CSC 252: Computer Organization Spring 2022: Lecture 14

Instructor: Yuhao Zhu

Department of Computer Science University of Rochester

- Programming assignment 3 is out
 - Details: <u>https://www.cs.rochester.edu/courses/252/spring2022/labs/</u> assignment3.html
 - Due on March 3, 11:59 PM
 - You (may still) have 3 slip days

13	14	15	16	17	18	19	
20	21	22	23	24	25	26	
20	L 1	L.L.	10	24	20	20	
27	28	Mar 1	2	3	4	5	
				Due			
		Today					
				wild-term			
						2	-

- Grades for Lab 2 are posted.
- Programming assignment 3 is in x86 assembly language. Seek help from TAs.
- TAs are best positioned to answer your questions about programming assignments!!!
- Programming assignments do NOT repeat the lecture materials. They ask you to synthesize what you have learned from the lectures and work out something new.
- Mid-term exam: March 3; online.
- Past exam & Problem set: <u>https://www.cs.rochester.edu/courses/</u> <u>252/spring2022/handouts.html</u>
- Exam will be electronic using Gradescope.

- Open book test: any sort of paper-based product, e.g., book, notes, magazine, old tests.
- Exams are designed to test your ability to apply what you have learned and not your memory (though a good memory could help).
- Nothing electronic (including laptop, cell phone, calculator, etc) other than the computer you use to take the exam.
- Nothing biological, including your roommate, husband, wife, your hamster, another professor, etc.
- "I don't know" gets15% partial credit. Must erase everything else.

Control Dependency

- **Definition**: Outcome of instruction A determines whether or not instruction B should be executed or not.
- 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?



- 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()
}
```

- 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

- 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

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

Iteration #1	0	1	2	3	4
Predicted Outcome	N	т	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

Predict with 1-bit	Ν	Т	Т	Т	Т
Actual Outcome	Т	Т	Т	Т	Ν
Predict with 2-bit	Ν	Ν	Т	Т	T

- 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

	-				-		-	-		
Predict with 1-bit	Ν	т	Т	Т	т	N	Т	Т	Т	Т
Actual Outcome	Т	Т	Т	Т	N	Т	Т	Т	Т	Ν
Predict with 2-bit	Ν	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

													_		_
Predict with 1-bit	N	Т	Т	Т	Т	N	Т	Т	Т	Т	Ν	Т	Т	Т	Т
Actual Outcome	Т	Т	Т	Т	N	Т	Т	Т	Т	N	Т	Т	т	Т	Ν
Predict with 2-bit	N	Ν	Т	Т	т	Т	Т	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

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

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</pre>

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

Return Instruction

```
0x000:
0x00a:
        call p
0x013:
        irmovq $5,%rsi
0x01d:
        halt
0x020: .pos 0x20
0x020: p: irmovq $-1,%rdi
0x02a:
        ret
0x02b:
        irmovq $1,%rax
0x035:
        irmovq $2,%rcx
0x03f:
        irmovq $3,%rdx
        irmovq $4,%rbx
0x049:
0x100: .pos 0x100
0x100: Stack:
```

procedure

- # Should not be executed

```
# Stack: Stack pointer
```

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

- Inserting Nops
- Stalling
- Out-of-order execution

- 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?



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.



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



- 0x00a: irmovq \$3,%rax
- 0x014: nop
- 0x015: addq %rdx,%rax
- 0x017: halt



addq still reads wrong %rdx and %rax

Data Dependencies: 2 Nop's



- 0x00a: irmovq \$3,%rax
- 0x014: nop
- 0x015: nop
- 0x016: addq %rdx,%rax
- 0x018: halt



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

Data Dependencies: 3 Nop's



- 0x00a: irmovq \$3,%rax
- 0x014: nop
- 0x015: nop
- 0x016: nop
- 0x017: addq %rdx,%rax
- 0x019: halt



addq reads the correct %rdx and %rax

Resolving Data Dependencies

- Software Mechanisms
 - Adding NOPs: requires compiler to insert nops, which also take memory space — not a good idea
- Hardware mechanisms
 - Stalling
 - Forwarding
 - Out-of-order execution






























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!!

Detecting Stall Condition



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



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



Limitation of Forwarding



Load-use dependency

- Value needed by end of decode stage in cycle 7
- Value read from memory in memory stage of cycle 8



Avoiding Load/Use Hazard



M_dstM = %rax

m_valM ← M[128] = 3

D

 $valA \leftarrow W_valE = 10$ $valB \leftarrow m valM = 3$

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

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



r0	=	r1	+	r2
r3	=	MEN	4[]	:0]
r7	=	r5	+	r1
		•••		
r4	=	r3	+	r6

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

Long-latency instruction. Forces the pipeline to stall.



r0	=	r1	+	r2
r3	=	MEN	4[]	:0]
r7	=	r5	+	r1
		•••		
r4	=	r3	+	r6

...

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

...



r0 = r1 + r2r3 = MEM[r0]r6 = r5 + r1...

r4 = r3 + **r6**

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

...



r0 = r1 + r2r3 = MEM[r0]**r6** = r5 + r1 ... r4 = r3 + **r6**

...

41

...



r4 = r3 + r6



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