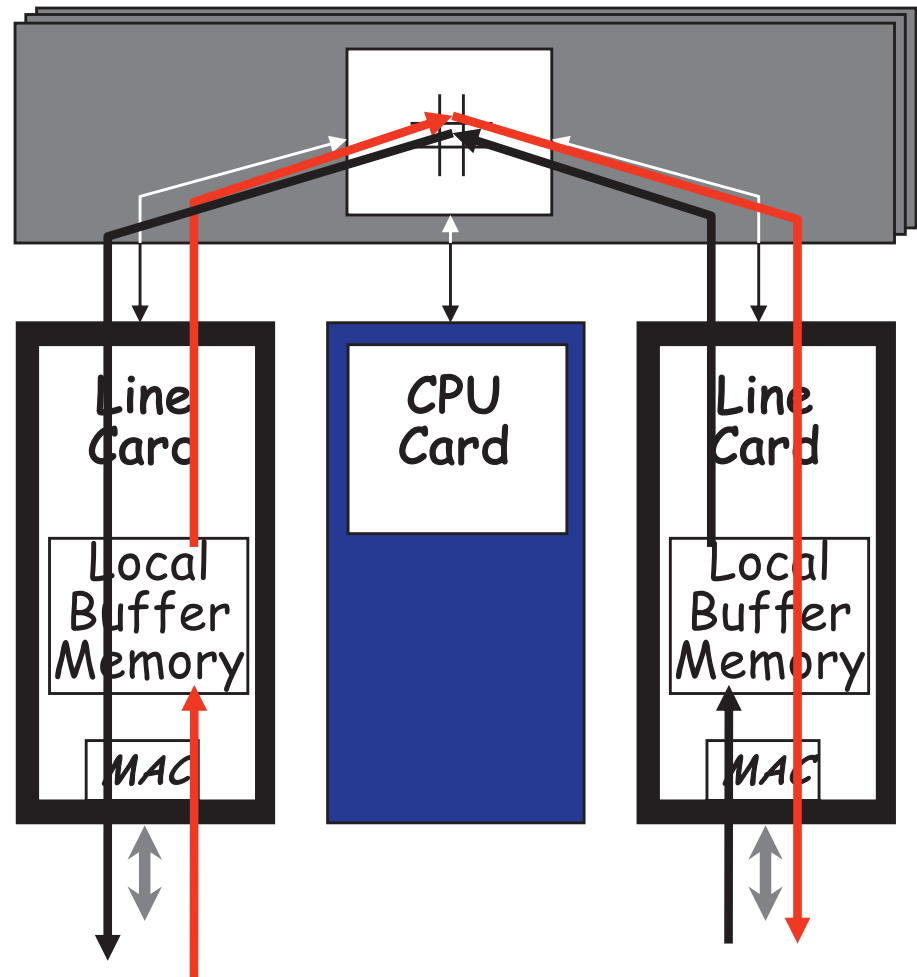
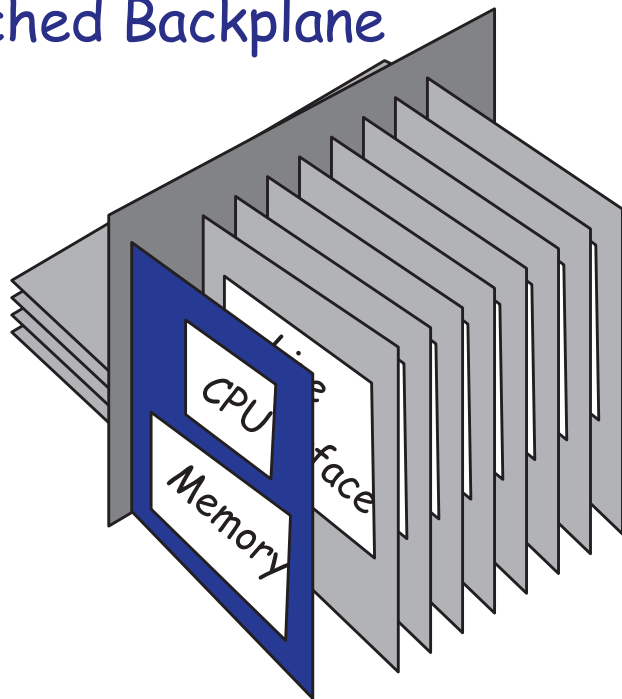


Second Generation Packet Switches

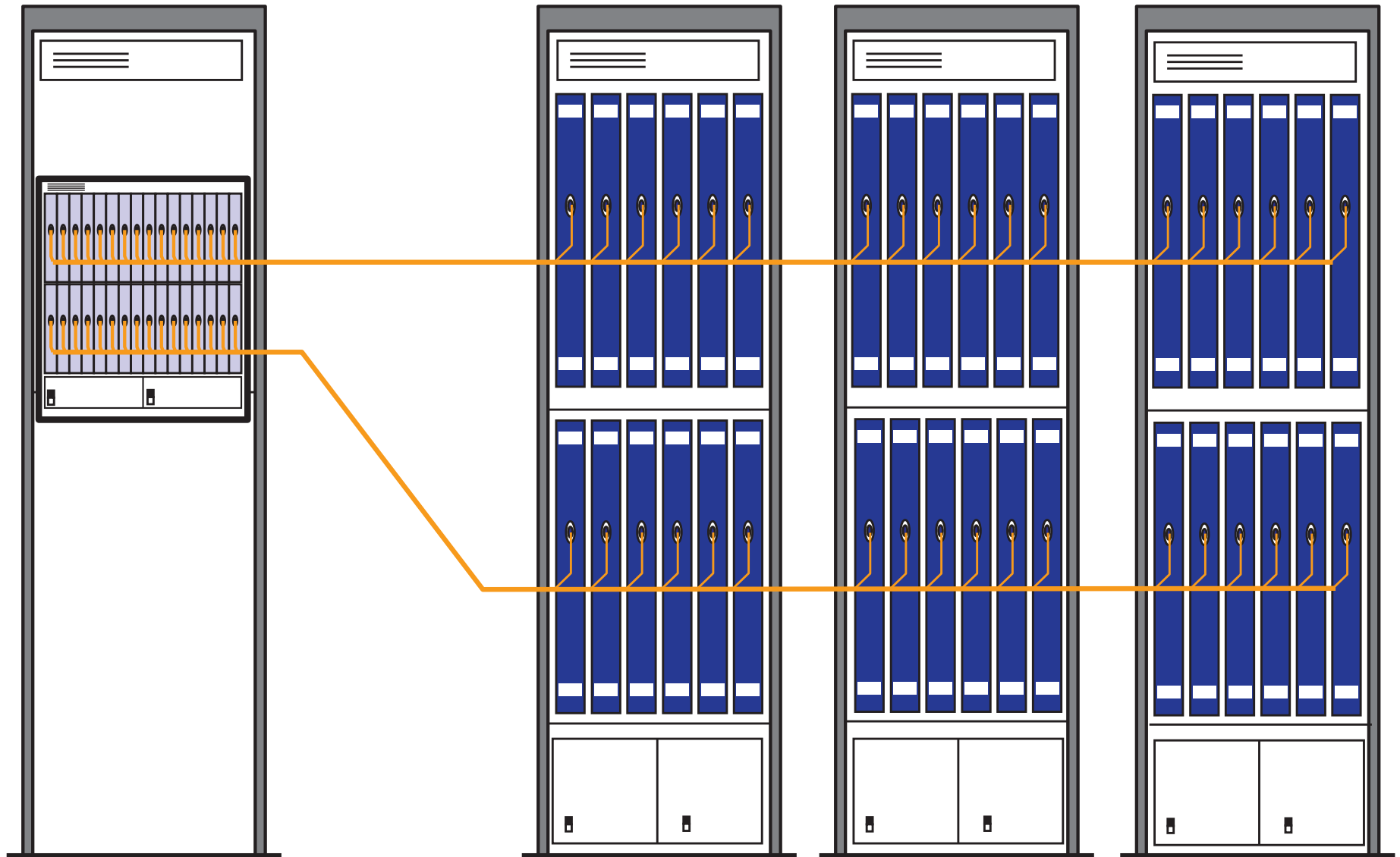


Third Generation Packet Switches

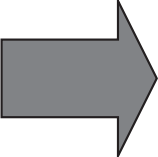
Switched Backplane



Fourth Generation Packet Switches

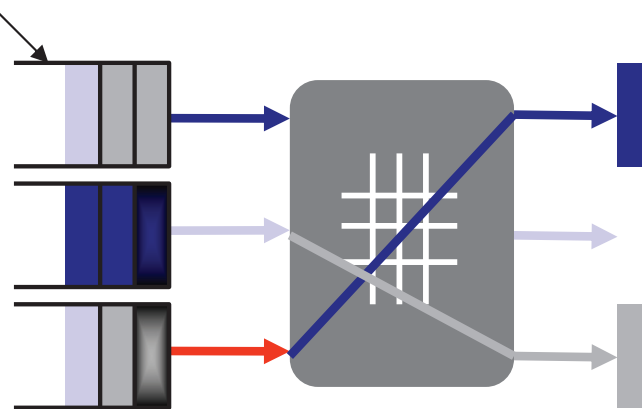


Outline

- Introduction
What is a packet-switch?
- • The Memory Bandwidth Problem
- Input-Queued Switches
Reducing memory bandwidth requirements
- Combined Input-Output Queued Switches
Making input-queued switches useful
- Parallel Packet Switches
Further reducing memory b/width requirements

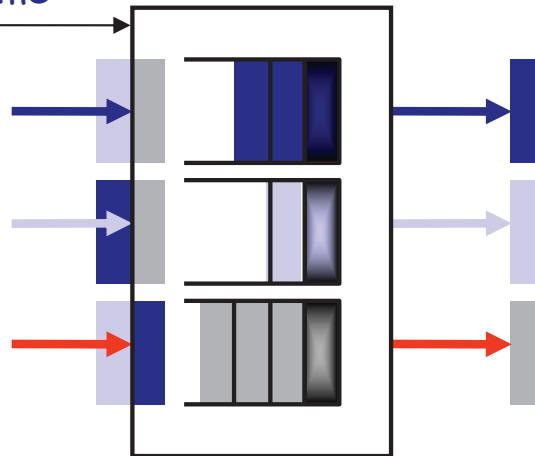
Two Basic Techniques

$1+1 = 2$ operations per cell time



Input-queued Crossbar

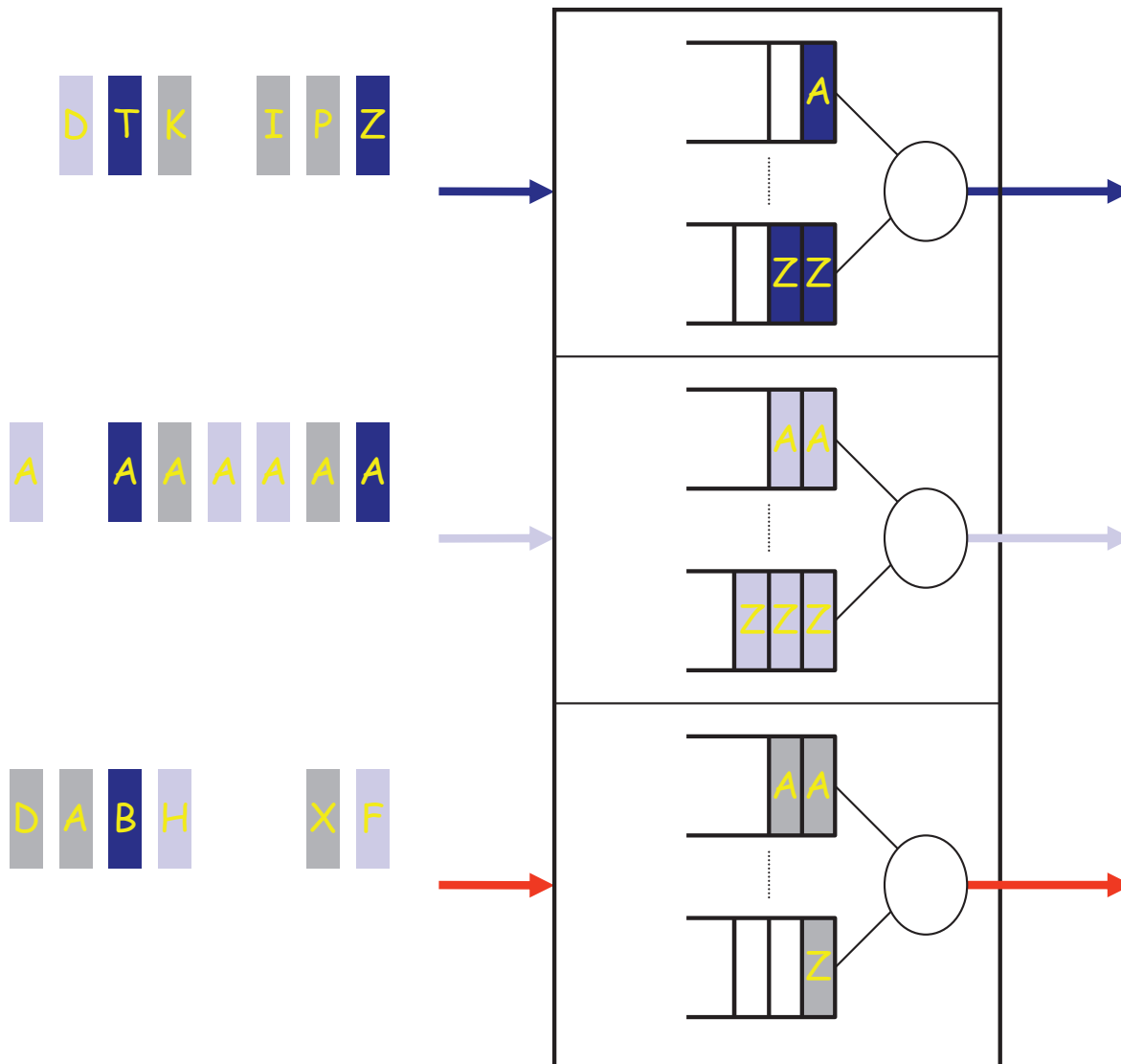
$N+N = 2N$ operations per cell time



Shared Memory

Shared Memory

The Ideal



Numerous work has proven and made possible:

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

A Comparison

Memory speeds for 32x32 switch

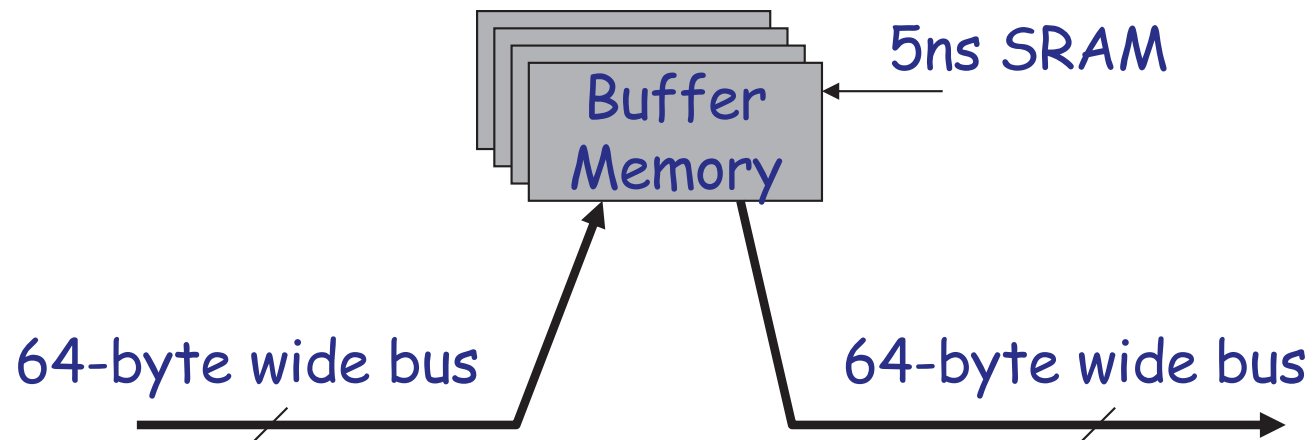
Shared-Memory

Input-queued

Line Rate	Memory BW	Access Time Per cell	Memory BW	Access Time
100 Mb/s	6.4 Gb/s	80 ns	200 Mb/s	2.12 μ s
1 Gb/s	64 Gb/s	8 ns	2 Gb/s	212 ns
2.5 Gb/s	160 Gb/s	3.2 ns	5 Gb/s	84.8 ns
10 Gb/s	640 Gb/s	0.8 ns	20 Gb/s	21.2 ns

Buffer Memory

How Fast Can I Make a Packet Buffer?

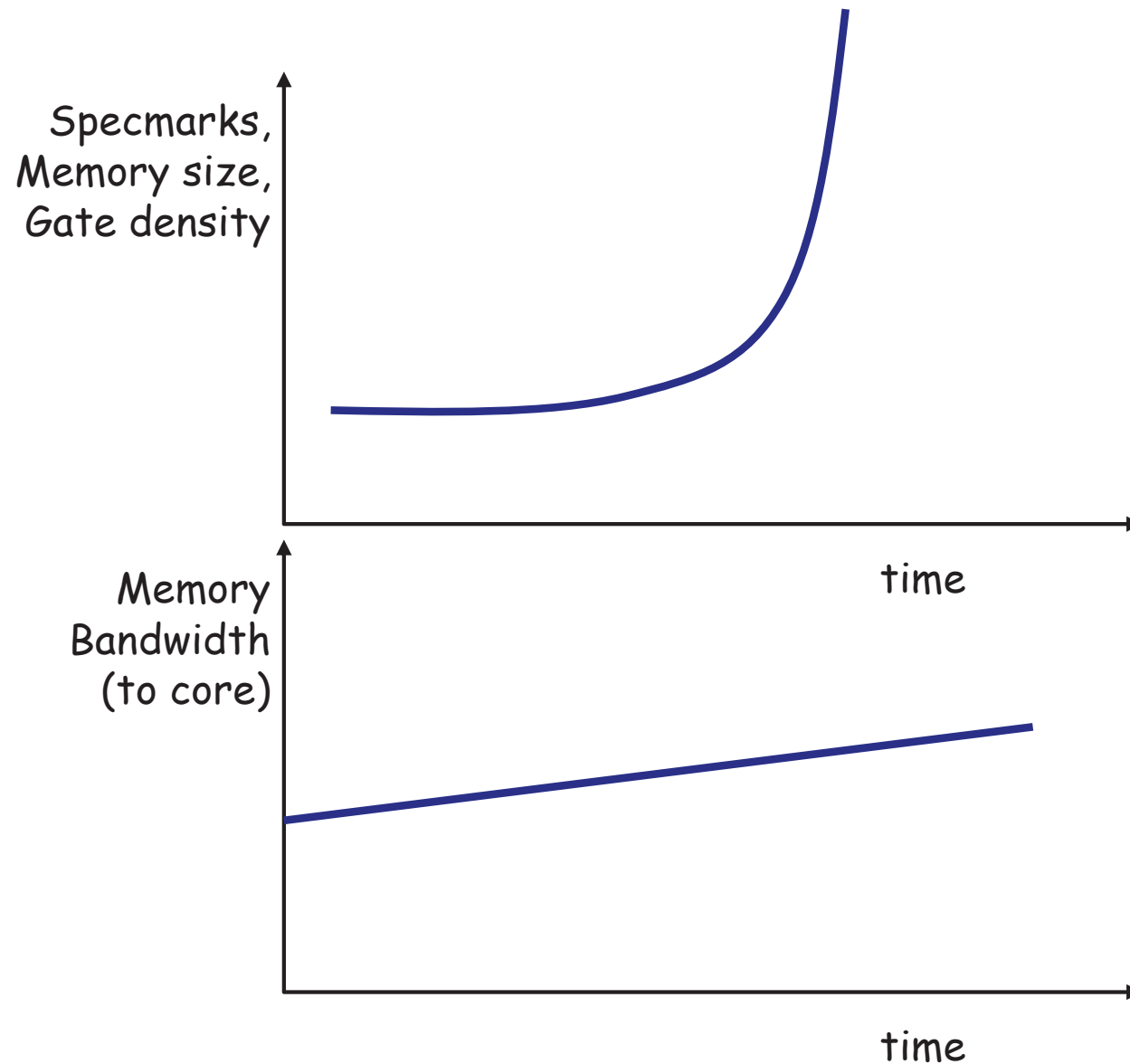


Rough Estimate:

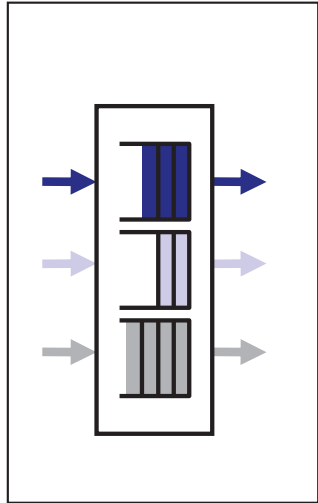
- 5ns per memory operation.
- Two memory operations per packet.
- Therefore, maximum 51.2Gb/s.
- In practice, closer to 40Gb/s.

Buffer Memory

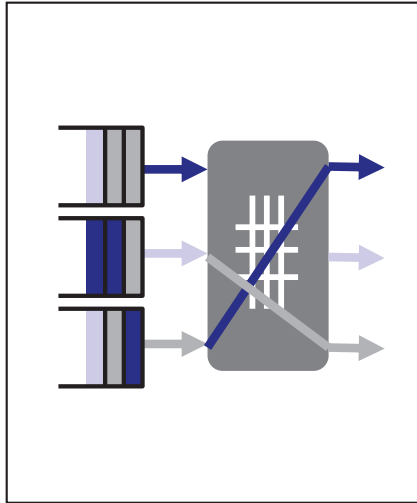
Is It Going to Get Better?



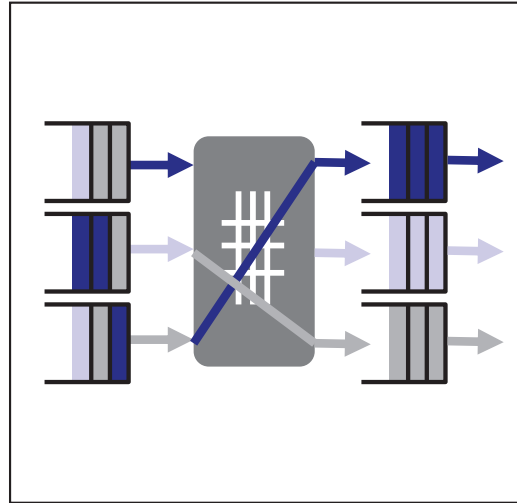
Progression



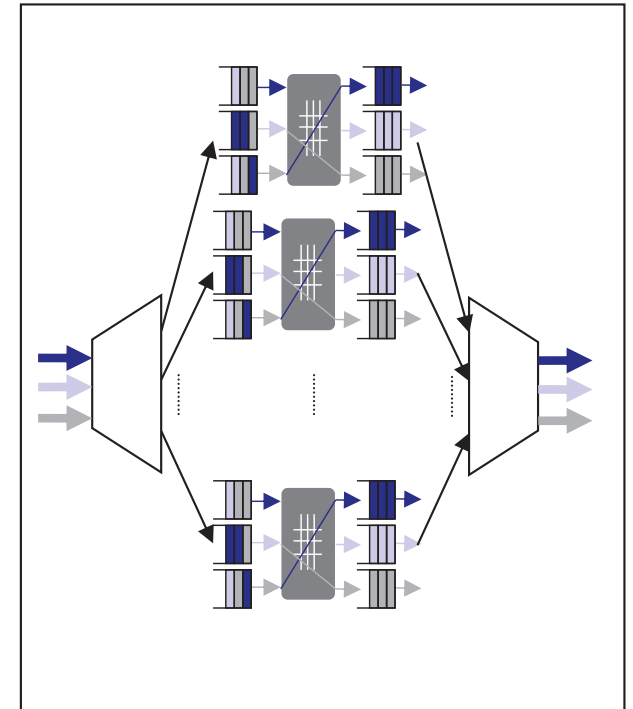
Shared
Memory



Input
Queued

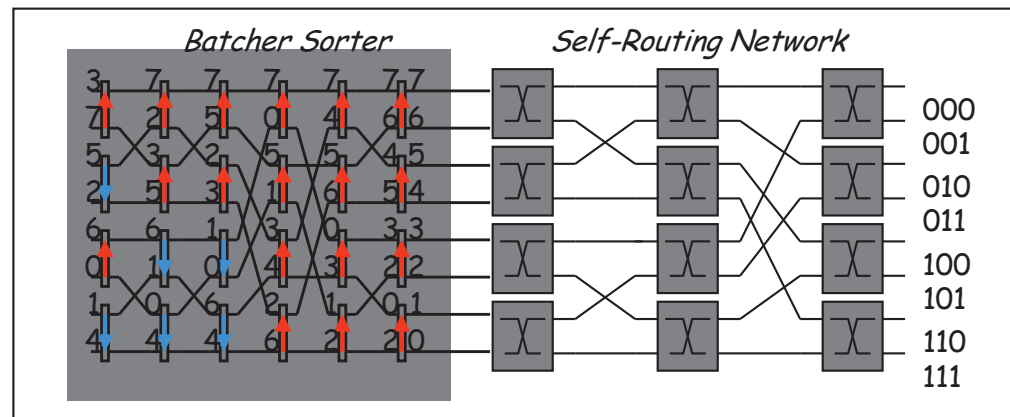


Combined
Input and
Output Queued

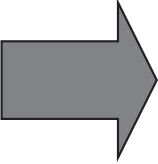


Parallel
Packet
Switches

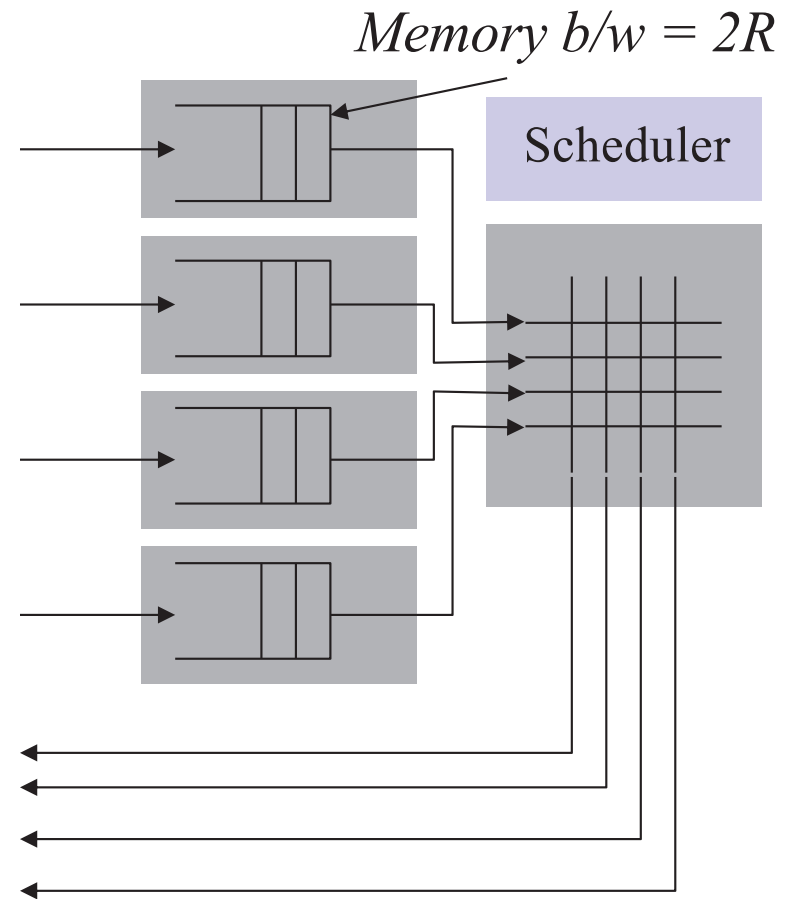
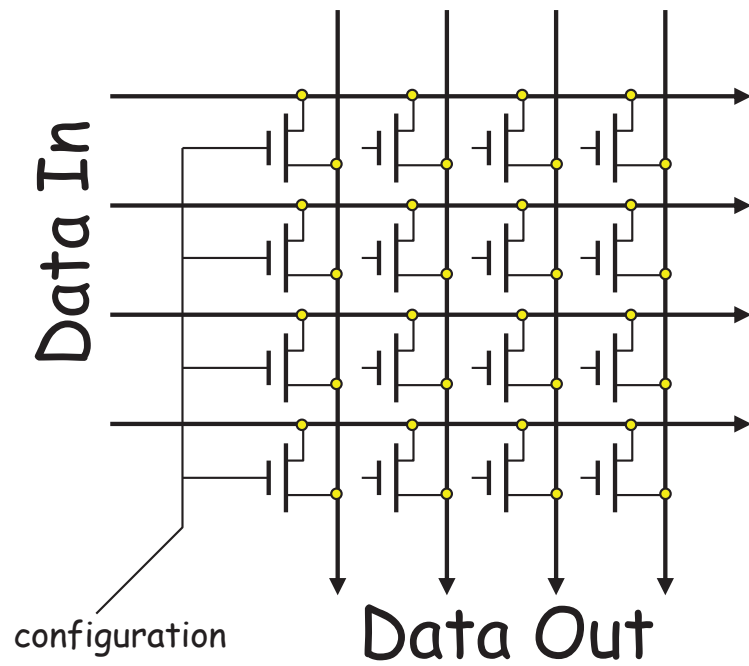
Multi
stage



Outline

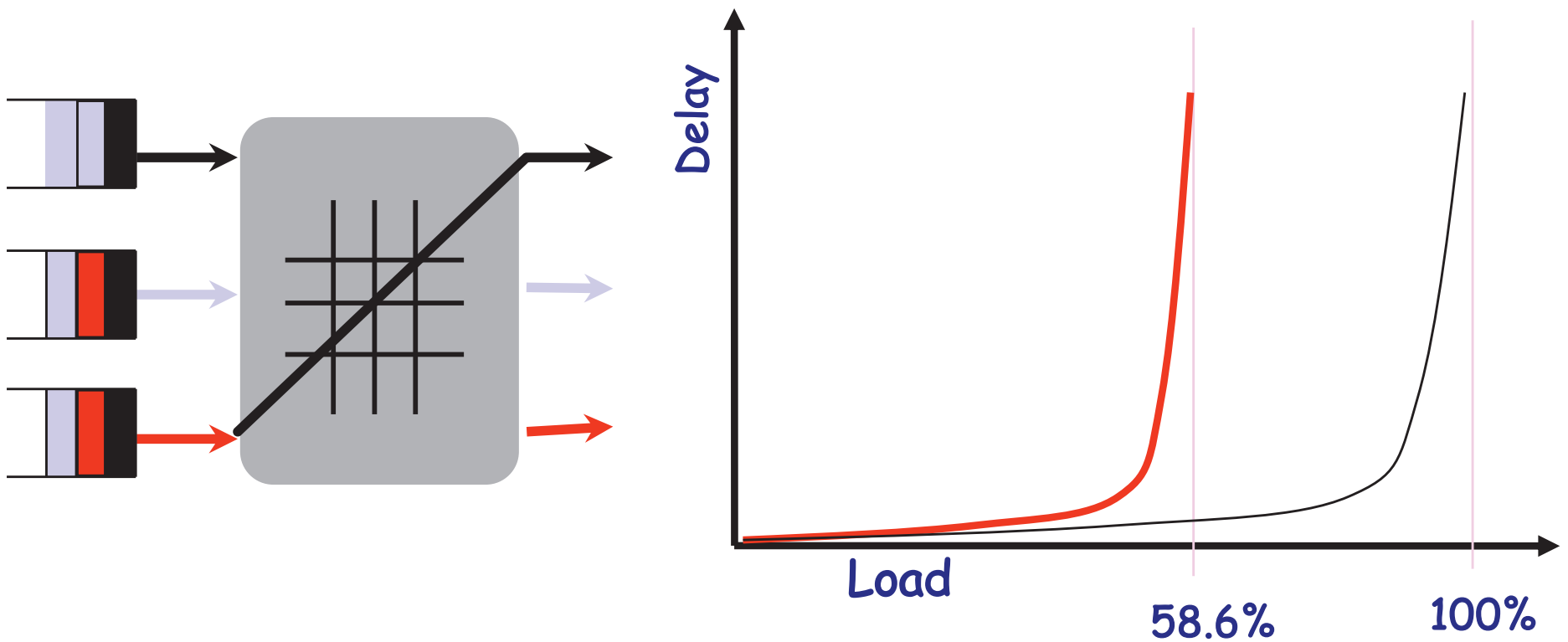
- Introduction
What is a packet-switch?
- The Memory Bandwidth Problem
- • Input-Queued Switches
Reducing memory bandwidth requirements
- Combined Input-Output Queued Switches
Making input-queued switches useful
- Parallel Packet Switches
Further reducing memory b/width requirements

Input Queueing

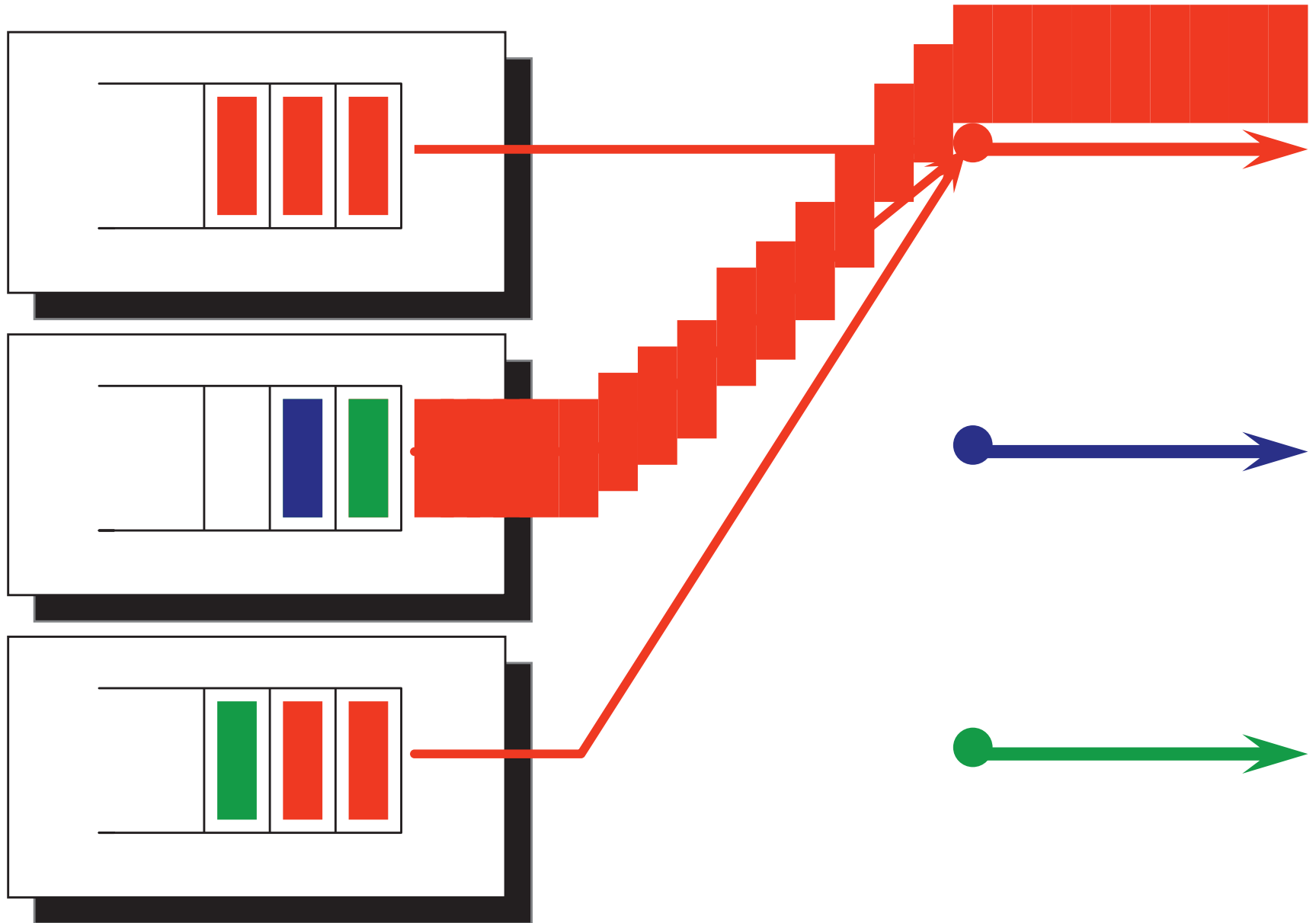


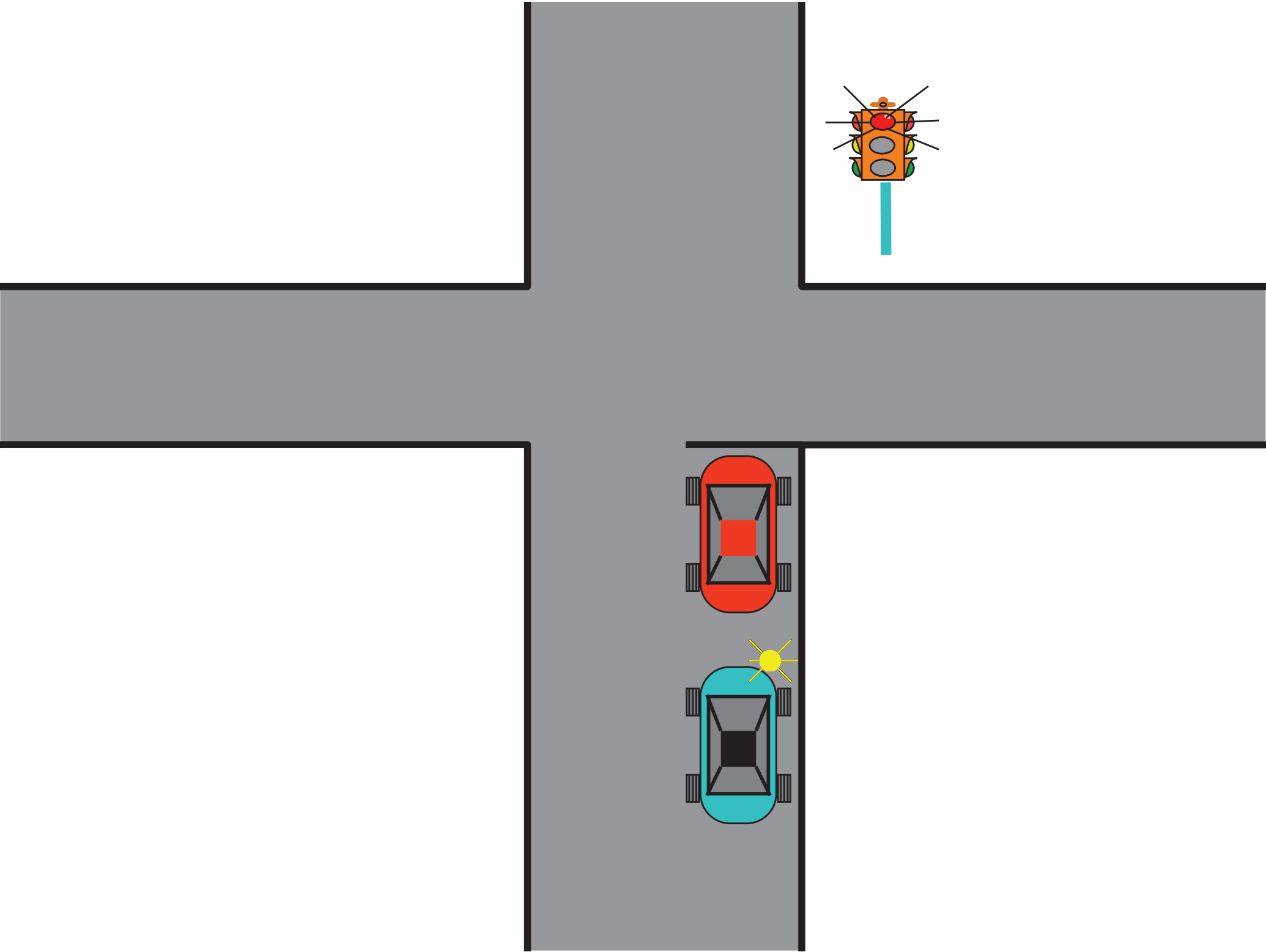
Input Queueing

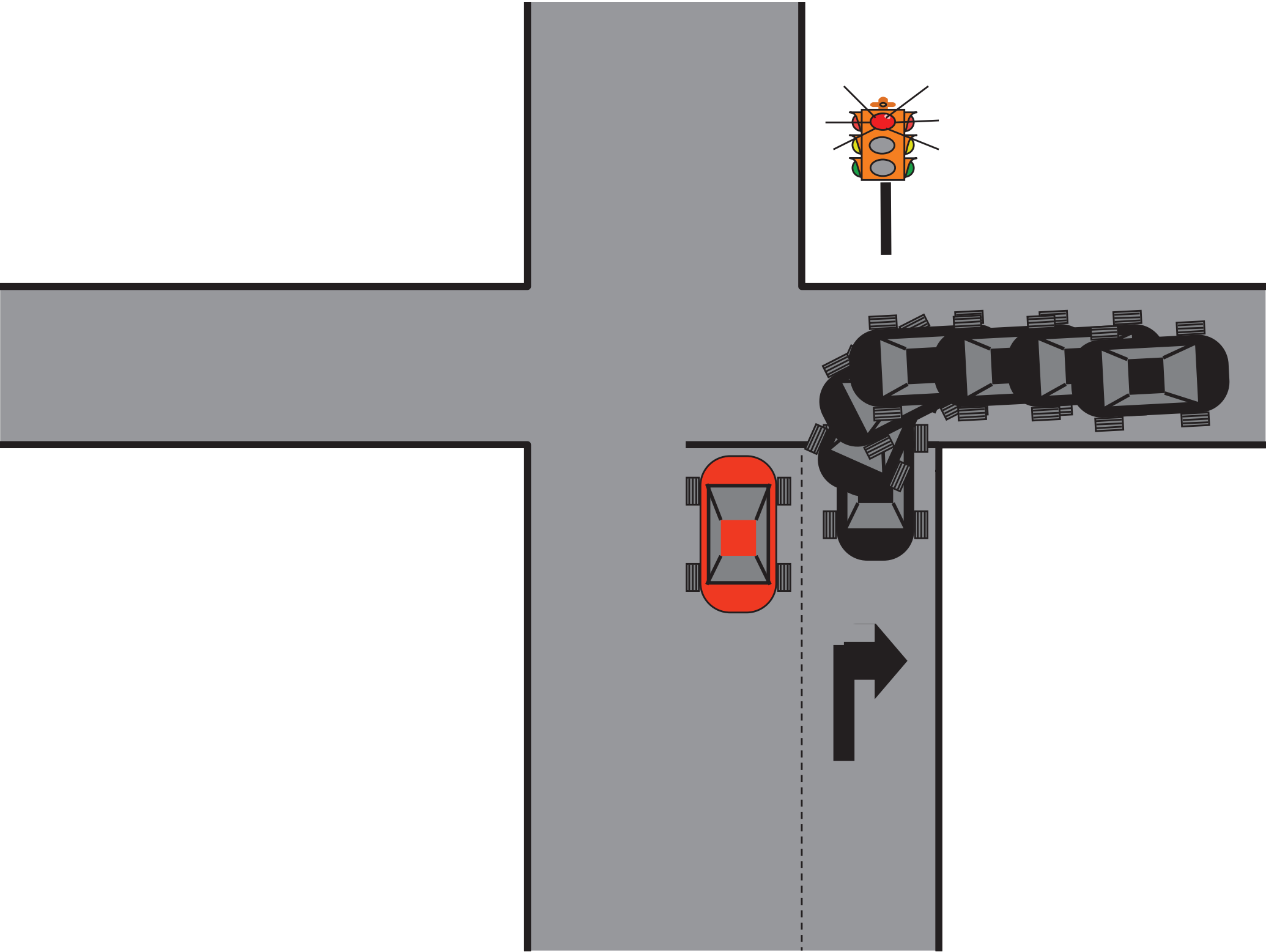
Head of Line Blocking



Head of Line Blocking

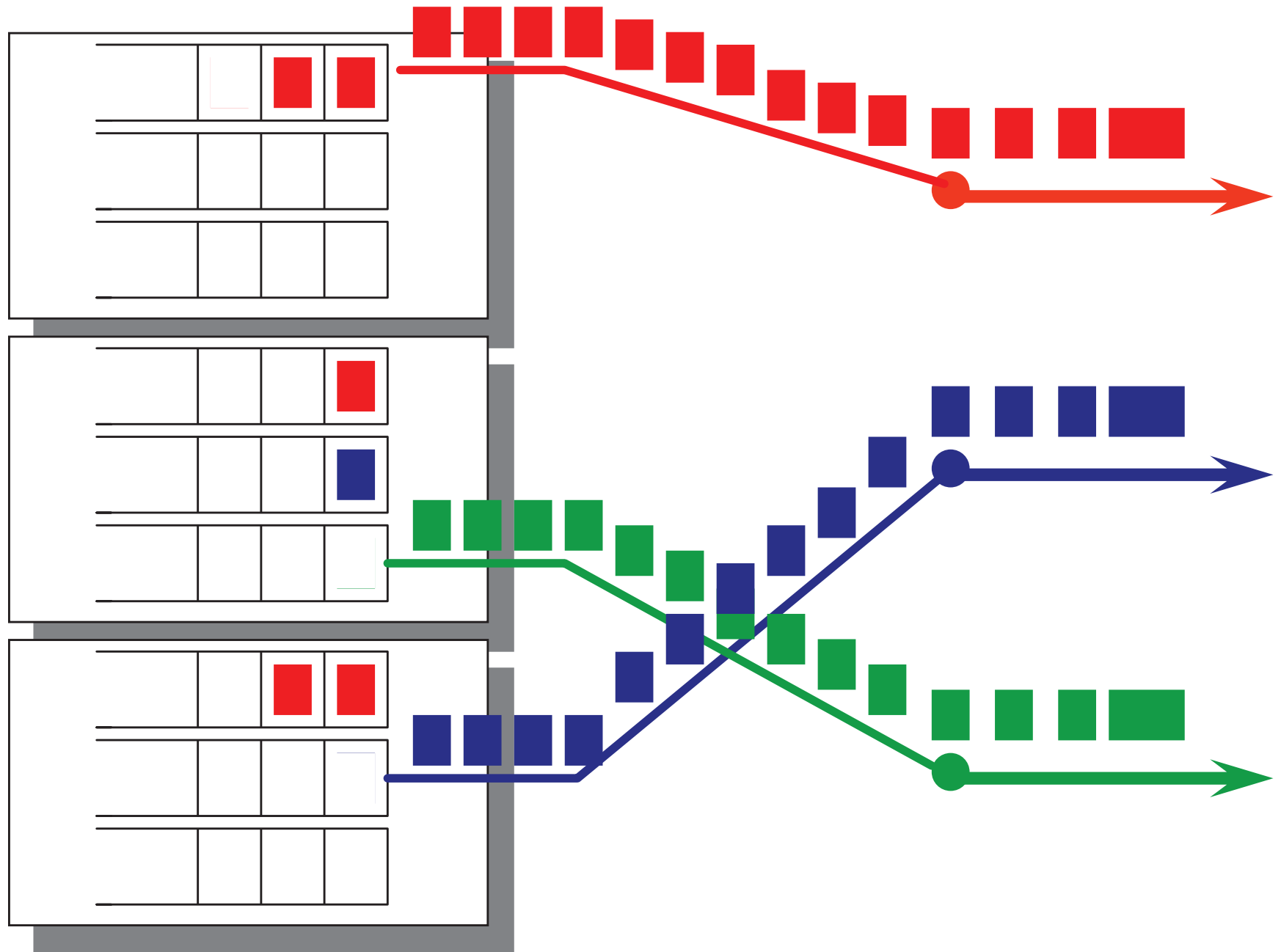






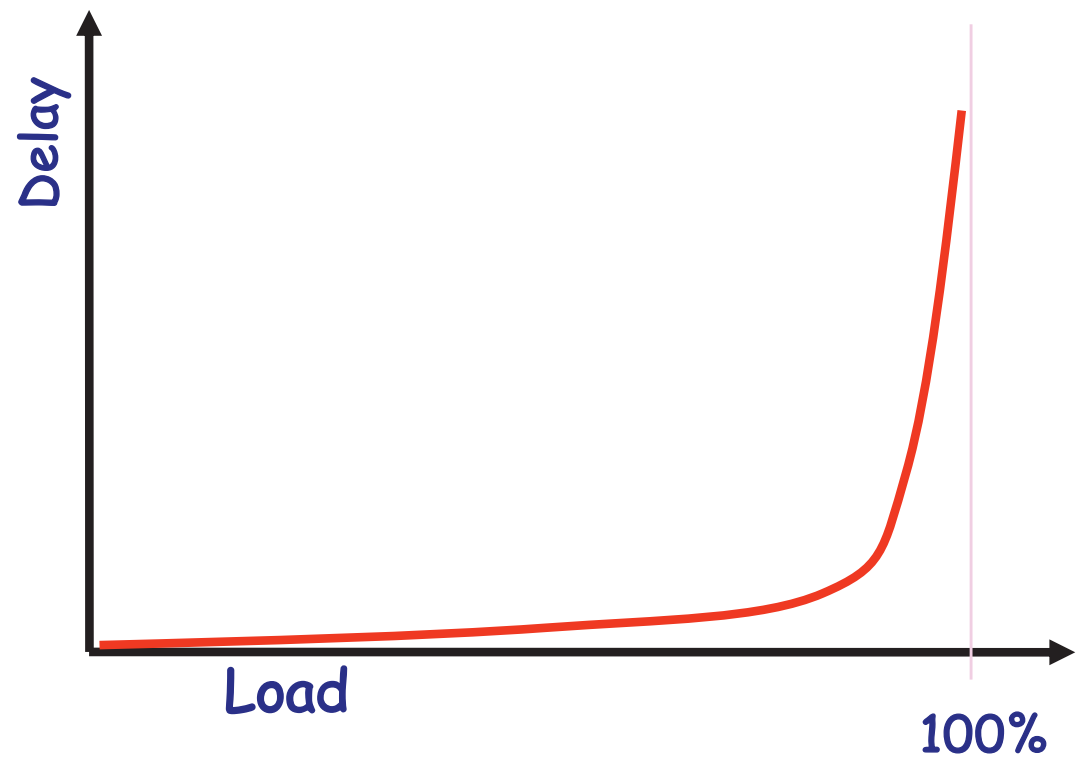
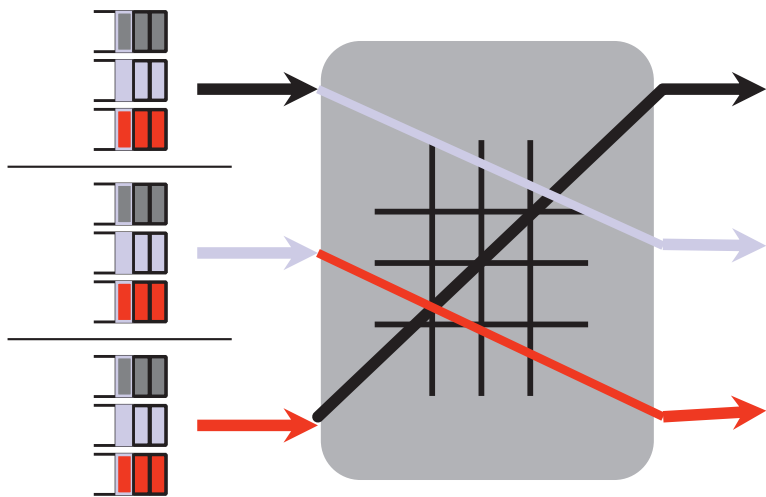
Input Queueing

Virtual output queues



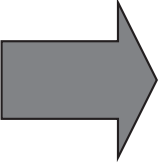
Input Queues

Virtual Output Queues



Proof by Lyapunov function

Outline

- Introduction
What is a packet-switch?
- The Memory Bandwidth Problem
- Input-Queued Switches
Reducing memory bandwidth requirements
-  • Combined Input-Output Queued Switches
Making input-queued switches useful
- Parallel Packet Switches
Further reducing memory b/width requirements

The Speedup Problem

Find a compromise: $1 < \text{Speedup} \ll N$

- to get the performance of a shared memory switch
- close to the cost of an IQ switch

Some Early Approaches

Probabilistic Analyses

- assume traffic models (Bernoulli, Markov-modulated, non-uniform loading, "friendly correlated")
- obtain mean throughput and delays, bounds on tails
- analyze different fabrics (crossbar, multistage, etc)

Numerical Methods

- use actual and simulated traffic traces
- run different algorithms
- set the "speedup dial" at various values

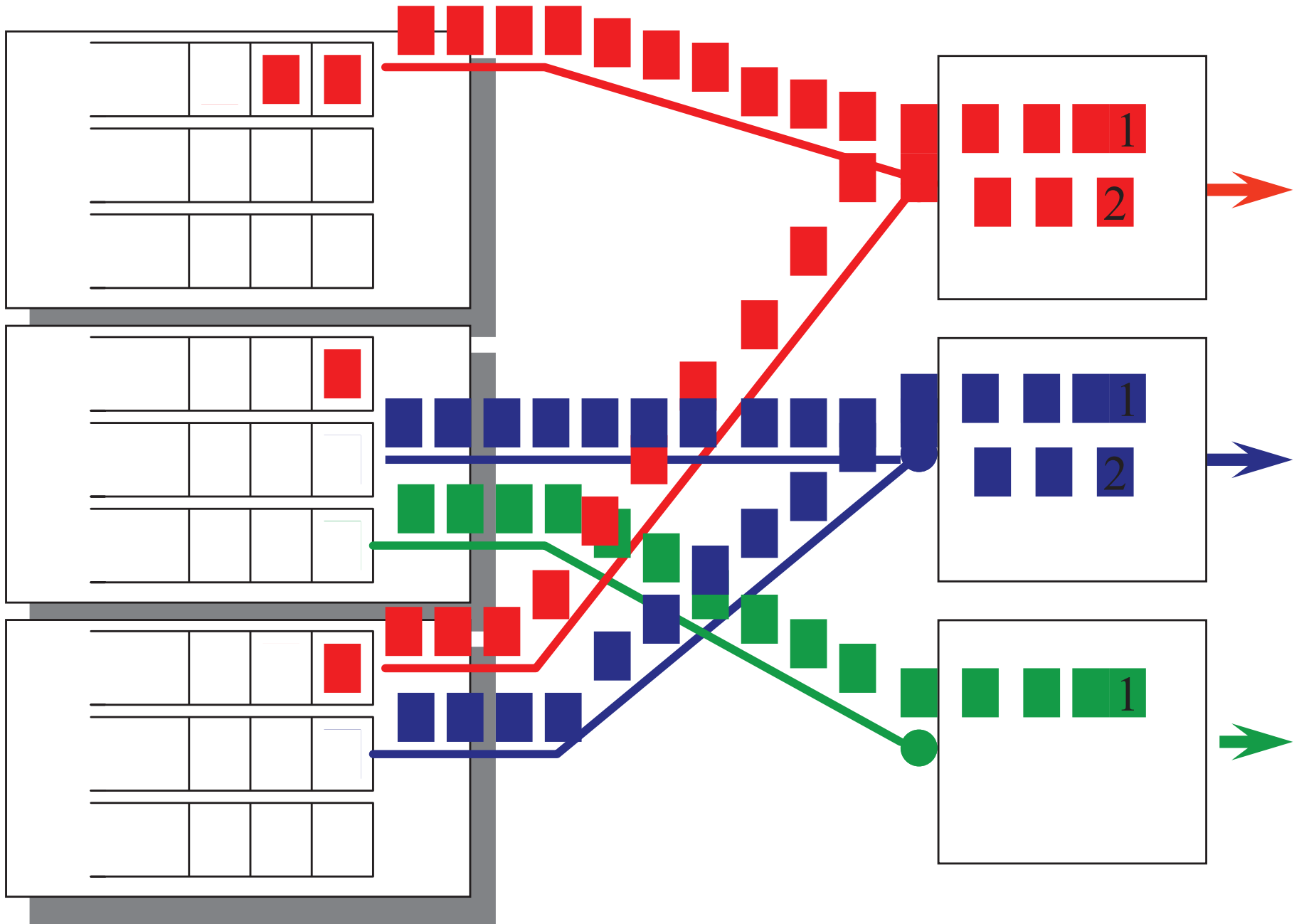
The findings

Very tantalizing ...

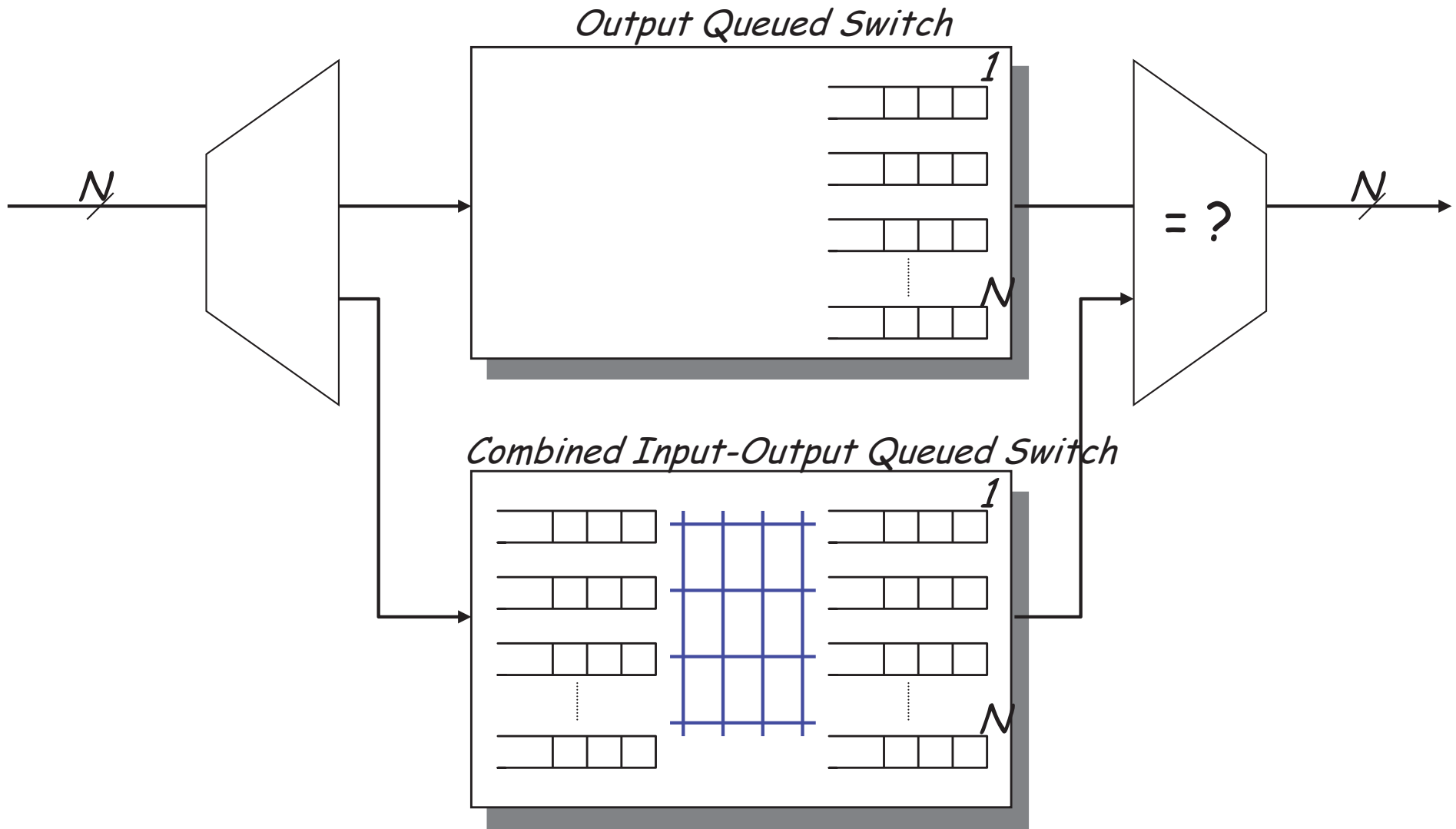
- under different settings
(traffic, loading, algorithm, etc)
- and even for varying switch sizes

A speedup of between 2 and 5 was sufficient!

Using Speedup



The Ideal Solution



Interesting Result

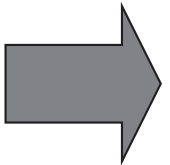
Theorem:

For a switch with combined input and output queueing to *exactly* mimic an output queued switch, for all types of traffic, a speedup of $2-1/N$ is necessary and sufficient.

Joint work with Balaji Prabhakar, Ashish Goel and Shang-tse Chuang.

Outline

- Introduction
What is a packet-switch?
- The Memory Bandwidth Problem
- Input-Queued Switches
Reducing memory bandwidth requirements
- Combined Input-Output Queued Switches
Making input-queued switches useful
- Parallel Packet Switches
Further reducing memory b/width requirements



Optical Physical Layers...

...are Going to Make Things "Worse"

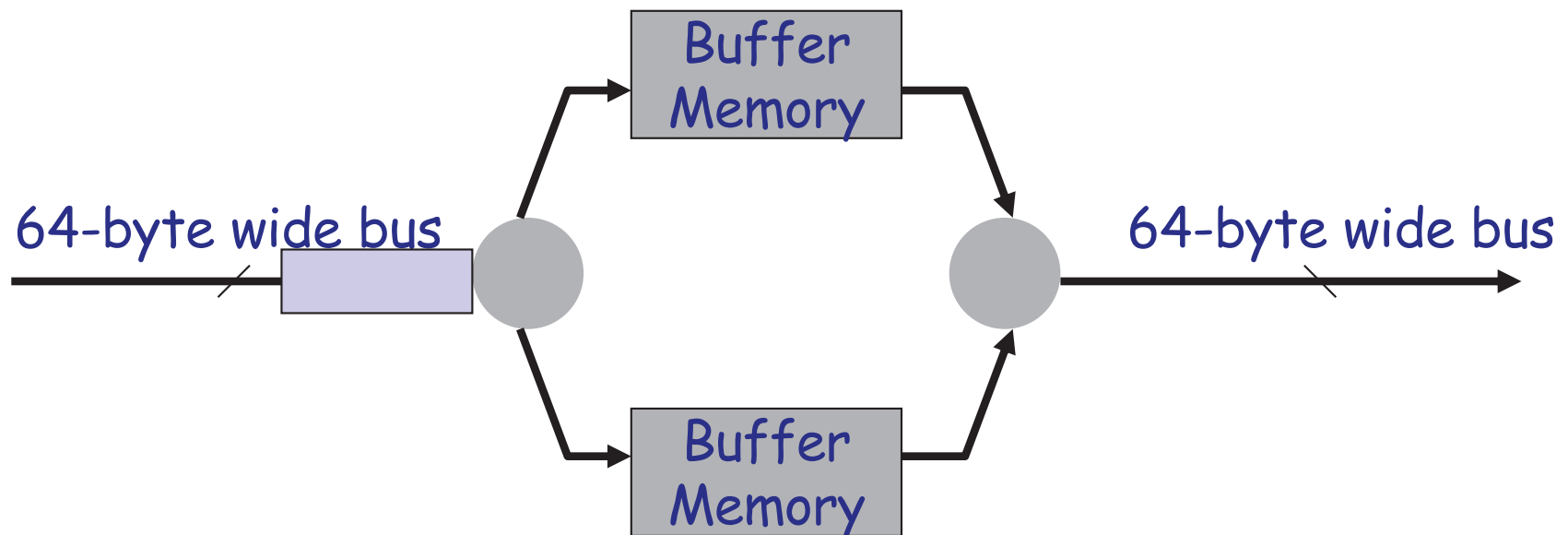
DWDM:

- More λ 's per fiber $\Rightarrow$ more ports per switch.
- # ports: 16, ..., 1000's.

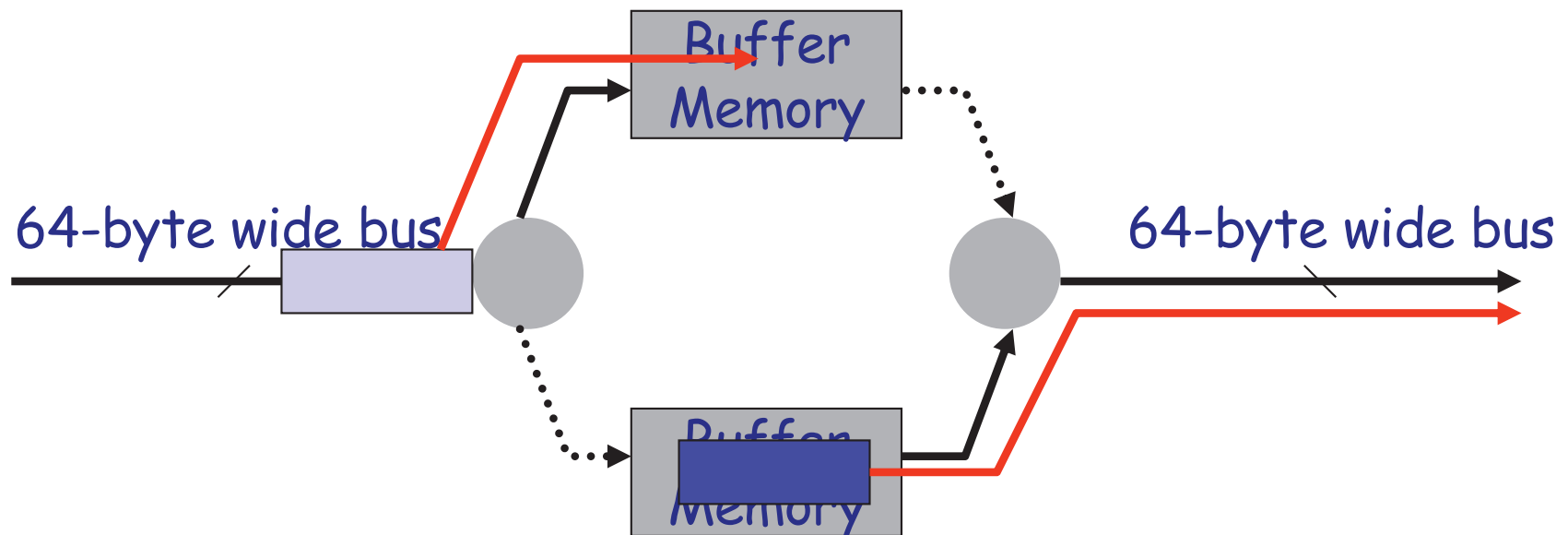
Data rate:

- More b/s per λ $\Rightarrow$ higher capacity.
- Data rates: 2.5Gb/s, 10Gb/s, 40Gb/s, 160Gb/s, ...

Approach #1: Ping-pong Buffering

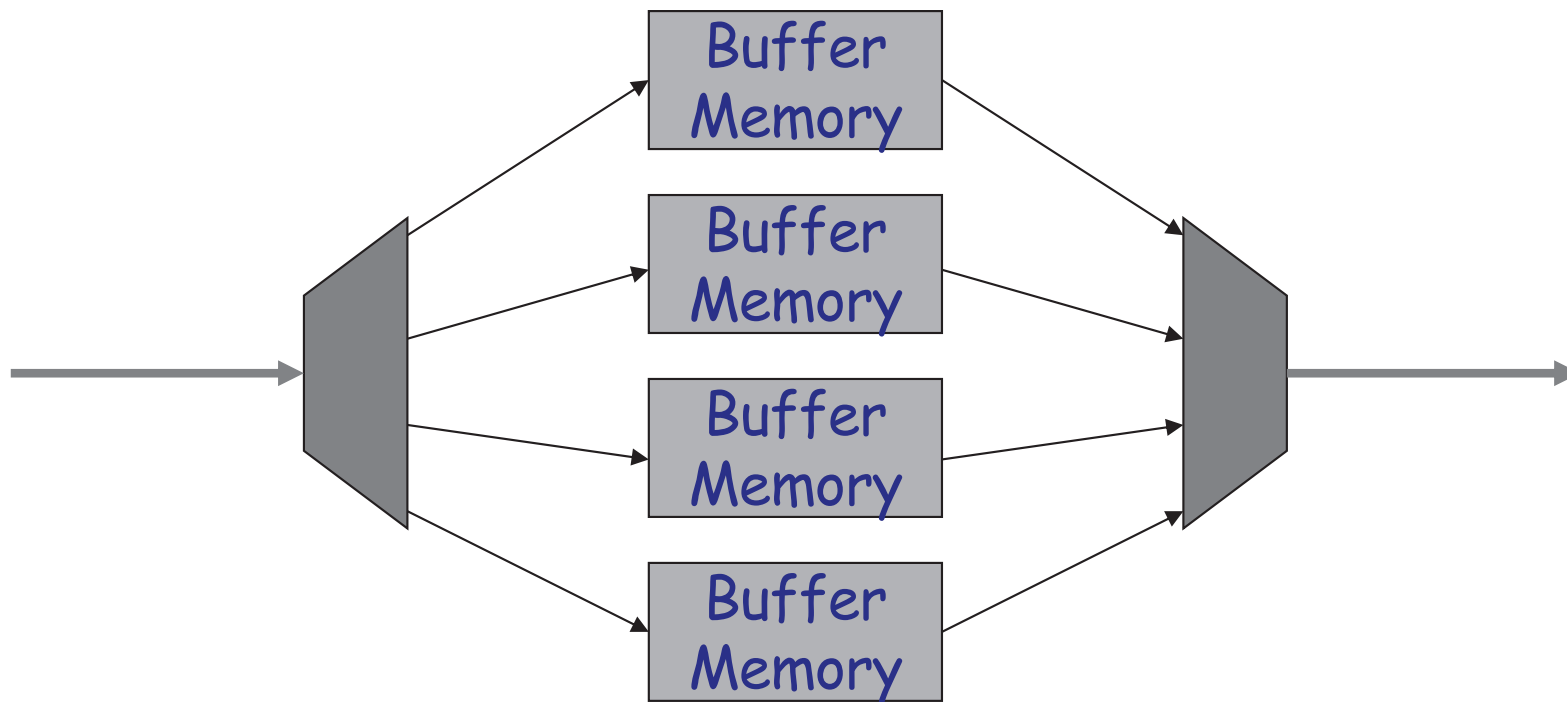


Approach #1: Ping-pong Buffering

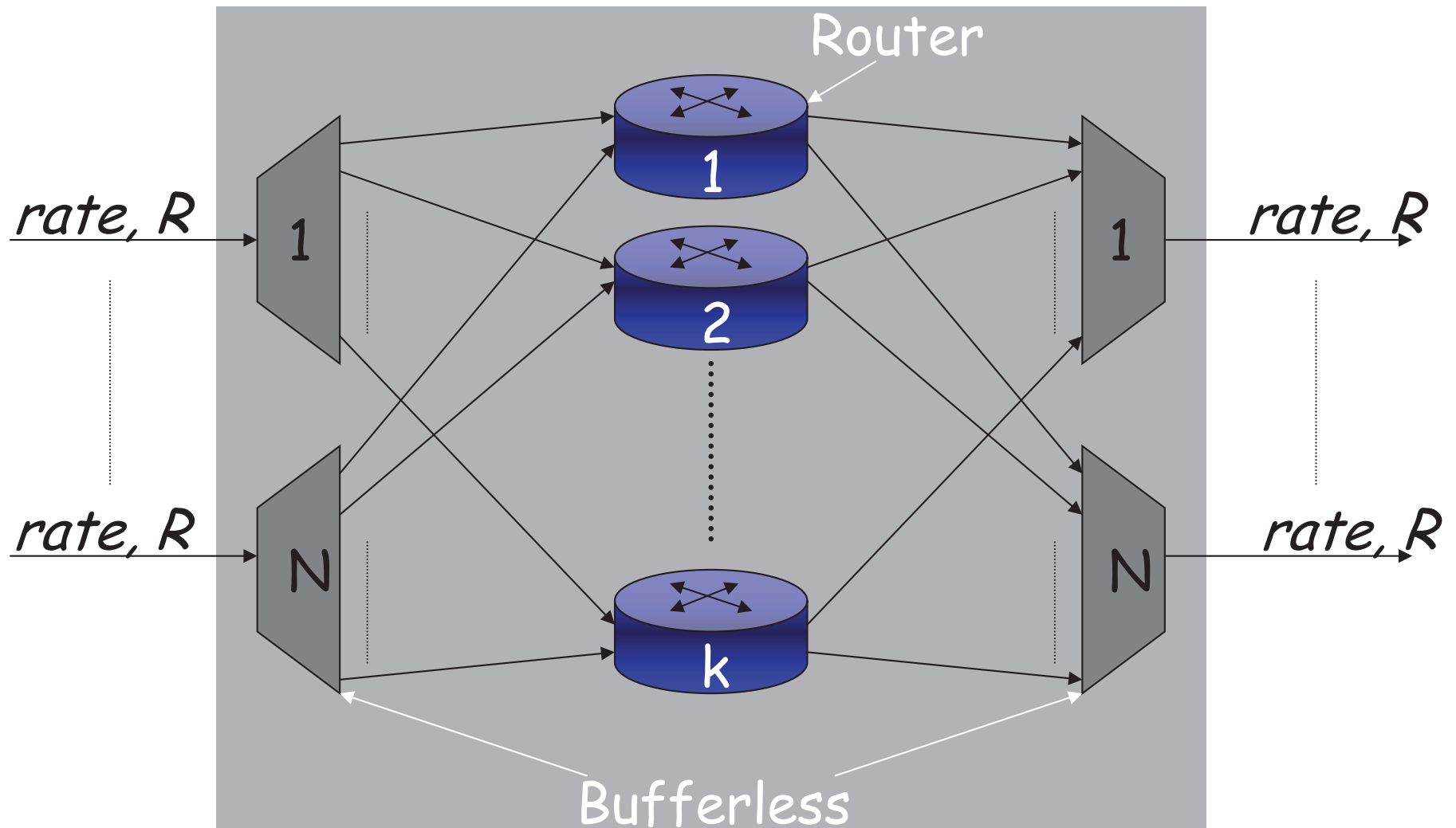


Memory bandwidth doubled to ~80 Gb/s

Approach #2: Multiple Parallel Buffers aka Banking, Interleaving



The Fork Join Router



The Fork Join Router

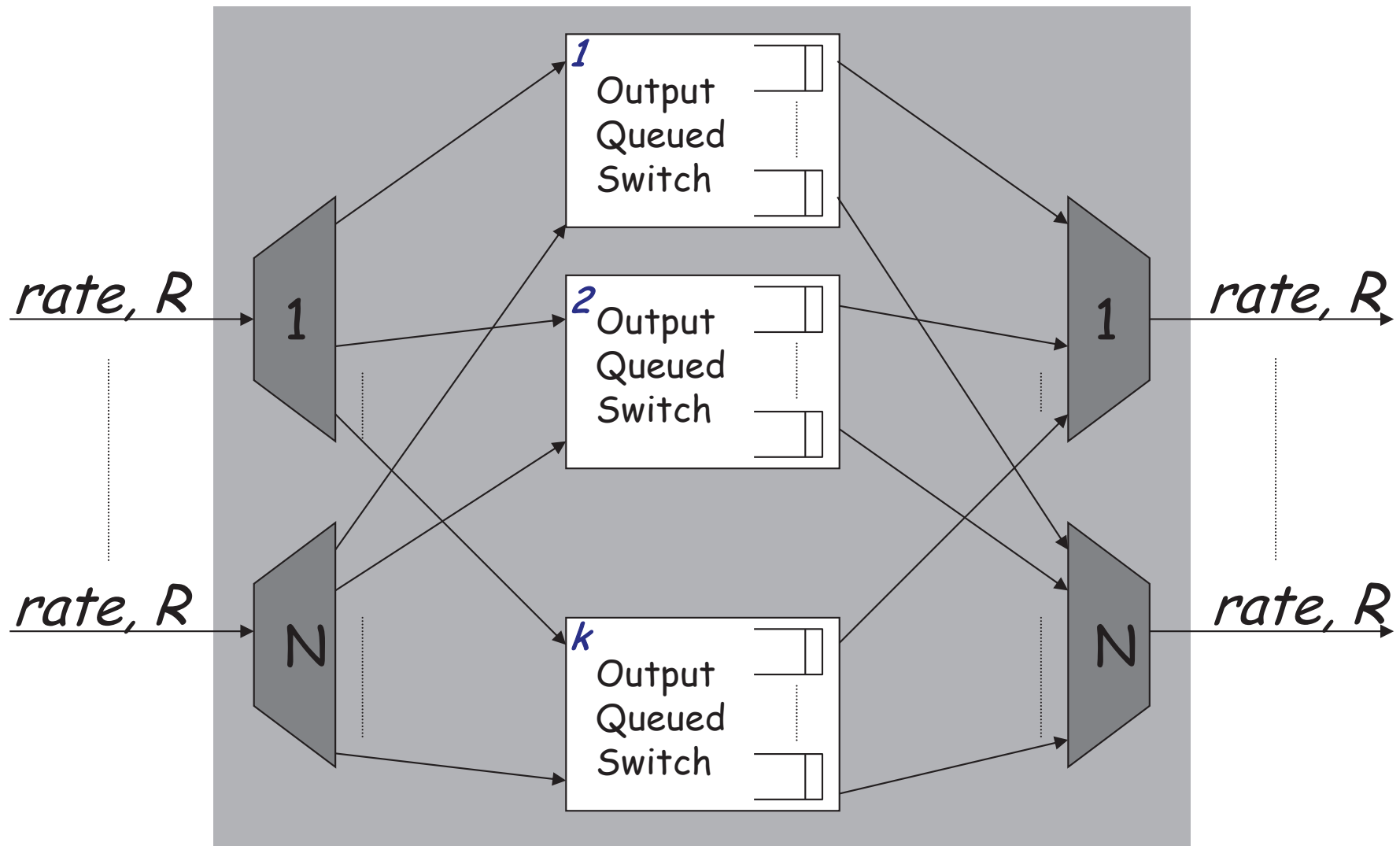
- Advantages

- $k \uparrow \Rightarrow$ memory bandwidth $\downarrow$
- $k \uparrow \Rightarrow$ lookup/classification rate $\downarrow$
- $k \uparrow \Rightarrow$ routing/classification table size $\downarrow$

- Problems

- How to demultiplex prior to lookup/classification?
- How does the system perform/behave?
- Can we predict/guarantee performance?

A Parallel Packet Switch



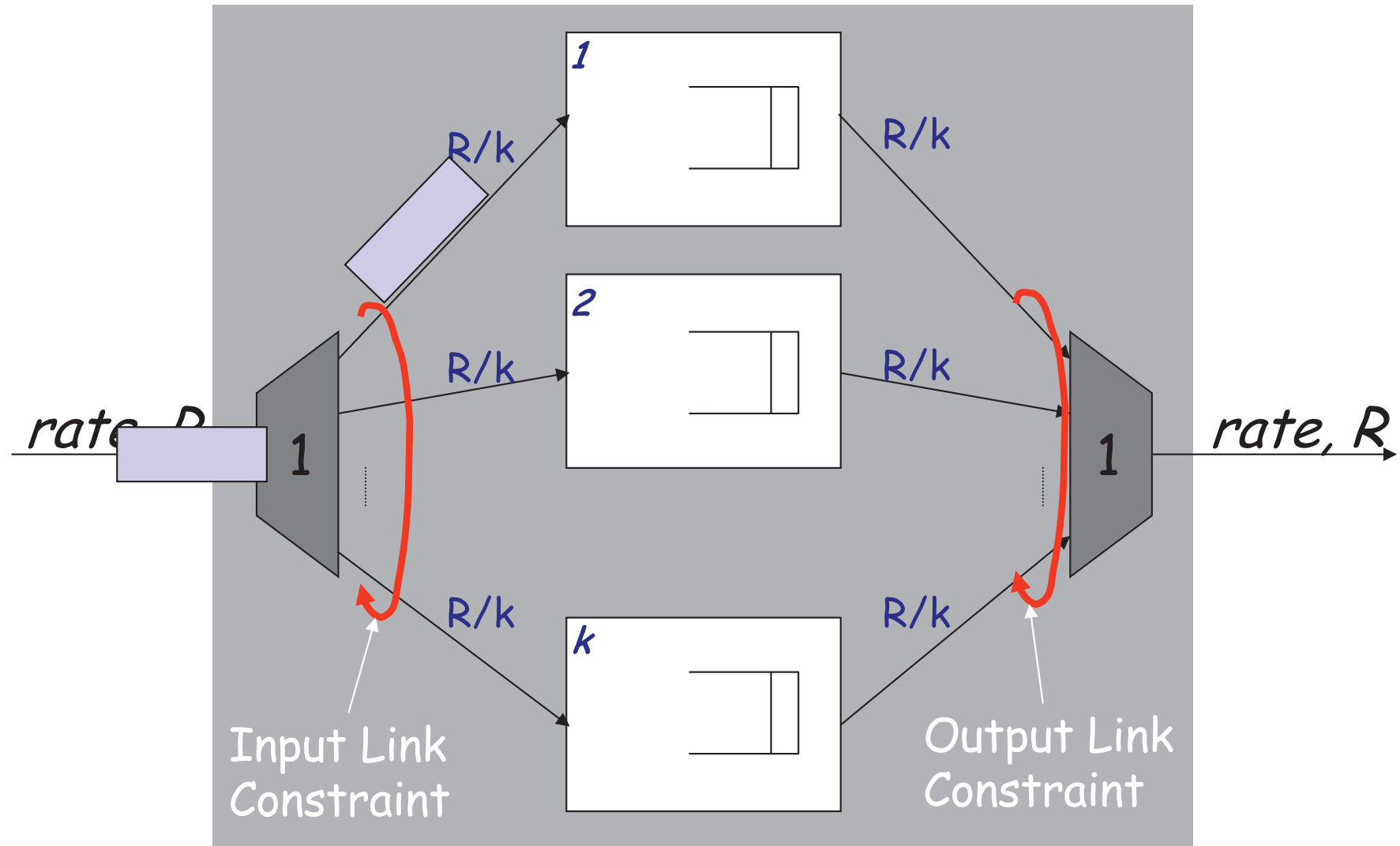
Parallel Packet Switch

Questions

1. Can it be work-conserving?
2. Can it emulate a single big shared memory switch?
3. Can it support delay guarantees, strict-priorities, WFQ, ...?

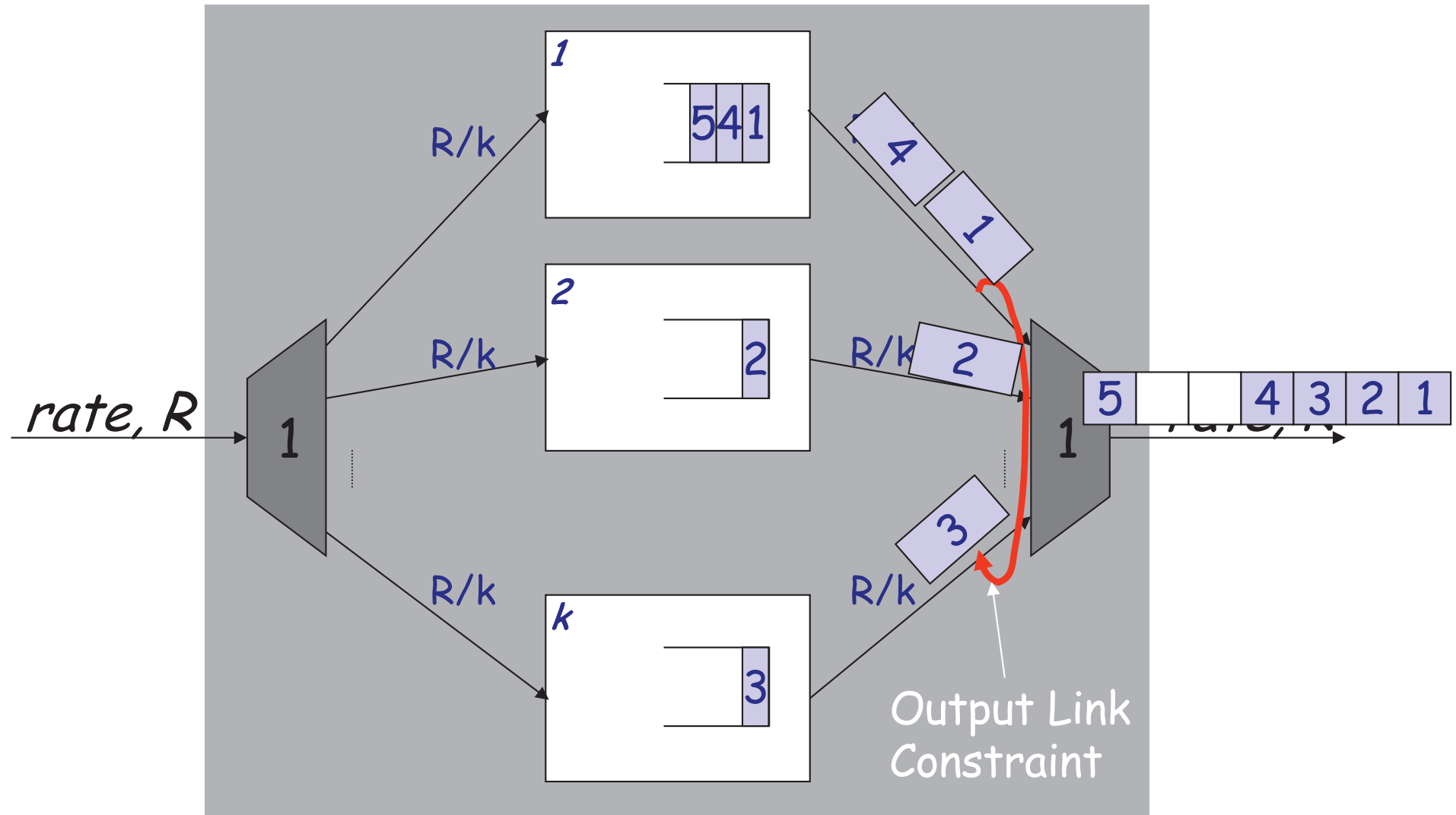
Parallel Packet Switch

Work Conservation



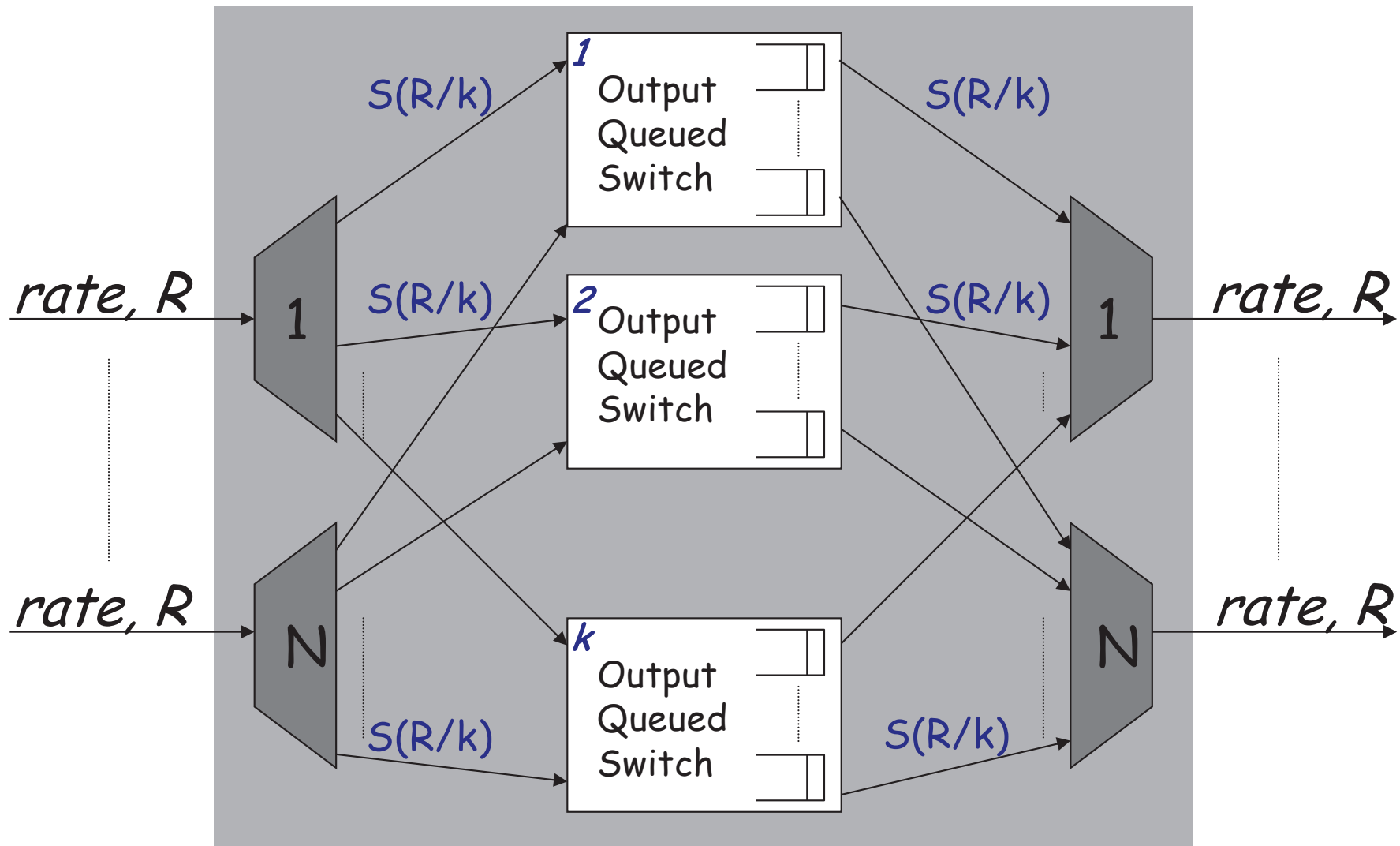
Parallel Packet Switch

Work Conservation



Parallel Packet Switch

Work Conservation



Parallel Packet Switch

Theorems

1. If $S > 2k/(k+2) \cong 2$ then a parallel packet switch can be work-conserving for all traffic.
2. If $S > 2k/(k+2) \cong 2$ then a parallel packet switch can *precisely* emulate a FCFS output-queued switch for all traffic.

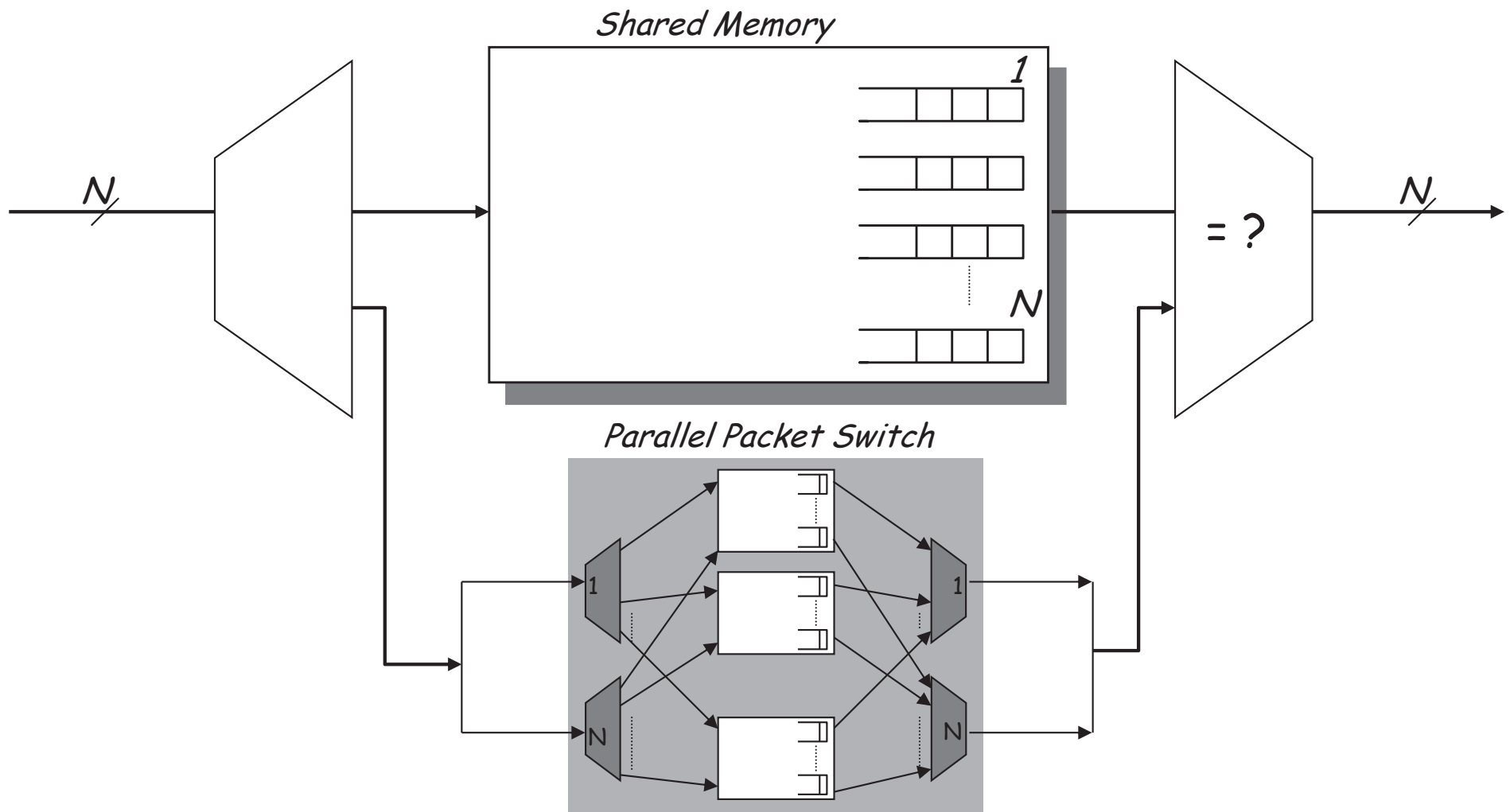
Parallel Packet Switch

Theorems

3. If $S > 3k/(k+3) \cong 3$ then a parallel packet switch can be precisely emulate a switch with WFQ, strict priorities, and other types of QoS, for all traffic.

With Sundar Iyer and Amr Awadallah

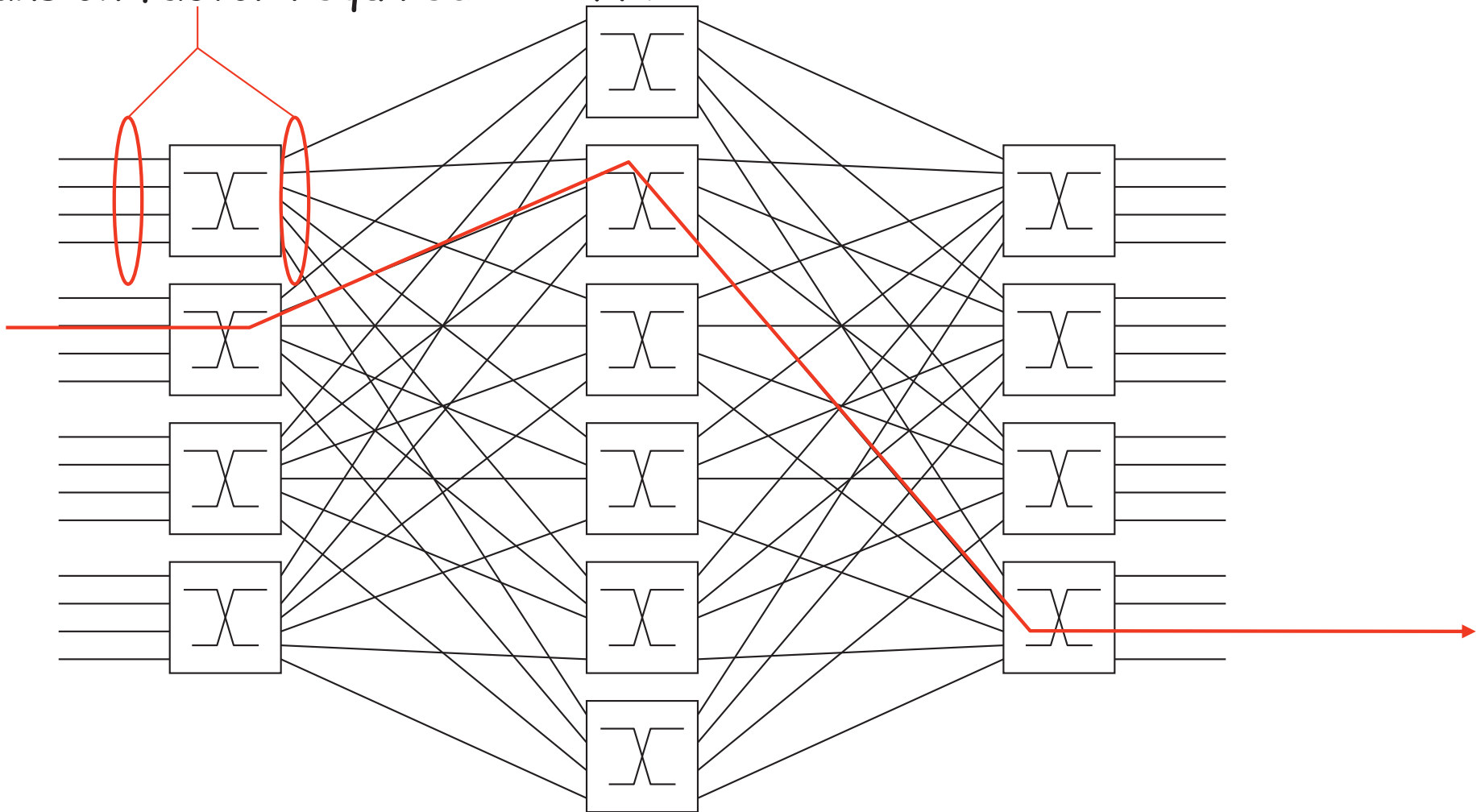
Precise Emulation of an FCFS Shared Memory Switch



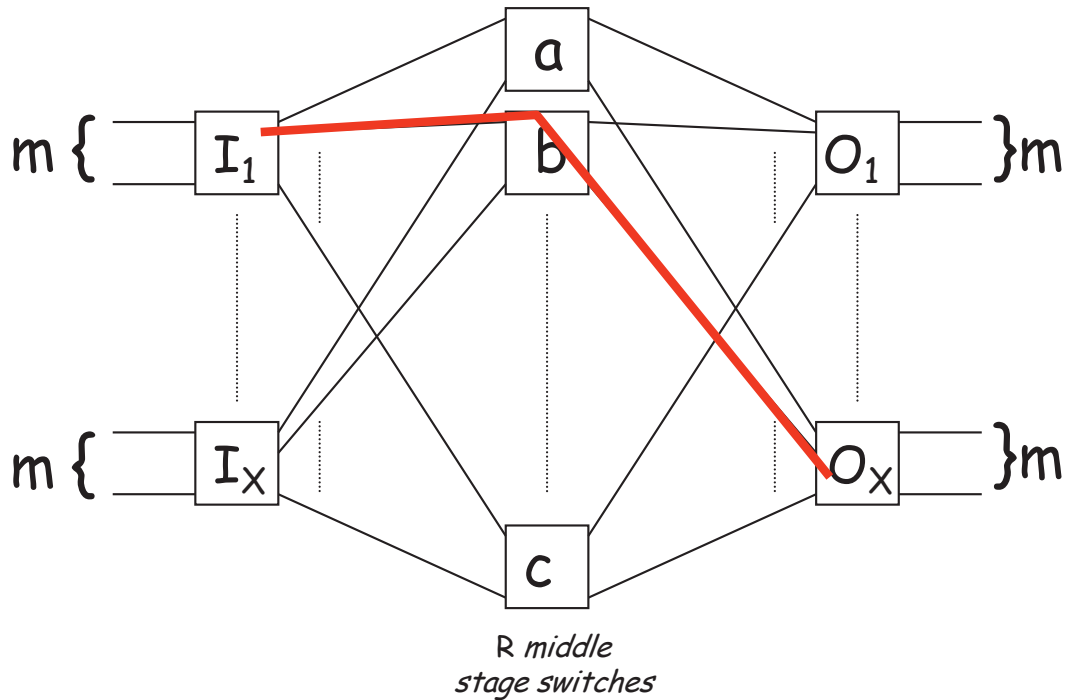
An aside

Unbuffered Clos Circuit Switch

Expansion factor required = $2 - 1/N$



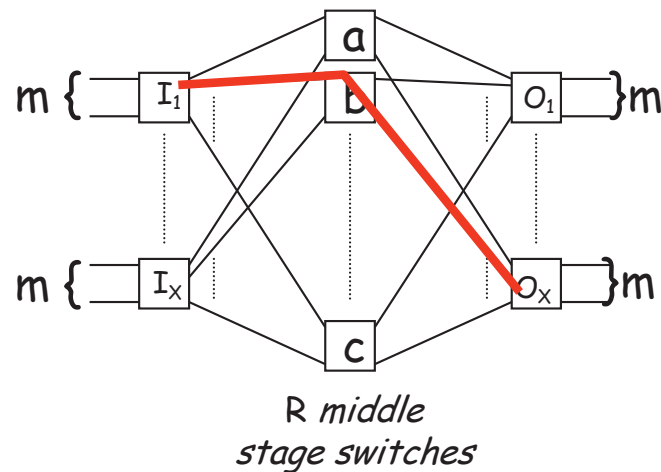
Clos Network



	O_1	O_2	O_3	$\dots$	O_x
I_1				$\dots$	b
I_2				$\dots$	
I_3				$\dots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_x				$\dots$	

$\leq \min(R, m)$ entries in each row
 $\leq \min(R, m)$ entries in each column

Clos Network



	O_1	O_2	O_3	O_x
I_1				b
I_2				
I_3				
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_x				

$\leq \min(R, m)$ entries in each row
 $\leq \min(R, m)$ entries in each column

Define: $UIL(I_i)$ = used links at switch I_i to connect to middle stages.
 $UOL(O_i)$ = used links at switch O_i to connect to middle stages.

If we wish to connect I_i to O_i :

When adding connection: $|UIL(I_i)| \leq m-1$ and $|UOL(O_i)| \leq m-1$

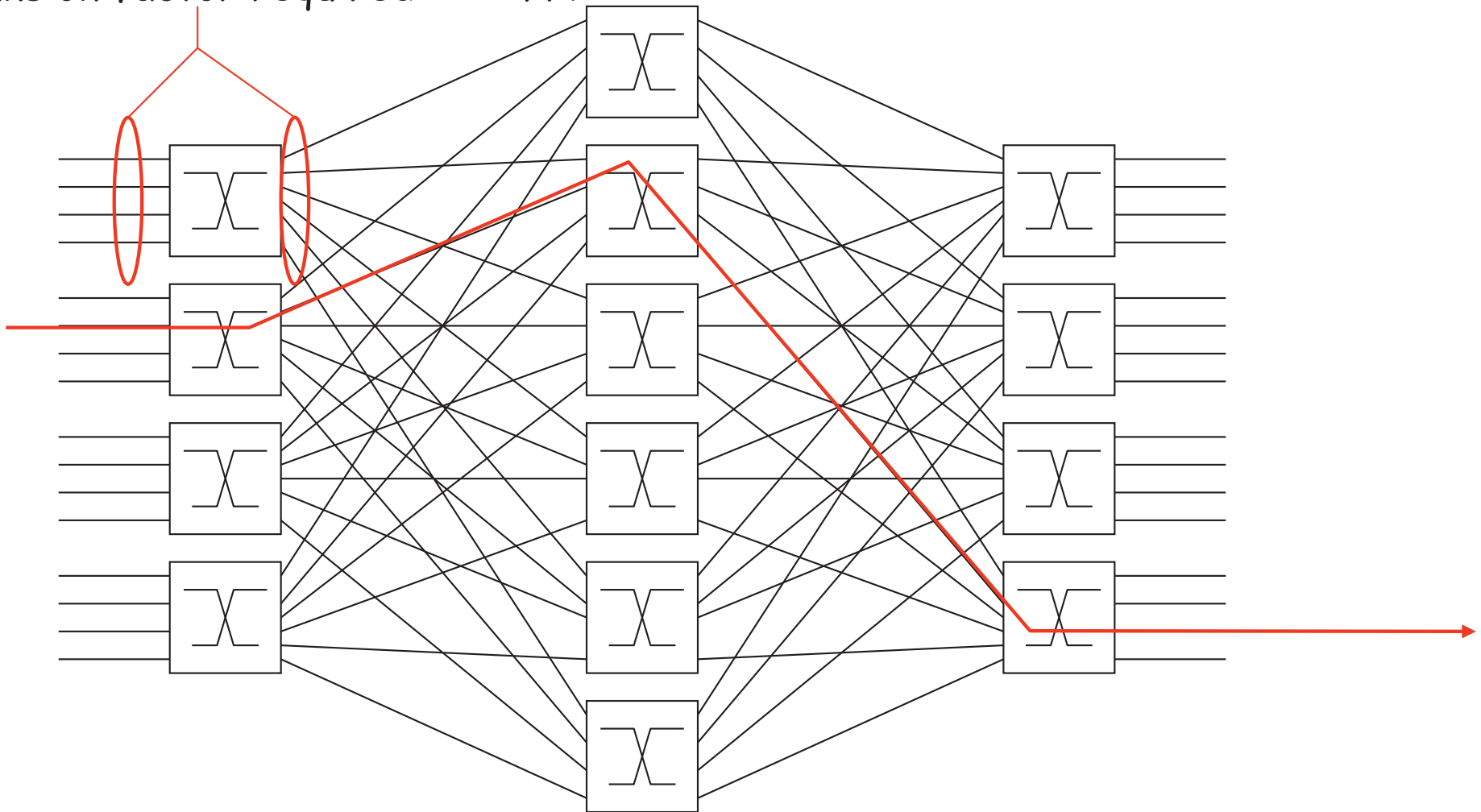
Worst-case: $|UIL(I_i) \cup UOL(O_i)| = 2m - 2$

Therefore, if $R \geq 2m-2$ there are always enough middle stages.

An aside

Unbuffered Clos Circuit Switch

Expansion factor required = $2 - 1/N$



Outline

- Introduction
What is a packet-switch?
- The Memory Bandwidth Problem
- Input-Queued Switches
Reducing memory bandwidth requirements
- Combined Input-Output Queued Switches
Making input-queued switches useful
- Parallel Packet Switches
Further reducing memory b/width requirements