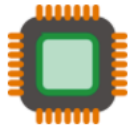
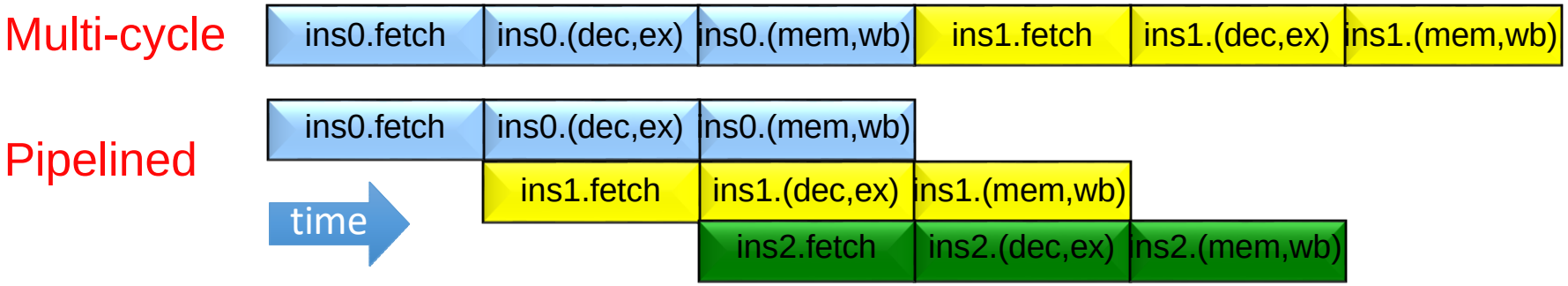


CS305: Computer Architecture

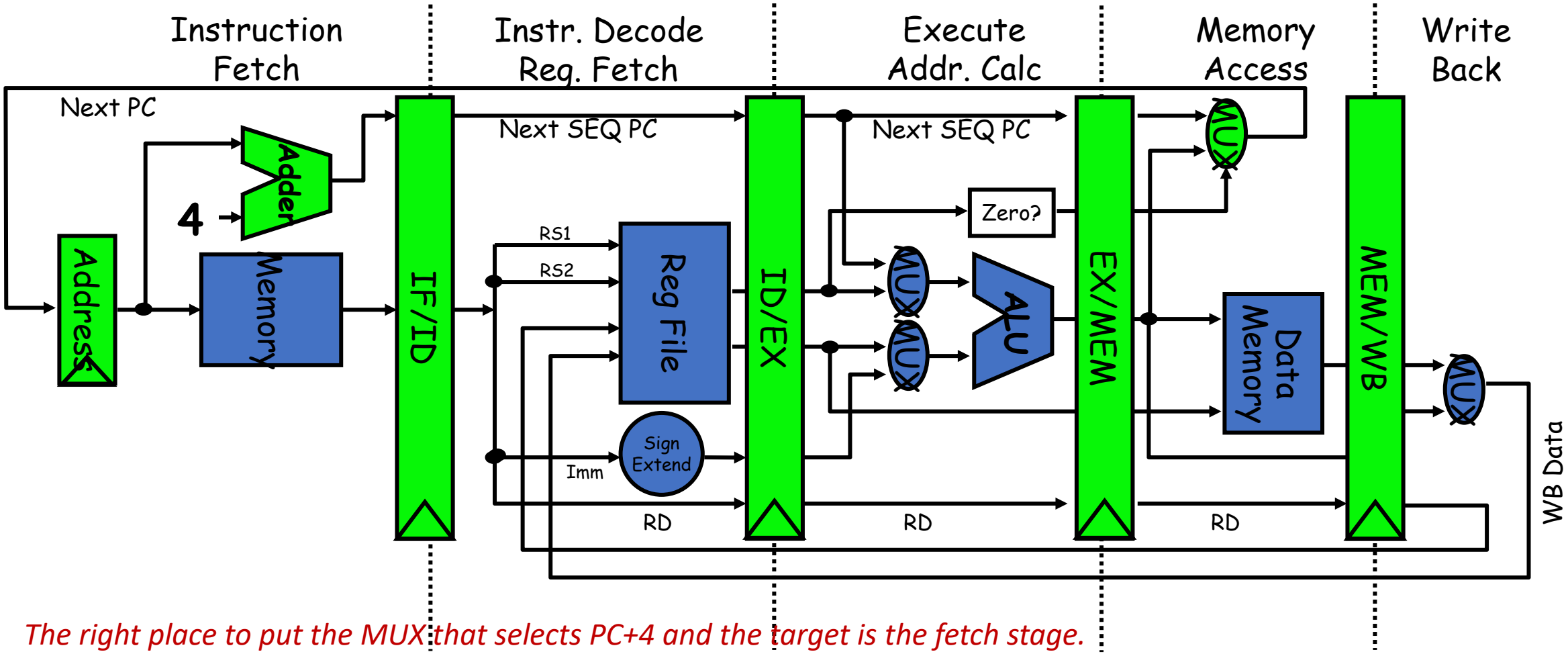
Instruction Pipelining-II

https://www.cse.iitb.ac.in/~biswa/courses/CS305/main.html

Multi-cycle vs Pipelined



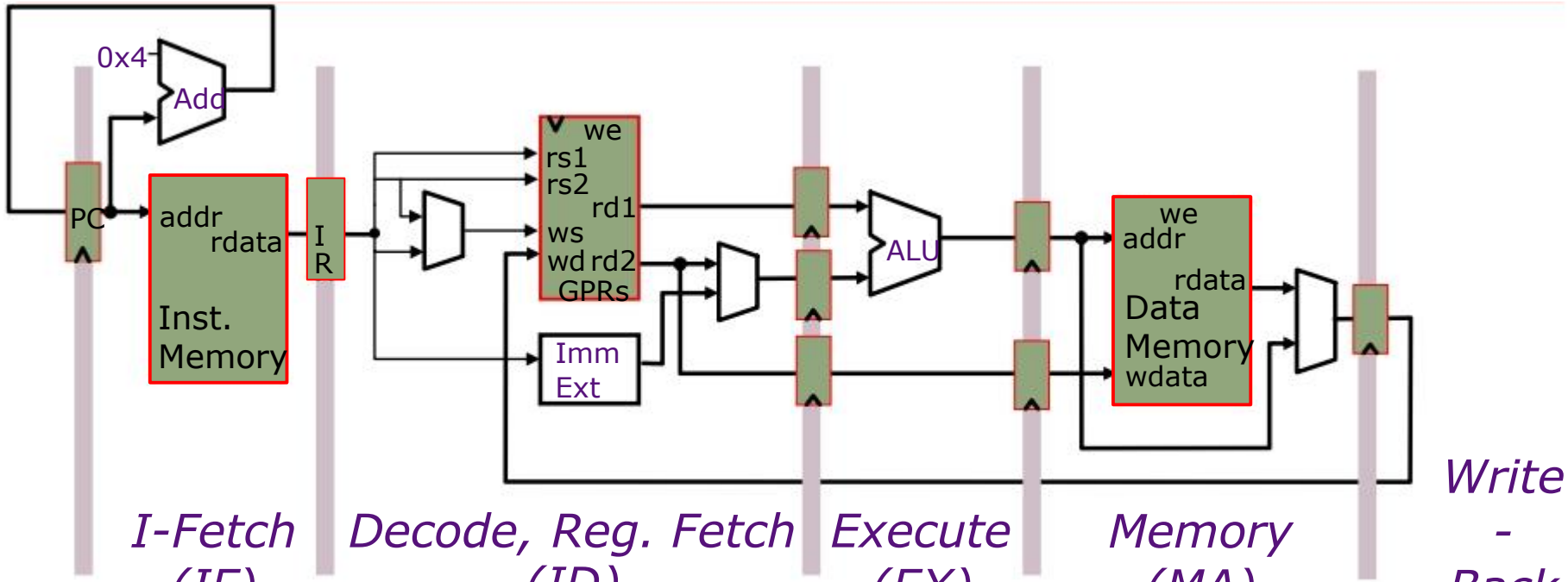
Vanilla 5-stage pipeline



The right place to put the MUX that selects PC+4 and the target is the fetch stage.

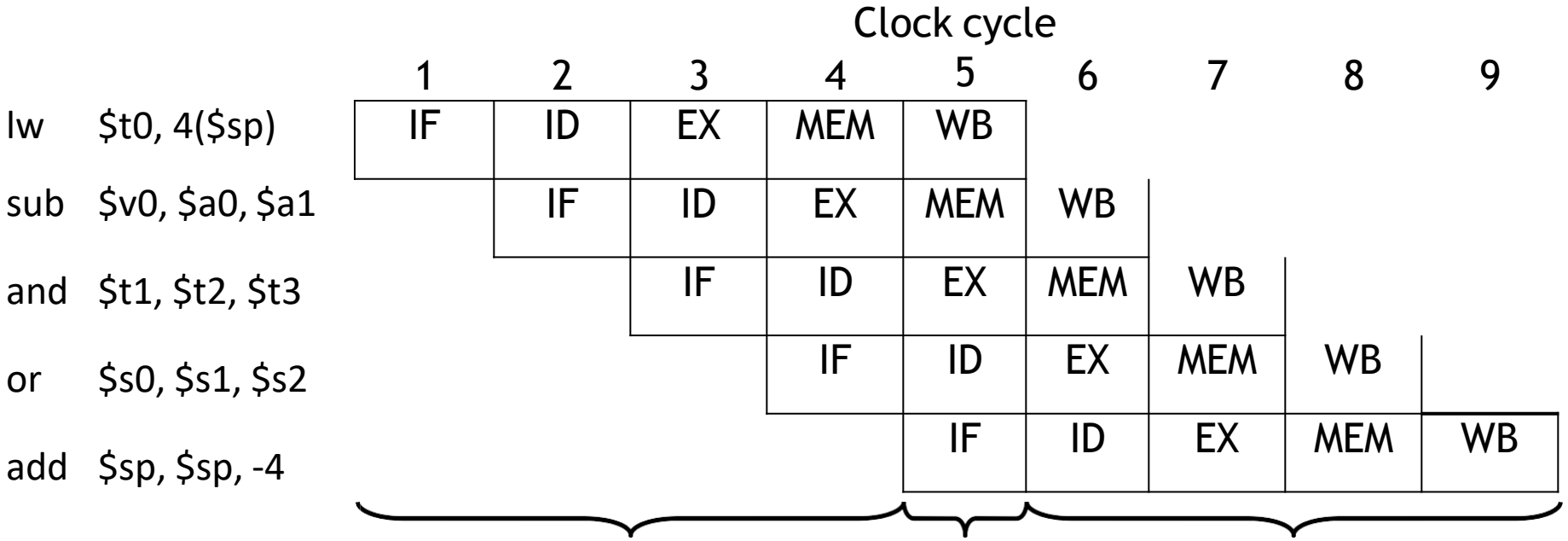
The slide shows a vanilla 5-stage pipeline if we just take a single cycle datapath and divide it into five stages.

Resource Utilization

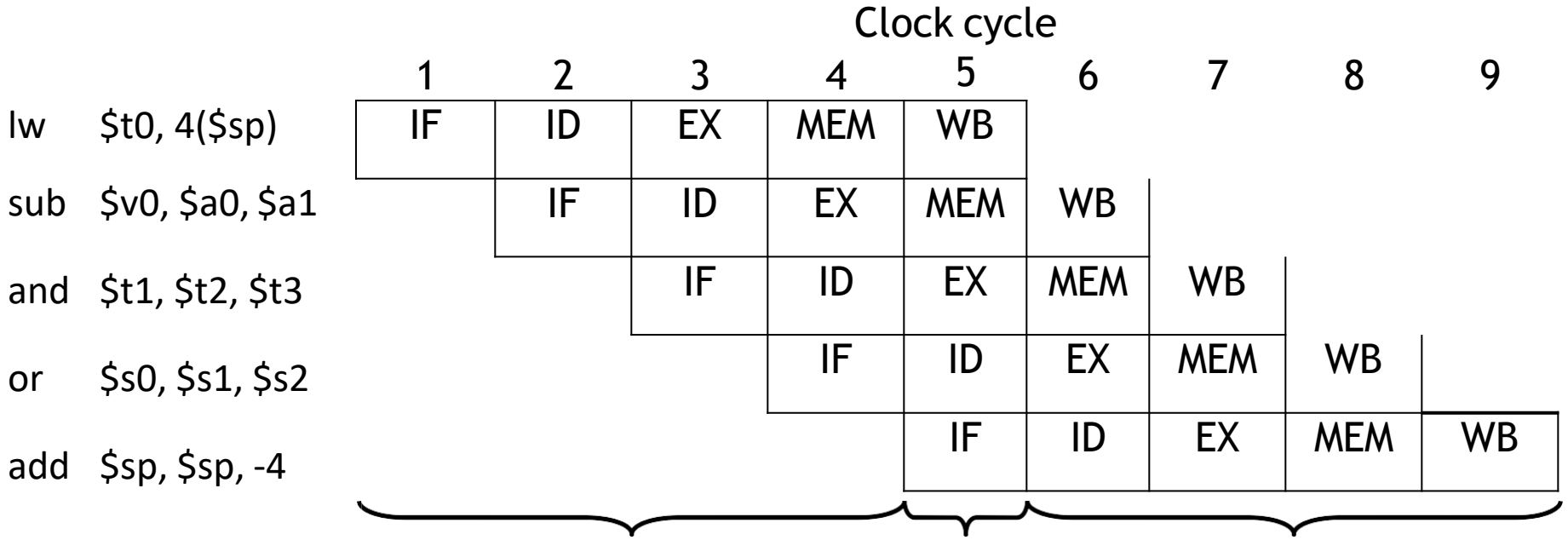


Resources	time	t0	t1	t2	t3	t4	t5	t6	t7	. . . .
	IF	I_1	I_2	I_3	I_4	I_5				
	ID		I_1	I_2	I_3	I_4	I_5			
	EX			I_1	I_2	I_3	I_4	I_5		
	MA				I_1	I_2	I_3	I_4	I_5	
	WB					I_1	I_2	I_3	I_4	I_5

Visualizing Pipeline



Visualizing Pipeline: Execution time



For a k-stage pipeline executing N instructions

first instruction: K cycles

Next N-1 instructions: N-1 cycles, total = K + (N-1) cycles

Latency and Bandwidth revisited

- Latency
  - time it takes to complete one instance

- Throughput
  - number of computations done per unit time

Let’s see how much throughput can be achieved?

Pipelined versus Single cycle CPU design

Instruction	Ifetch	Decode	Execute	Memory	Writeback	Total time
LOAD	200ns	100	200	200	100	800ns
STORE	200	100	200	200		700ns
ADD	200	100	200		100	600ns
BRANCH	200	100	200			500ns

Total latency in single cycle CPU: 3200 ns

Total latency in pipelined CPU (200ns clock cycle):

1000ns (1st instruction) + 3 X 200 ns (for next three) = 1600 ns

What's the big deal

Speedup = 3200ns/1600ns = 2X

What if we have a billion instructions?

Single cycle = 1 billion X 800ns = 800 seconds

Pipelined = 1000ns + (1 billion -1) X 200ns ~ 200 seconds

Speedup = 4X ☺

Let’s include latch latency too

Inter-stage latch = 10ns

New clock cycle time in the pipelined design = 210ns

First instruction will get completed by 1040ns (five stages X 200 ns + four inter-stage latches X 10ns)

New Speedup = 800s/210s ~ 3.8X

How to Divide the Datapath?

Suppose memory is significantly slower than other stages. For example, suppose

t_{IM} = 10 units
t_{DM} = 10 units
t_{ALU} = 5 units
t_{RF} = 1 unit
t_{RW} = 1 unit

Since the slowest stage determines the clock, it may be possible to combine some stages without any loss of performance

#Stages and Speedup

	Assumptions	Unpipelined t_C	Pipelined t_C	Speedup
1.	$t_{IM} = t_{DM} = 10$, $t_{ALU} = 5$, $t_{RF} = t_{RW} = 1$, 4-stage pipeline	27	10	2.7
2.	$t_{IM} = t_{DM} = t_{ALU} = t_{RF} = t_{RW} = 5$, 4-stage pipeline	25	10	2.5
3.	$t_{IM} = t_{DM} = t_{ALU} = t_{RF} = t_{RW} = 5$, 5-stage pipeline	25	5	5.0

Tashakkor Mikonam