RAZOR: CIRCUIT-LEVEL CORRECTION OF TIMING ERRORS FOR LOW-POWER OPERATION

Shohaib Aboobacker
TU München

22nd March 2011

Based on

Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation

Dan Ernst, Trevor Mudge , Shidhartha Das, Sanjay Pant, Rajeev Rao, Toan Pham,Conrad Ziesler, David Blaauw, Todd Austin, Krisztian Flautner (ARM), and Nam Sung Kim (Intel)

Published in the 36th Annual International Symposium on Microarchitecture (MICRO-36), December 2003.

Outline

- Introduction
- Critical supply voltage
- Razor approach
- Error correction and detection
- Circuit level implementation issues
- Pipeline error recovery mechanisms
- Supply Voltage control
- Summary

Introduction

- Need for high performance with low power budget
- Dynamic power scales quadratically with supply voltage
- Reducing supply voltage increases delay and limits the maximum frequency

  Supply voltage ↓
  Dynamic Power ↓
  Propagation delays ↑
  Maximum frequency ↓

- Dynamic Voltage Scaling
  - Adapting voltage to meet performance demands of workload

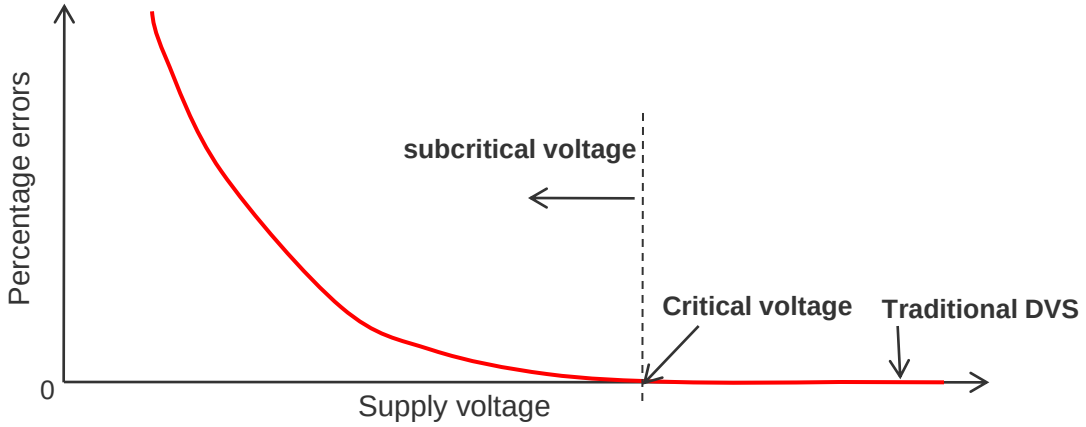
Critical supply voltage

- Critical supply voltage
  - Minimum supply voltage that ensures correct operation
- Affected by environmental and process-related variabilities
  - Voltage drops in power supply network
  - Temperature fluctuations
  - Changes in doping concentration
  - Cross-coupling noise
- Traditional approach to find critical voltage too conservative
  - Pessimistic approach
  - Worst case corner conditions highly improbable

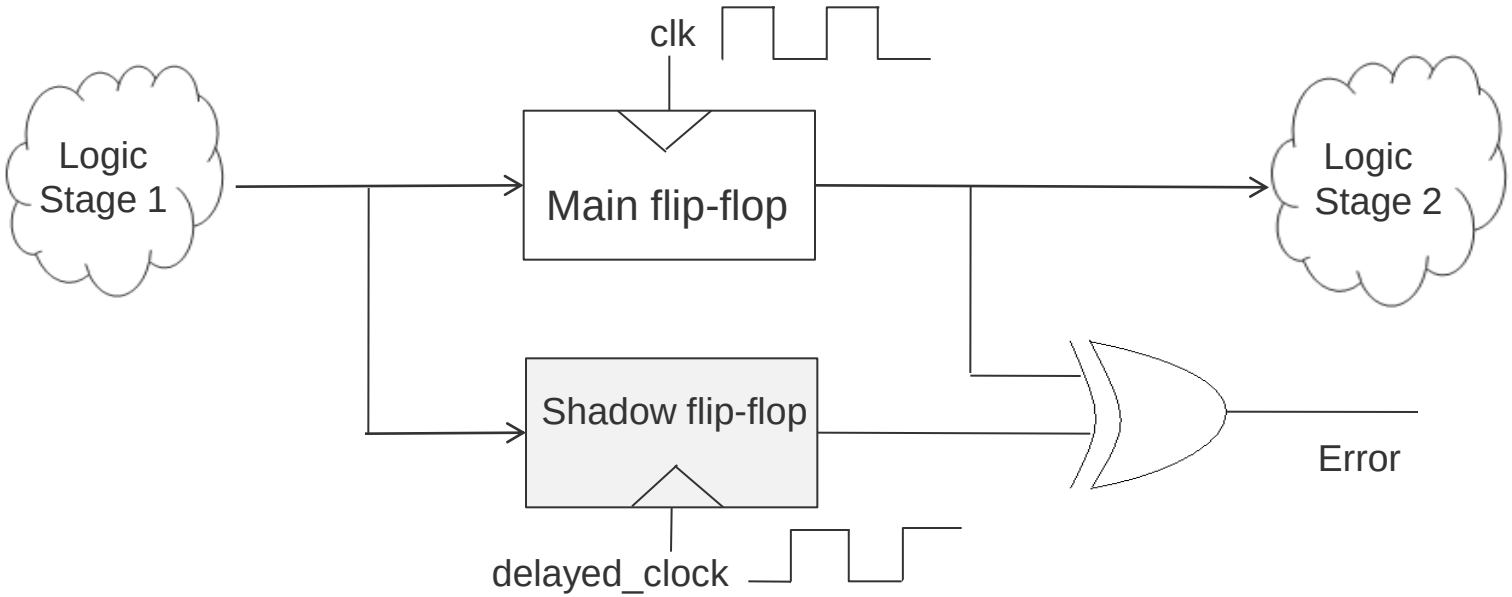
Razor approach

- Developed in the Electrical Engineering and Computer Science Department at the University of Michigan

- Operation at subcritical supply voltages
- Monitor error rate during operation
- Dynamic detection and correction of delay failures
- Power penalty of correction vs Voltage power savings

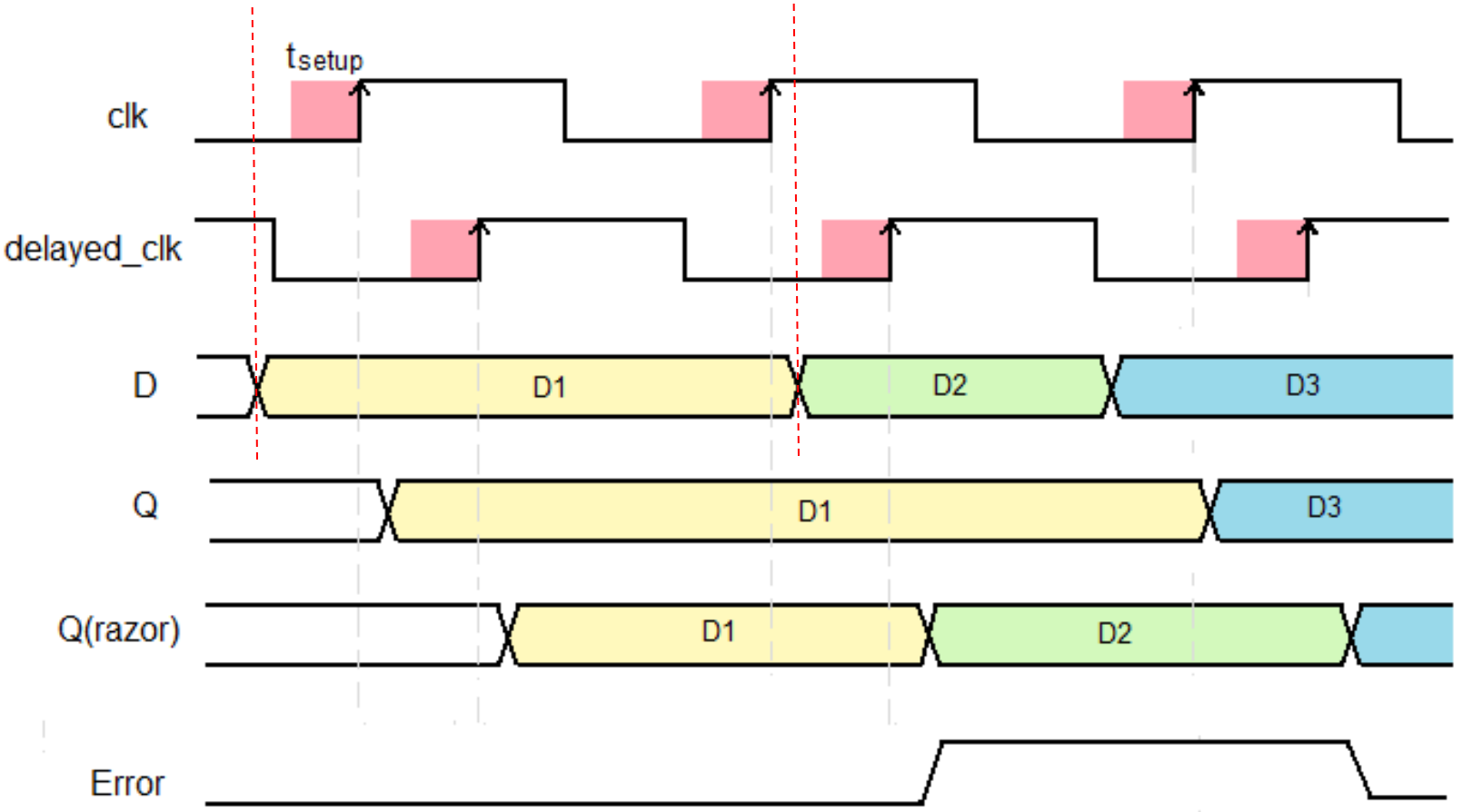


Error detection

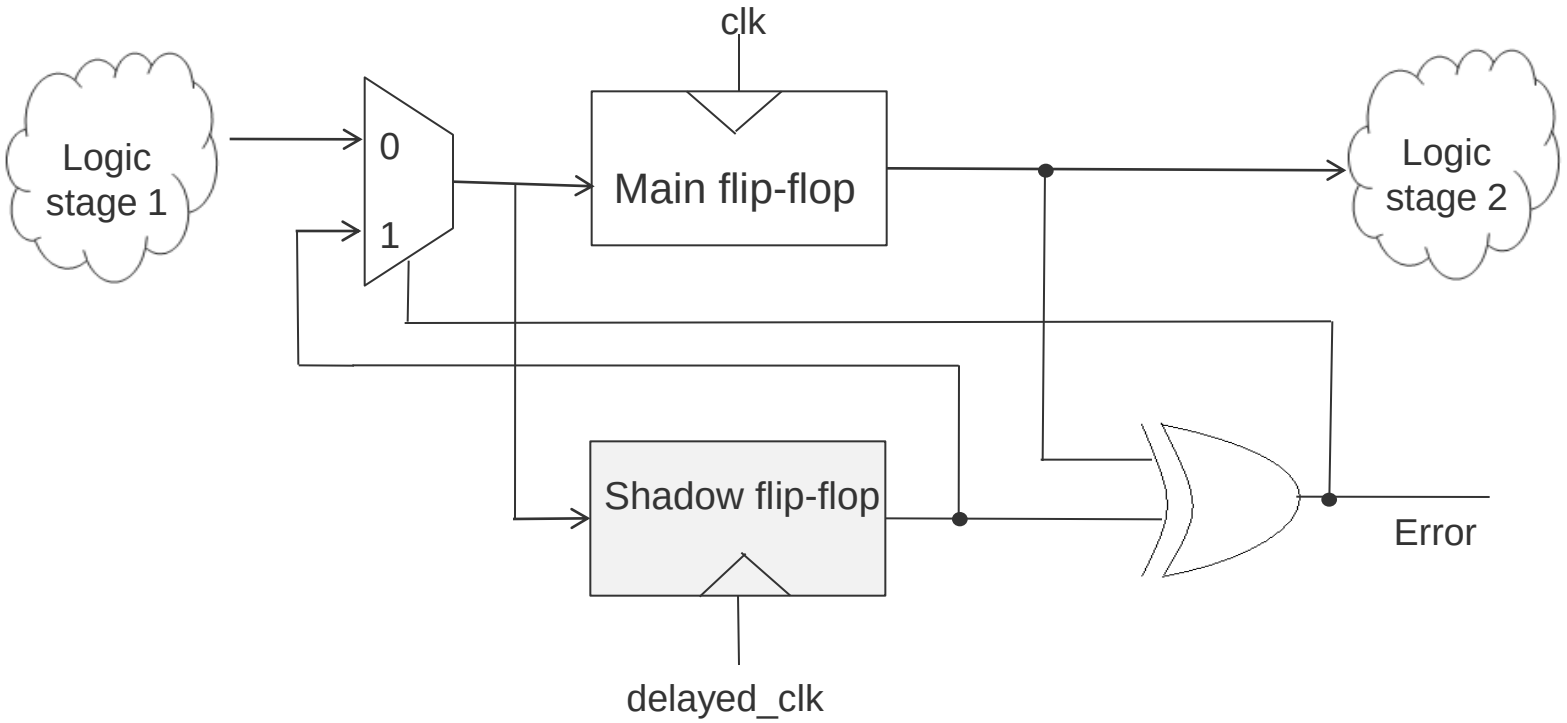


- **Shadow flip-flop** with delayed clock with every flip-flop
- Operating voltage constrained such that the worst-case delay is guaranteed to meet the shadow flip-flop setup time.
- No error if logic stage 1 meets setup time for main flip-flop
- Otherwise, main FF will latch wrong value, and shadow flip-flop will latch late correct value

Error detection timing diagram

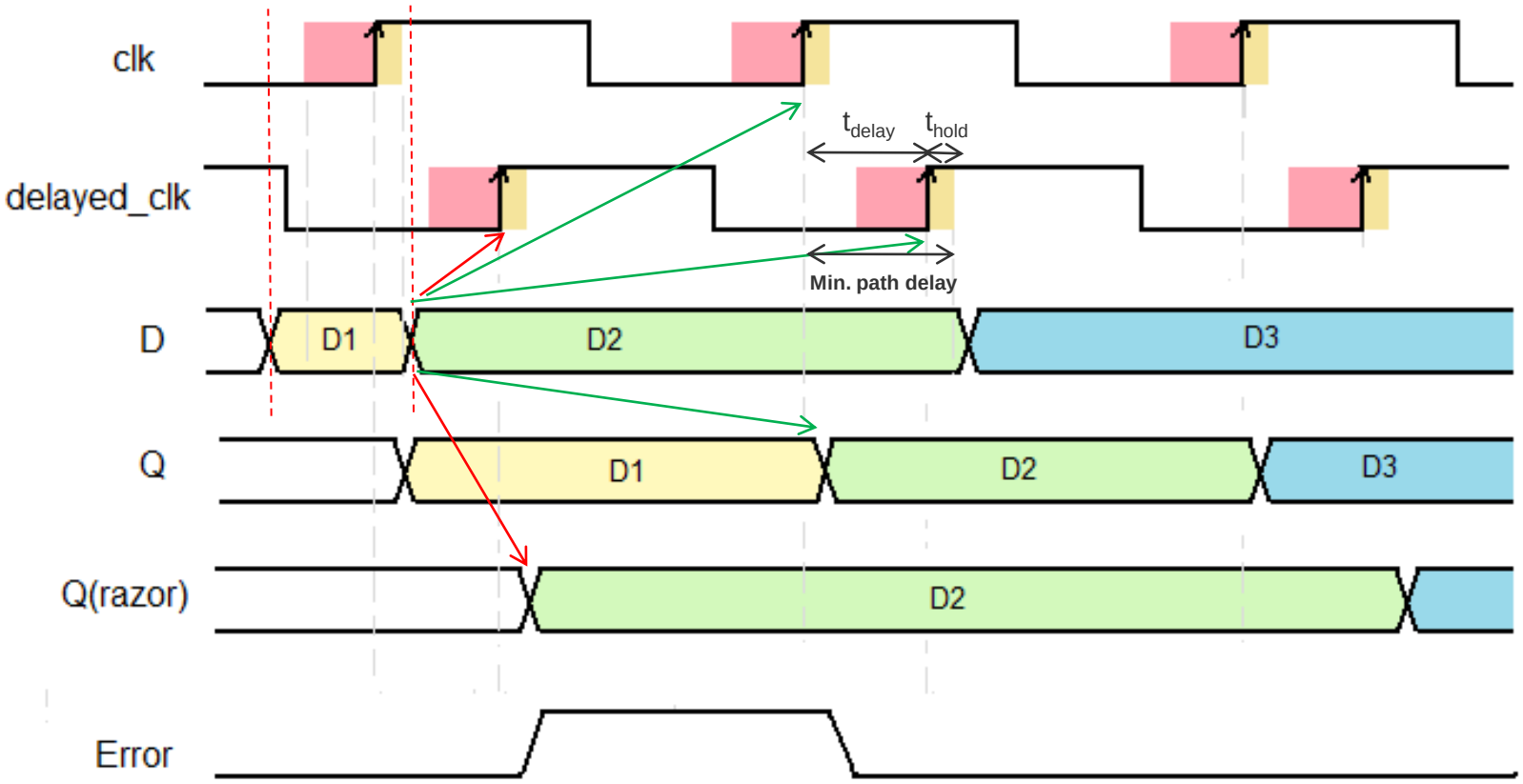


Error correction



- If error is high, correct value from shadow flip-flop is restored to input of main flip-flop
- No error if logic stage 1 meets setup time for main flip-flop

Short path constraint

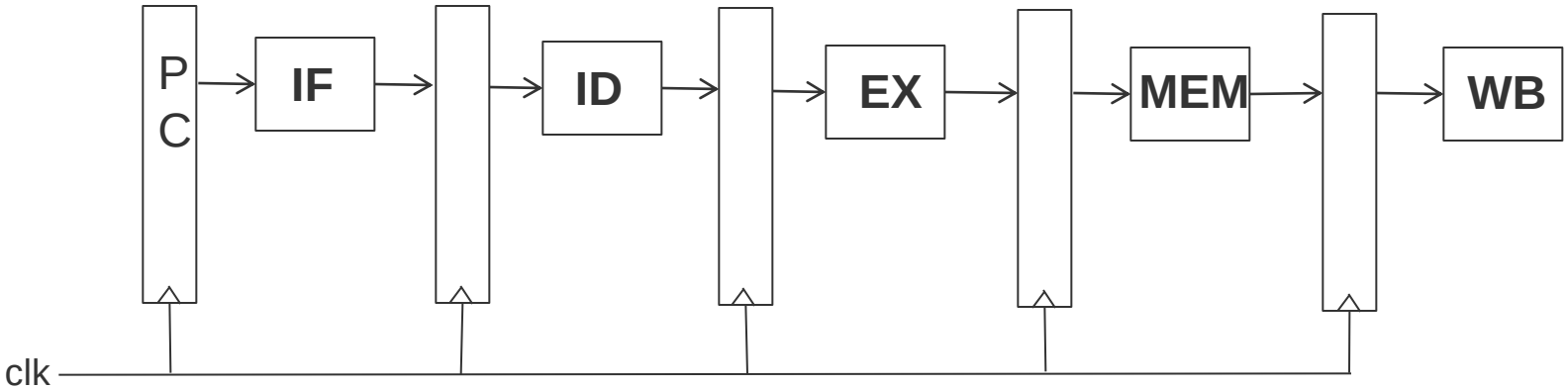


Short path constraint

$$\text{Minimum Path delay} = t_{delay} + t_{hold}$$

- Large clock delay increases power overhead and need for buffers
- Small clock delay reduces margin

Pipelined processing

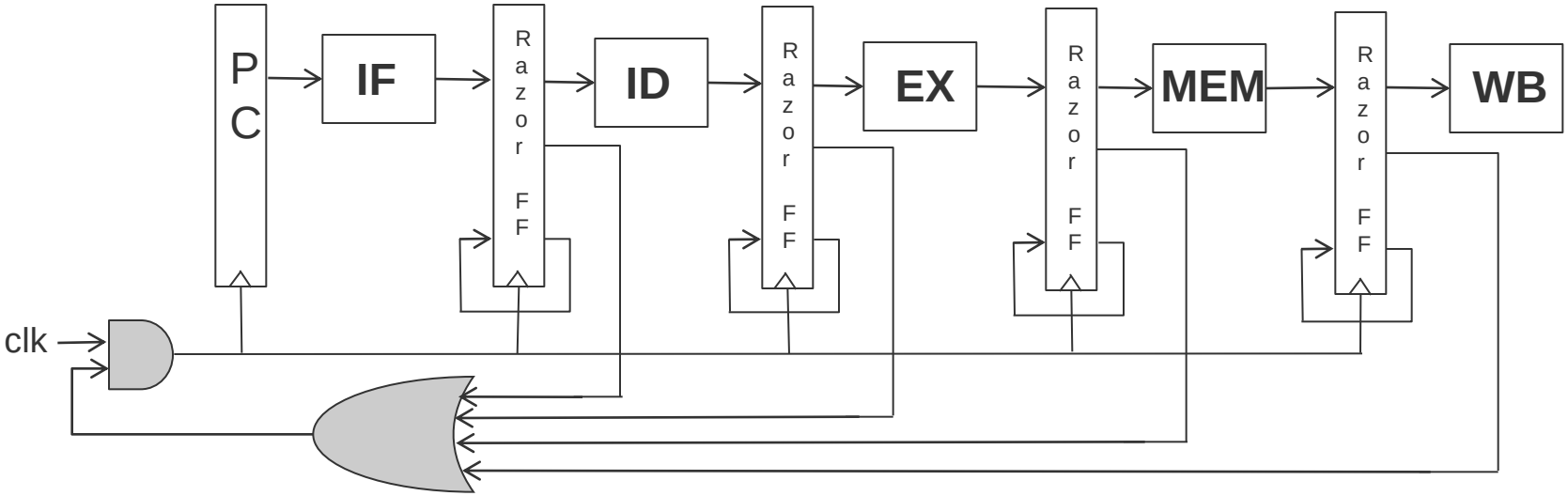


Instr. \ Cycle	1	2	3	4	5	6	7	8	9
1	IF	ID	EX	MEM	WB				
2		IF	ID	EX	MEM	WB			
3			IF	ID	EX	MEM	WB		
4				IF	ID	EX	MEM	WB	
5					IF	ID	EX	MEM	WB

Pipeline error recovery mechanisms

- Clock gating
  - Stall pipeline for one cycle in case of error
  - Recompute result of every stage in extra period using shadow flip-flop as input
  - A single cycle can tolerate any number of errors

Pipeline recovery with clock gating

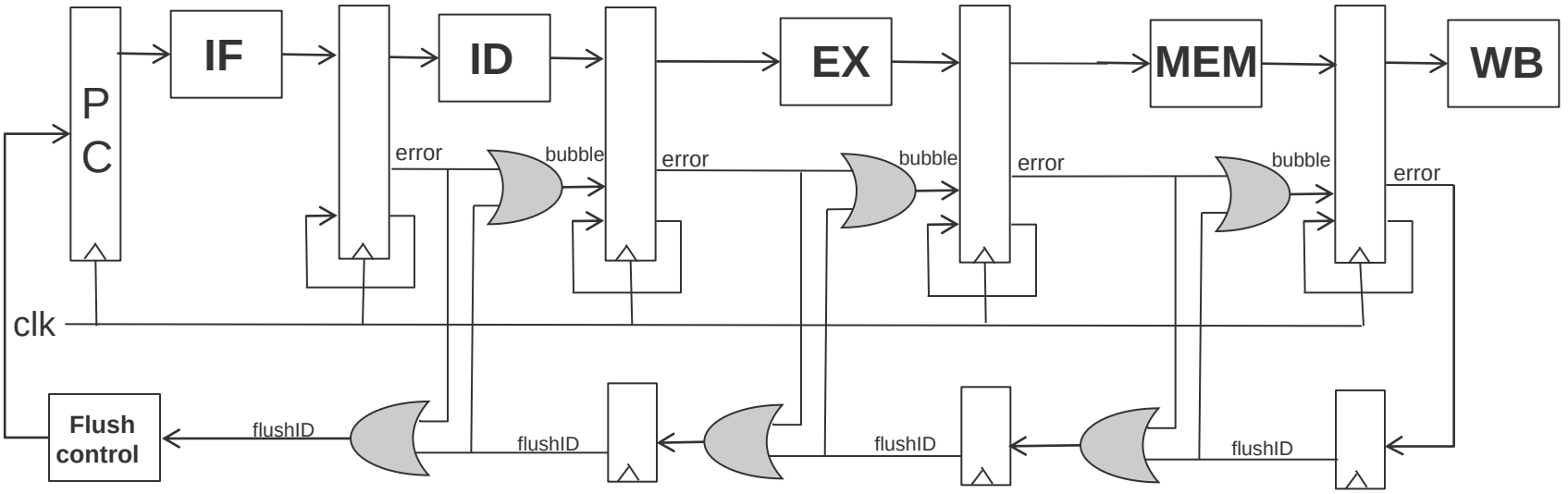


Instr. \ Cycle	1	2	3	4	5	6	7	8	9	10
1	IF	ID	EX	MEM	stall	WB				
2		IF	ID*	EX*	EX	MEM	WB			
3			IF	ID	stall	EX	MEM	WB		
4				IF	stall	ID	EX	MEM	WB	
5					stall	IF	ID	EX	MEM	WB

Pipeline error recovery mechanisms

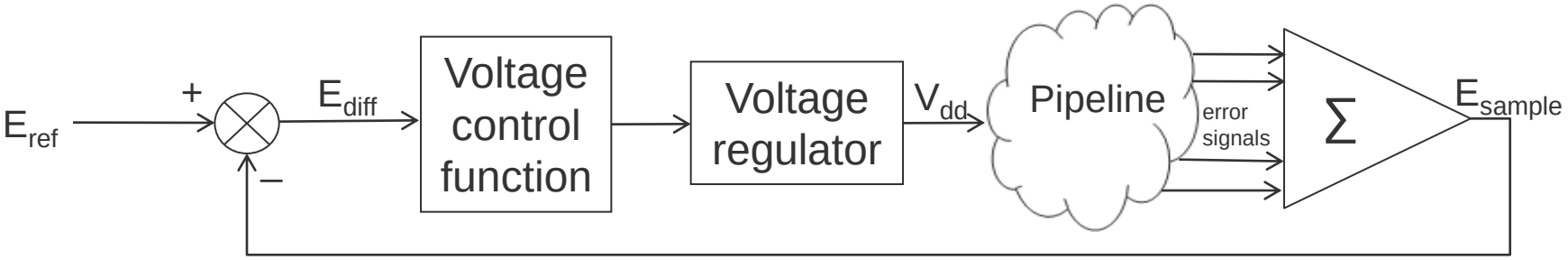
- Counterflow pipelining
  - Uses bubble signal to invalidate following instructions
  - Error propagation is pipelined
  - Flush train propagates in the opposite direction
  - When flush reaches the start, PC restarts execution.

Pipeline recovery with counterflow



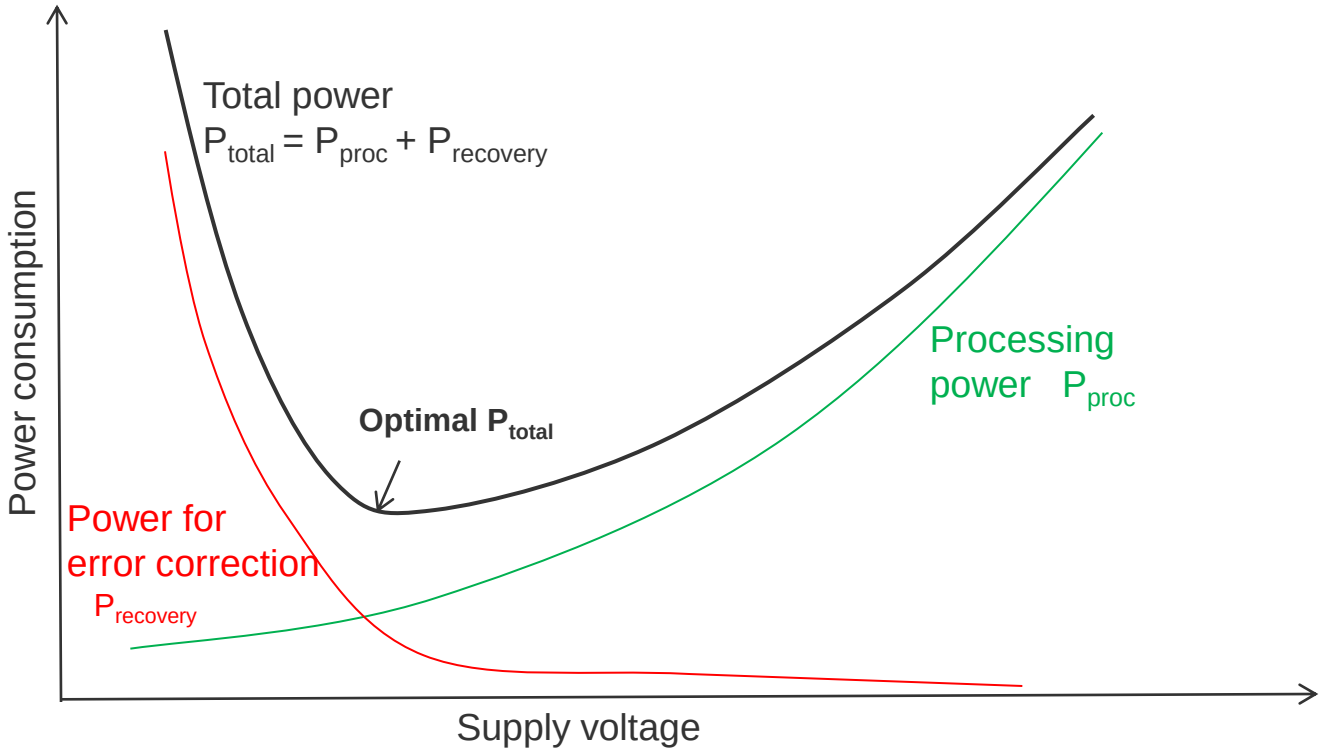
Instr. \ Cycle	1	2	3	4	5	6	7	8
1	IF	ID	EX	MEM	WB			
2		IF	ID*	EX*	bubble	MEM	WB	
3			IF	ID	flushID	flushIF	IF	ID
4				IF				IF

Supply voltage control



- Supply voltage adjusted based on monitored error rates.
- Low error rates means voltage can be lowered further
- Increasing error rates indicate failing timing constraints and voltage should be increased
- Find optimal non-zero error rate

Power savings by Razor DVS



Summary

- Purposely operate at subcritical voltages to capture data-dependent latency margins
- Tolerate some errors and correct them
- In-circuit error detection and correction using Shadow flip-flop
- Tune voltage based on error rate
- Pipeline initiates recovery after timing error
- Tradeoff between power savings from lower voltage and overhead of correction