

Predictable Communication Performance in On-Chip Networks

Axel Jantsch

June 22, 2011

Overview

Introduction

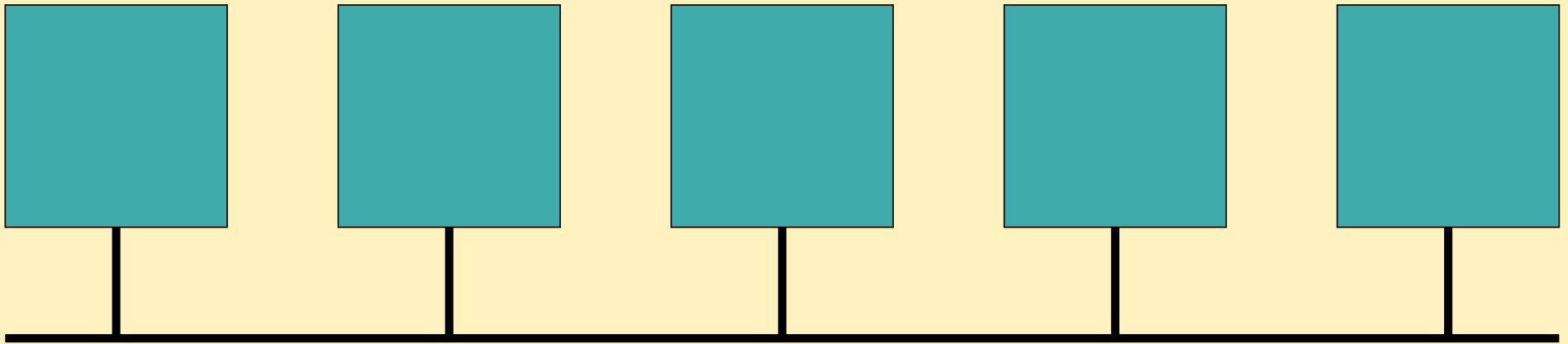
Circuit Switching

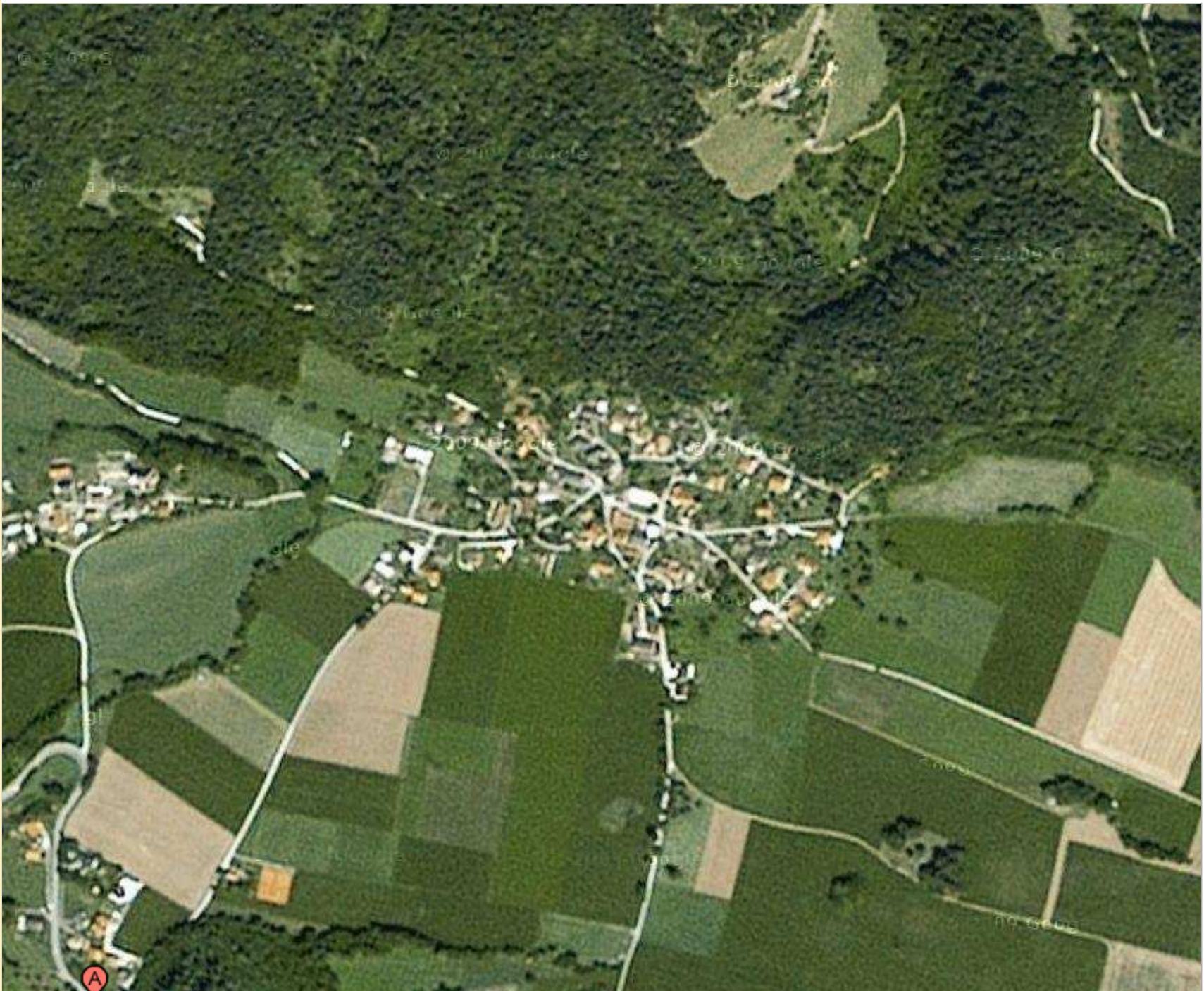
Time Division Multiplexing

Aggregate Resource Allocation

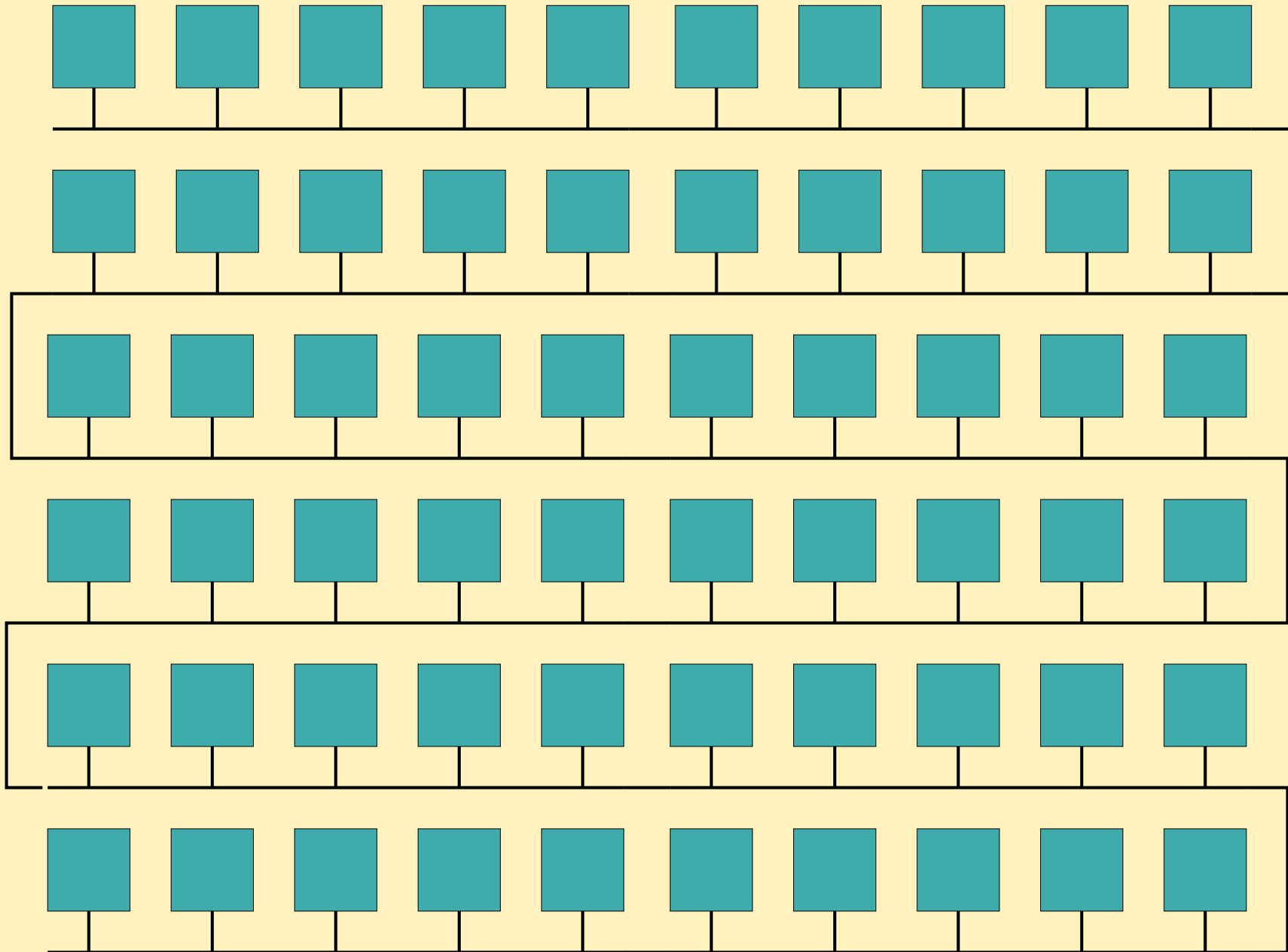
Summary

Buses are Efficient for Small Systems

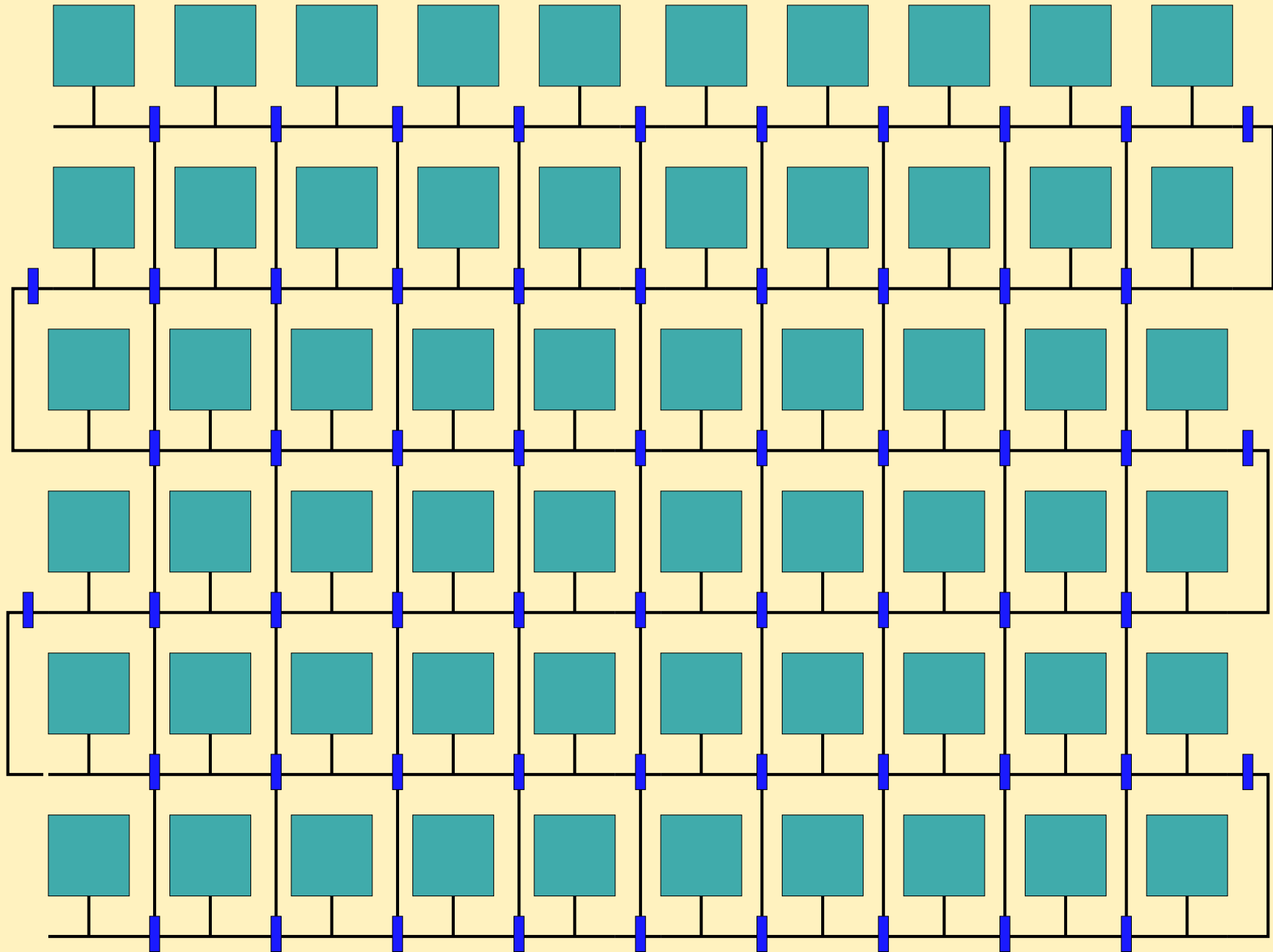




Buses Do Not Scale



Buses + Pipelining + Path Diversity



Phases of Communication

■ Allocation Phase

- ◆ Allocate the resources
- ◆ May not offer guarantees
- ◆ Design time - Start up time - Run time

■ Transmission Phase

- ◆ Guarantee for delivery
- ◆ Guarantee for minimum bandwidth
- ◆ Guarantee for maximum latency

■ Deallocation Phase

Minimal Guarantees require Minimal Provision

Circuit Switching

Introduction

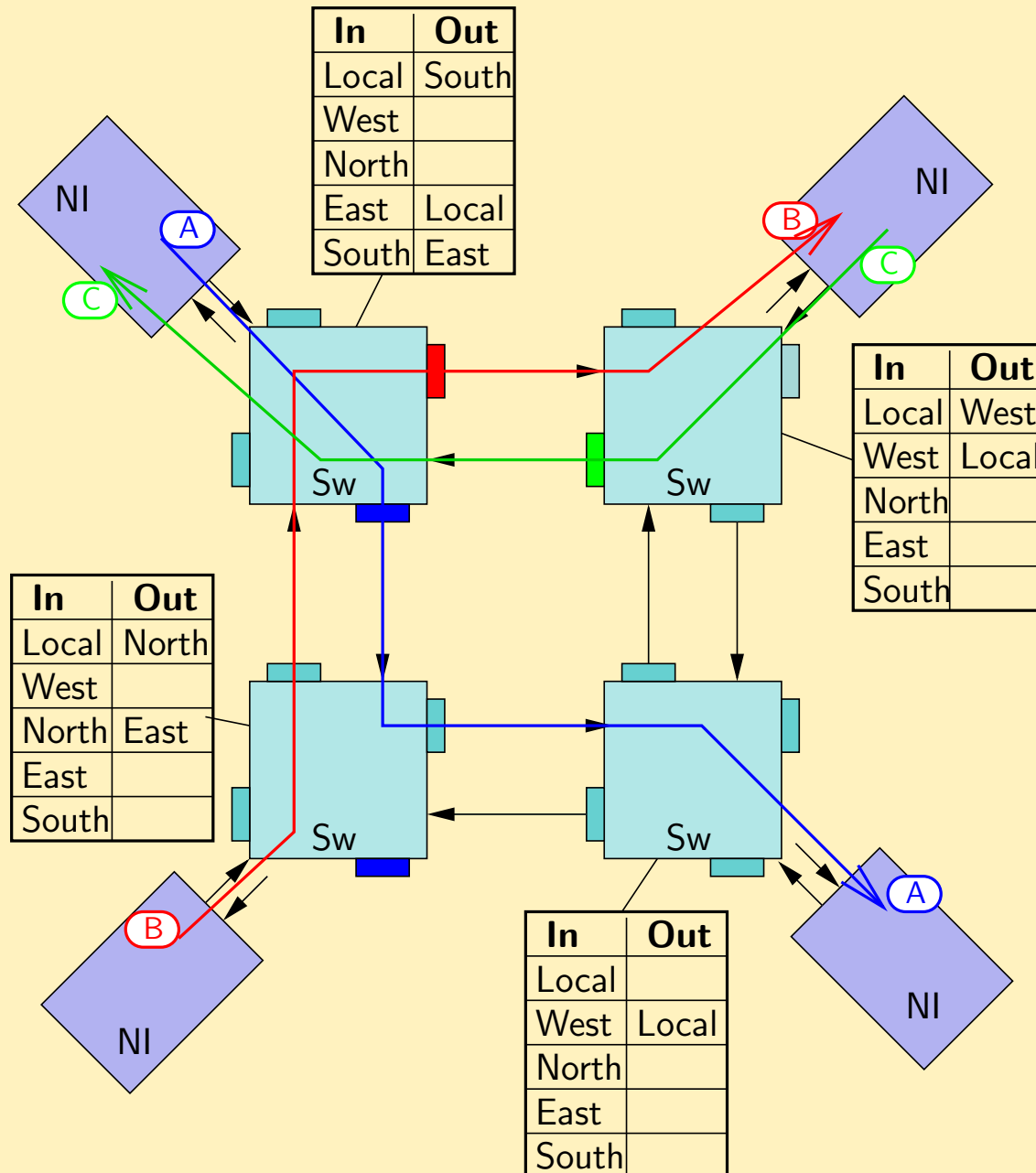
Circuit Switching

Time Division Multiplexing

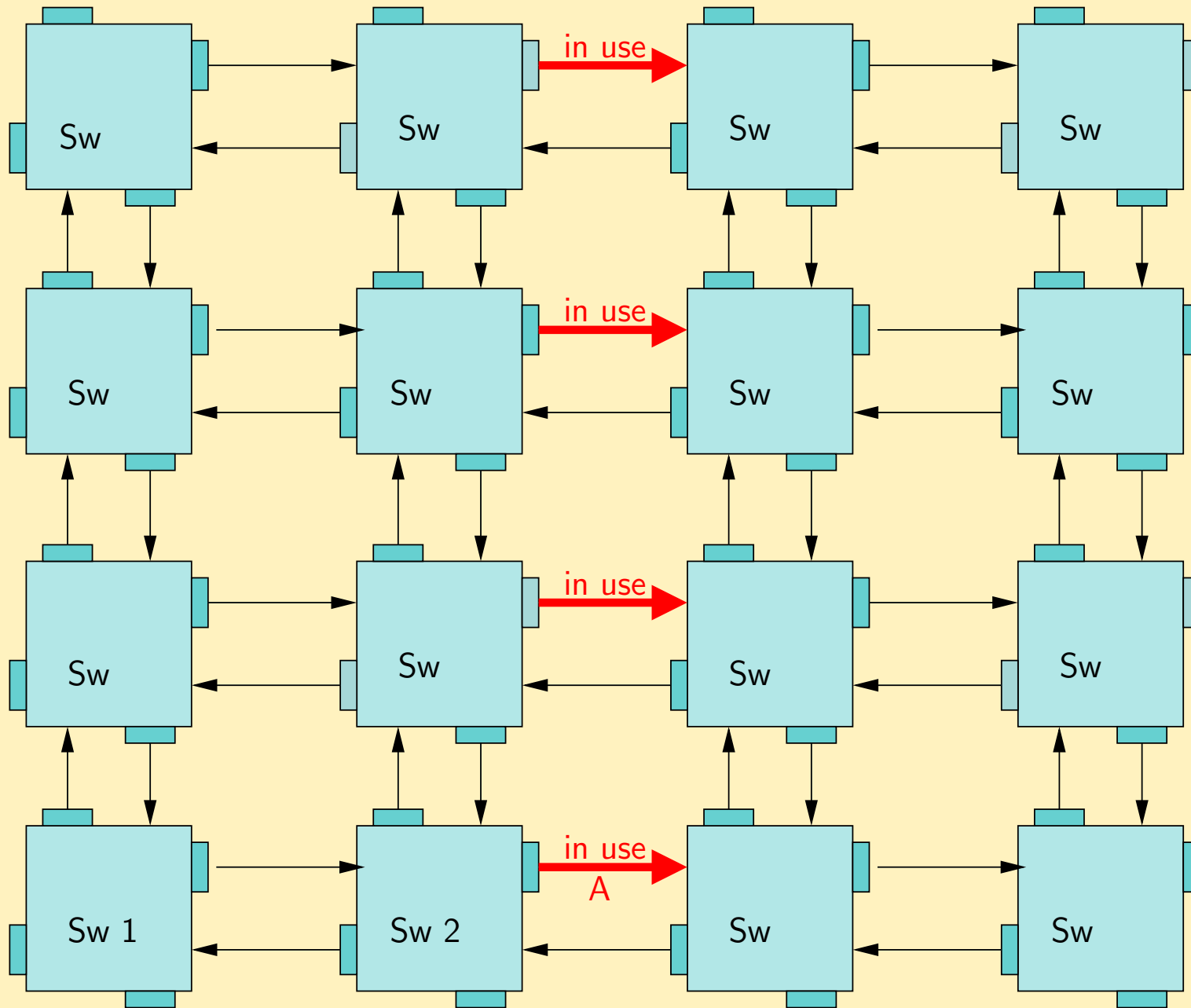
Aggregate Resource Allocation

Summary

Exclusive Resource Allocation



Circuit Switching Inflexibilities



Circuit Switching Pros & Cons

■ Disadvantages:

- ◆ Exclusive allocation of resources
- ◆ Long setup phase

■ Advantages:

- ◆ High performance - throughput and latency
- ◆ Low power consumption
- ◆ Low overhead during transmission phase
- ◆ Predictable transmission

Time Division Multiplexing

Introduction

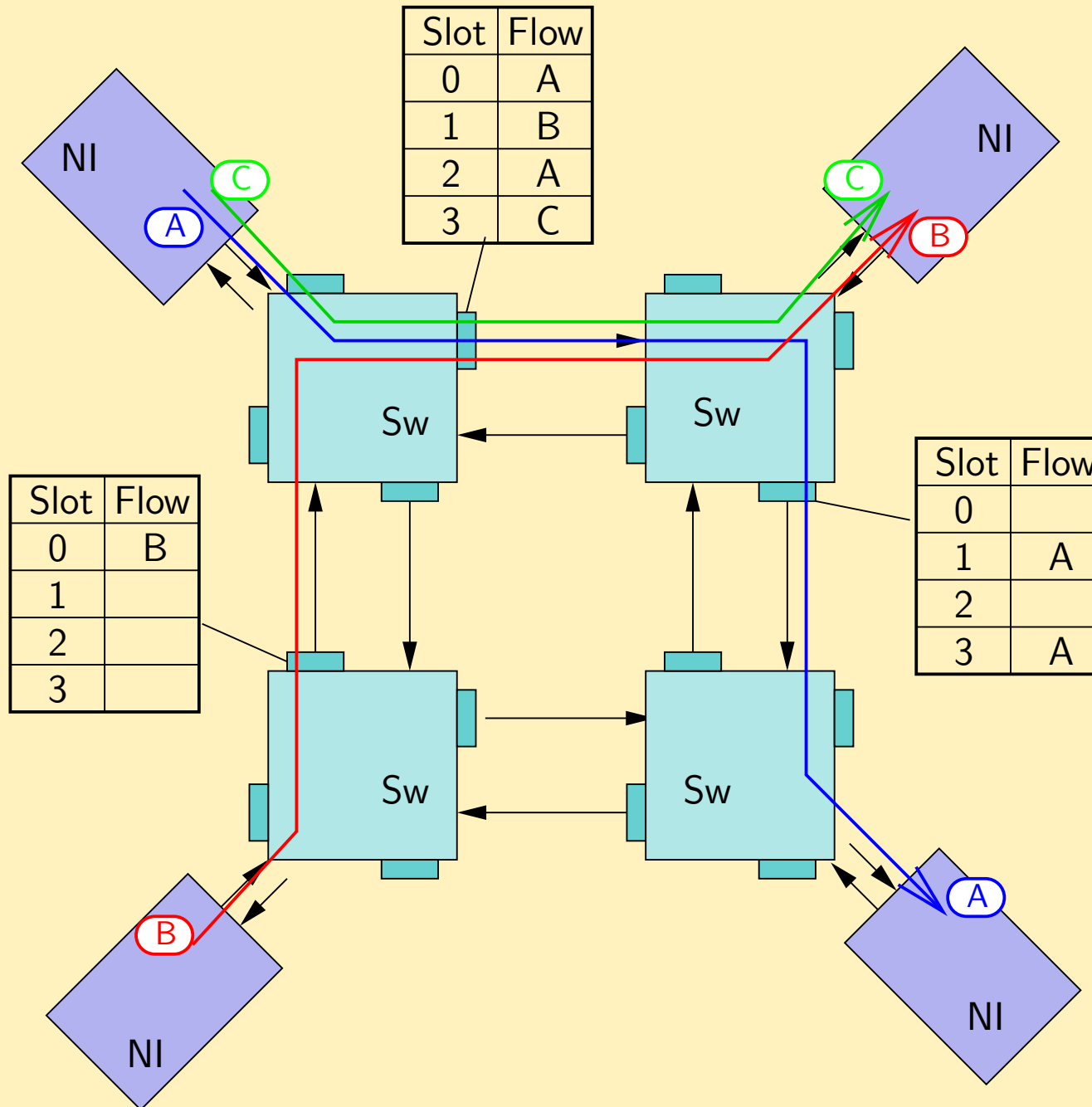
Circuit Switching

Time Division Multiplexing

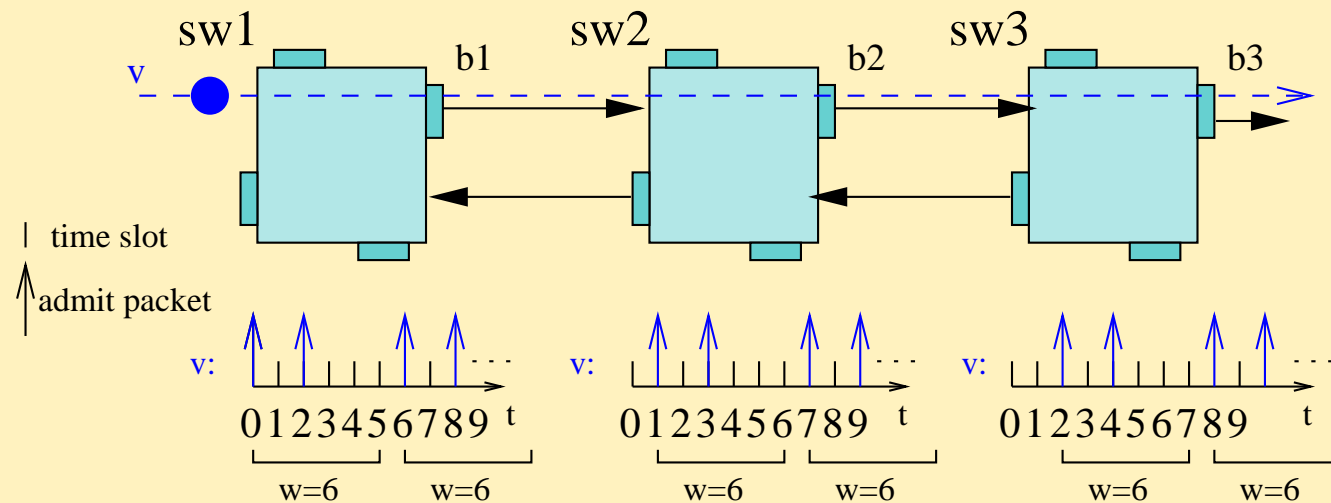
Aggregate Resource Allocation

Summary

Time Slot Based Resource Allocation



Time Slot Based Resource Allocation



- Network is synchronized by a global time
- Connections are defined by routing tables in switches
- Repetitive traffic patterns/window
- Latency and bandwidth guarantees are linked
- Setup
 - ◆ Constrained optimization problem
 - ◆ Path selection
 - ◆ Slot allocation

Aggregate Resource Allocation

Introduction

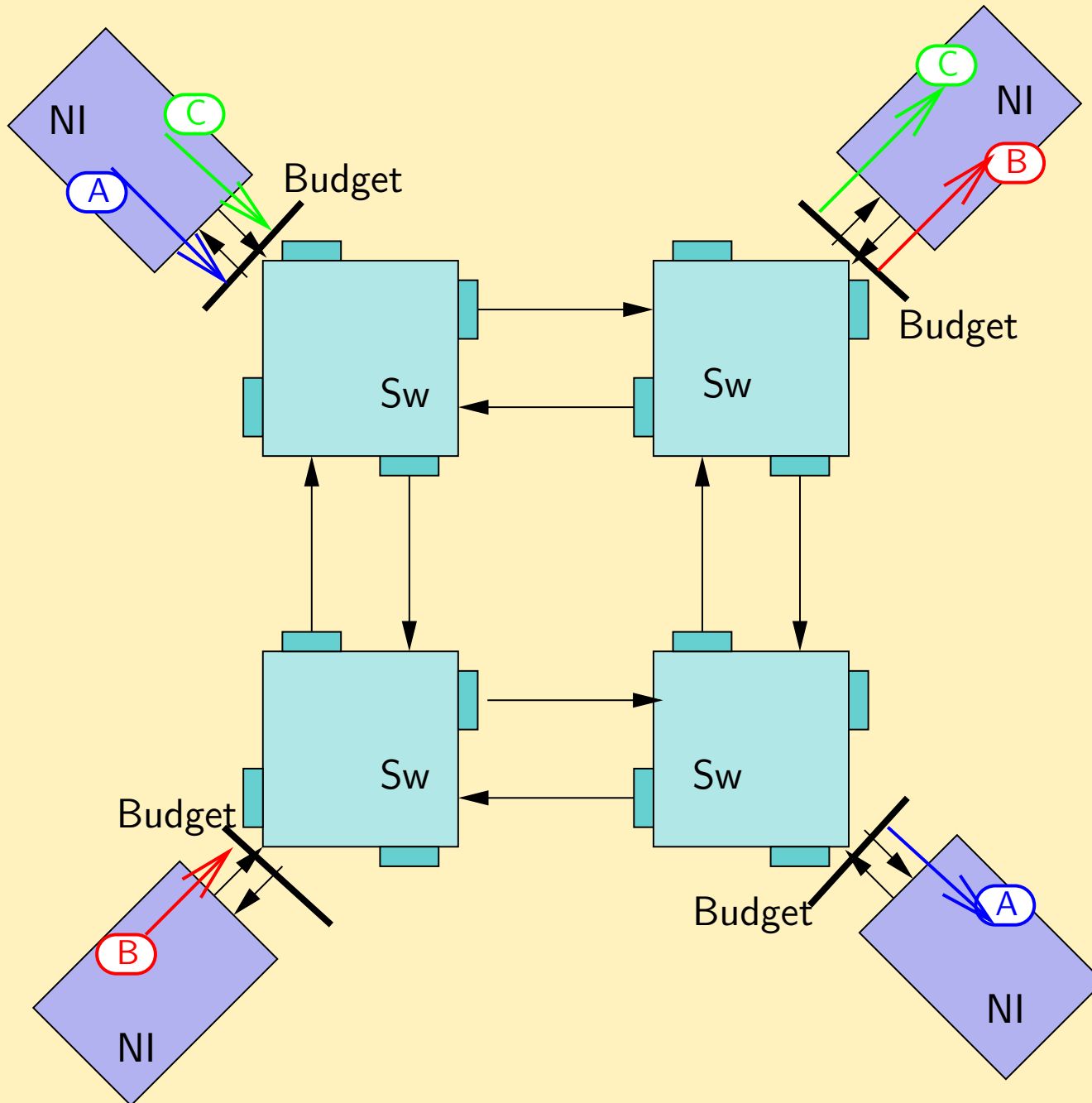
Circuit Switching

Time Division Multiplexing

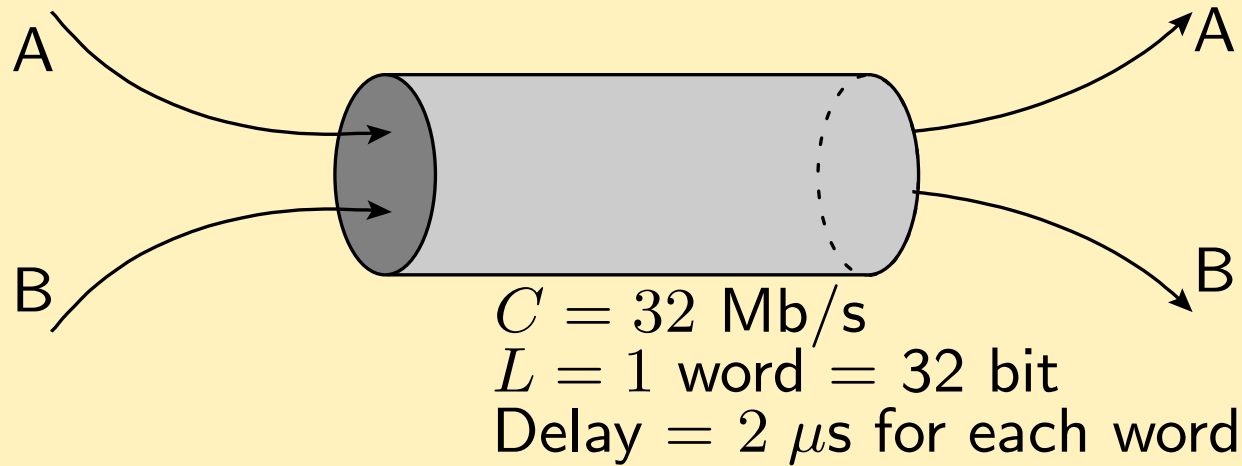
Aggregate Resource Allocation

Summary

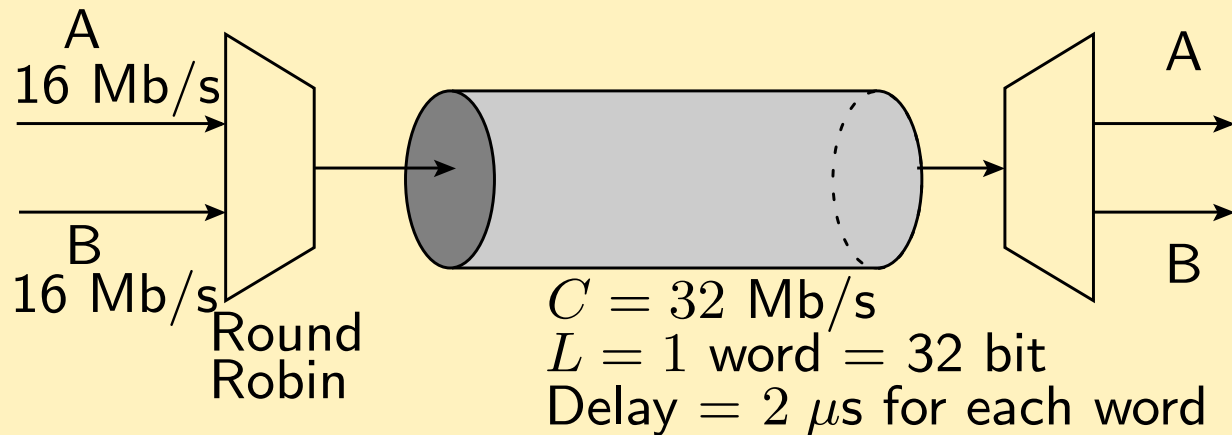
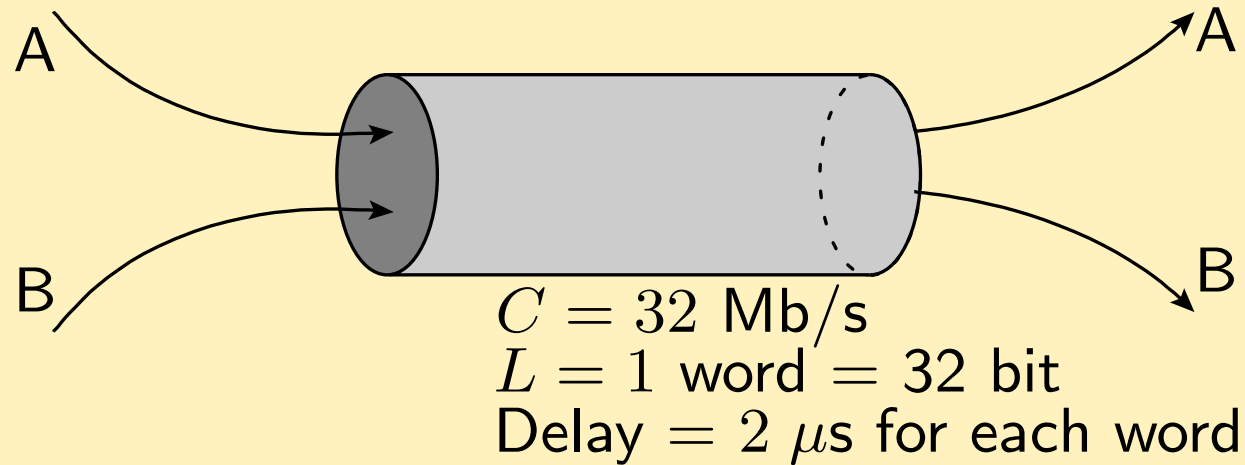
Aggregate Resource Allocation



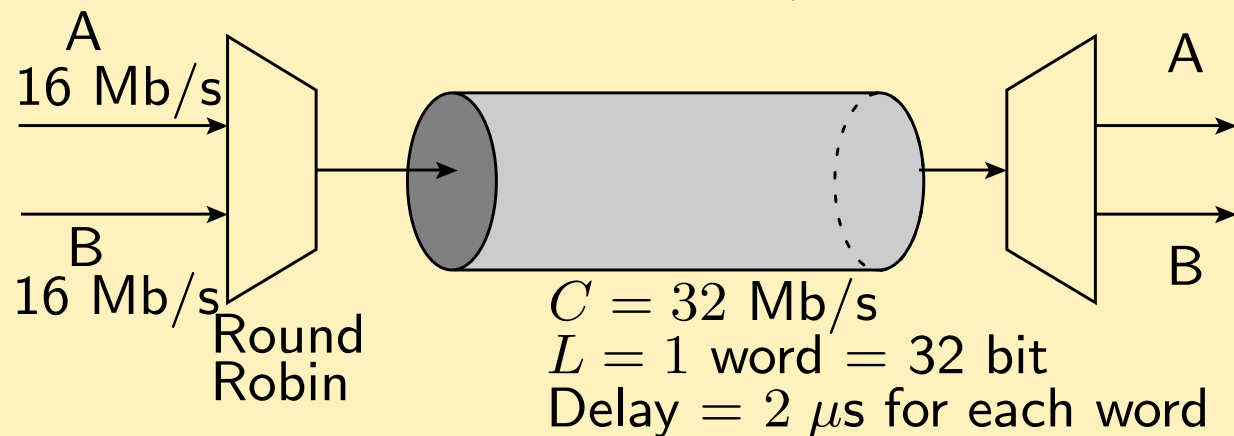
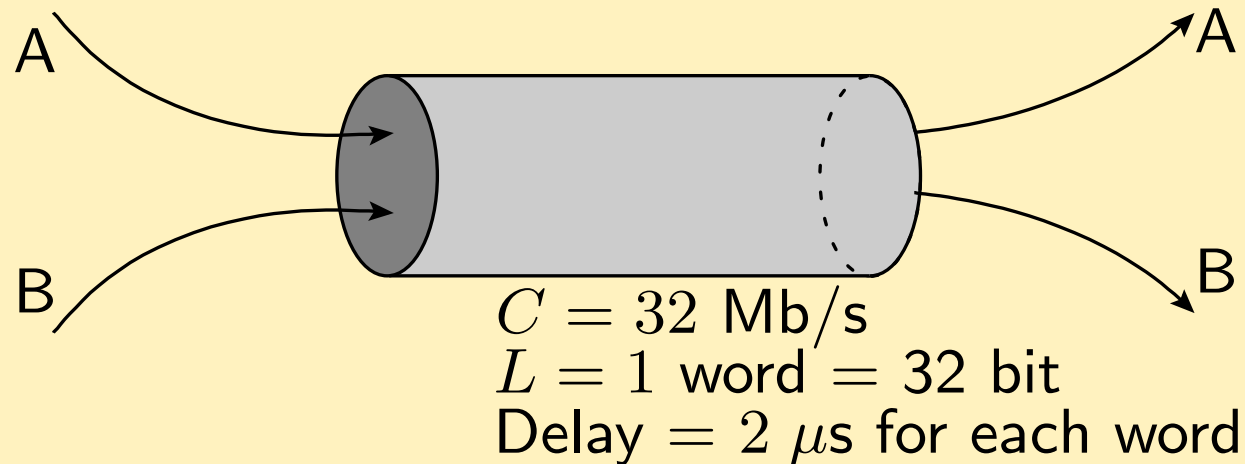
Aggregate Allocation of a Channel



Aggregate Allocation of a Channel

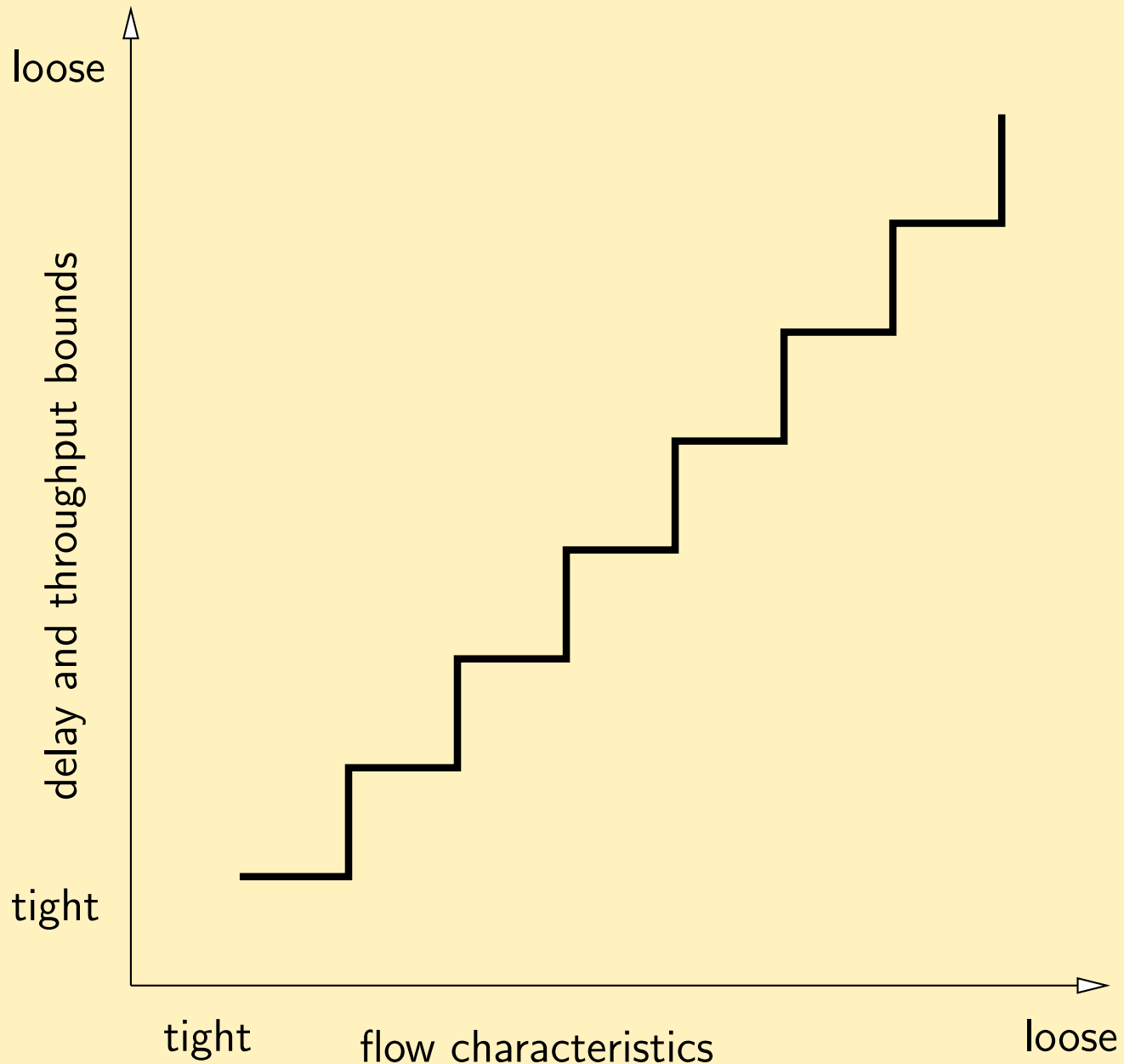


Aggregate Allocation of a Channel

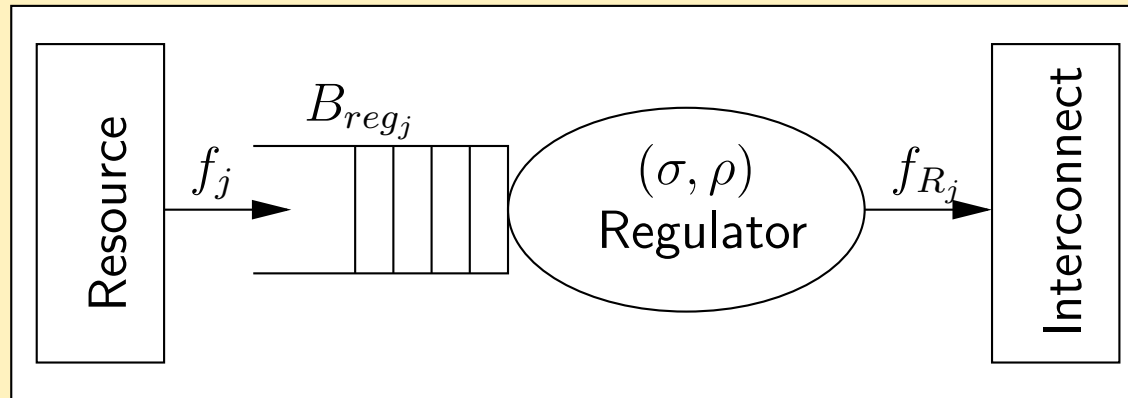


- Characteristics of channel
- Characteristics of flows
- Arbitration policy for channel access

Bounds for Aggregate Resource Allocation

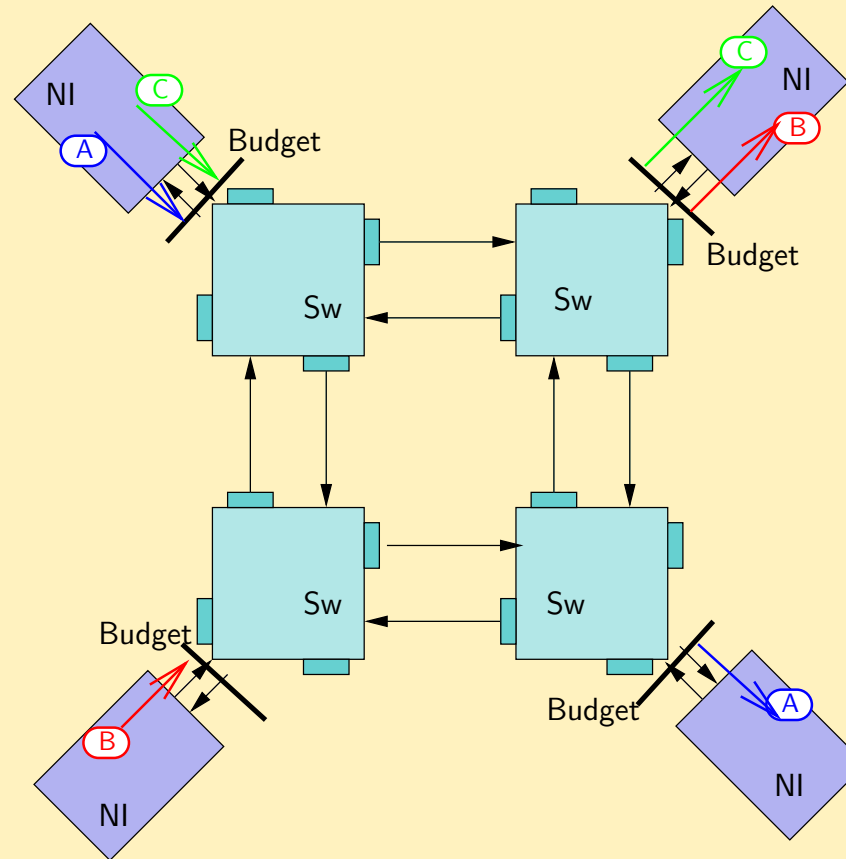


Service Control by Traffic Regulation



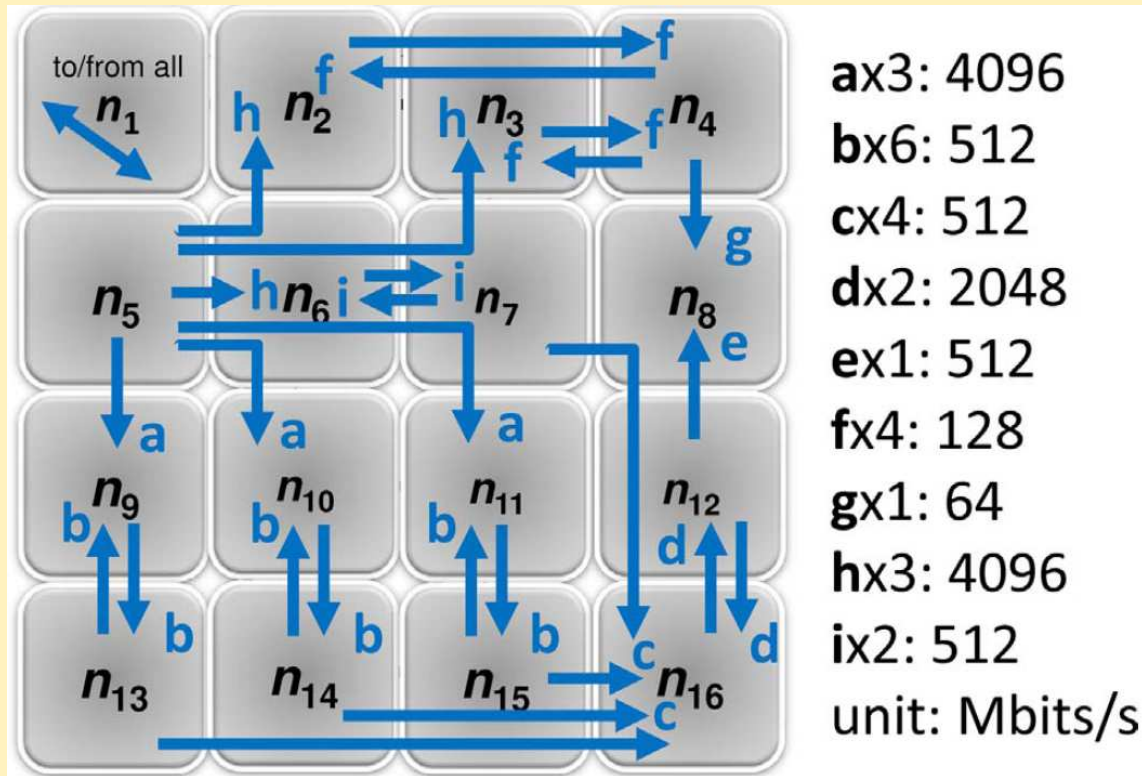
- Based on Network Calculus Theory
- Traffic regulator in each NI
- Regulation parameters:
 - ◆ Burstiness
 - ◆ Peak rate
- Optimization problem:
 - ◆ Given a set of flows with bandwidth and delay requirements
 - ◆ Minimize overall buffer size
 - ◆ Guarantee max delay and min throughput

Aggregate Resource Allocation by Contract



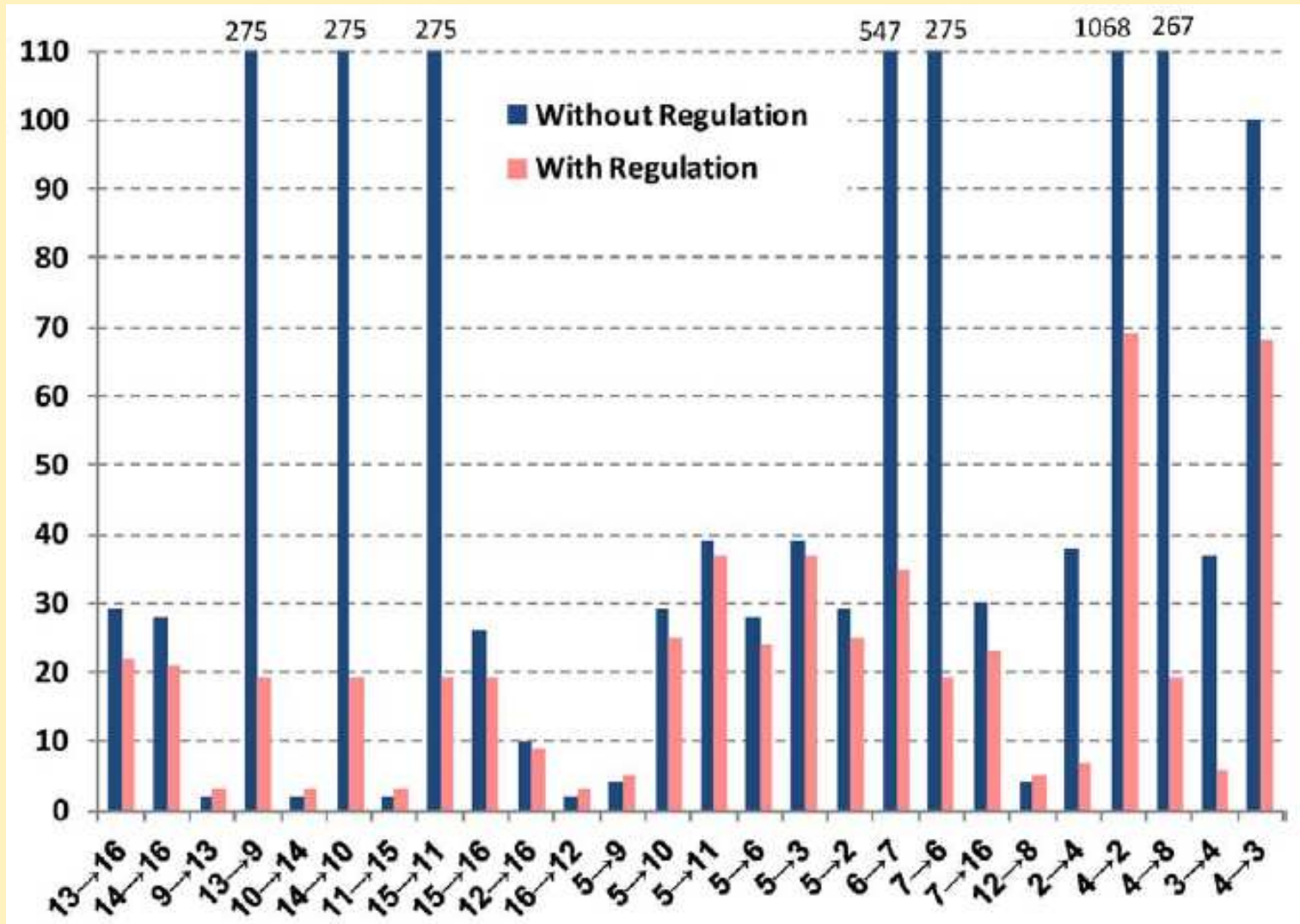
- Granularity of resource allocation \implies Granularity of contracts
- Compositionality is based on compliance with contracts
- Predictability is based on known correlation between allocated resources and performance

Experiment with a Baseband Processing Subsystem



- Minimal throughput requirements
- Maximal latency requirements

Experiment with a Baseband Processing Subsystem: Max delay (cycles)



Experiments: Buffer Size in flit

		in Network	in Regulator	Total
Baseband	Without regulation	404	0	404
	Minimize size	118	28	146
	Minimize variance			192
	Both objectives			150
Hot spot	Without regulation	361	0	361
	Both objectives	144	53	197
Bit complement	Without regulation	254	0	254
	Both objectives	112	16	128

Summary

Introduction

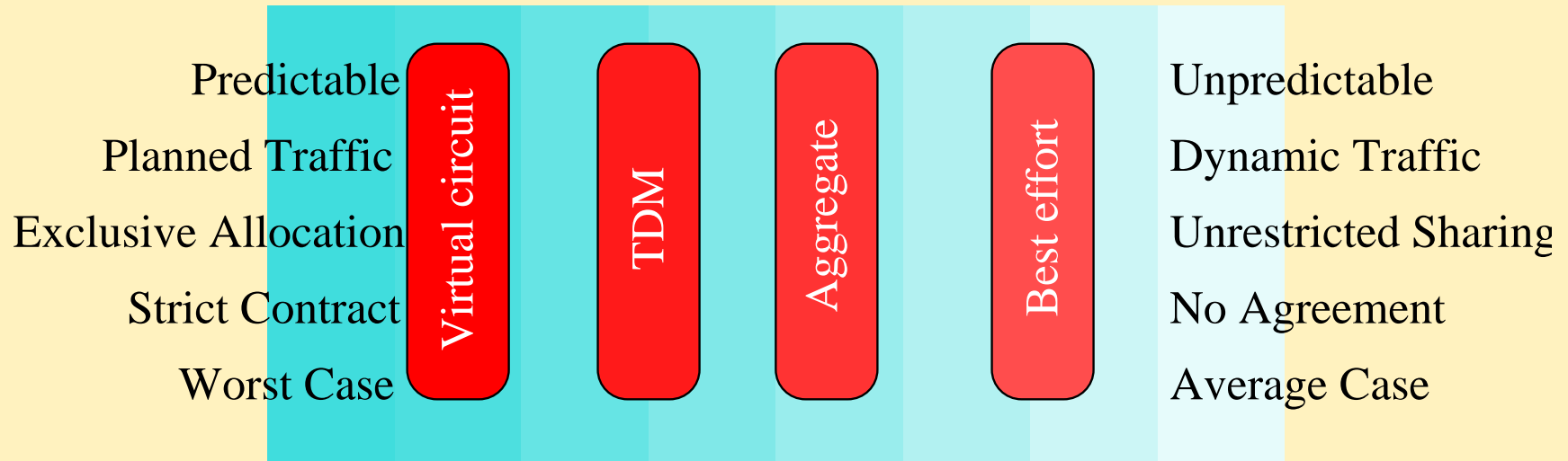
Circuit Switching

Time Division Multiplexing

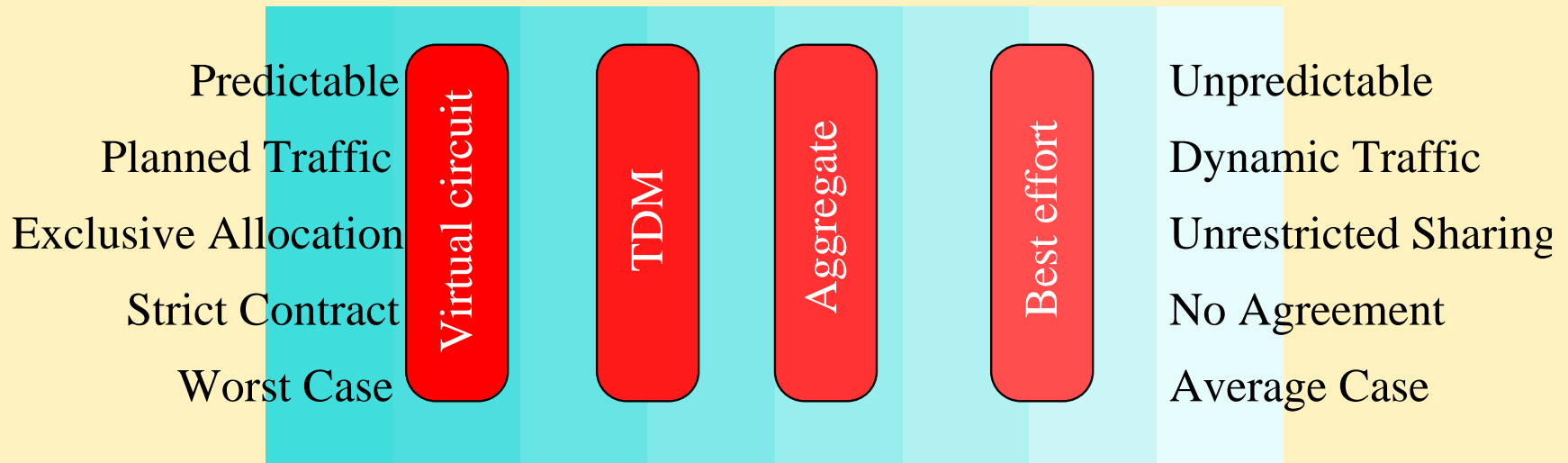
Aggregate Resource Allocation

Summary

Trade-offs in the Design Space



Trade-offs in the Design Space



Design space is defined by

- Granularity of resource allocation
- Granularity of traffic planning

Trade-off parameters:

- Resource allocation policy
- Traffic shaping policy

Next Steps

- Dynamic resource allocation
- Integrate average case and worst case analysis
- Integrate predictable performance with fault tolerance