Revisiting Branch Hazard Solutions

- Stall
- Predict Not Taken
- Predict Taken
- Branch Delay Slot

Predict Not Taken

Branch | IF | ID | EX | MEM | WB
---|---|---|---|----|---
I+1 | IF | ID | EX | MEM | WB
I+2 | IF | ID | EX | MEM | WB
I+3 | IF | ID | EX | MEM | WB

Branch | IF | ID | EX | MEM | WB
---|---|---|---|----|---
I+1 | IF (bubble) (bubble) (bubble) (bubble)
I+2 | IF | ID | EX | MEM | WB
Branch Target | IF | ID | EX | MEM | WB
T+1 | IF | ID | EX | MEM | WB

Delayed Branch

Branch | IF | ID | EX | MEM | WB
---|---|---|---|----|---
I+1 (delay slot) | IF | ID | EX | MEM | WB
I+2 | IF | ID | EX | MEM | WB
I+3 | IF | ID | EX | MEM | WB

Branch | IF | ID | EX | MEM | WB
---|---|---|---|----|---
I+1 (delay slot) | IF | ID | EX | MEM | WB
Branch Target | IF | ID | EX | MEM | WB
T+1 | IF | ID | EX | MEM | WB

Filling the delay slot (e.g., in the compiler)

Can be done when?
Improves performance when?

lw R1, 10000(R7)
add R5, R6, R1
beq R5, label:
sub R8, R1, R3
add R4, R8, R9
and R2, R4, R8

label: add R2, R5, R8
Problems filling delay slot

1. need to predict __________ of branch to be most effective
2. limited by ______________ restriction

Branch Likely

- Branch likely
- I+1 (delay slot)
- I+2
- I+3

Branch Target

- T+1

Delay Slot Utilization

- 18% of delay slots left empty
- 11% of delay slots (1) use canceling branches and (2) end up getting canceled
Branch Performance

CPI = BCPI + pipeline stalls from branches per instruction
= 1.0 + branch frequency * branch penalty
assume 20% branches, 67% taken:

<table>
<thead>
<tr>
<th>branch</th>
<th>taken</th>
<th>not taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>scheme</td>
<td>penalty</td>
<td>penalty</td>
</tr>
<tr>
<td>stall</td>
<td>predict taken</td>
<td>predict not taken</td>
</tr>
</tbody>
</table>

Delay Slots, the scorecard

- Pros
- Cons

Static Branch Prediction

- Static branch prediction takes place at compile time, dynamic branch prediction during program execution
- static bp done by software, dynamic bp done in hardware
- How to make static branch predictions?
- Static branch prediction enables
  - more effective code scheduling around hazards (how?)
  - more effective use of delay slots

MIPS Integer Pipeline Performance

- Only stalls for load hazards and branch hazards, both of which can be reduced (but not eliminated) by software
But now, the real world interrupts...

- Pipelining is not as easy as we have made it seem so far...
  - interrupts and exceptions
  - long-latency instructions

Exceptions and Interrupts

- Transfer of control flow (to an exception handler) without an explicit branch or jump
- are often unpredictable
- examples
  - I/O device request
  - OS system call
  - arithmetic overflow/underflow
  - FP error
  - page fault
  - memory-protection violation
  - hardware error
  - undefined instruction

Classes of Exceptions

- synchronous vs. asynchronous
- user-initiated vs. coerced
- user maskable vs. nonmaskable
- within instruction vs. between instructions
- resume vs. terminate

when the pipeline can be stopped just before the faulting instruction, and can be restarted from there (if necessary), the pipeline supports precise exceptions

Basic Exception Methodology

- turn off writes for faulting instruction and following
- force a trap into the pipeline at the next IF
- save the PC of the faulting instruction (not quite enough for delayed branches)
Exceptions Can Occur In Several Places in the pipeline

- IF -- page fault on memory access, misaligned memory access, memory-protection violation
- ID -- illegal opcode
- EX -- arithmetic exception
- MEM -- page fault, misaligned access, memory-protection violation
- WB -- none

(and, of course, asynchronous can happen anytime)

Simplifying Exceptions in the ISA

1. Each instruction changes machine state only once
   1. autoincrement
   2. string operations
   3. condition codes
2. Each instruction changes machine state at the end of the pipeline (when you know it will not cause an exception)

But now, the real world interrupts...

- Pipelining is not as easy as we have made it seem so far...
  - interrupts and exceptions
  - long-latency instructions

Handling Multicycle Operations

- Unrealistic to expect that all operations take the same amount of time to execute
- ___, some _____________ will take longer
- This violates some of the assumptions of our simple pipeline
New problems

- structural hazards
  - divide unit
  - WB stage
- WAW hazards are possible
- out-of-order completion
- WAR hazards still not possible

Hazard Detection in the ID stage

- An instruction can only issue (proceed past the ID stage) when:
  - there are no structural hazards (divide unit is free, WB port will be free when needed)
  - no RAW data hazards (that forwarding can’t handle)
  - no WAW hazards with instructions in long pipes

structural hazards and WAW hazards

- structural hazards
  - divide unit
  - WB stage

<table>
<thead>
<tr>
<th>FU</th>
<th>Latency</th>
<th>Initiation interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integer</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Memory</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>FP add</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>FP multiply</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>FP divide</td>
<td>24</td>
<td>24</td>
</tr>
</tbody>
</table>

- WAW hazards

<table>
<thead>
<tr>
<th>ADDD F8, ...</th>
<th>IF</th>
<th>ID</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>MEM</th>
<th>WB</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD F8, ...</td>
<td>IF</td>
<td>ID</td>
<td>EX</td>
<td>MEM</td>
<td>WB</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Key Points

- Data Hazards can be significantly reduced by forwarding
- Branch hazards can be reduced by early computation of condition and target, branch delay slots, branch prediction
- Data hazard and branch hazard reduction require complex compiler support
- Exceptions are hard, precise exceptions are really hard
- Variable-length instructions introduce structural hazards, WAW hazards, more RAW hazards