Computer Architecture
ELE 475 / COS 475
Slide Deck 8A: Branch Prediction

David Wentzlaff
Department of Electrical Engineering
Princeton University
Agenda

• Branch Cost Motivation

• Branch Prediction
  – Outcome
    • Static
    • Dynamic
  – Target Address
Longer Frontends Means More Control
Flow Penalty

Penalty includes instructions in IQ
Longer Pipeline Frontends Amplify Branch Cost

<table>
<thead>
<tr>
<th>Basic Pentium III Processor Misprediction Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Fetch</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Basic Pentium 4 Processor Misprediction Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 TC Nxt IP</td>
</tr>
</tbody>
</table>

Pentium 3: 10 cycle branch penalty
Pentium 4: 20 cycle branch penalty

Dual Issue and Branch Cost
### Superscalars Multiply Branch Cost

<table>
<thead>
<tr>
<th>Operation</th>
<th>Format</th>
<th>Destination</th>
<th>Instruction</th>
<th>A0</th>
<th>A1</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEQZ</td>
<td>F</td>
<td>D</td>
<td>I</td>
<td>A0</td>
<td>A1</td>
<td>W</td>
</tr>
<tr>
<td>OpA</td>
<td>F</td>
<td>D</td>
<td>I</td>
<td>B0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OpB</td>
<td>F</td>
<td>D</td>
<td>I</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OpC</td>
<td>F</td>
<td>D</td>
<td>I</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OpD</td>
<td>F</td>
<td>D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OpE</td>
<td>F</td>
<td>D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OpF</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OpG</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OpH</td>
<td>F</td>
<td>D</td>
<td>I</td>
<td>A0</td>
<td>A1</td>
<td>W</td>
</tr>
<tr>
<td>OpI</td>
<td>F</td>
<td>D</td>
<td>I</td>
<td>B0</td>
<td>B1</td>
<td>W</td>
</tr>
</tbody>
</table>

How much work is lost if pipeline doesn’t follow correct instruction flow?

\[ \sim \text{pipeline width} \times \text{branch penalty} \]
Agenda

• Branch Cost Motivation

• Branch Prediction
  – Outcome
    • Static
    • Dynamic
  – Target Address
Branch Prediction

• Essential in modern processors to mitigate branch delay latencies

Two types of Prediction
1. Predict Branch Outcome
2. Predict Branch/Jump Address
Where is the Branch Information Known?

Know branch outcome
Know target address for JR and JALR

Know target address for branches, J, and JAL
Agenda

• Branch Cost Motivation
• Branch Prediction
  – Outcome
    • Static
    • Dynamic
  – Target Address
Branch Delay Slots
(expose control hazard to software)

- Change the ISA semantics so that the instruction that follows a jump or branch is always executed
  - gives compiler the flexibility to put in a useful instruction where normally a pipeline bubble would have resulted.

| I_1  | 096 | ADD |
| I_2  | 100 | BEQZ r1 +200 |
| I_3  | 104 | ADD |
| I_4  | 108 | ADD |
| I_5  | 304 | ADD |

Delay slot instructions executed regardless of branch outcome
Static Branch Prediction

Overall probability a branch is taken is ~60-70% but:

- **backward**: 90%
- **forward**: 50%
Static Software Branch Prediction

• Extend ISA to enable compiler to tell microarchitecture if branch is likely to be taken or not (Can be up to 80% accurate)

<table>
<thead>
<tr>
<th>BR.T</th>
<th>F</th>
<th>D</th>
<th>X</th>
<th>M</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpA</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Targ</td>
<td>F</td>
<td>D</td>
<td>X</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>BR.NT</td>
<td>F</td>
<td>D</td>
<td>X</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>OpB</td>
<td>F</td>
<td>D</td>
<td>X</td>
<td>M</td>
<td>W</td>
</tr>
<tr>
<td>OpC</td>
<td>F</td>
<td>D</td>
<td>X</td>
<td>M</td>
<td>W</td>
</tr>
</tbody>
</table>

What if hint is wrong?

<table>
<thead>
<tr>
<th>BR.T</th>
<th>F</th>
<th>D</th>
<th>X</th>
<th>M</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpA</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Targ</td>
<td>F</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OpA</td>
<td>F</td>
<td>D</td>
<td>X</td>
<td>M</td>
<td>W</td>
</tr>
</tbody>
</table>
Static Hardware Branch Prediction

1. Always Predict Not-Taken
   – What we have been assuming
   – Simple to implement
   – Know fall-through PC in Fetch
   – Poor Accuracy, especially on backward branches

2. Always Predict Taken
   – Difficult to implement because don’t know target until Decode
   – Poor accuracy on if-then-else

3. Backward Branch Taken, Forward Branch Not Taken
   – Better Accuracy
   – Difficult to implement because don’t know target until Decode
Agenda

• Branch Cost Motivation

• Branch Prediction
  – Outcome
    • Static
    • Dynamic
  – Target Address
Dynamic Hardware Branch Prediction: Exploiting Temporal Correlation

- Exploit structure in program: The way a branch resolves may be a good indicator of the way it will resolve the next time it executes (Temporal Correlation)

1-bit Saturating Counter

![Diagram of 1-bit Saturating Counter]
### 1-bit Saturating Counter

For Backward branch in loop
- Assume 4 Iterations
- Assume is executed multiple times

Always 2 Mispredicts

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Prediction</th>
<th>Actual</th>
<th>Mispredict?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NT</td>
<td>T</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>NT</td>
<td>Y</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>NT</td>
<td>T</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>NT</td>
<td>Y</td>
</tr>
</tbody>
</table>
### 2-bit Saturating Counter

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Prediction</th>
<th>Actual</th>
<th>Mispredict?</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NT</td>
<td>T</td>
<td>Y</td>
<td>Strong NT</td>
</tr>
<tr>
<td>2</td>
<td>NT</td>
<td>T</td>
<td>Y</td>
<td>Weak NT</td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>T</td>
<td></td>
<td>Weak T</td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>NT</td>
<td>Y</td>
<td>Strong T</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>T</td>
<td>T</td>
<td></td>
<td>Weak T</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>T</td>
<td></td>
<td>Strong T</td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>T</td>
<td></td>
<td>Strong T</td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>NT</td>
<td>Y</td>
<td>Strong T</td>
</tr>
</tbody>
</table>

Only 1 Mispredict at end of loop
Other 2-bit FSM Branch Predictors

Jump directly to strong from weak

Predict T

Predict NT

Predict T

Predict NT

Weak Taken

Strong Taken

Weak Not Taken

Strong Not Taken
Branch History Table (BHT)

4K-entry BHT, 2 bits/entry, ~80-90% correct predictions
Exploiting Spatial Correlation

Yeh and Patt, 1992

if (x[i] < 7) then
  y += 1;
if (x[i] < 5) then
  c -= 4;

If first condition false, second condition also false

Branch History Register, BHR, records the direction of the last N branches executed by the processor (Shift Register)
Pattern History Table (PHT)
Two-Level Branch Predictor

Branch Outcome (T/NT) → BHR → Indexes PHT → PHT 0 → PHT 2^(k-1) → FSM Output Logic → Prediction (T/NT)

PC → k

BHR

PC

PHT 0

PHT 2^(k-1)
Generalized Two-Level Branch Predictor

Branch Outcome (T/NT)

BHR

PC

Indexes PHT

PHT 0

PHT 2^(k-1)

For non-trivial m and k, > 97% accuracy

FSM Output Logic

Prediction (T/NT)
Tournament Predictors (ex: Alpha 21264)

- Choice predictor learns whether best to use local or global branch history in predicting next branch
- Global history is speculatively updated but restored on mispredict
- Claim 90-100% success on range of applications
Agenda

• Branch Cost Motivation
• Branch Prediction
  – Outcome
    • Static
    • Dynamic
  – Target Address
Predicting Target Address

Even with best possible prediction of branch outcome, still have to wait for branch target address to be determined
Branch Target Buffer (BTB)

Put BTB in Fetch Stage in parallel with PC+4 Speculation logic
BTB is only for Control Instructions

• BTB contains useful information for branch and jump instructions only
  – Do not update it for other instructions

• For all other instructions the next PC is PC+4!

*How to achieve this effect without decoding the instruction?*

*When do we update BTB information?*
Uses of Jump Register (JR)

• Switch statements (jump to address of matching case)
  BTB works well if same case used repeatedly

• Dynamic function call (jump to run-time function address)
  BTB works well if same function usually called, (e.g., in C++ programming, when objects have same type in virtual function call)

• Subroutine returns (jump to return address)
  BTB works well if usually return to the same place
  ⇒ Often one function called from many distinct call sites!

How well does BTB work for each of these cases?
Subroutine Return Stack

Small structure to accelerate JR for subroutine returns, typically much more accurate than BTBs.

```c
fa() { fb(); }
fb() { fc(); }
fc() { fd(); }
```

Push call address when function call executed

Pop return address when subroutine return decoded

k entries
(typically k=8-16)
Acknowledgements

• These slides contain material developed and copyright by:
  – Arvind (MIT)
  – Krste Asanovic (MIT/UCB)
  – Joel Emer (Intel/MIT)
  – James Hoe (CMU)
  – John Kubiatowicz (UCB)
  – David Patterson (UCB)
  – Christopher Batten (Cornell)

• MIT material derived from course 6.823
• UCB material derived from course CS252 & CS152
• Cornell material derived from course ECE 4750