

DESIGN OF DIGITAL CIRCUITS AND SYSTEMS

Pipelining

Instructor: Justin Hsia

Teaching Assistants:

Colton Carroll

Hemil Patel

Rasya Fawwaz

Grace Zhou

Quinlyn Donohue

Rose Maresh

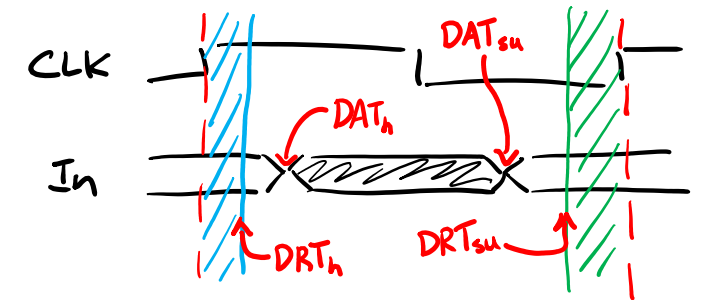
Relevant Course Information

- ❖ Quiz 3 starts at 11:50 AM
- ❖ Lab 4 due Friday (5/8), demos next week
- ❖ Homework 5 released today, due next Friday (5/15)
 - Static Timing Analysis and Pipelining
- ❖ Lab 5 released today, due in two weeks (5/22)
 - Hardest lab for many students
 - You will need to use the VGA interface on LabsLand
 - There's a creative component and opportunity for extra credit

Review: Timing Closure

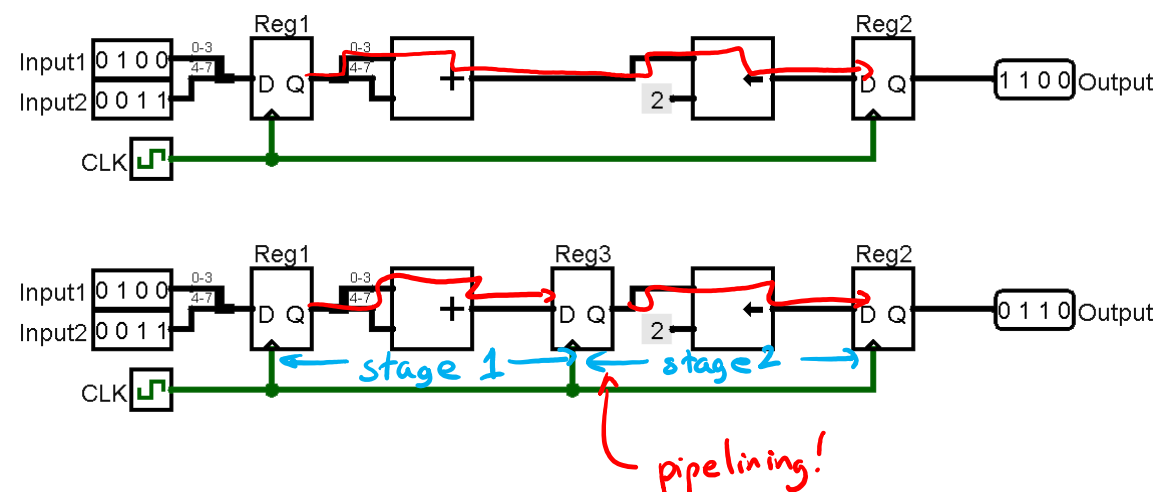
(setup and hold slack)

- ❖ Fixing hold violations: caused by fast data path and destination register's clock latency
 - Add delay in the data path with buffers or pairs of inverters (done automatically by Quartus)
- ❖ Fixing setup violations: data arrives too late compared to the destination register's clock speed
 - Slow down the clock (undesirable)
 - Tell fitter to try harder or confine logic to a smaller area
 - Rewrite code to simplify logic
 - ↳ ■ Add *pipelining* (today!)



Pipelining

- ❖ **Pipelining** is a set of data processing elements connected in series with buffer storage inserted between
 - In digital systems, the buffer storage are FFs & registers and data processing elements are "stages" of combinational logic
 - In its simplest form, can be thought of as adding registers in the middle of a computation to reduce our clock period



Performance

- ❖ What does it mean to say X performs better than Y?
- ❖ Silly example: a Tesla vs. a school bus
 - 2026 Tesla Model S Plaid
 - 5 passengers, 9.23 secs in quarter mile
 - 2011 Type D school bus
 - Up to 90 passengers, quarter mile time?



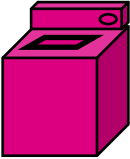
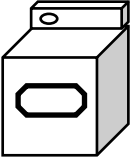


Measurements of Performance

- ❖ *Latency* (or *response time* or *execution time*)
 - Time to complete one task
- ❖ *Throughput* (or *bandwidth*)
 - Tasks completed per unit time

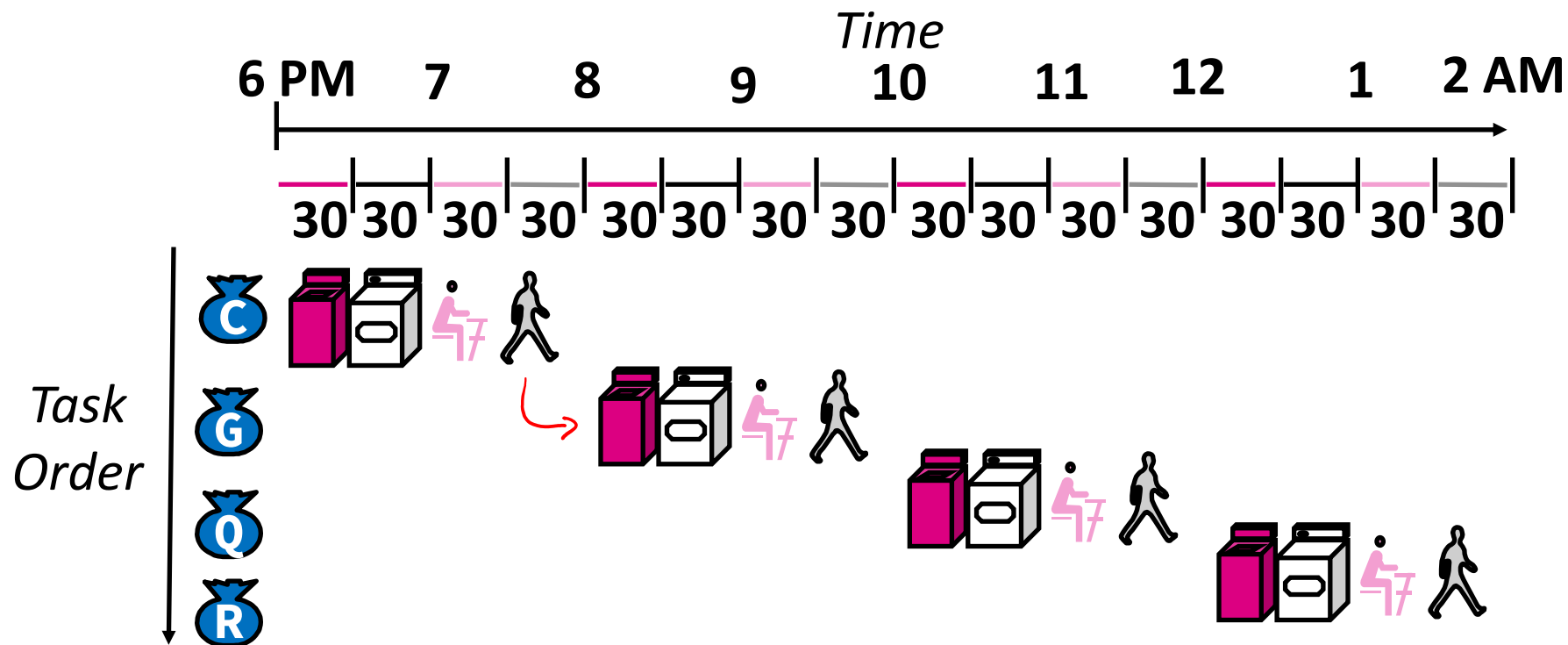
Analogy: Doing Laundry



- ❖ Colton, Grace, Quin, and Rose each have one load of clothes to wash, dry, fold, and put away:

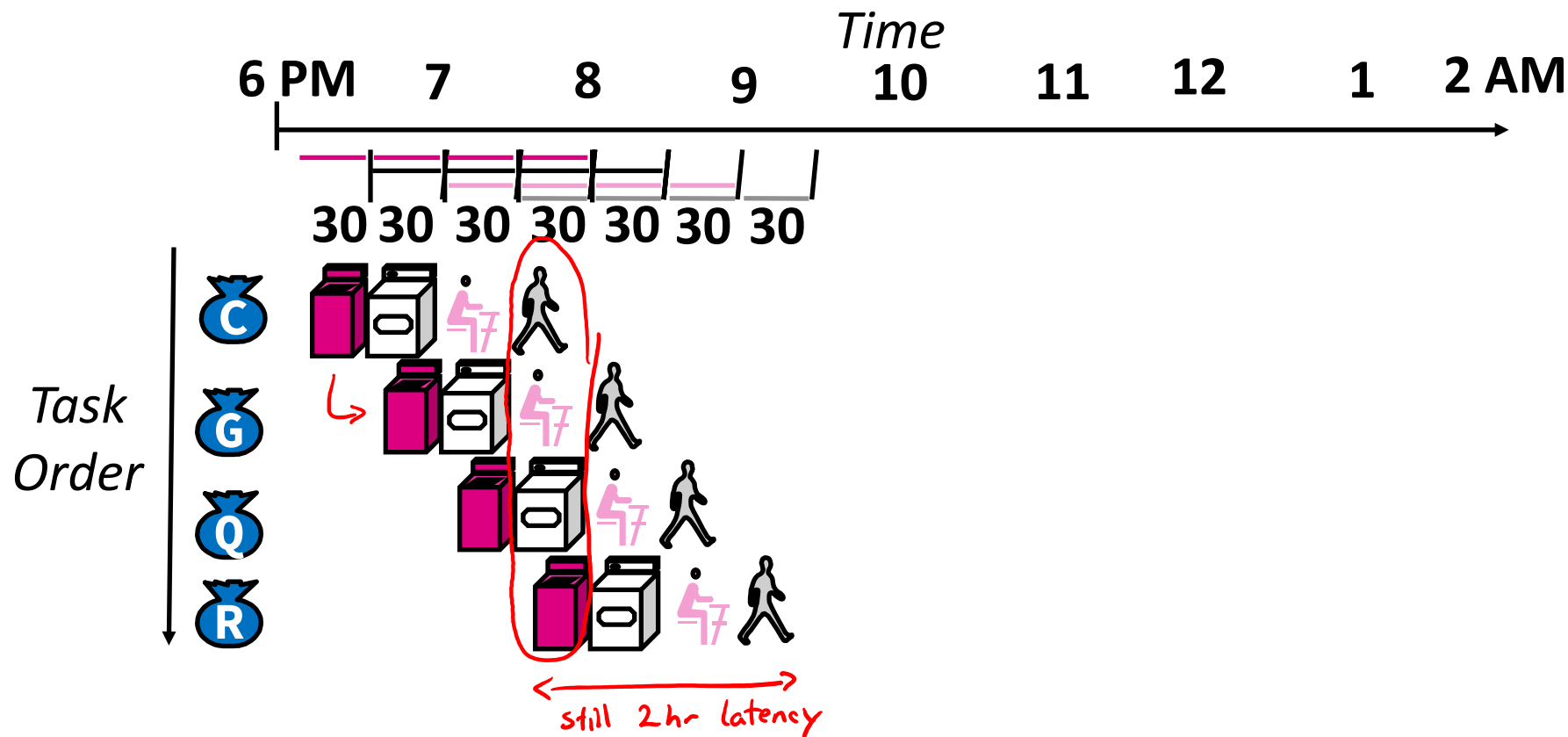
- Washer  takes 30 minutes
- Dryer  takes 30 minutes
- “Folder”  takes 30 minutes
- “Stasher”  takes 30 minutes to put clothes into drawers

Sequential Laundry



- Sequential laundry takes 8 hours for 4 loads
only 1 person in laundry room at a time!

Pipelined Laundry



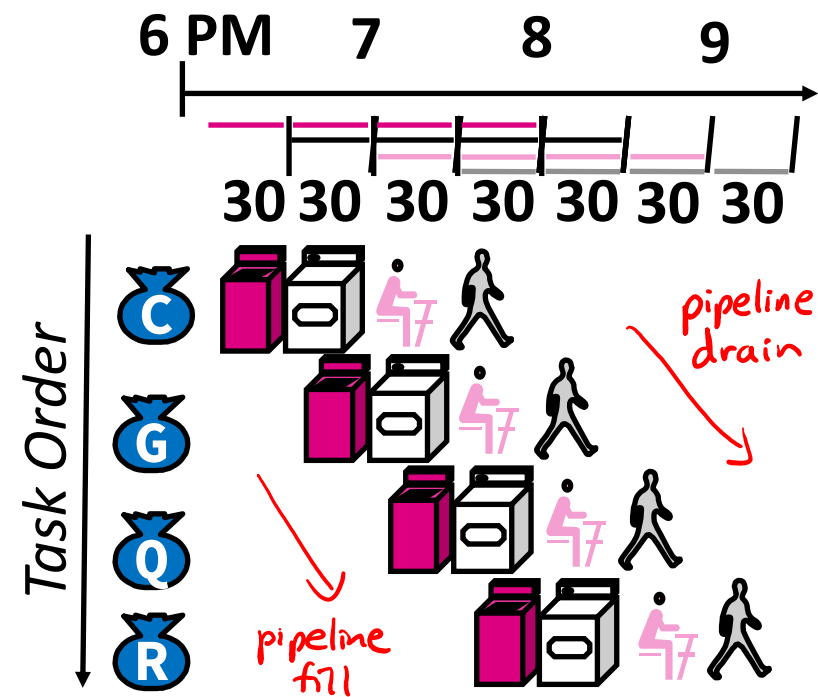
- Pipelined laundry takes 3.5 hours for 4 loads!

up to 4 people in the laundry room at a time!

at most 1 person using any station at a time

Pipelining Notes

- ❖ Pipelining helps throughput of overall workload, but not latency of single task
 - Reduction in critical pathway allows for shorter clock period
- ❖ *Multiple* tasks operating simultaneously using different resources
 - Executing different parts of multiple computations at the same time using the same hardware – like an assembly line
 - Greater utilization of logic resources

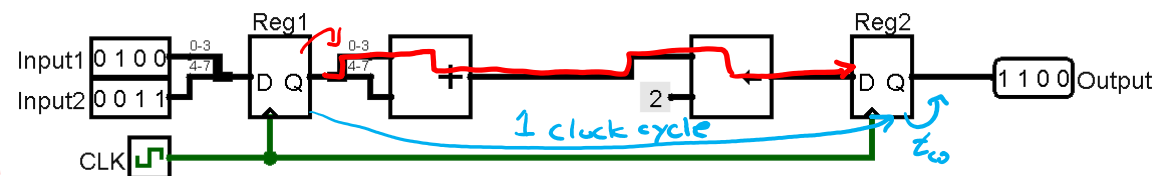


Pipelined Performance Example

- ❖ Assume $t_{CO} = 10\text{ ns}$, $t_{add} = 90\text{ ns}$, $t_{shl} = 50\text{ ns}$
 - For simplicity, assume $t_{clk} = t_{wire} = t_h = t_{su} = 0$
- ❖ Solve for the minimum clock period for each circuit
 - Given this min clock period, solve for the latency and throughput of each circuit

■ Circuit 1:

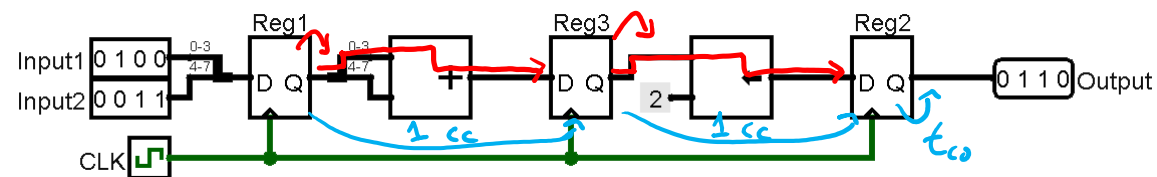
$T_{min} = 150\text{ ns}$
 throughput = $1/T = 1/(150\text{ ns})$
 latency = $T + t_{co} = 160\text{ ns}$



critical path = $t_{co} + t_{add} + t_{shl} \leq T - t_{su}$

■ Circuit 2:

$T_{min} = 100\text{ ns}$
 throughput = $1/T = 1/(100\text{ ns})$
 latency = $2T + t_{co} = 210\text{ ns}$



critical path = $\max(t_{co} + t_{add}, t_{co} + t_{shl}) \leq T - t_{su}$

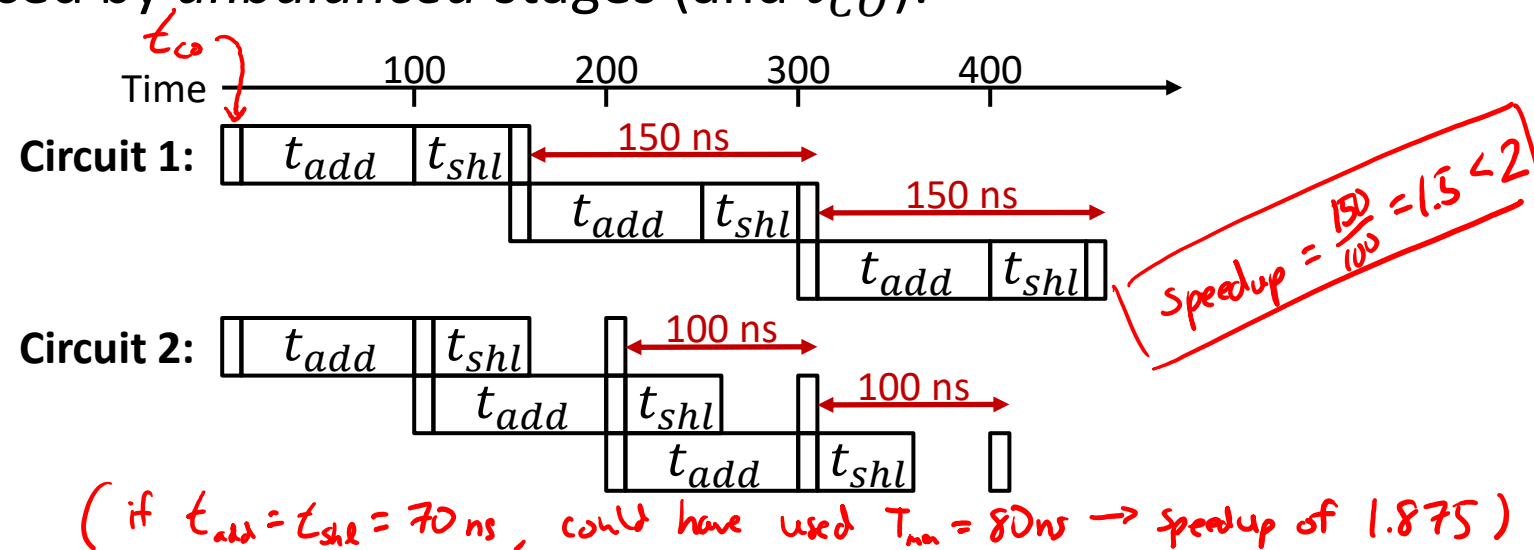
Pipeline Performance

❖ In theory, can measure “speedup” as the ratio in time per completion (TC) of computations

- $speedup = \frac{TC_{original}}{TC_{pipelined}}$

- $speedup_{max} = \# \text{ of pipeline stages}$

- Speedup is reduced by *unbalanced* stages (and t_{CO}):

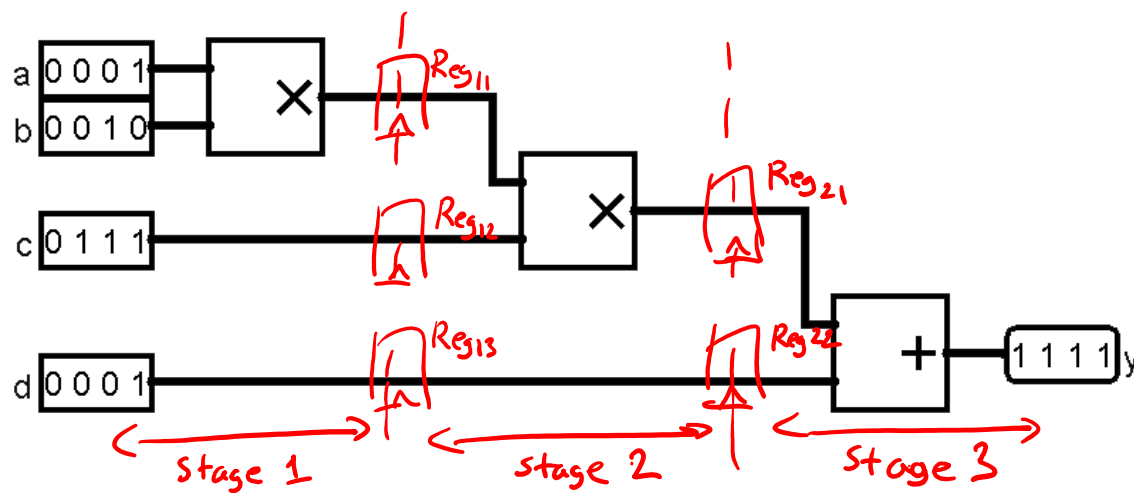


SHORT TECH

BREAK

Pipeline Registers (1/2)

- ❖ Where to add pipeline registers?
 - For a given computation, all paths from any input to output must pass through the *same number* of pipeline registers
- ❖ Example: $y_i = (a_i \times b_i) \times c_i + d_i$

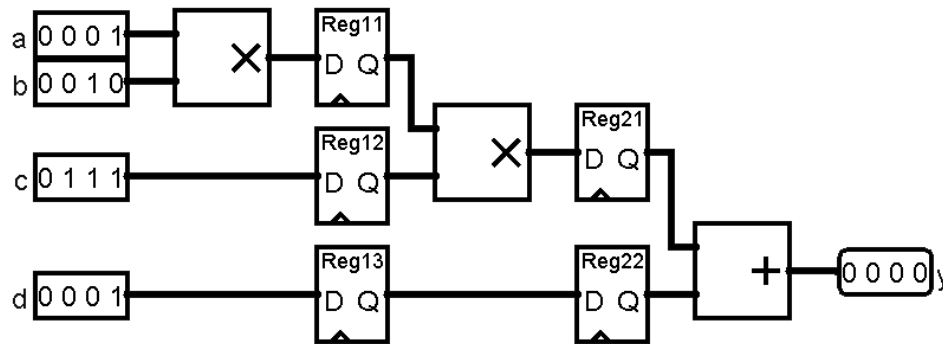


Pipeline Registers (2/2)

- ❖ Where to add pipeline registers?
 - For a given computation, all paths from any input to output must pass through the *same number* of pipeline registers

❖ Example: $y_i = (a_i \times b_i) \times c_i + d_i$

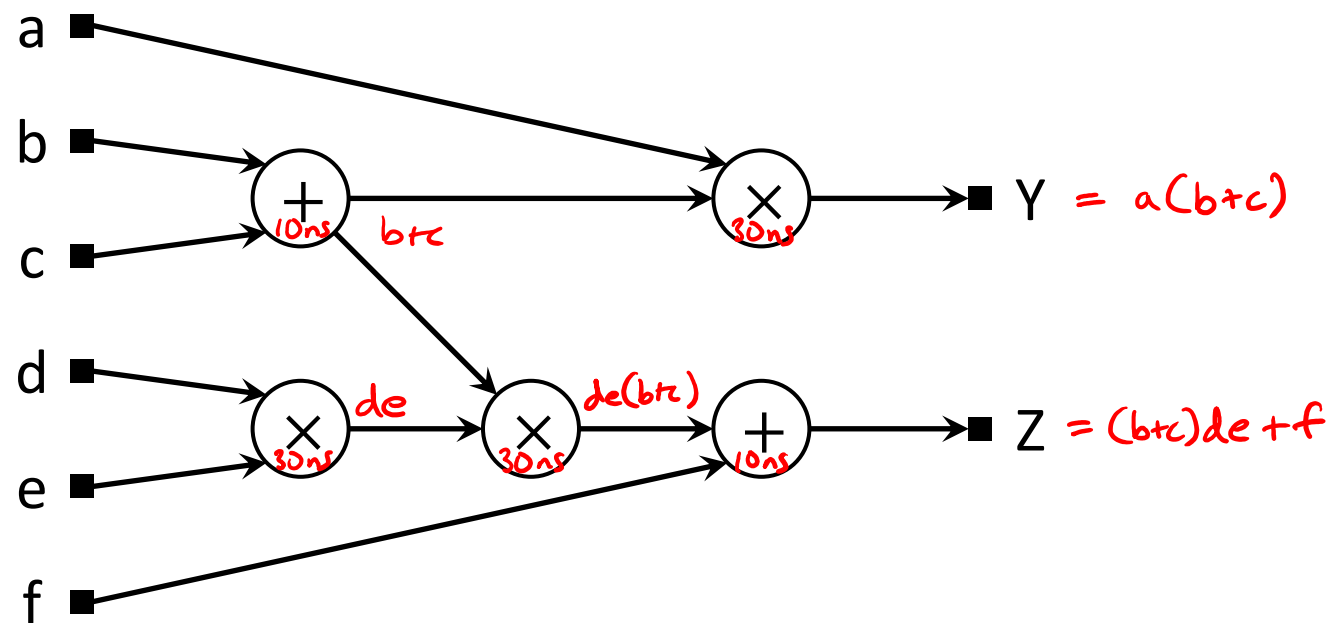
■ Signal flow:



Cycle	1	2	3	4
Reg11	X	$a_1 b_1$	$a_2 b_2$	$a_3 b_3$
Reg12	X	c_1	c_2	c_3
Reg13	X	d_1	d_2	d_3
Reg21	X	X	$a_1 b_1 c_1$	$a_2 b_2 c_2$
Reg22	X	X	d_1	d_2

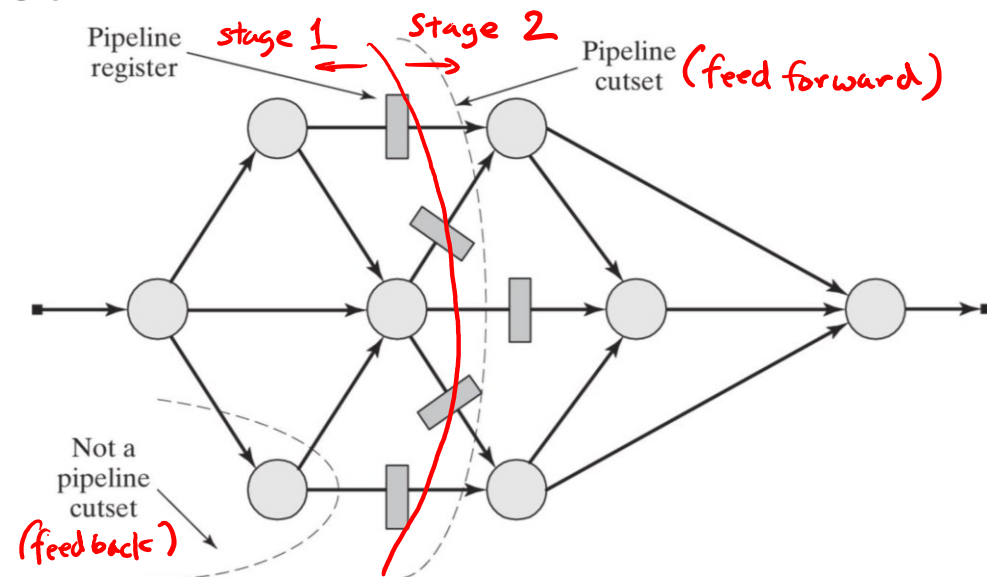
Data Flow Graph

- ❖ A **data flow graph** (DFG) is a visualization tool that can be used to simplify circuits into *directed graphs*
 - Nodes are computations (and their delays)
 - Edges represent data dependencies



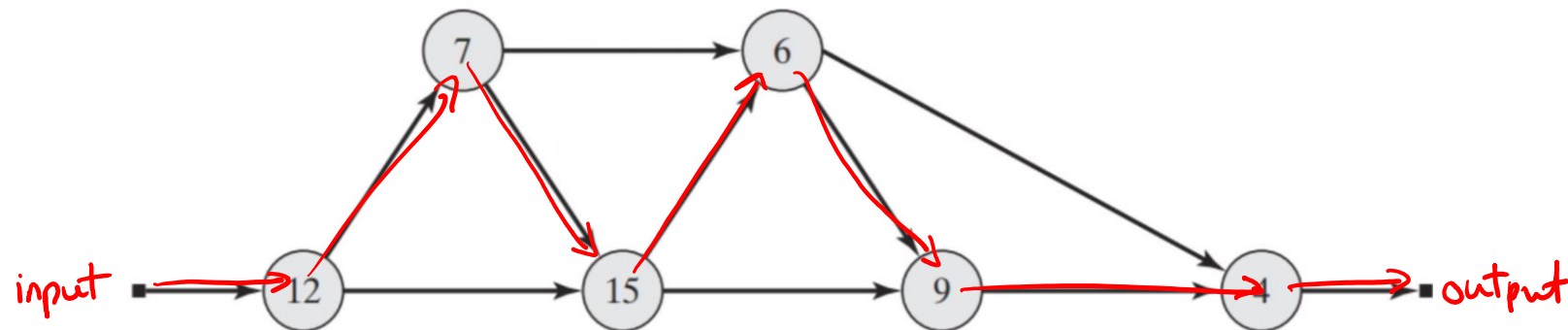
Pipeline Cutset

- ❖ A **cutset** is a set of edges that form two disjoint graphs when removed/cut
 - *Feedforward* cutset: data travels only forward in the cutset
 - *Feedback* cutset: data travels in both directions in the cutset
- ❖ Pipelining is done by placing a register along every edge in a **pipeline (feedforward) cutset**:



Pipeline Cutset Example: Original

- ❖ The following data flow graph shows the propagation delay in each node
 - For simplicity, assume $t_{CO} = 0$
 - Original (non-pipelined) performance:

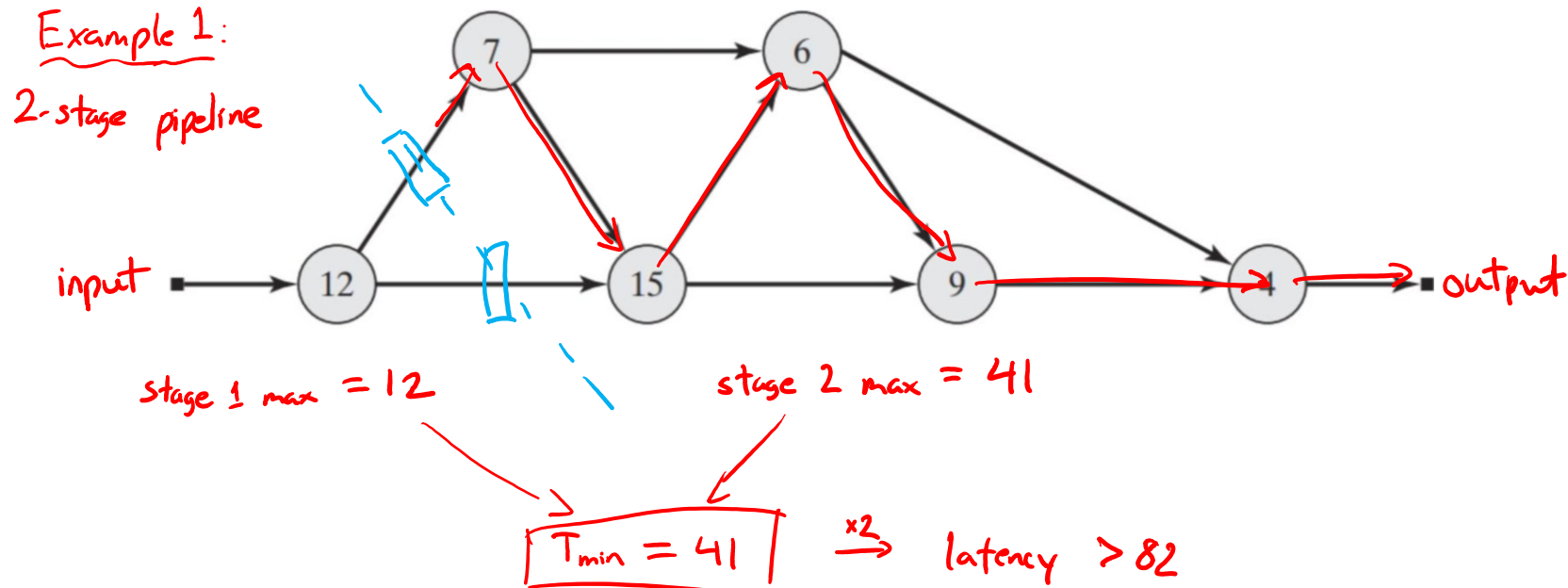


critical path has delay of $12+7+15+6+9+4 = 53$

$$\boxed{T_{mm} = 53} \xrightarrow{\times 1} \text{latency} > 53$$

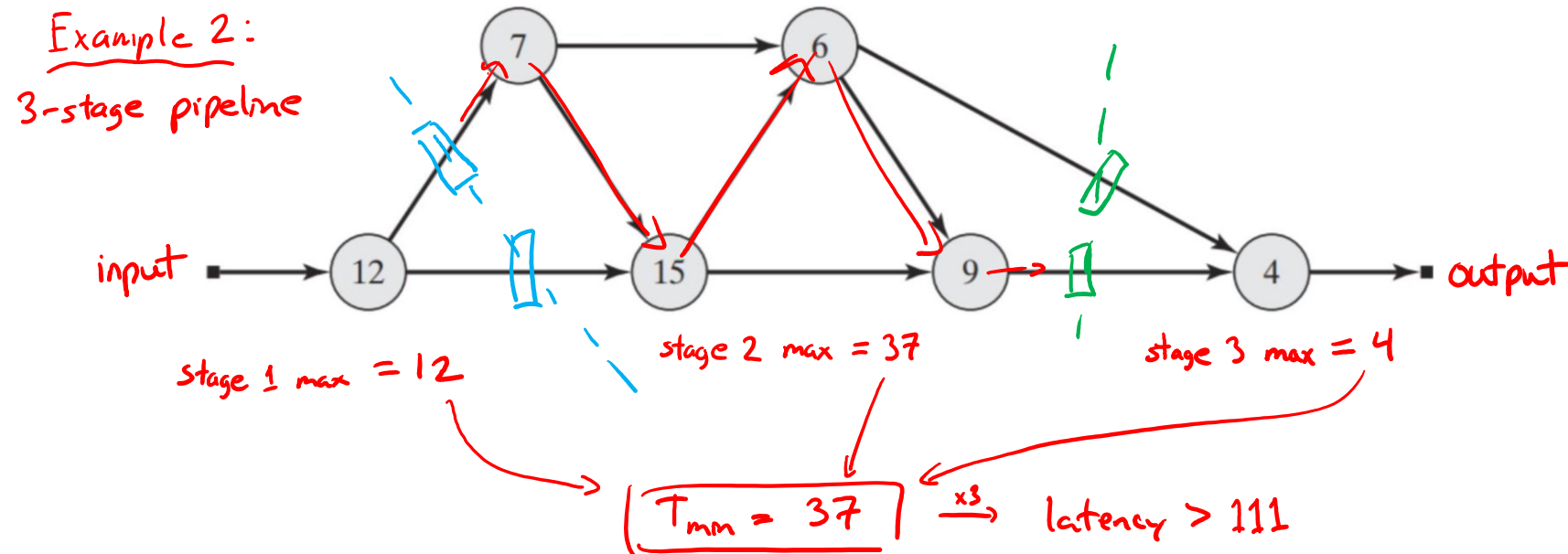
Pipeline Cutset Example: Version 1

- ❖ The following data flow graph shows the propagation delay in each node
 - Create 2-3 different pipelined versions of this DFG and compute the maximum delay of each stage and minimum clock period for the pipelined computation
 - For simplicity, assume $t_{CO} = 0$



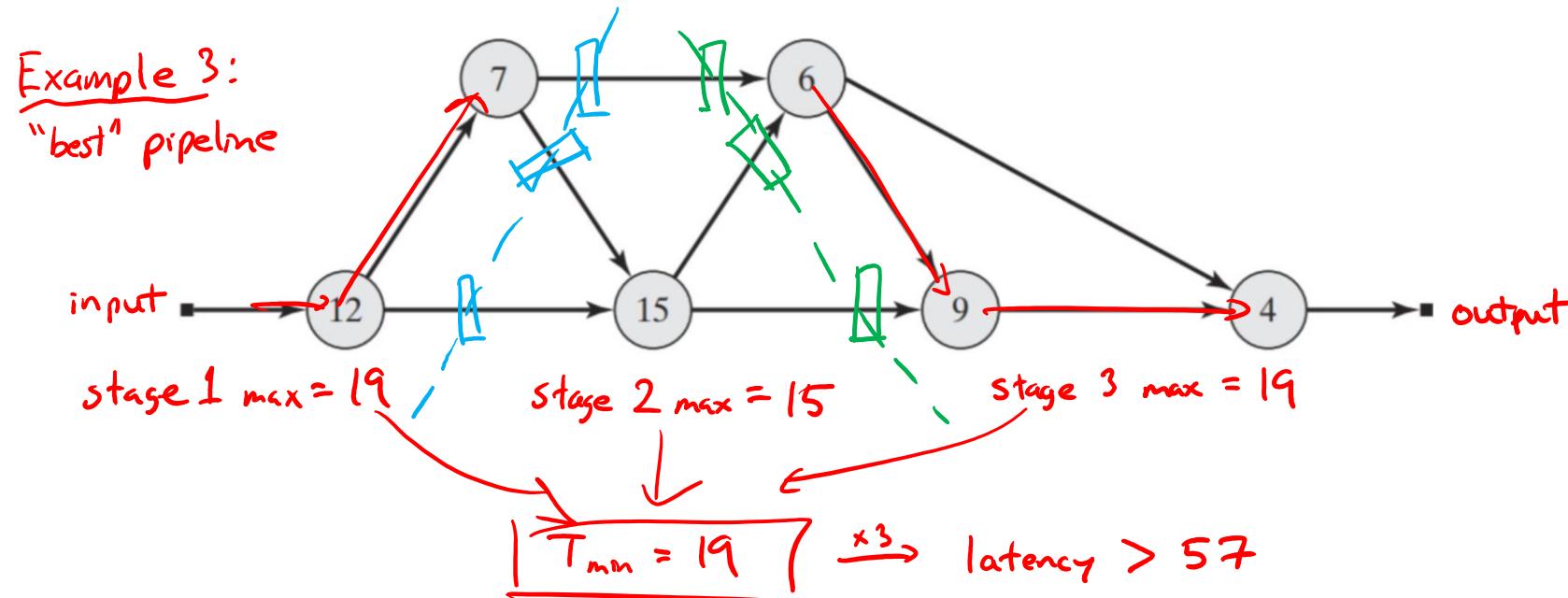
Pipeline Cutset Example: Version 2

- ❖ The following data flow graph shows the propagation delay in each node
 - Create 2-3 different pipelined versions of this DFG and compute the maximum delay of each stage and minimum clock period for the pipelined computation
 - For simplicity, assume $t_{CO} = 0$



Pipeline Cutset Example: Version 3

- ❖ The following data flow graph shows the propagation delay in each node
 - Create 2-3 different pipelined versions of this DFG and compute the maximum delay of each stage and minimum clock period for the pipelined computation
 - For simplicity, assume $t_{CO} = 0$



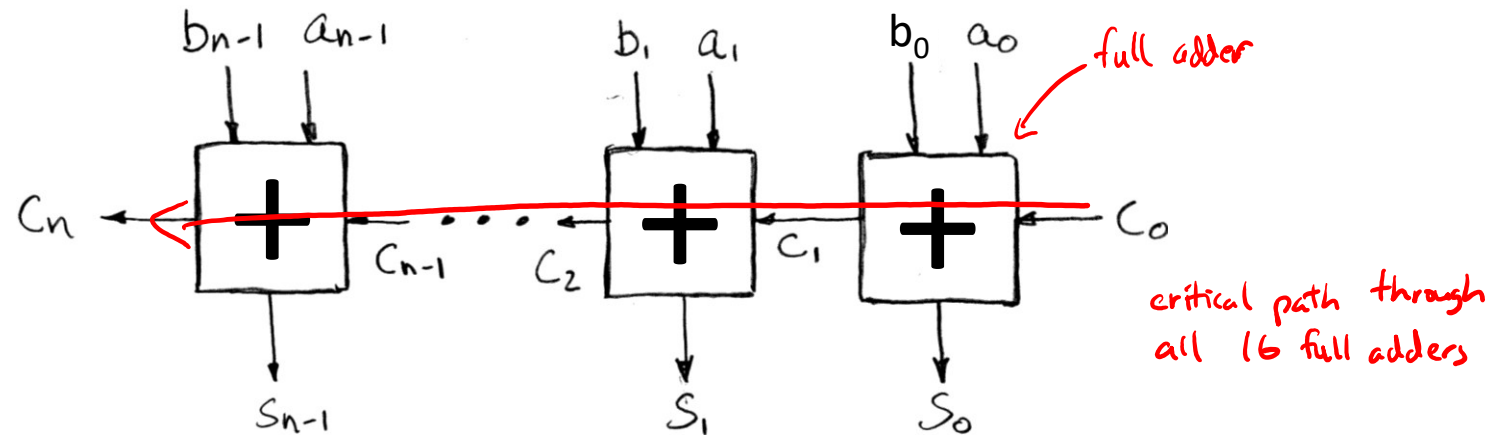
Pipeline Design Questions

- ❖ When should I add pipelining?
 - Check if it is possible first (*i.e.*, a pipeline cutset must exist)
 - Want to reduce the **critical path** in your computation/system
 - Your system can afford the increase in latency and hardware

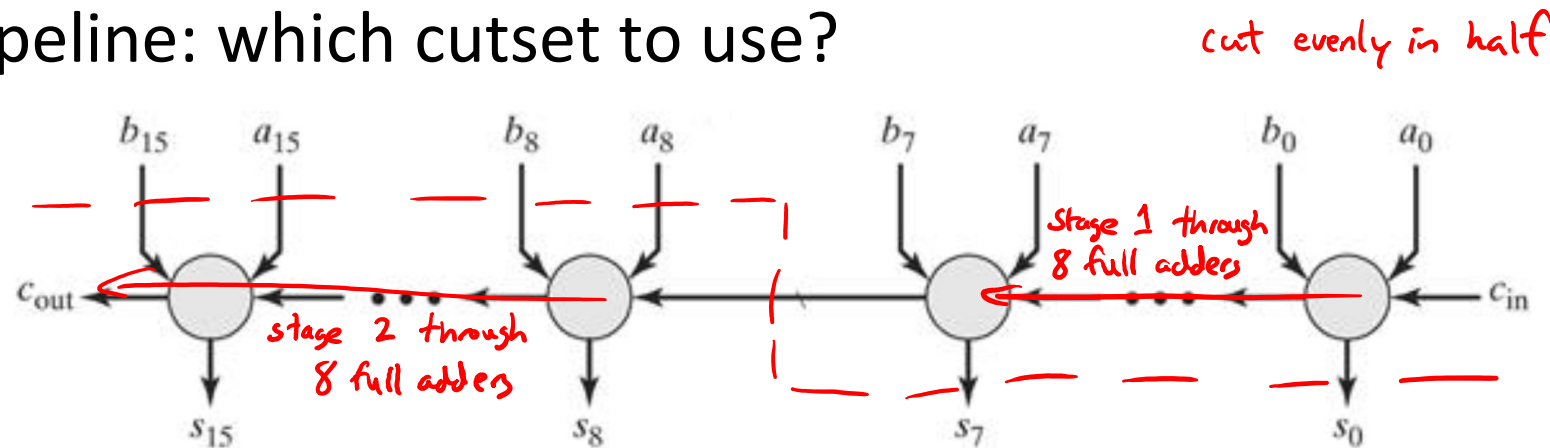
- ❖ Where do the pipeline registers go?
 - Must be placed at proper pipeline cutsets
 - Want to make pipeline stages as balanced as possible to maximize speedup

Design Example: 16-bit Ripple-Carry Adder

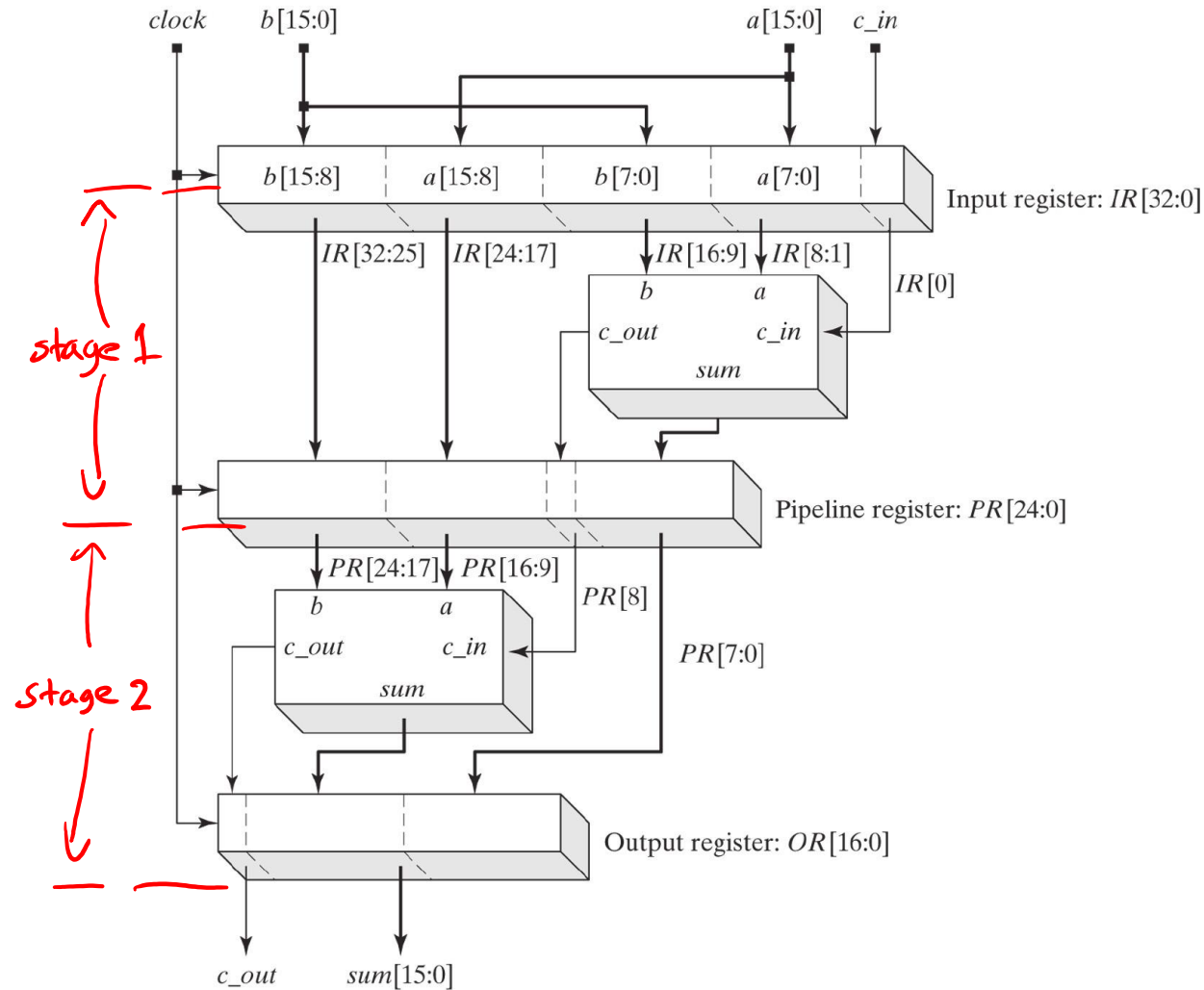
- ❖ Problem: C_n takes a long time to compute!



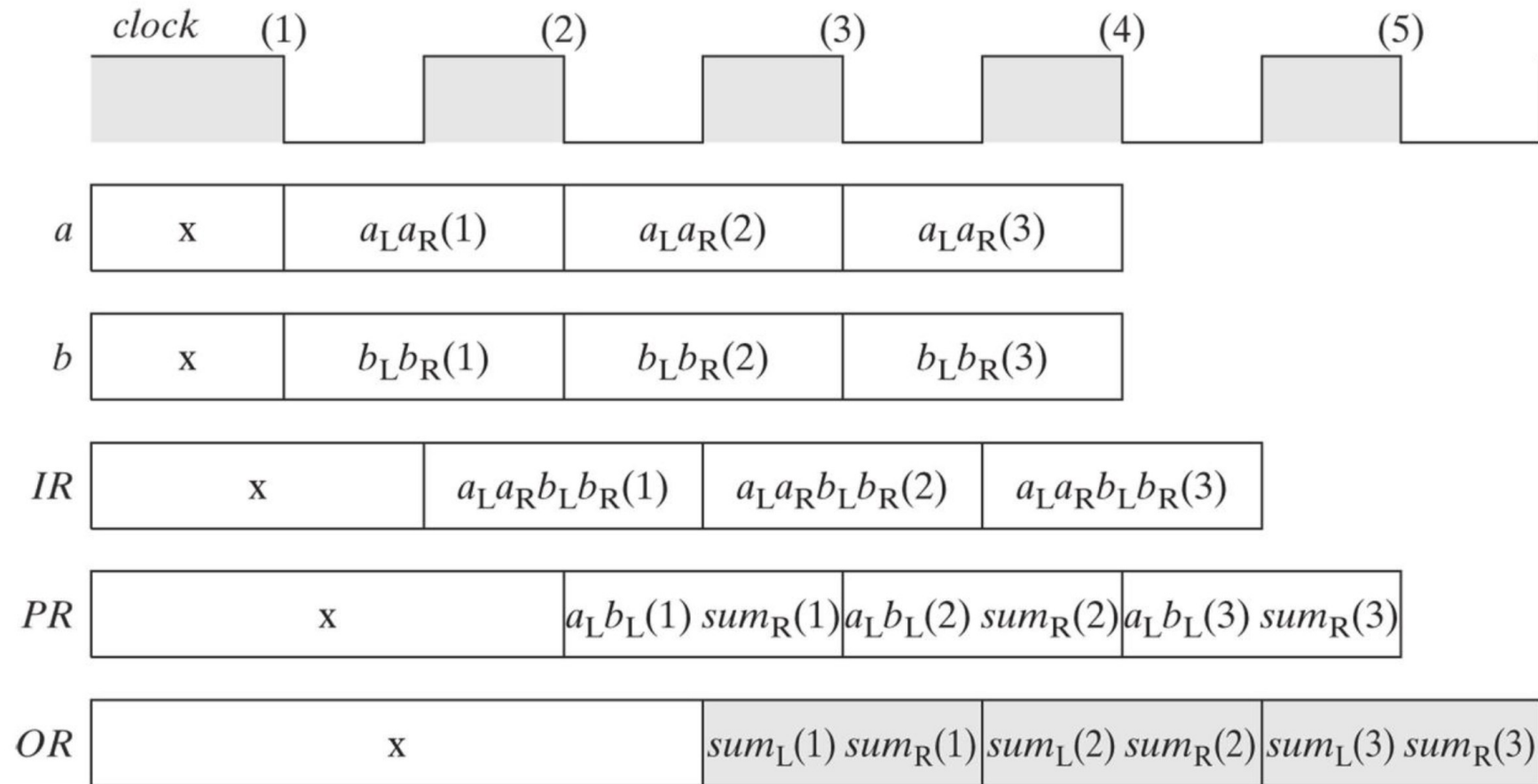
- ❖ 2-stage pipeline: which cutset to use?



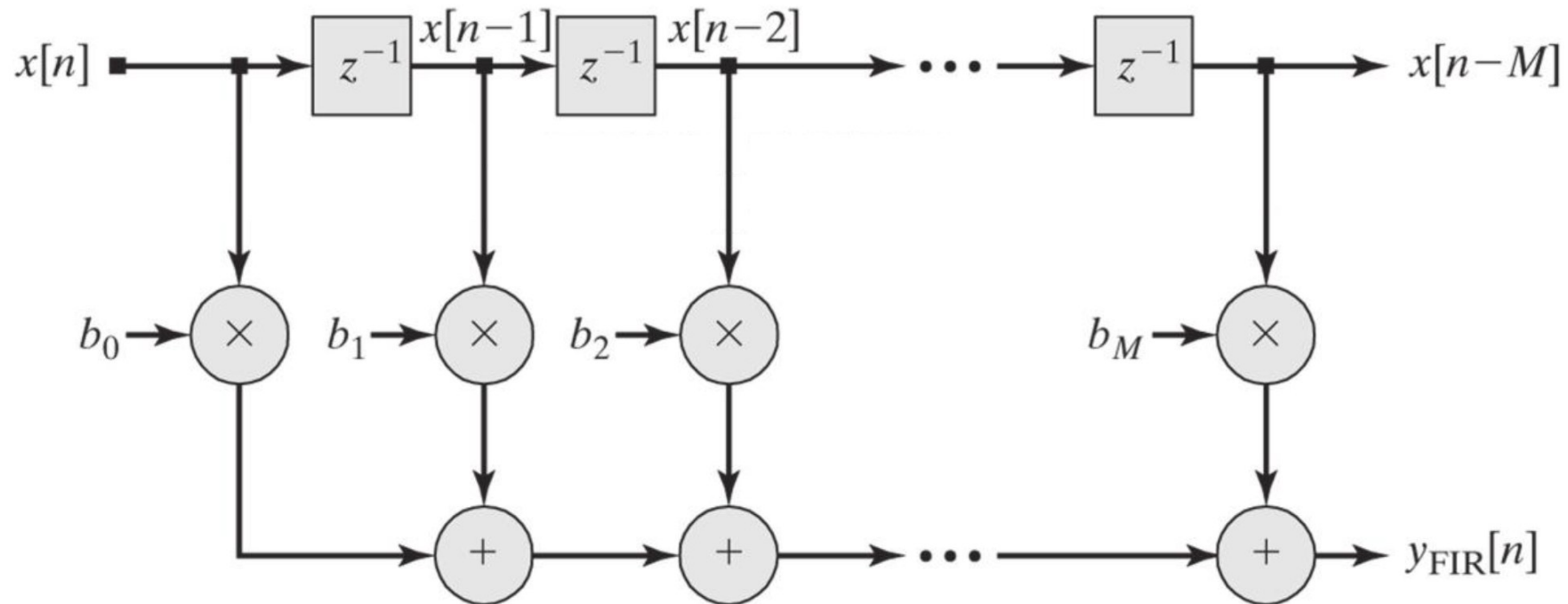
Design Example: 16-bit Pipelined Adder (1/2)



Design Example: 16-bit Pipelined Adder (2/2)



Design Example: FIR Filter



Design Example: Pipelined FIR Filter

