# CSC 252/452: Computer Organization Fall 2024: Lecture 14

Instructor: Yanan Guo

Department of Computer Science
University of Rochester

## **Announcements**

- Programming assignment 3 is out
  - Details: <a href="https://www.cs.rochester.edu/courses/252/fall2024/labs/assignment3.html">https://www.cs.rochester.edu/courses/252/fall2024/labs/assignment3.html</a>
  - Due on **Oct. 25th**, 11:59 PM

## Pipeline Stages

### **Fetch**

- Use PC to read instruction
- Compute new PC for nonjump instructions

#### Decode

Read program registers

#### Execute

- Operate ALU
- Compute new PC for jump instructions

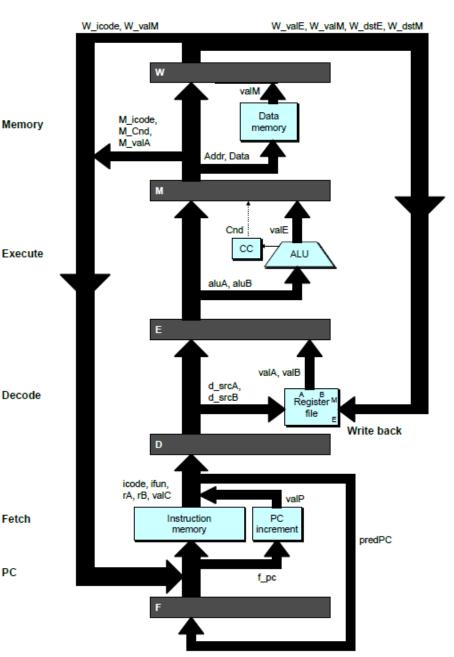
### Memory

Read or write data memory

PC

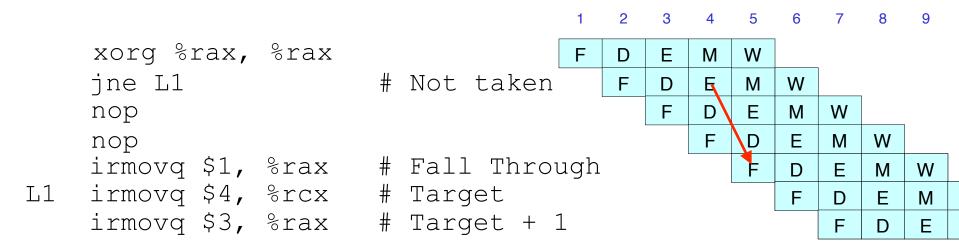
#### Write Back

Update register file



## **Control Dependency**

- **Definition**: Outcome of instruction A determines whether or not instruction B should be executed.
- Jump instruction example below:
  - jne L1 determines whether irmovq \$1, %rax should be executed
  - But jne doesn't know its outcome until after its Execute stage



## **Better Pipelining**

#### **Fetch**

- Use PC to read instruction
- Compute new PC for nonjump instructions

#### Decode

- Read program registers
- Compute new PC for jump instructions

### Execute

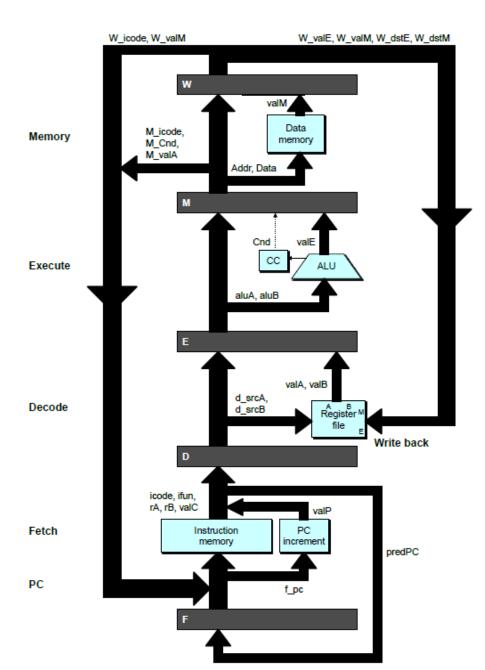
· Operate ALU

### Memory

Read or write data memory

#### Write Back

Update register file



## **Better Pipelining**

#### **Fetch**

- Use PC to read instruction
- Compute new PC for nonjump instructions

#### Decode

- Read program registers
- Compute new PC for jump instructions

#### Execute

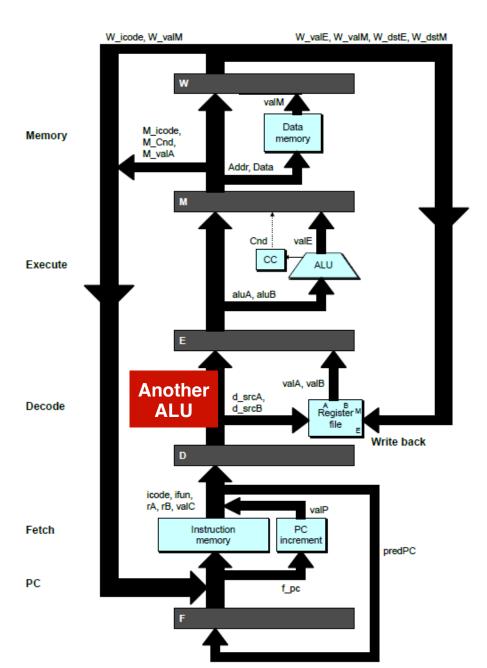
Operate ALU

### Memory

Read or write data memory

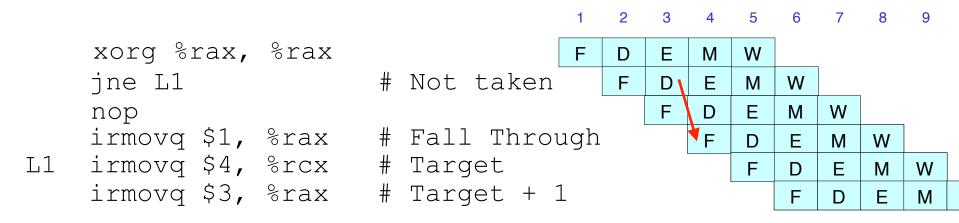
### Write Back

Update register file



## Saving One Cycle

- **Definition**: Outcome of instruction A determines whether or not instruction B should be executed.
- Jump instruction example below:
  - jne L1 determines whether irmovq \$1, %rax should be executed
  - But jne doesn't know its outcome until after its Execute stage



## **Resolving Control Dependencies**

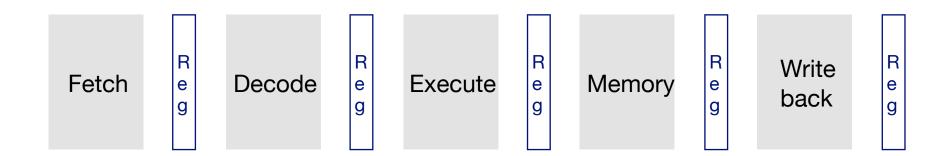
### Software Mechanisms

- Adding NOPs: requires compiler to insert nops, which also take memory space — not a good idea
- Delay slot: insert instructions that do not depend on the effect of the preceding instruction. These instructions will execute even if the preceding branch is taken — old RISC approach

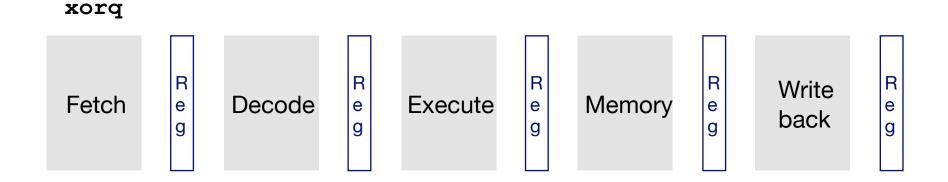
### Hardware mechanisms

- Stalling (Think of it as hardware automatically inserting nops)
- Branch Prediction
- Return Address Stack

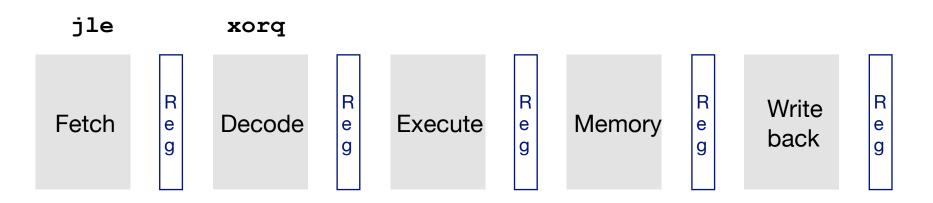
- Stall: the pipeline register shouldn't be written
- Bubble : signals correspond to a nop
- Why is it good for the hardware to do so anyways?



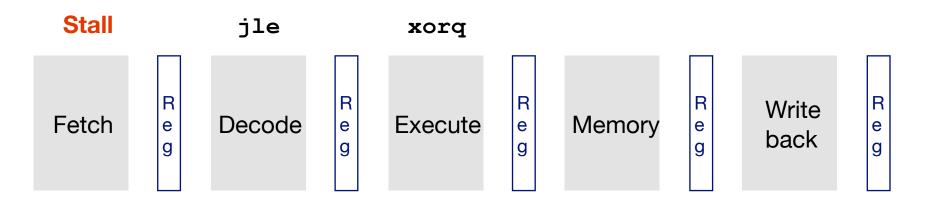
- Stall: the pipeline register shouldn't be written
- Bubble: signals correspond to a nop
- Why is it good for the hardware to do so anyways?



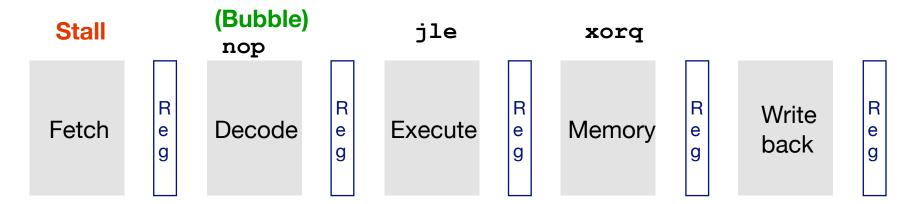
- Stall: the pipeline register shouldn't be written
- Bubble : signals correspond to a nop
- Why is it good for the hardware to do so anyways?



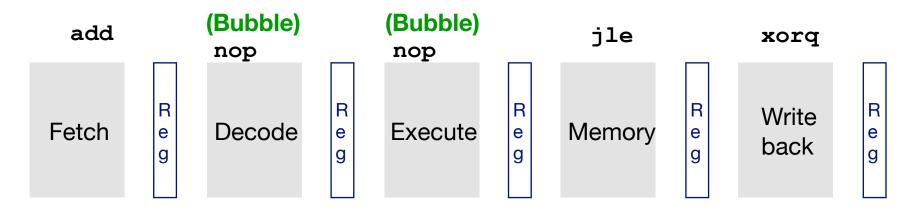
- Stall: the pipeline register shouldn't be written
- Bubble : signals correspond to a nop
- Why is it good for the hardware to do so anyways?



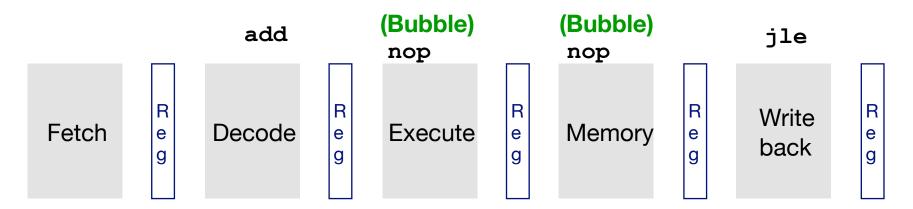
- Stall: the pipeline register shouldn't be written
- Bubble : signals correspond to a nop
- Why is it good for the hardware to do so anyways?



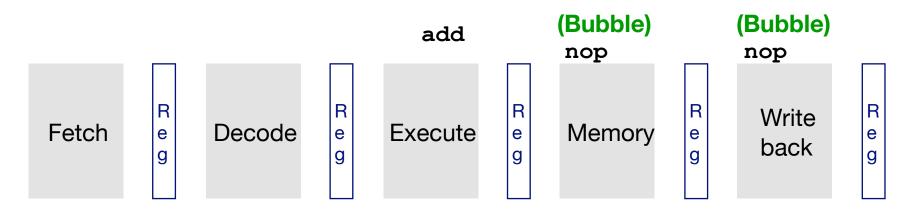
- Stall: the pipeline register shouldn't be written
- Bubble : signals correspond to a nop
- Why is it good for the hardware to do so anyways?



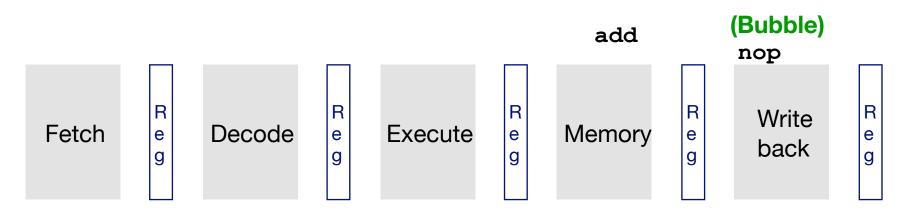
- Stall: the pipeline register shouldn't be written
- Bubble: signals correspond to a nop
- Why is it good for the hardware to do so anyways?



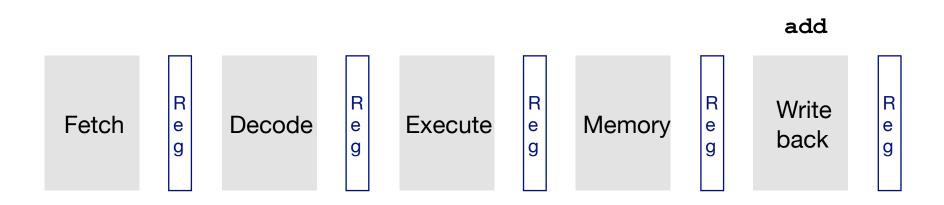
- Stall: the pipeline register shouldn't be written
- Bubble: signals correspond to a nop
- Why is it good for the hardware to do so anyways?

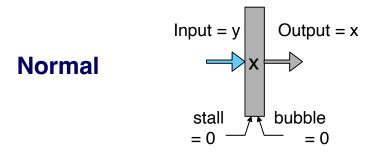


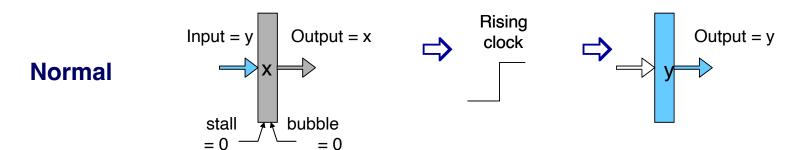
- Stall: the pipeline register shouldn't be written
- Bubble: signals correspond to a nop
- Why is it good for the hardware to do so anyways?

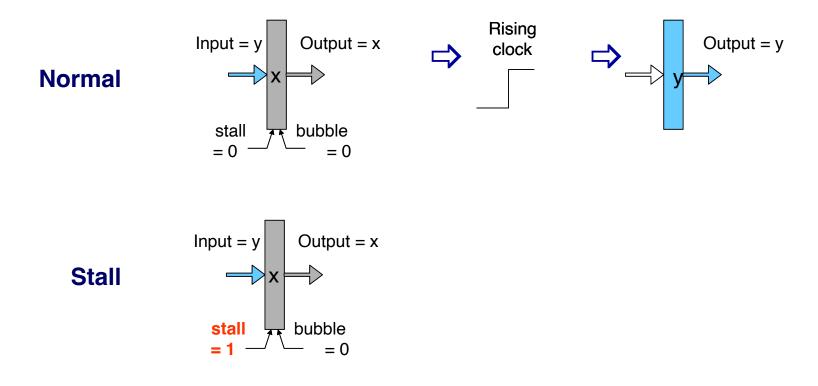


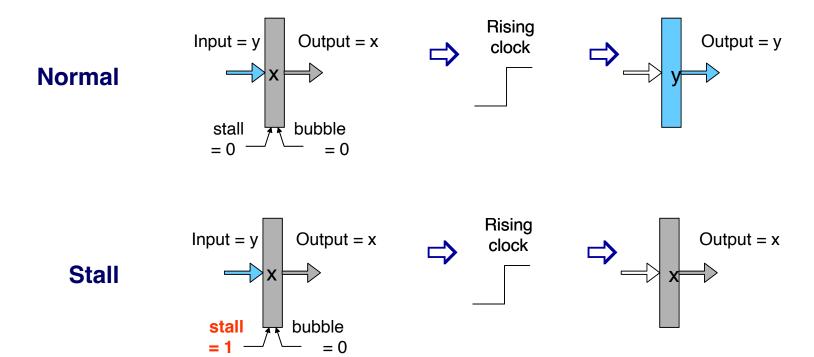
- Stall: the pipeline register shouldn't be written
- Bubble : signals correspond to a nop
- Why is it good for the hardware to do so anyways?

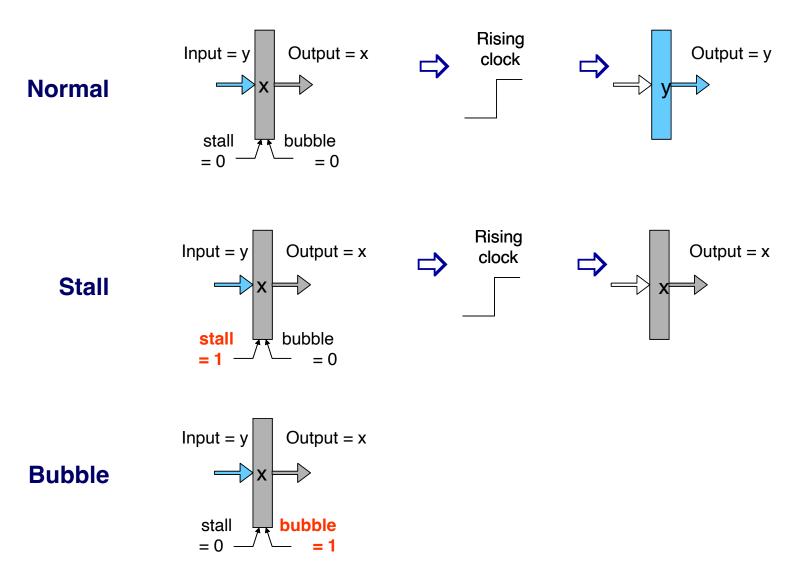


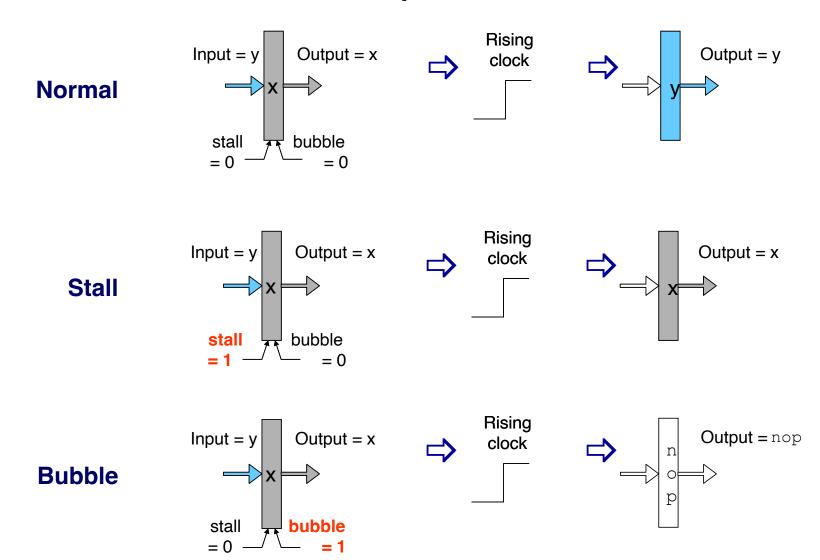




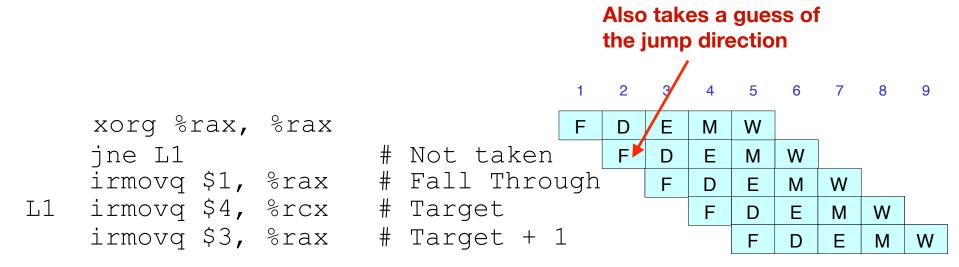






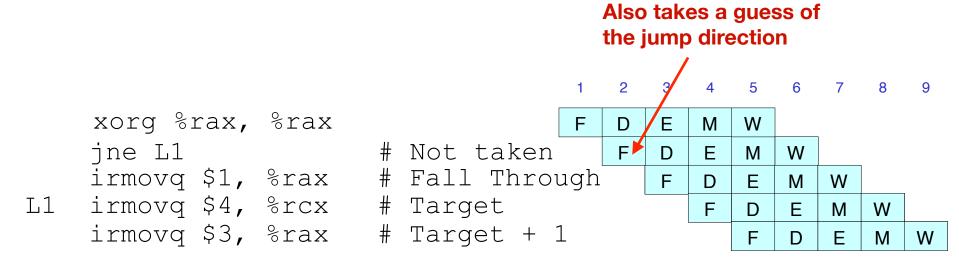


Idea: instead of waiting, why not just guess the direction of jump?



Idea: instead of waiting, why not just guess the direction of jump?

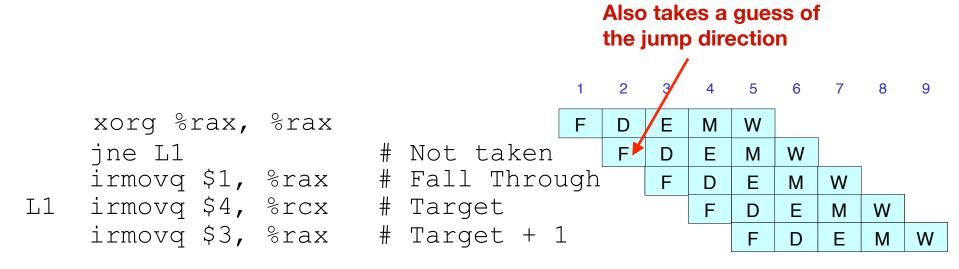
If prediction is correct: pipeline moves forward without stalling



Idea: instead of waiting, why not just guess the direction of jump?

If prediction is correct: pipeline moves forward without stalling

If mispredicted: kill mis-executed instructions, start from the correct target



Idea: instead of waiting, why not just guess the direction of jump?

If prediction is correct: pipeline moves forward without stalling

If mispredicted: kill mis-executed instructions, start from the correct target

#### Static Prediction

- Always Taken
- Always Not-taken

### **Dynamic Prediction**

Dynamically predict taken/not-taken for each specific jump instruction

- People use jumps to check corner cases. These branches are mostly not taken because corner cases are rare.
- People use jumps to implement loops. These branches are mostly taken because a loop takes multiple iterations.

- People use jumps to check corner cases. These branches are mostly not taken because corner cases are rare.
- People use jumps to implement loops. These branches are mostly taken because a loop takes multiple iterations.

```
cmpq %rsi,%rdi
  jle .corner_case
  <do_A>
.corner_case:
    <do_B>
    ret
```

- People use jumps to check corner cases. These branches are mostly not taken because corner cases are rare.
- People use jumps to implement loops. These branches are mostly taken because a loop takes multiple iterations.

- People use jumps to check corner cases. These branches are mostly not taken because corner cases are rare.
- People use jumps to implement loops. These branches are mostly taken because a loop takes multiple iterations.

- People use jumps to check corner cases. These branches are mostly not taken because corner cases are rare.
- People use jumps to implement loops. These branches are mostly taken because a loop takes multiple iterations.

### Observation (Assumption really): Two uses of jumps

- People use jumps to check corner cases. These branches are mostly not taken because corner cases are rare.
- People use jumps to implement loops. These branches are mostly taken because a loop takes multiple iterations.

### Strategy:

- Forward jumps (i.e., if-else): always predict not-taken
- Backward jumps (i.e., loop): always predict taken

#### **Static Prediction**

Knowing branch prediction strategy helps us write faster code

- Any difference between the following two code snippets?
- What if you know that hardware uses the always non-taken branch prediction?

```
if (cond) {
   do_A()
   do_B()
} else {
   do_B()
   do_A()
}
```

- Simplest idea:
  - If last time taken, predict taken; if last time not-taken, predict not-taken
  - Called 1-bit branch predictor
  - Works nicely for loops

- Simplest idea:
  - If last time taken, predict taken; if last time not-taken, predict not-taken
  - Called 1-bit branch predictor
  - Works nicely for loops

```
for (i=0; i < 5; i++) {...}
```

#### Simplest idea:

- If last time taken, predict taken; if last time not-taken, predict not-taken
- Called 1-bit branch predictor
- Works nicely for loops

for 
$$(i=0; i < 5; i++) {...}$$

Iteration #1	0	1	2	3	4
Predicted Outcome	N	Т	Т	Т	Т
Actual Outcome	Т	Т	Т	Т	N

#### Simplest idea:

- If last time taken, predict taken; if last time not-taken, predict not-taken
- Called 1-bit branch predictor
- Works nicely for loops

for 
$$(i=0; i < 5; i++) {...}$$

Iteration #1	0	1	2	3	4
Predicted Outcome	N	Ţ	Т	Т	Т
Actual Outcome	Т	Т	Т	Т	N

- With 1-bit prediction, we change our mind instantly if mispredict
- Might be too quick. Thus 2-bit branch prediction: we have to mispredict twice in a row before changing our mind

- With 1-bit prediction, we change our mind instantly if mispredict
- Might be too quick. Thus 2-bit branch prediction: we have to mispredict twice in a row before changing our mind

for 
$$(i=0; i < 5; i++) {...}$$

Predict with 1-bit	N	Т	Т	Т	Т
Actual Outcome	Т	Т	Т	Т	N
Predict with 2-bit	N	Z	Т	Т	Т

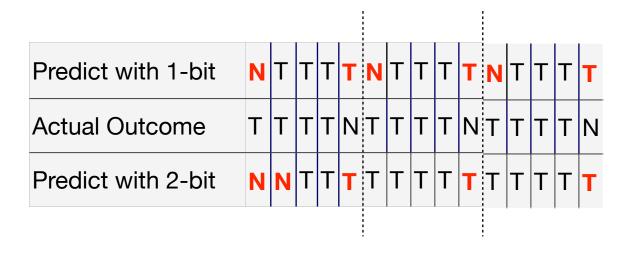
- With 1-bit prediction, we change our mind instantly if mispredict
- Might be too quick. Thus 2-bit branch prediction: we have to mispredict twice in a row before changing our mind

for 
$$(i=0; i < 5; i++) {...}$$

						: : :				
Predict with 1-bit	N	Т	Т	Т	Т	N	Т	Т	Т	Т
Actual Outcome	Т	Т	Т	Т	N	Т	Т	Т	Т	N
Predict with 2-bit	N	N	Т	Т	т	Т	Т	Т	Т	Т
		•	•							

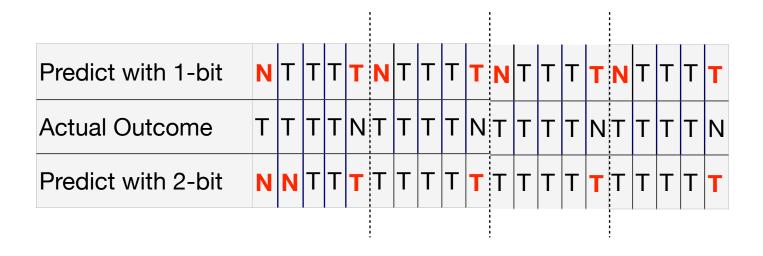
- With 1-bit prediction, we change our mind instantly if mispredict
- Might be too quick. Thus 2-bit branch prediction: we have to mispredict twice in a row before changing our mind





- With 1-bit prediction, we change our mind instantly if mispredict
- Might be too quick. Thus 2-bit branch prediction: we have to mispredict twice in a row before changing our mind

for 
$$(i=0; i < 5; i++) {...}$$



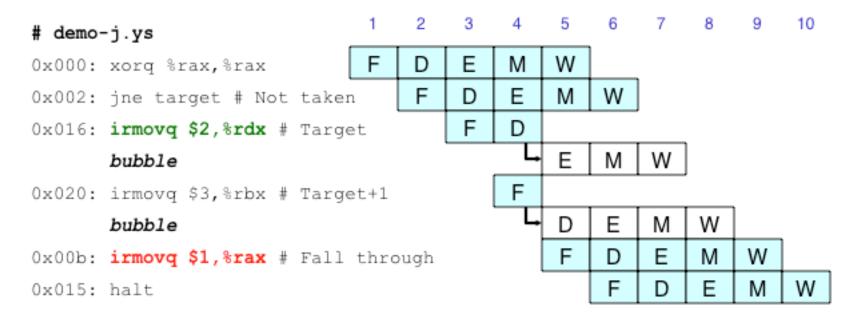
## More Advanced Dynamic Prediction

- Look for past histories across instructions
- Branches are often correlated
  - Direction of one branch determines another

cond1 branch nottaken means (x <=0) branch taken

$$x = 0$$
  
if (cond1)  $x = 3$   
if (cond2)  $y = 19$   
if (x <= 0)  $z = 13$ 

## What Happens If We Mispredict?



#### Cancel instructions when mispredicted

- Assuming we detect branch not-taken in execute stage
- On following cycle, replace instructions in execute and decode by **bubbles**
- No side effects have occurred yet

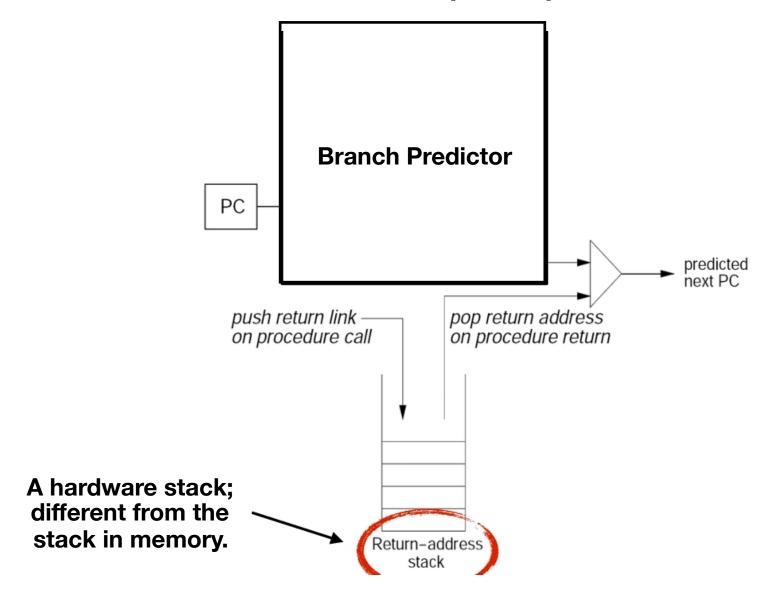
#### Stalling for Return

- As ret passes through pipeline, stall at fetch stage
  - While in decode, execute, and memory stage
- Inject bubble into decode stage
- Release stall when reach write-back stage

## Return Address Stack (RAS)

- Stalling for return is silly since we know where exactly we need to jump to, except the jump target is retrieved later in the memory stage.
- Can we get that sooner? Where should we get it?

### Return Address Stack (RAS)



# Today: Making the Pipeline Really Work

- Control Dependencies
  - Inserting Nops
  - Stalling
  - Delay Slots
  - Branch Prediction
- Data Dependencies
  - Inserting Nops
  - Stalling
  - Out-of-order execution

```
1 irmovq $50, %rax
2 addq %rax, %rbx
3 mrmovq 100(%rbx), %rdx
```

```
1 irmovq $50, %rax
2 addq %rax, %rbx
3 mrmovq 100(%rbx), %rdx
```

```
1 irmovq $50, %rax
2 addq %rax, %rbx
3 mrmovq 100(%rbx), %rdx
```

```
1 irmovq $50, %rax
2 addq %rax, %rbx
3 mrmovq 100(%rbx), %rdx
```

- Result from one instruction used as operand for another
  - Read-after-write (RAW) dependency
- Very common in actual programs
- Must make sure our pipeline handles these properly
  - Get correct results
  - Minimize performance impact

#### A Subtle Data Dependency

- Jump instruction example below:
  - jne L1 determines whether irmovq \$1, %rax should be executed
  - But jne doesn't know its outcome until after its Execute stage.
     Why?

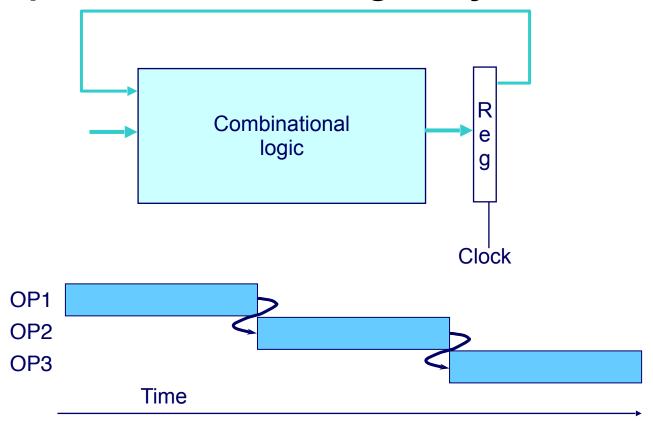
```
xorg %rax, %rax
jne L1  # Not taken
irmovq $1, %rax  # Fall Through
L1 irmovq $4, %rcx  # Target
irmovq $3, %rax  # Target + 1
```

#### A Subtle Data Dependency

- Jump instruction example below:
  - jne L1 determines whether irmovq \$1, %rax should be executed
  - But jne doesn't know its outcome until after its Execute stage.
     Why?
- There is a data dependency between xorg and jne. The "data" is the status flags.

```
xorg %rax, %rax
jne L1  # Not taken
irmovq $1, %rax  # Fall Through
L1 irmovq $4, %rcx  # Target
irmovq $3, %rax  # Target + 1
```

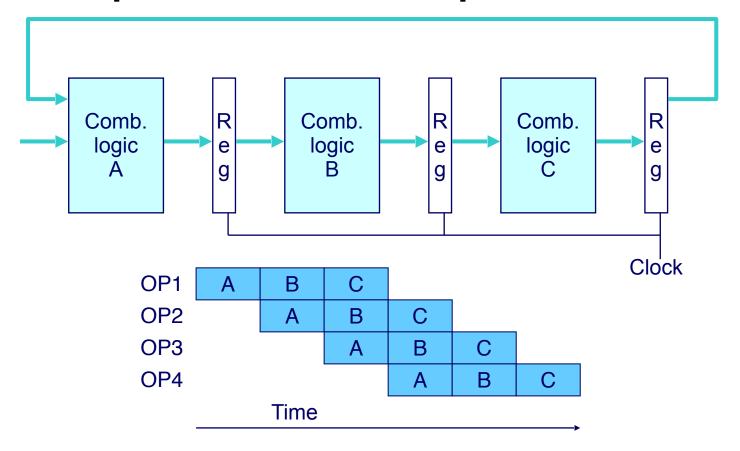
#### Data Dependencies in Single-Cycle Machines



#### In Single-Cycle Implementation:

Each operation starts only after the previous operation finishes.
 Dependency always satisfied.

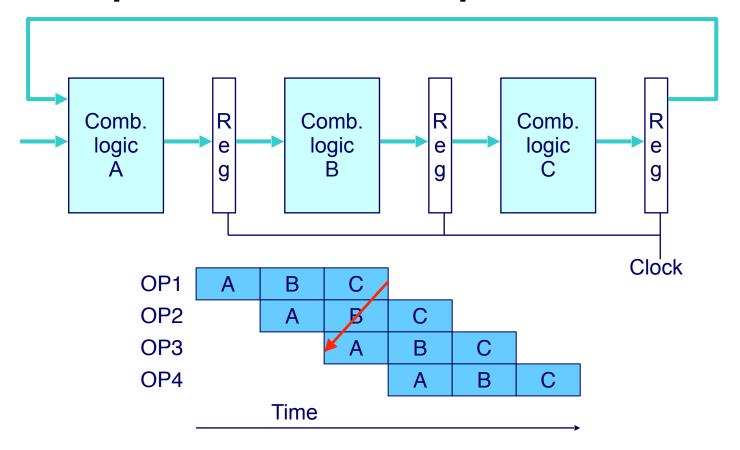
#### Data Dependencies in Pipeline Machines



Data Hazards happen when:

Result does not feed back around in time for next operation

#### Data Dependencies in Pipeline Machines



Data Hazards happen when:

Result does not feed back around in time for next operation

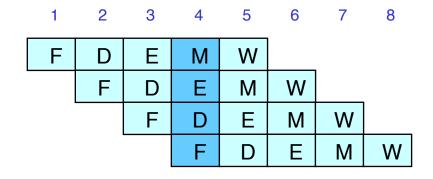
#### Data Dependencies: No Nop

0x000: irmovq \$10,%rdx

0x00a: irmovq \$3,%rax

0x014: addq %rdx,%rax

0x016: halt



Remember registers get updated in the Write-back stage

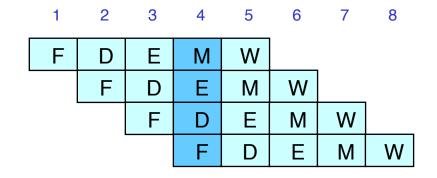
#### Data Dependencies: No Nop

0x000: irmovq \$10,%rdx

0x00a: irmovq \$3,%rax

0x014: addq %rdx,%rax

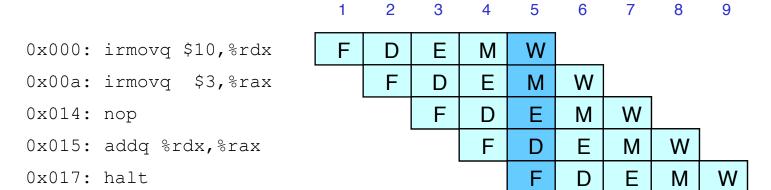
0x016: halt



Remember registers get updated in the Write-back stage

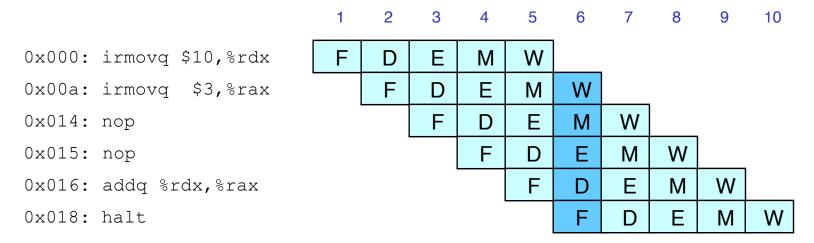
addq reads wrong %rdx and %rax

#### Data Dependencies: 1 Nop



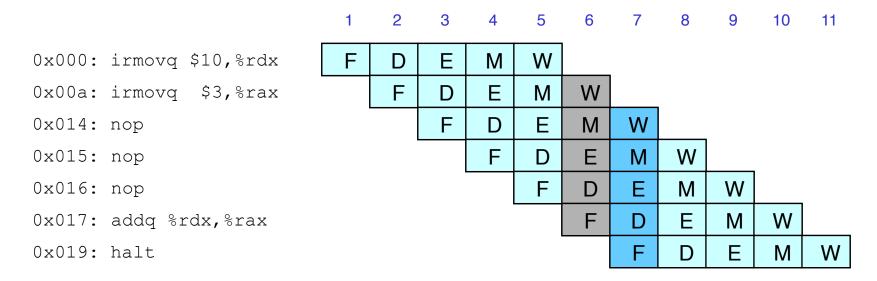
addq still reads wrong %rdx and %rax

#### Data Dependencies: 2 Nop's



addq reads the correct %rdx, but %rax still wrong

#### Data Dependencies: 3 Nop's



addq reads the correct %rdx and %rax

Can we have the hardware automatically generates a nop?

Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g
-------	-------------	--------	-------------	---------	-------------	--------	-------------	---------------	-------------

Can we have the hardware automatically generates a nop?

Why is it good for the hardware to do so anyways?

#### irmov

Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g
-------	-------------	--------	-------------	---------	-------------	--------	-------------	---------------	-------------

Can we have the hardware automatically generates a nop?

irmov		irmov							
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g

Can we have the hardware automatically generates a nop?

add		irmov		irmov					
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g

Can we have the hardware automatically generates a nop?

Inst3		add		irmov		irmov			
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g

Can we have the hardware automatically generates a nop?

Inst3		Stall add		irmov		irmov			
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g

Can we have the hardware automatically generates a nop?

Inst3		Stall add		bubble (nop)		irmov		irmov		
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g	

Can we have the hardware automatically generates a nop?

Stall Inst3			bubble (nop)			irmov		irmov		
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g	

Can we have the hardware automatically generates a nop?

Stall Inst3				bubble (nop)		bubble (nop)		irmov		
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g	

Can we have the hardware automatically generates a nop?

Inst4	Inst3			add		bubble (nop)		bubble (nop)	
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g

Can we have the hardware automatically generates a nop?

		Inst4		Inst3		add		bubble (nop)	
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g

Can we have the hardware automatically generates a nop?

				Inst4		Inst3		add	
Fetch	R e g	Decode	R e g	Execute	R e g	Memory	R e g	Write back	R e g

## **Detecting Stall Condition**

- Using a "scoreboard". Each register has a bit.
- Every instruction that writes to a register sets the bit.
- Every instruction that reads a register would have to check the bit first.
  - If the bit is set, then generate a bubble
  - · Otherwise, free to go!!

## **Data Forwarding**

#### Naïve Pipeline

- Register isn't written until completion of write-back stage
- Source operands read from register file in decode stage
- The decode stage can't start until the write-back stage finishes

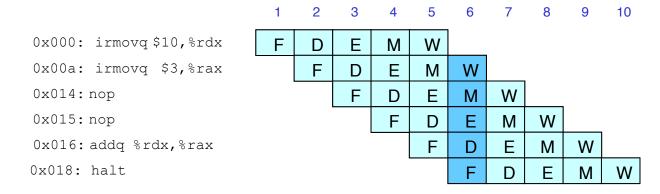
#### Observation

Value generated in execute or memory stage

#### Trick

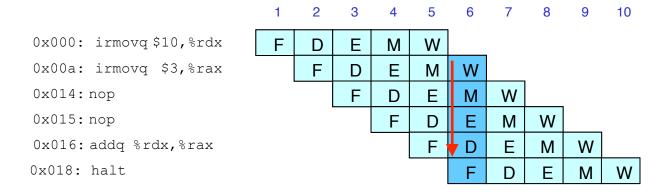
- Pass value directly from generating instruction to decode stage
- Needs to be available at end of decode stage

## Data Forwarding Example



- irmovq writes %rax to the register file at the end of the write-back stage
- But the value of %rax is already available at the beginning of the writeback stage
- Forward %rax to the decode stage of addq.

## Data Forwarding Example



- irmovq writes %rax to the register file at the end of the write-back stage
- But the value of %rax is already available at the beginning of the writeback stage
- Forward %rax to the decode stage of addq.

## Data Forwarding Example #2

0x000: irmovq \$10,%rdx

0x00a: irmovq \$3,%rax

0x014: addq %rdx, %rax

0x016: halt

1	2	3	4	5	О	/	8	
F	D	Е	М	W		_		
	H	D	Ш	М	W		_	
		L	D	Ш	М	W		
			F	D	Е	М	W	

#### Register %rdx

Forward from the memory stage

#### Register %rax

Forward from the execute stage

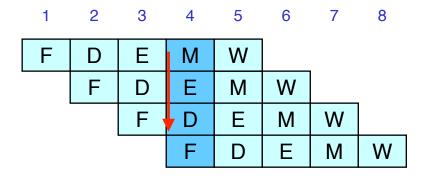
## Data Forwarding Example #2

0x000: irmovq \$10,%rdx

0x00a: irmovq \$3,%rax

0x014: addq %rdx,%rax

0x016: halt



#### Register %rdx

Forward from the memory stage

#### Register %rax

Forward from the execute stage

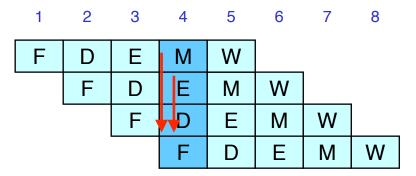
## Data Forwarding Example #2

0x000: irmovq \$10,%rdx

0x00a: irmovq \$3,%rax

0x014: addq %rdx,%rax

0x016: halt



#### Register %rdx

Forward from the memory stage

#### Register %rax

Forward from the execute stage

## **Bypass Paths**

Memory

#### **Decode Stage**

- Forwarding logic selects valA
- Normally from register file
- Forwarding: get valA or valB fr

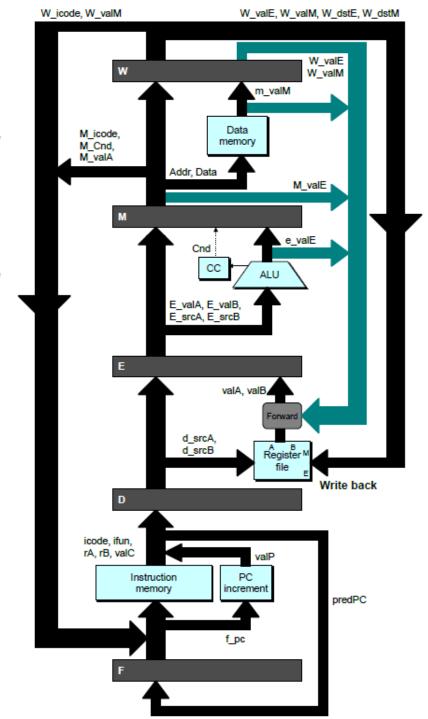
#### Forwarding Sources

- Execute: valE
- · Memory: valE, valM
- Write back: valE, valM

Decode

Fetch

PC



- Compiler could do this, but has limitations
- Generally done in hardware

# Long-latency instruction. Forces the pipeline to stall.

```
r0 = r1 + r2
r3 = MEM[r0]
r4 = r3 + r6
r7 = r5 + r1
...
r4 = r3 + r6
r4 = r5 + r1
r7 = r5 + r1
```

- Compiler could do this, but has limitations
- Generally done in hardware

# Long-latency instruction. Forces the pipeline to stall.

```
r0 = r1 + r2
r3 = MEM[r0]
r4 = r3 + r6
r6 = r5 + r1
```

$$r0 = r1 + r2$$
 Is this correct?  $r0 = r1 + r2$   
 $r3 = MEM[r0]$   
 $r4 = r3 + r6$   
 $r6 = r5 + r1$   
...
 $r4 = r3 + r6$ 





$$r4 = r3 + r6$$

"Tomasolu Algorithm" is the algorithm that is most widely implemented in modern hardware to get out-of-order execution right.