T13 Advanced Processors:
Branch Prediction

School of Electrical and Computer Engineering
Cornell University

revision: 2016-11-29-19-25

1 Branch Prediction Overview

2 Software-Based Branch Prediction
  2.1. Static Software Hints
  2.2. Branch Delay Slots
  2.3. Predication

3 Hardware-Based Branch Prediction
  3.1. Fixed Branch Predictor
  3.2. Branch History Table (BHT) Predictor
  3.3. Two-Level Predictor For Temporal Correlation
  3.4. Two-Level Predictor For Spatial Correlation
  3.5. Generalized Two-Level Predictors
  3.6. Tournament Predictors
  3.7. Branch Target Buffers (BTBs) Predictor
1. Branch Prediction Overview

Assume incorrect branch prediction in dual-issue I2OL processor.

<table>
<thead>
<tr>
<th>bne</th>
<th>opA</th>
<th>opB</th>
<th>opC</th>
<th>opD</th>
<th>opE</th>
<th>opF</th>
<th>opG</th>
<th>opTARG</th>
</tr>
</thead>
</table>

Assume correct branch prediction in dual-issue I2OL processor.

<table>
<thead>
<tr>
<th>bne</th>
<th>opA</th>
<th>opTARG</th>
<th>opX</th>
<th>opY</th>
<th>opZ</th>
</tr>
</thead>
</table>

Three critical pieces of information we need to predict control flow:

- (1) Is this instruction a control flow instruction?
- (2) What is the target of this control flow instruction?
- (3) Do we redirect control flow to the target or next instr?
When do we know these critical pieces of information?

<table>
<thead>
<tr>
<th></th>
<th>jal</th>
<th>jr</th>
<th>bne</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Is this instruction a control flow instruction?</td>
<td>D</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>(2) What is the target of this control flow instruction?</td>
<td>D</td>
<td>X</td>
<td>D</td>
</tr>
<tr>
<td>(3) Do we redirect ctrl flow to the target or next instr?</td>
<td>D</td>
<td>D</td>
<td>X</td>
</tr>
</tbody>
</table>

What do we need to predict in F stage vs. D stage?

<table>
<thead>
<tr>
<th></th>
<th>jal</th>
<th>jr</th>
<th>bne</th>
</tr>
</thead>
<tbody>
<tr>
<td>F stage</td>
<td>predict 1,2,3</td>
<td>predict 1,2,3</td>
<td>predict 1,2,3</td>
</tr>
<tr>
<td>D stage</td>
<td>no prediction</td>
<td>predict 2</td>
<td>predict 3</td>
</tr>
</tbody>
</table>

2. Software-Based Branch Prediction

- Static software hints
- Branch delay slots
- Predication
2.1. Static Software Hints

Software provides hints about whether a control flow instruction is likely to be taken or not taken. These hints are part of the instruction and thus are available earlier in the pipeline (e.g., in the D stage).

What if the hint is wrong?

| bne.t | opA | opTARG | bne.nt | opY | opZ |

| bne.t | opA | opTARG | bne.nt | opA | opB |
2.2. Branch Delay Slots

Without branch delay slots must squash fall through instructions if branch is taken.

<table>
<thead>
<tr>
<th>bne</th>
<th>opA</th>
<th>opB</th>
<th>targ</th>
</tr>
</thead>
</table>

With branch delay slots compiler can put useful work in the slots. Instructions in the delay slots are always executed regardless of branch condition.

| bne | opA | opB | targ |
2.3. Predication

Not really “prediction”. Idea is to turn control flow into dataflow completely eliminating the control hazard.

**Conditional move instructions** conditionally move a source register to a destination register.

- `movn rd, rs1, rs2` if `(R[rs2] != 0)` `R[rd] ← R[rs1]`
- `movz rd, rs1, rs2` if `(R[rs2] == 0)` `R[rd] ← R[rs1]`

Pseudocode

<table>
<thead>
<tr>
<th>Pseudocode</th>
<th>w/o Predication</th>
<th>w/ Predication</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>if ( a &lt; b )</code></td>
<td><code>slt x1, x2, x3</code></td>
<td><code>slt x1, x2, x3</code></td>
</tr>
<tr>
<td><code>x = a</code></td>
<td><code>beq x1, x0, L1</code></td>
<td><code>movz x4, x2, x1</code></td>
</tr>
<tr>
<td><code>else</code></td>
<td><code>addi x4, x2, x0</code></td>
<td><code>movn x4, x3, x1</code></td>
</tr>
<tr>
<td><code>x = b</code></td>
<td><code>jal x0, L2</code></td>
<td></td>
</tr>
<tr>
<td></td>
<td><code>L1:</code></td>
<td><code>L1:</code></td>
</tr>
<tr>
<td></td>
<td><code>addi x4, x3, x0</code></td>
<td></td>
</tr>
<tr>
<td></td>
<td><code>L2:</code></td>
<td></td>
</tr>
</tbody>
</table>

**Full predication** enables almost all instructions to be executed under a predicate. If predicate is false, instruction should turn into a NOP.

<table>
<thead>
<tr>
<th>Pseudocode</th>
<th>w/ Predication</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>if ( a &lt; b )</code></td>
<td><code>slt.p p1, x2, x3</code></td>
</tr>
<tr>
<td><code>opA</code></td>
<td><code>(!p1) opC</code></td>
</tr>
<tr>
<td><code>opB</code></td>
<td><code>(!p1) opD</code></td>
</tr>
<tr>
<td><code>else</code></td>
<td></td>
</tr>
<tr>
<td><code>opC</code></td>
<td></td>
</tr>
<tr>
<td><code>opD</code></td>
<td></td>
</tr>
</tbody>
</table>

- What if both sides of branch have many instructions?
- What if one side of branch has many more than the other side?
3. Hardware-Based Branch Prediction

- Fixed branch predictor
- Branch history table (BHT) predictor
- Two-level predictor for temporal correlation
- Two-level predictor for temporal correlation
- Generalized two-level predictors
- Tournament predictor
- Branch target buffer (BTB) predictor

3.1. Fixed Branch Predictor

- Always predict not taken
  - What we have been assuming so far
  - Simple to implement and can perform prediction in F
  - Poor accuracy, especially on very important backwards branch in loops

- Always predict taken
  - Difficult to implement: we don’t know if this is a branch until D
  - Difficult to implement: we don’t know target until at least D
  - Could predict not taken in F, and then adjust in D
  - Poor accuracy, especially on if/then/else

- Predict taken for backward branches and predict not taken for forward branches
  - Difficult to implement: we don’t know if this is a branch until D
  - Difficult to implement: we don’t know target until at least D
  - Could predict not taken in F, and then adjust in D
  - Better accuracy
3. Hardware-Based Branch Prediction  
3.1. Fixed Branch Predictor

```assembly
loop: <------------------.
lw x1, 0(x2)         | backward
lw x3, 0(x4)         | branches
slt x5, x1, x3       | taken on avg
beq x5, x0, L1       --. forward | 90%
addi x6, x1, x0      | branches |
jal x0, L2            | taken on avg |
L1: <-' 50%          |
    addi x6, x3, x0   |
L2:                   |
    sw x6, 0(x7)      |
    addi x2, x2, 4    |
    addi x4, x4, 4    |
    addi x7, x7, 4    |
    addi x8, x8, -1   |
    bne x8, x0, loop  -------------------'
```

- For now let’s focus on conditional branches as opposed to unconditional jumps
- Let’s assume we always predict not-taken in the F stage
- In the D stage, we know if the instruction is a branch and we know the target of the branch
- So key goal is to predict whether or not we need to redirect the control flow, i.e., to predict the branch outcome in the D stage instead of waiting until the X stage
- By doing this prediction in the D stage we can reduce the branch misprediction penalty by several cycles although it is still not zero if we predict the branch is taken
3.2. Branch History Table (BHT) Predictor

How can we do better? Exploit structure in the program, namely temporal correlation: the outcomes of specific static branch in the past may be a good indicator of the outcomes of future dynamic instances of the same static branch.

One-Bit Saturating Counter

Remember the previous outcome of a specific static branch and predict the outcome will be the same for the next dynamic instance of the same branch.

Consider how this saturating counter would be have for a backwards branch in a loop with four iterations. Assume the entire loop is executed several times.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Prediction</th>
<th>Actual</th>
<th>Mispredict?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Exploiting temporal correlation works well, but a one-bit saturating counter will always mispredicts the backwards branch in a loop twice. Loops are very common!

**Two-Bit Saturating Counter**

Remember the last two outcomes of a specific static branch. Require two consecutive “counter examples” before changing the prediction.

Consider how this saturating counter would be have for a backwards branch in a loop with four iterations. Assume the entire loop is executed several times.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Prediction</th>
<th>Actual</th>
<th>Mispredict?</th>
<th>ST</th>
<th>WT</th>
<th>WNT</th>
<th>SNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

What if start state is strongly taken?
Other Two-Bit FSM Branch Predictors

See Fig 5.8 in [source](#) for other alternatives.
3. Hardware-Based Branch Prediction

3.2. Branch History Table (BHT) Predictor

**Branch History Table**

- So far we have focused on a simple FSM that exploits temporal correlation to make a prediction for a *specific static branch*.

- To make predictions for many different static branches, we need to keep track of a *dedicated* FSM per static branch.

- A branch history table (BHT) is a table where each entry is the state of the FSM for a different static branch.

- Two PC’s can “alias” to the same entry in BHT.

- Aliasing is similar to a cache conflict.

- We could store the PC as a tag along with the FSM state to make sure we don’t mix up the FSM state across two static branches.

- Storing the PC is too expensive though, so we can just let branches alias and this just reduces the branch prediction accuracy.

- Can reduce aliasing with larger BHT.

BHT with 4k entries and 2bits/entry = 80–90% accuracy

How do we continue to improve prediction accuracy? Exploit even more complicated temporal correlation.
Often a branch exhibits more complicated patterns than just “always taken” or “always not taken”. Could develop a more complicated FSM, but then patterns vary per branch. We want per branch customized FSMs.

```c
void convolve(int B[], int A[], int size) {
    for (int i = 2; i < size-2; i++) {
        for (int j = 0; j < 5; j++) {
        }
    }
}
```

Can we predict that every fifth dynamic instance of the backwards loop branch will be not taken?
3.3. Two-Level Predictor For Temporal Correlation

When a branch is taken or not taken we shift in either a one (taken) or a zero (not taken) into the least significant bit of the corresponding BHSR.

<table>
<thead>
<tr>
<th>Index</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>0111</td>
<td>ST</td>
</tr>
<tr>
<td>1000</td>
<td>WT</td>
</tr>
<tr>
<td>1001</td>
<td>WT</td>
</tr>
<tr>
<td>1010</td>
<td>WT</td>
</tr>
<tr>
<td>1011</td>
<td>ST</td>
</tr>
<tr>
<td>1100</td>
<td>WT</td>
</tr>
<tr>
<td>1101</td>
<td>ST</td>
</tr>
<tr>
<td>1110</td>
<td>ST</td>
</tr>
<tr>
<td>1111</td>
<td>SNT</td>
</tr>
</tbody>
</table>

- BHSR captures temporal pattern for that branch
- We use the BHSR to index into the PHT. A BHT has an entry per branch, but a PHT has an entry per branch pattern.
- The PHT says for a given pattern over the past n executions of a branch, should I take or not take the next execution of this branch?
- Once the two-level predictor is warmed up for previous nested loop example, the state of the PHT would be what is shown on the left
- Need at least four bits of “history” to learn this pattern and perfectly predict this branch
3. Hardware-Based Branch Prediction  
3.3. Two-Level Predictor For Temporal Correlation

Problem: Multiple branches with same history might need different predictions. In other words, aliasing in the PUT can lose accuracy.

Solution: Add multiple PUTs, use bits from PC to choose which PUT to use.

[Diagram showing two-level predictor structure with PC, PUTs, and FSM output logic]
HW Based: Exploiting Spatial Correlation

The way one branch is resolved may be a good indicator of the way a later (different) branch will resolve.

\[
\begin{align*}
\text{if } (x < 7) & \quad r_1, r_2, c_1, 7 \\
\text{then } & \quad \text{branch } \phi \\
\text{if } (x < 5) & \quad r_3, r_4, 1 \\
\text{then } & \quad \text{branch } 1
\end{align*}
\]

If branch \( \phi \) is taken (i.e. \( x \geq 7 \))
then branch 1 is always taken (i.e. \( x \) must be \( \geq 5 \)).

So whether branch \( \phi \) is taken or not taken can be used to predict if we should take branch 1.

For above example, \( \text{Bus 1} \) will capture history — so we will know when first branch is taken, and that value(s) of\( \text{Bus 1} \) will point to an entry in the \( \text{Put Table} \) that predicts taken.
As before, multiple Puts can help avoid aliasing in Put.
3. Hardware-Based Branch Prediction

3.5. Generalized Two-Level Predictors

**Generalized Two-Level Tbits**

Combined approach to exploit from complex temporal correlation and spatial correlation.

Difference from discussion on complex temporal correlation is that we purposely choose a smaller \( m \) to cause aliasing in the Tbits, since this aliasing allows us to capture spatial correlation.

\[
\begin{array}{c|cccc}
\text{One Bhit} & \text{One Bhit} & \text{One Bhit for each PC} & \text{One Bhit} & \text{One Bhit} \\
\text{Put} & k=0 & 0<k<30 & k=30 & 0<k<30 \\
\text{MAC} & G_Ag & G_As & G_Ap & G_Ag \\
\text{MAC} & P_Ag & P_As & P_Ap & P_Ag \\
\text{MAC} & S_Ag & S_As & S_Ap & S_Ag \\
\end{array}
\]

97% Accuracy
3. Hardware-Based Branch Prediction

3.5. Generalized Two-Level Predictors

Instead of concatenating 30 bits from PC with BUSH, helps avoid aliasing in the put/out more effectively.
3.6. Tournament Predictors

Different predictors are better at predicting different types of branches:
- One-level 2-bit counting counter
- Two-level gshare
- Loops
- Irregular code

Branch Predictor Selection Table
- Predicts which branch predictor we should use
Branch Target Buffers (BTBs) Predictor

Even with best possible prediction of branch outcome, still need to wait for target address to be determined.

**HW BRANCH: PREDICTING TARGET ADDRESS**

Find output logic

**Branch Target Buffer**

- **PC**
- **Predicted FSM**
- **Target State**

Hit?

New PC = Predicted Target PC
- If hit and predicted taken

Can put BTB in fetch stage
- Predicting if PC points to a branch
- Predicting target of branch
- Predicting if branch is taken

Sometimes $b=0$, if hit then assume predict taken
3. Hardware-Based Branch Prediction

3.7. Branch Target Buffers (BTBs) Predictor

```
COMBINE BTB AND BUT

BTB is much more expensive than BUT, but BTB earlier in pipeline can accelerate JR

COMBINE BTB W/ few entries with BUT W/ many entries

RETURN ADDRESS STACK PREDICTOR

BTB only works for JR function call returns if always call function from same place (not realistic)

STACK PREDICTOR
- Push target address on stack for JAL/JALR
- Pop off target address for JR to predict target

Move stack predictor into fetch and predict which PC's are JR.

USE TOURNAMENT PREDICTOR TO CHOOSE BETWEEN BTB AND STACK PREDICTOR.
```