This Set

Material from Section 4.3
This set under construction.

Outline
- Branch Prediction Overview
- One-Level Predictor
- Two-Level Correlating Predictor
- Other topics to be added.
- Sample Problems

Branch Prediction

Methods Covered

One-level predictor, a.k.a. bimodal.

Two-Level Predictors

GAg, a.k.a. Global History.
gshare.

Local History, a.k.a. PAg.

Branch and Target Prediction

Motivation

Branches occur frequently in code.
At best, one cycle of branch delay; more with dependencies.
Therefore, impact on CPI is large.

Techniques

Branch Prediction:
Predict outcome of branch. (Taken or not taken.)

Branch Target Prediction:
Predict branch or other CTI’s target address.

Branch Folding:
Replace branch or other CTI with target instruction.

Branch Prediction Idea

Idea: Predict using past behavior.

Example:

```
LOOP:
  lw  r1, 0(r2)  # Read random number, either 0 or 1.
  addi r2, r2, 4
  slt  r6, r2, r7
  beq  r1, r0 SKIP
  nop
  addi r3, r3, 1
SKIP:
  bneq  r6, r0 LOOP  # Loop executes 100 iterations.
  nop
```

Second branch, bneq, taken 99 out of 100 executions.
Pattern for bneq: T T T T T T T T T T T T
First branch shows no pattern.
Prediction Accuracy

SPEC9 benchmarks on IBM POWER (predecessor to PowerPC).

FIGURE 4.14 Prediction accuracy of a 4096-entry two-bit prediction buffer for the SPEC9 benchmarks.

Branch Prediction Terminology

**Outcome:** [of a branch instruction execution].
The outcome of the execution of a branch instruction.

T:
A taken branch.

NT: or N
A branch that is not taken.

**Prediction:** [made by branch prediction hardware].
The predicted outcome of a branch.

**Misprediction:**
An incorrectly predicted outcome.

**Prediction Accuracy:** [of a branch prediction scheme].
The number of correct predictions divided by the number of predictions.

Speculative Execution:
The execution of instructions following a predicted branch.

Misprediction Recovery:
Undoing the effect of speculatively executed instructions ... ... and re-starting instruction fetch at the correct address.
One-Level Branch Predictor

Idea: maintain a *branch history* for each branch instruction.

**Branch History:**
Information about past behavior of the branch.

Branch histories stored in a *branch history table (BHT).*

Often, branch history is sort of number of times branch taken . . . minus number of times not taken.

Other types of history possible.

Branch history read to make a prediction.

Branch history updated when branch outcome known.

Branch History Counter

If a counter used, branch history incremented when branch taken . . . and decremented when branch not taken.

Symbol *n* denotes number of bits for branch history.

To save space and for performance reasons . . . branch history limited to a few bits, usually *n* = 2.

Branch history updated using a *saturating counter*.

A saturating counter is an arithmetic unit that can add or subtract one . . .

in which . . .

. . .

. . .

. . .

For an *n*-bit counter, predict taken if counter > 2^n - 1.

One-Level Branch Predictor Hardware

Illustrated for Chapter-3 DLX implementation . . .

. . . even though prediction not very useful.

Branch Prediction Steps

1: Predict.

Read branch history, available in ID.

2: Determine Branch Outcome

Execute predicted branch in usual way.

3: Recover (If necessary.)

Undo effect of speculatively executing instructions, start fetching from correct path.

4: Update Branch History

Branch History Table

Stores branch histories.

Implemented using a memory device.

Address (called index) is *hash* of branch address (PC).

For 2^m-entry BHT, hash is *m* lowest bits of branch PC skipping alignment.

<table>
<thead>
<tr>
<th>BHT Addr</th>
<th>Align.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Branch address:</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>m+2</td>
</tr>
</tbody>
</table>

Data input and output of BHT is branch history.
Outcomes for individual branches, categorized by pattern, sorted by frequency.

Branches running \TeX text formatter compiled for SPARC (Solaris).

<table>
<thead>
<tr>
<th>Pattern</th>
<th># Branches</th>
<th>gshre</th>
<th>corr</th>
<th>Local History</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: fe7f</td>
<td>0.0004</td>
<td>1397</td>
<td>0.912</td>
<td>Local History</td>
</tr>
<tr>
<td>1: ffff</td>
<td>0.0004</td>
<td>1323</td>
<td>0.924</td>
<td>Local History</td>
</tr>
<tr>
<td>2: fcff</td>
<td>0.0004</td>
<td>1245</td>
<td>0.910</td>
<td>Local History</td>
</tr>
<tr>
<td>3: ff9f</td>
<td>0.0003</td>
<td>1235</td>
<td>0.955</td>
<td>Local History</td>
</tr>
<tr>
<td>4: f9ff</td>
<td>0.0003</td>
<td>1188</td>
<td>0.926</td>
<td>Local History</td>
</tr>
<tr>
<td>5: ffcf</td>
<td>0.0003</td>
<td>1163</td>
<td>0.973</td>
<td>Local History</td>
</tr>
<tr>
<td>6: fccf</td>
<td>0.0003</td>
<td>1159</td>
<td>0.955</td>
<td>Local History</td>
</tr>
<tr>
<td>7: 80</td>
<td>0.0003</td>
<td>1158</td>
<td>0.949</td>
<td>Local History</td>
</tr>
<tr>
<td>8: c0</td>
<td>0.0003</td>
<td>1155</td>
<td>0.944</td>
<td>Local History</td>
</tr>
</tbody>
</table>

**Phase Change, pat 26, br 48190, 0.8453 0.9040 0.8470 (0.01295)**

<table>
<thead>
<tr>
<th>Pattern</th>
<th># Branches</th>
<th>gshre</th>
<th>corr</th>
<th>Local History</th>
</tr>
</thead>
<tbody>
<tr>
<td>0: e000</td>
<td>0.0009</td>
<td>4554</td>
<td>0.653</td>
<td>Local History</td>
</tr>
<tr>
<td>1: f000</td>
<td>0.0008</td>
<td>2942</td>
<td>0.756</td>
<td>Local History</td>
</tr>
<tr>
<td>2: fff0</td>
<td>0.0008</td>
<td>2878</td>
<td>0.908</td>
<td>Local History</td>
</tr>
<tr>
<td>3: fff0</td>
<td>0.0007</td>
<td>2642</td>
<td>0.786</td>
<td>Local History</td>
</tr>
<tr>
<td>4: 80</td>
<td>0.0007</td>
<td>2572</td>
<td>0.968</td>
<td>Local History</td>
</tr>
<tr>
<td>5: fe00</td>
<td>0.0006</td>
<td>2435</td>
<td>0.815</td>
<td>Local History</td>
</tr>
<tr>
<td>6: f000</td>
<td>0.0006</td>
<td>2225</td>
<td>0.836</td>
<td>Local History</td>
</tr>
<tr>
<td>7: ff00</td>
<td>0.0006</td>
<td>2140</td>
<td>0.856</td>
<td>Local History</td>
</tr>
<tr>
<td>8: ff00</td>
<td>0.0006</td>
<td>2061</td>
<td>0.854</td>
<td>Local History</td>
</tr>
</tbody>
</table>

**Two-Level Correlating Predictors**

Idea: Base branch decision on …

… the address of the branch instruction (as in the one-level scheme) …

… and the most recent branch outcomes.

**History:**

The outcome (taken or not taken) of the most recent branches. Usually stored as a bit vector with 1 indicating taken.

**Pattern History Table (PHT):**

Memory for 2-bit counters, indexed (addressed) by some combination of history and the branch instruction address.
Some Types of Two-Level Predictors

*Global*, a.k.a. *GAg*.

History is global (same for all branches), stored in a *global history register* (*GHR*).

PHT indexed using history only.

*gshare*

History is global (same for all branches), stored in a *global history register* (*GHR*).

PHT indexed using history exclusive-ored with branch address.

*gselect*

History is global (same for all branches), stored in a *global history register* (*GHR*).

PHT indexed using history concatenated with branch address.

---

**Global History Example**

```
! Loop always iterates 4 times.
! Branch below never taken.
   bneq r2, SKIP N N
   addi f0, f0, f2
   SKIP:
      addi r1, r0, #4
   LOOP:
      multd f0, f0, f2
      subi r1, r1, #1
   bneq r1, LOOP T T T N ...
! Cycle 10 20 30 40 50 110 120 130 140 150
!
! Global History (m=4), X: depends on earlier branches.
! 10 XXXN Human would predict taken.
! 10 XXNT Human would predict taken.
! 10 XNTT Human would predict taken.
! 10 NTTT Human would predict not taken.
! 50 TTTN
```

---

**Steps in BHT Use on a Dynamically Scheduled Proc.**

Register r1 not available until cycle ten.

Cycle 1: When branch in ID, read BHT and make prediction.

Cycle 1: (Optional) Backup (checkpoint) register map (if present).

Cycle 10: Execute branch in usual way and check prediction.

Cycle 11: If prediction correct, update BHT when branch commits.

Cycle 11: If pred. wrong, start recovery process (does not occur here).

---

1 *Predict not taken, not taken.*

Cycle: 0 1 2 3 10 11 12 13
   bneq r1, TARGET IF ID Q ...
   xor r2, r3, r4 IF ID Q EX WB C
   ...
   TARGET:
      and r5, r6, r7

---

1 Perhaps due to a cache miss, or maybe it depended on a long-latency floating-point operation, the reason is not important.
BHT use when branch taken, correctly predicted.

Register r1 not available until cycle 10.

Cycle 1: When branch in ID, compute target, read BHT and make prediction.
Cycle 10: Execute branch in usual way.
Cycle 11: Check outcome. Correctly predicted.
Cycle 23: Commit branch after div.

BHT use when branch taken, incorrectly predicted, register map not backed up.

Register r1 not available until cycle 10.

Cycle 1: When branch in ID, compute target, read BHT and make prediction.
Cycle 10: Compute branch condition.
Cycle 11: Misprediction “discovered.” Because register map not backed up, recovery must wait until commit.
Cycle 23: Start recovery: Squash instructions in reorder buffer, start fetching correct path.

Global History and Dynamic Scheduling

Global history must be accurate.

Why that’s a problem:

Global history includes assumption that bneq not taken.
First branch: Predict not taken, taken. Register map backed up.

Cycle: 0 1 2 3 10 11 12 13 ... 21 22 23

div f0, f2, f4 IF ID Q DIV M
bneq r1, TARGET IF ID Q ... B WB C
beqz r2, SKIP IF ID Q B ...
xor r2, r3, r4 IF ID Q EX...

TARGET:
and r5, r6, r7 IF ID Q ...
beqz r4, LINE1 IF ID ...

Cycle: 0 1 2 3 10 11 12 13 ... 21 22 23

Global History and Dynamic Execution

Global History in Two-Level Predictor with Dynamic Execution

Global history backed up (checkpointed) at each branch.

Predicted outcome shifted into global history.

If misprediction discovered, global history restored from backup ...

just as the register map can be.

Target Prediction and Folding

Target Prediction

Predicting the outcome and target of a branch.

Branch Target Buffer:

A table indexed by branch address holding a predicted target address.

Target Prediction

Put BTB in IF stage.

Use PC to read an entry from BTB.

If valid entry found, replace PC with predicted target.

With target correctly predicted, zero branch delay.

Target Prediction Example

Static scheduled system (for clarity).

<table>
<thead>
<tr>
<th>Cycle: 0 1 2 3 4 10 11 12</th>
<th>13 14</th>
<th>bneq r1, TARGET IF ID EX MEM WB</th>
<th>IF ID EX MEM WB</th>
<th>xor r2, r3, r4 IF ID EX</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARGET:</td>
<td></td>
<td>and r5, r6, r7 IF ID EX MEM WB</td>
<td>IF X</td>
<td></td>
</tr>
</tbody>
</table>

Cycle 0

BTB lookup and prediction. Predict taken.

Target from BTB will be clocked into PC.
Target Prediction Example, continued.

Static scheduled system (for clarity).

<table>
<thead>
<tr>
<th>Cycle</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>bneq r1, TARGET</td>
<td>IF</td>
<td>ID</td>
<td>EX</td>
<td>MEM</td>
<td>WB</td>
<td>IF</td>
<td>ID</td>
<td>EX</td>
<td>MEM</td>
<td>WB</td>
</tr>
<tr>
<td>xor r2, r3, r4</td>
<td>IF</td>
<td>ID</td>
<td>EX</td>
<td>MEM</td>
<td>WB</td>
<td>IF</td>
<td>ID</td>
<td>EX</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TARGET:

| and r5, r6, r7 | IF | ID | EX | MEM | WB | IF | X |

Cycle 1

Start fetching predicted target.

Execute branch instruction (in ID).

Check predicted outcome and predicted target.

Correct predictions, continue execution.

Target Prediction for Register-Indirect CTI

What BTB predicts for branch instructions:

That instruction will be a CTI.

If CTI is a branch, that branch is taken.

CTI target.

For branches and non-indirect jumps (j, jal)...

... predicting target is easy, since target always same.

bneq r1, LOOP  ! Target always PC + 4 + 4 * LOOP
j LINEJ  ! Target always PC + 4 + 4 * LINEJ

For register-indirect jumps (jr, jalr)...

... prediction depends on predictable behavior.

jr r1  ! Target is in r1. Can be different each time.
jalr r1  ! Target is in r1. Can be different each time.

Behavior of Register-Indirect Jumps

Predictability depends on how jumps used.

Major Uses

- Procedure Passed as Parameter
  
  For example, function passed to the C library's qsort.

  These rarely change so target is predictable.

- Case Statements
  
  These change, and so prediction more difficult.
Indirect Jump Target Prediction

Separate techniques used for procedure returns and other indirect jumps.

Return Address Prediction
- Keep a stack of (what appear to be) return addresses.

Other Indirect Jumps Prediction
- Predict last target.
- Use global branch history to index BTB.

Predict Return Address

Used for return instruction. (An instruction used for a procedure return, which may not have the mnemonic `return`).

Operation
- Hardware keeps a stack of return addresses.
- BTB stores whether instruction is a return.
- When a call instruction encountered push return address on stack.
- When BTB identifies instruction as a return target address is popped off stack.

Effectiveness
- Works fairly well.
- Can be confused when returns skipped (as with long jumps).
- Costly to implement precisely with dynamic scheduling.

Predict Last Target

Can be used for everything except return instructions.

Last time instruction executed target address stored in BTB.

If entry found and predicted taken (for a branch), last target address used.

Effectiveness:
- Perfect for non-indirect jumps and branches (if taken).
- Reasonably effective on indirect branches.
- Use Global History
- Can be used for everything except return instructions.
- Much more effective on than last target.

Target Prediction Example

Consider code for C `switch` statement:

```c
! Possible code for a switch statement.
! switch( r2 ) { case 0: foo(); break; case 1: bar(); break; ... }
! Set r1 to base of switch address table.
  lhi  r1, #0x1234
  ori  r1, r1, #0x5670
! Multiply switch index by stride of table (4 bytes per address).
  slli r3, r2, #2
! Get address of case code address.
  add  r1, r1, r3
! Get case code address.
  lw   r4, 0(r1)
! Jump to case code.
  jr  r4
```

If `r2` rarely changes, `jr` predictable.
Possible BTB Contents

Target address.

History information (replaces BHT).

Tag, to detect collisions.