Reduction of Control Hazards (Branch) Stalls with Dynamic Branch Prediction

• So far we have dealt with control hazards in instruction pipelines by:
  – Assuming that the branch is not taken (i.e. stall when branch is taken).
  – Reducing the branch penalty by resolving the branch early in the pipeline
    • Branch penalty if branch is taken = stage resolved - 1
  – Branch delay slot and canceling branch delay slot. (ISA support needed)
  – Compiler-based static branch prediction encoded in branch instructions
    • Prediction is based on program profile or branch direction
    • ISA support needed.

How to further reduce the impact of branches on pipeline processor performance?

• Dynamic Branch Prediction:
  – Hardware-based schemes that utilize run-time behavior of branches to make dynamic predictions:
    • Information about outcomes of previous occurrences of branches are used to dynamically predict the outcome of the current branch.

• Branch Target Buffer (BTB): (Goal: zero stall taken branches)
  – A hardware mechanism that aims at reducing the stall cycles resulting from correctly predicted taken branches to zero cycles.

(Static Prediction in Chapter 4.2 Dynamic Prediction in Chapter 3.4, 3.5)
Static Conditional Branch Prediction

- Branch prediction schemes can be classified into static (at compilation time) and dynamic (at runtime) schemes.
- Static methods are carried out by the compiler. They are static because the prediction is already known before the program is executed.
- Static Branch prediction is encoded in branch instructions using one prediction (or branch direction hint) bit: 0 = Not Taken, 1 = Taken.
  - Must be supported by ISA, Ex: HP PA-RISC, PowerPC, UltraSPARC
- Two basic methods to statically predict branches at compile time:
  - Use the direction of a branch to base the prediction on. Predict backward branches (branches which decrease the PC) to be taken (e.g. loops) and forward branches (branches which increase the PC) not to be taken.
  - Profiling can also be used to predict the outcome of a branch.
    - A number runs of the program are used to collect program behavior information (i.e. if a given branch is likely to be taken or not)
    - This information is included in the opcode of the branch (one bit branch direction hint) as the static prediction.

(Static Prediction in Chapter 4.2, covered in lecture 2)
Static Profile-Based Compiler Branch Misprediction Rates for SPEC92

Misprediction rate for a profile-based predictor varies widely but is generally better for the FP programs, which have an average misprediction rate of 9% with a standard deviation of 4%, than for the integer programs, which have an average misprediction rate of 15% with a standard deviation of 5%.

(repeated here from lecture2)
Dynamic Conditional Branch Prediction

- Dynamic branch prediction schemes are different from static mechanisms because they utilize hardware-based mechanisms that use the run-time behavior of branches to make more accurate predictions than possible using static prediction.

- Usually information about outcomes of previous occurrences of branches (branching history) is used to dynamically predict the outcome of the current branch. Some of the proposed dynamic branch prediction mechanisms include:
  - One-level or Bimodal: Uses a Branch History Table (BHT), a table of usually two-bit saturating counters which is indexed by a portion of the branch address (low bits of address). (First proposed mid 1980s)
  - Two-Level Adaptive Branch Prediction. (First proposed early 1990s),
  - MCFarling’s Two-Level Prediction with index sharing (gshare, 1993).
  - Hybrid or Tournament Predictors: Uses a combinations of two or more (usually two) branch prediction mechanisms (1993).

- To reduce the stall cycles resulting from correctly predicted taken branches to zero cycles, a Branch Target Buffer (BTB) that includes the addresses of conditional branches that were taken along with their targets is added to the fetch stage.

(Dynamic Branch Prediction in Chapter 3.4, 3.5)
Branch Target Buffer (BTB)

- Effective branch prediction requires the target of the branch at an early pipeline stage. (resolve the branch early in the pipeline)
- One can use additional adders to calculate the target, as soon as the branch instruction is decoded. This would mean that one has to wait until the ID stage before the target of the branch can be fetched, taken branches would be fetched with a one-cycle penalty (this was done in the enhanced MIPS pipeline Fig A.24).
- To avoid this problem one can use a Branch Target Buffer (BTB). A typical BTB is an associative memory where the addresses of taken branch instructions are stored together with their target addresses.
- Some designs store \( n \) prediction bits as well, implementing a combined BTB and Branch history Table (BHT).
- Instructions are fetched from the target stored in the BTB in case the branch is predicted-taken and found in BTB. After the branch has been resolved the BTB is updated. If a branch is encountered for the first time a new entry is created once it is resolved as taken.
- Branch Target Instruction Cache (BTIC): A variation of BTB which caches also the code of the branch target instruction in addition to its address. This eliminates the need to fetch the target instruction from the instruction cache or from memory.

Goal of BTB: Zero stall taken branches
Basic Branch Target Buffer (BTB)

IF

- Fetch instruction from instruction memory (I-L1 Cache)
- Branch Addresses
  - 0 = NT = Not Taken
  - 1 = T = Taken

Goal of BTB: Zero stall taken branches

BTB is accessed in Instruction Fetch (IF) cycle

A branch-target buffer.
The steps involved in handling an instruction with a branch-target buffer.

BTB Operation

Send PC to memory and branch-target buffer

Entry found in branch-target buffer?

Send out predicted PC

Taken branch?

Branch correctly predicted; continue with execution with no stalls

Mispredicted branch; kill fetched instruction; restart fetch at other target; delete entry from target buffer

Enter branch PC and next PC into branch target buffer

Is instruction a taken branch?

No

Normal instruction execution

One more stall to update BTB

Penalty = 1 + 1 = 2 cycles
Branch Penalty Cycles
Using A Branch-Target Buffer (BTB)

Base Pipeline Taken Branch Penalty = 1 cycle

<table>
<thead>
<tr>
<th>Instruction in buffer</th>
<th>Prediction</th>
<th>Actual branch</th>
<th>Penalty cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Taken</td>
<td>Taken</td>
<td>0</td>
</tr>
<tr>
<td>Yes</td>
<td>Taken</td>
<td>Not taken</td>
<td>2</td>
</tr>
<tr>
<td>No</td>
<td>Taken</td>
<td>Taken</td>
<td>2</td>
</tr>
</tbody>
</table>

Assuming one more stall cycle to update BTB
Penalty = 1 + 1 = 2 cycles

Penalties for all possible combinations of whether the branch is in the buffer and what it actually does, assuming we store only taken branches in the buffer.
Basic Dynamic Branch Prediction

- **Simplest method: (One-Level)**
  - A branch prediction buffer or Branch History Table (BHT) indexed by low address bits of the branch instruction.
  - Each buffer location (or BHT entry) contains one bit indicating whether the branch was recently taken or not.
    - e.g. 0 = not taken, 1 = taken
  - Always mispredicts in first and last loop iterations.

- **To improve prediction accuracy, two-bit prediction is used:**
  - A prediction must miss twice before it is changed.
    - Thus, a branch involved in a loop will be mispredicted only once when encountered the next time as opposed to twice when one bit is used.
  - Two-bit prediction is a specific case of n-bit saturating counter incremented when the branch is taken and decremented when the branch is not taken.
    - The counter (predictor) used is updated after the branch is resolved.
  - Two-bit prediction counters are usually always used based on observations that the performance of two-bit BHT prediction is comparable to that of n-bit predictors.
One-Level (Bimodal) Branch Predictors

- One-level or bimodal branch prediction uses only one level of branch history.
- These mechanisms usually employ a table which is indexed by lower N bits of the branch address.
- Each table entry (or predictor) consists of \( n \) history bits, which form an n-bit automaton or saturating counters.
- Smith proposed such a scheme, known as the Smith Algorithm, that uses a table of two-bit saturating counters. (1985)
- One rarely finds the use of more than 3 history bits in the literature.
- Two variations of this mechanism:
  - Pattern History Table: Consists of directly mapped entries.
  - Branch History Table (BHT): Stores the branch address as a tag. It is associative and enables one to identify the branch instruction during IF by comparing the address of an instruction with the stored branch addresses in the table (similar to BTB).
One-Level Bimodal Branch Predictors
Pattern History Table (PHT)

Sometimes referred to as Decode History Table (DHT) or Branch History Table (BHT).

Table has \(2^N\) entries (also called predictors).

Example:

For \(N=12\)

Table has \(2^N = 2^{12}\) entries

\[ = 4096 = 4k \text{ entries}\]

Number of bits needed = \(2 \times 4k = 8k\) bits

Most common one-level implementation

High bit determines branch prediction

\(0 = NT = \text{Not Taken}\)

\(1 = T = \text{Taken}\)

Update counter after branch is resolved:
- Increment counter used if branch is taken
- Decrement counter used if branch is not taken
One-Level Bimodal Branch Predictors

Branch History Table (BHT)

- N: Low Bits of Branch Address
- 2-bit saturating counters
- Address Tag
- Prediction Bits

<table>
<thead>
<tr>
<th>Address Tag</th>
<th>Prediction Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X   X</td>
</tr>
<tr>
<td></td>
<td>X   X</td>
</tr>
<tr>
<td></td>
<td>X   X</td>
</tr>
<tr>
<td></td>
<td>X   X</td>
</tr>
<tr>
<td></td>
<td>X   X</td>
</tr>
<tr>
<td></td>
<td>X   X</td>
</tr>
</tbody>
</table>

- High bit determines branch prediction:
  - 0 = NT = Not Taken
  - 1 = T = Taken

Not a common one-level implementation
Basic Dynamic Two-Bit Branch Prediction:

- Two-bit Predictor State Transition Diagram (in textbook):
  - Predict taken (11) -> Taken (T)
  - Not taken (00) -> Taken (T)
  - Predict not taken (01) -> Not taken (NT)
  - Predict not taken (00) -> Not taken (NT)

- Or Two-bit saturating counter predictor state transition diagram (Smith Algorithm):
  - Not Taken (NT)
  - Predict not taken (01) -> Predict not taken (00) -> Not Taken (NT)
  - Predict taken (10) -> Predict taken (11) -> Taken (T)

The two-bit predictor used is updated after the branch is resolved.
Prediction Accuracy of a 4096-entry two-bit prediction buffer for the SPEC89 benchmarks.

<table>
<thead>
<tr>
<th>SPEC89 benchmarks</th>
<th>Frequency of mispredictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integer</td>
<td>N=12</td>
</tr>
<tr>
<td>gcc</td>
<td>12%</td>
</tr>
<tr>
<td>espresso</td>
<td>5%</td>
</tr>
<tr>
<td>eqntott</td>
<td>18%</td>
</tr>
<tr>
<td>li</td>
<td>10%</td>
</tr>
<tr>
<td>nasa7</td>
<td>1%</td>
</tr>
<tr>
<td>matrix300</td>
<td>0%</td>
</tr>
<tr>
<td>tomcatv</td>
<td>1%</td>
</tr>
<tr>
<td>doduc</td>
<td>5%</td>
</tr>
<tr>
<td>spice</td>
<td>9%</td>
</tr>
<tr>
<td>fpppp</td>
<td>9%</td>
</tr>
</tbody>
</table>

Misprediction Rate:
- Integer average 11%
- FP average 4%

(Lower misprediction rate due to more loops)
### From The Analysis of Static Branch Prediction:

#### MIPS Performance Using Canceling Delay Branches

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>% conditional branches</th>
<th>% conditional branches with empty slots</th>
<th>% conditional branches that are cancelling</th>
<th>% cancelling branches that are cancelled</th>
<th>% branches with cancelled delay slots</th>
<th>Total % branches with empty or cancelled delay slot</th>
</tr>
</thead>
<tbody>
<tr>
<td>compress</td>
<td>14%</td>
<td>18%</td>
<td>31%</td>
<td>43%</td>
<td>13%</td>
<td>31%</td>
</tr>
<tr>
<td>eqntott</td>
<td>24%</td>
<td>24%</td>
<td>50%</td>
<td>24%</td>
<td>12%</td>
<td>36%</td>
</tr>
<tr>
<td>espresso</td>
<td>15%</td>
<td>29%</td>
<td>19%</td>
<td>21%</td>
<td>4%</td>
<td>33%</td>
</tr>
<tr>
<td>gcc</td>
<td>15%</td>
<td>16%</td>
<td>33%</td>
<td>34%</td>
<td>11%</td>
<td>27%</td>
</tr>
<tr>
<td>li</td>
<td>15%</td>
<td>20%</td>
<td>55%</td>
<td>48%</td>
<td>26%</td>
<td>46%</td>
</tr>
<tr>
<td>Integer average</td>
<td>17%</td>
<td>21%</td>
<td>38%</td>
<td>34%</td>
<td>13%</td>
<td>35%</td>
</tr>
<tr>
<td>dduc</td>
<td>8%</td>
<td>33%</td>
<td>12%</td>
<td>62%</td>
<td>8%</td>
<td>41%</td>
</tr>
<tr>
<td>ear</td>
<td>10%</td>
<td>37%</td>
<td>36%</td>
<td>14%</td>
<td>5%</td>
<td>42%</td>
</tr>
<tr>
<td>hydro2d</td>
<td>12%</td>
<td>0%</td>
<td>69%</td>
<td>24%</td>
<td>16%</td>
<td>17%</td>
</tr>
<tr>
<td>mdljdp2</td>
<td>9%</td>
<td>0%</td>
<td>86%</td>
<td>10%</td>
<td>8%</td>
<td>8%</td>
</tr>
<tr>
<td>su2cor</td>
<td>3%</td>
<td>7%</td>
<td>17%</td>
<td>57%</td>
<td>10%</td>
<td>17%</td>
</tr>
<tr>
<td>FP average</td>
<td>8%</td>
<td>16%</td>
<td>44%</td>
<td>34%</td>
<td>9%</td>
<td>25%</td>
</tr>
<tr>
<td>Overall average</td>
<td>12%</td>
<td>18%</td>
<td>41%</td>
<td>34%</td>
<td>11%</td>
<td>30%</td>
</tr>
</tbody>
</table>

Delayed and cancelling delay branches for MIPS allow branch hazards to be hidden 70% of the time on average for these 10 SPEC benchmarks.

70% Static Branch Prediction Accuracy

(repeated here from lecture2)
Prediction Accuracy of Basic One-Level Two-Bit Branch Predictors:

4096-entry buffer Vs. An Infinite Buffer Under SPEC89
Correlating Branches

Recent branches are possibly correlated: The behavior of recently executed branches affects prediction of current branch.

Example:

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>if (aa==2)</td>
<td>DSUBUI R3, R1, #2</td>
<td>R3 = R1 - 2</td>
<td>BNEZ R3, L1</td>
<td>B1 (aa!=2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>aa=0; (not taken)</td>
<td>DADD R1, R0, R0</td>
<td>aa==0</td>
<td>B1 not taken</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>if (bb==2)</td>
<td>DSUBUI R3, R2, #2</td>
<td>R3 = R2 - 2</td>
<td>BNEZ R3, L2</td>
<td>B2 (bb!=2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>bb=0; (not taken)</td>
<td>DADD R2, R0, R0</td>
<td>bb==0</td>
<td>B2 not taken</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>if (aa!==bb)</td>
<td>DSUBUI R3, R1, R2</td>
<td>R3 = aa-bb</td>
<td>DSUBUI R3, L3</td>
<td>B3 (aa==bb)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BEQZ R3, L3</td>
<td>B3 taken if aa=bb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Branch B3 is correlated with branches B1, B2. If B1, B2 are both not taken, then B3 will be taken. Using only the behavior of one branch cannot detect this behavior.
Correlating Two-Level Dynamic GAp Branch Predictors

- Improve branch prediction by looking not only at the history of the branch in question but also at that of other branches using two levels of branch history.
- Uses two levels of branch history:
  - **First level (global):**
    - Record the global pattern or history of the m most recently executed branches as taken or not taken. Usually an m-bit shift register.
  - **Second level (per branch address):**
    - $2^m$ prediction tables, each table entry has n bit saturating counter.
    - The branch history pattern from first level is used to select the proper branch prediction table in the second level.
    - The low N bits of the branch address are used to select the correct prediction entry (predictor) within a the selected table, thus each of the $2^m$ tables has $2^N$ entries and each entry is 2 bits counter.
    - Total number of bits needed for second level = $2^m \times n \times 2^N$ bits
- In general, the notation: **GAp (m,n) predictor means:**
  - Record last m branches to select between $2^m$ history tables.
  - Each second level table uses n-bit counters (each table entry has n bits).
- Basic two-bit single-level Bimodal BHT is then a (0,2) predictor.
Organization of A Correlating Two-level GAp (2,2) Branch Predictor

First Level
Branch History Register (BHR)
(2 bit shift register)  (N= 4)

Selects correct table
2-bit global branch history

Second Level
Pattern History Tables (PHTs)

Selects correct Entry (predictor) in table

High bit determines branch prediction
0 = Not Taken
1 = Taken

Low 4 bits of address
Selects correct table
(predictor)

m = # of branches tracked in first level = 2
Thus 2^m = 2^2 = 4 tables in second level

N = # of low bits of branch address used = 4
Thus each table in 2nd level has 2N = 24 = 16 entries

n = # number of bits of 2nd level table entry = 2

Number of bits for 2nd level = 2^m x n x 2^N
= 4 x 2 x 16 = 128 bits

A (2,2) branch-prediction buffer uses a two-bit global history to choose from among four predictors for each branch address.

GAp (m,n) here m= 2 n =2 Thus Gap (2, 2)
Dynamic Branch Prediction: Example

if (d==0)
  d=1;
if (d==1)
  ...

Possible execution sequences for a code fragment.

<table>
<thead>
<tr>
<th>Initial value of d</th>
<th>d==0?</th>
<th>b1</th>
<th>Value of d before b2</th>
<th>d==1?</th>
<th>b2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Yes</td>
<td>Not taken</td>
<td>1</td>
<td>Yes</td>
<td>Not taken</td>
</tr>
<tr>
<td>1</td>
<td>No</td>
<td>Taken</td>
<td>1</td>
<td>Yes</td>
<td>Not taken</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Taken</td>
<td>2</td>
<td>No</td>
<td>Taken</td>
</tr>
</tbody>
</table>

Behavior of a one-bit predictor initialized to not taken.

<table>
<thead>
<tr>
<th>d=?</th>
<th>b1 prediction</th>
<th>b1 action</th>
<th>New b1 prediction</th>
<th>b2 prediction</th>
<th>b2 action</th>
<th>New b2 prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>NT</td>
<td>T</td>
<td>T</td>
<td>NT</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>0</td>
<td>T</td>
<td>NT</td>
<td>NT</td>
<td>T</td>
<td>NT</td>
<td>NT</td>
</tr>
<tr>
<td>2</td>
<td>NT</td>
<td>T</td>
<td>T</td>
<td>NT</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>0</td>
<td>T</td>
<td>NT</td>
<td>NT</td>
<td>T</td>
<td>NT</td>
<td>NT</td>
</tr>
</tbody>
</table>

One Level with one-bit table entries (predictors): NT = 0 = Not Taken
T = 1 = Taken
Dynamic Branch Prediction: Example (continued)

if (d==0) 
  d=1;
if (d==1) 

BNEZ R1, L1 ; branch b1 (d!=0)
DADDIU R1, R0, #1 ; d==0, so d=1
L1: DADDIU R3, R1, # -1
BNEZ R3, L2 ; branch b2 (d!=1)

... 
L2:

Combinations and meaning of the taken/not taken prediction bits.

<table>
<thead>
<tr>
<th>Initial value of d</th>
<th>d==0?</th>
<th>b1</th>
<th>Value of d before b2</th>
<th>d==1?</th>
<th>b2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Yes</td>
<td>Not taken</td>
<td>1</td>
<td>Yes</td>
<td>Not taken</td>
</tr>
<tr>
<td>1</td>
<td>No</td>
<td>Taken</td>
<td>1</td>
<td>Yes</td>
<td>Not taken</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Taken</td>
<td>2</td>
<td>No</td>
<td>Taken</td>
</tr>
</tbody>
</table>

The action of the one-bit predictor with one bit of correlation, initialized to not taken/not taken.

<table>
<thead>
<tr>
<th>d=?</th>
<th>b1 prediction</th>
<th>b1 action</th>
<th>New b1 prediction</th>
<th>b2 prediction</th>
<th>b2 action</th>
<th>New b2 prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>T/NT</td>
<td>T</td>
<td>T/NT</td>
<td>NT/NT</td>
<td>T</td>
<td>NT/T</td>
</tr>
<tr>
<td>0</td>
<td>T/NT</td>
<td>NT</td>
<td>T/NT</td>
<td>NT/NT</td>
<td>NT</td>
<td>NT/T</td>
</tr>
<tr>
<td>2</td>
<td>T/NT</td>
<td>T</td>
<td>T/NT</td>
<td>NT/NT</td>
<td>NT</td>
<td>NT/T</td>
</tr>
<tr>
<td>0</td>
<td>T/NT</td>
<td>NT</td>
<td>T/NT</td>
<td>NT/NT</td>
<td>NT</td>
<td>NT/T</td>
</tr>
</tbody>
</table>

Two level GAp(1,1)

m=1

n=1
Prediction Accuracy of Two-Bit Dynamic Predictors Under SPEC89
MCFarling's gshare Predictor

gshare = global history with index sharing

- McFarling noted (1993) that using global history information might be less efficient than simply using the address of the branch instruction, especially for small predictors.
- He suggests using both global history (BHR) and branch address by hashing them together. He proposed using the XOR of global branch history register (BHR) and branch address since he expects that this value has more information than either one of its components. The result is that this mechanism outperforms GAp scheme by a small margin.
- This mechanism uses less hardware than GAp, since both branch history (first level) and pattern history (second level) are kept globally.
- The hardware cost for k history bits is \( k + 2 \times 2^k \) bits, neglecting costs for logic.

gshare is one of the most widely implemented two-level dynamic branch prediction schemes
gshare Predictor

Branch and pattern history are kept globally. History and branch address are XORed and the result is used to index the pattern history table.

First Level:

2-bit saturating counters (predictors)

Second Level:

One Pattern History Table (PHT) with $2^k$ entries (predictors)

Here: $m = N = k$

Index the second level

Prediction

gshare = global history with index sharing
gshare Performance

- gshare
- GAp
- One Level

Conditional Branch Prediction Accuracy (%)

Predictor Size (bytes)

- gshare-best
- gselect-best (Gap)
- global (One Level)
Hybrid Predictors

*(Also known as tournament or combined predictors)*

- Hybrid predictors are simply combinations of two or more branch prediction mechanisms.
- This approach takes into account that different mechanisms may perform best for different branch scenarios.
- McFarling presented (1993) a number of different combinations of two branch prediction mechanisms.
- He proposed to use an additional 2-bit counter selector array which serves to select the appropriate predictor for each branch.
- One predictor is chosen for the higher two counts, the second one for the lower two counts.
- If the first predictor is wrong and the second one is right the counter is decremented, if the first one is right and the second one is wrong, the counter is incremented. No changes are carried out if both predictors are correct or wrong.
A Generic Hybrid Predictor

Usually only two predictors are used (i.e. \( n = 2 \))
e.g. As in Alpha, IBM POWER 4, 5
MCFarling’s Hybrid Predictor Structure

The hybrid predictor contains an additional counter array (selector array) with 2-bit up/down saturating counters. Which serves to select the best predictor to use. Each counter in the selector array keeps track of which predictor is more accurate for the branches that share that counter.

Specifically, using the notation $P_1c$ and $P_2c$ to denote whether predictors $P_1$ and $P_2$ are correct respectively, the selector counter is incremented or decremented by $P_1c - P_2c$ as shown.

<table>
<thead>
<tr>
<th>P1c</th>
<th>P2c</th>
<th>P1c-P2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(no change)  (decrement counter)  (increment counter)  (no change)

Here two predictors are combined

(Current example implementations: IBM POWER4, POWER5)
MCFarling’s Hybrid Predictor Performance by Benchmark

- **bimodal** (Single Level)
- **gshare**
- **bimodal/gshare** (Combined)

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>bimodal</th>
<th>gshare</th>
<th>bimodal/gshare</th>
</tr>
</thead>
<tbody>
<tr>
<td>doduc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eqntott</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>espress</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fpppp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gcc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>li</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mat300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nasa7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tomatcv</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conditional Branch Prediction Accuracy (%)
## Processor Branch Prediction Examples

<table>
<thead>
<tr>
<th>Processor</th>
<th>Released</th>
<th>Accuracy</th>
<th>Prediction Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyrix 6x86</td>
<td>early '96</td>
<td>ca. 85%</td>
<td>BHT associated with BTB</td>
</tr>
<tr>
<td>Cyrix 6x86MX</td>
<td>May '97</td>
<td>ca. 90%</td>
<td>BHT associated with BTB</td>
</tr>
<tr>
<td>AMD K5</td>
<td>mid '94</td>
<td>80%</td>
<td>BHT associated with I-cache</td>
</tr>
<tr>
<td>AMD K6</td>
<td>early '97</td>
<td>95%</td>
<td>2-level adaptive associated with BTIC and ALU</td>
</tr>
<tr>
<td>Intel Pentium</td>
<td>late '93</td>
<td>78%</td>
<td>BHT associated with BTB</td>
</tr>
<tr>
<td>Intel P6</td>
<td>mid '96</td>
<td>90%</td>
<td>2-level adaptive with BTB</td>
</tr>
<tr>
<td>PowerPC750</td>
<td>mid '97</td>
<td>90%</td>
<td>BHT associated with BTIC</td>
</tr>
<tr>
<td>MC68060</td>
<td>mid '94</td>
<td>90%</td>
<td>BHT associated with BTIC</td>
</tr>
<tr>
<td>DEC Alpha</td>
<td>early '97</td>
<td>95%</td>
<td>Hybrid 2-level adaptive associated with I-cache</td>
</tr>
<tr>
<td>HP PA8000</td>
<td>early '96</td>
<td>80%</td>
<td>BHT associated with BTB</td>
</tr>
<tr>
<td>SUN UltraSparc</td>
<td>mid '95</td>
<td>88%int</td>
<td>BHT associated with I-cache</td>
</tr>
<tr>
<td></td>
<td></td>
<td>94%FP</td>
<td>BHT = One Level</td>
</tr>
</tbody>
</table>

**S+D**: Uses both static (ISA supported) and dynamic branch prediction
Intel Pentium

- It uses a single-level 2-bit Smith algorithm BHT associated with a four way associative BTB which contains the branch history information.
- The Pentium does not fetch non-predicted targets and does not employ a return address stack (RAS) for subroutine return addresses.
- It does not allow multiple branches to be in flight at the same time.
- Due to the short Pentium pipeline the misprediction penalty is only three or four cycles, depending on what pipeline the branch takes.
Intel P6, II, III

- Like Pentium, the P6 uses a BTB that retains both branch history information and the predicted target of the branch. However the BTB of P6 has 512 entries reducing BTB misses. Since the
- The average misprediction penalty is 15 cycles. Misses in the BTB cause a significant 7 cycle penalty if the branch is backward.
- To improve prediction accuracy a two-level branch history algorithm is used.
- Although the P6 has a fairly satisfactory accuracy of about 90%, the enormous misprediction penalty should lead to reduced performance. Assuming a branch every 5 instructions and 10% mispredicted branches with 15 cycles per misprediction the overall penalty resulting from mispredicted branches is 0.3 cycles per instruction. This number may be slightly lower since BTB misses take only seven cycles.
AMD K6

- Uses a two-level adaptive branch history algorithm implemented in a BHT (gshare) with 8192 entries (16 times the size of the P6).
- However, the size of the BHT prevents AMD from using a BTB or even storing branch target address information in the instruction cache. Instead, the branch target addresses are calculated on-the-fly using ALUs during the decode stage. The adders calculate all possible target addresses before the instruction is fully decoded and the processor chooses which addresses are valid.
- A small branch target cache (BTC) is implemented to avoid a one cycle fetch penalty when a branch is predicted taken.
- The BTC supplies the first 16 bytes of instructions directly to the instruction buffer.
- Like the Cyrix 6x86 the K6 employs a return address stack (RAS) for subroutines.
- The K6 is able to support up to 7 outstanding branches.
- With a prediction accuracy of more than 95% the K6 outperformed all other microprocessors when introduced in 1997 (except the Alpha).
The K6 Instruction Buffer

32-K byte Level-One Instruction Cache

16 Bytes

Branch Target Cache 16x16 Bytes

16 Bytes

2:1

Branch Target Address Adders

Return Address Stack 16x16 Bytes

Fetch Unit

16 Instruction Bytes plus
16 Sets of Predecode Bits

Instruction Buffer
Motorola PowerPC 750

- A **dynamic branch prediction** algorithm is combined with **static branch prediction** which enables or disables the dynamic prediction mode and predicts the outcome of branches when the dynamic mode is disabled.
- Uses a **single-level Smith** algorithm 512-entry BHT and a 64-entry Branch Target Instruction Cache (BTIC), which contains the most recently used branch target instructions, typically in pairs. When an instruction fetch does not hit in the BTIC the branch target address is calculated by adders.
- The return address for subroutine calls is also calculated and stored in user-controlled special purpose registers.
- The PowerPC 750 supports up to two branches, although instructions from the second predicted instruction stream can only be fetched but not dispatched.
The HP PA 8000

- The HA PA 8000 uses static branch prediction combined with dynamic branch prediction.
- The static predictor can turn the dynamic predictor on and off on a page-by-page basis. It usually predicts forward conditional branches as not taken and backward conditional branches as taken.
- It also allows compilers to use profile based optimization and heuristic methods to communicate branch probabilities to the hardware.
- Dynamic bench prediction is implemented by a single-level 256-entry BHT where each entry is a three bit shift register which records the outcome of the last three branches instead of saturated up and down counters. The outcome of a branch (taken or not taken) is shifted in the register as the branch instruction retires.
- To avoid a taken branch penalty of one cycle the PA 8000 is equipped with a Branch Target Address Cache (BTAC) which has 32 entries.
The HP PA 8000 Branch Prediction Algorithm

Not Taken

Taken
The SUN UltraSparc

- Uses a dynamic single-level BHT Smith algorithm.
- It employs a static prediction which is used to initialize the state machine (saturated up and down counters).
- However, the UltraSparc maintains a large number of branch history entries (up to 2048 or every other line of the I-cache).
- To predict branch target addresses a branch following mechanism is implemented in the instruction cache. The branch following mechanism also allows several levels of speculative execution.
- The overall claimed performance of UltraSparc is 94% for FP applications and 88% for integer applications.
The Alpha 21264

- The Alpha 21264 uses a two-level adaptive hybrid method combining two algorithms (a global history and a per-branch history scheme) and chooses the best according to the type of branch instruction encountered.

- The prediction table is associated with the lines of the instruction cache. An I-cache line contains 4 instructions along with a next line and a set predictor.

- If an I-cache line is fetched that contains a branch the next line will be fetched according to the line and set predictor. For lines containing no branches or unpredicted branches the next line predictor point simply to the next sequential cache line.

- This algorithm results in zero delay for correct predicted branches but wastes I-cache slots if the branch instruction is not in the last slot of the cache line or the target instruction is not in the first slot.

- The misprediction penalty for the alpha is 11 cycles on average and not less than 7 cycles.

- The resulting prediction accuracy is about 95%.

- Supports up to 6 branches in flight and employs a 32-entry return address stack (RAS) for subroutines.
The Basic Alpha 21264 Pipeline

Transit  Calc PC

Mispredicted branch penalty: 7 cycles minimum

Possible delay in instr queue

Fetch  Transit  Map  Queue  Register  Execute  Write
Access I-cache  Send to decoder  Rename registers  Insert in queue  Read operands  Integer calc  Write result

Load-use penalty: 2 cycles minimum

Register  Address  Cache1  Cache2  Write
Read operands  Calculate address  Access D-cache  Get result, check tags  Write result

Register  FP1  FP2  FP3  FP4  Write
Read operands  Start FP op  FP op  FP op  Round result  Write result
Alpha 21264 Branch Hybrid Prediction

Program Counter

P1

Local History Table
1,024 x 10 bits

Local Predict
1,024 x 3 bits

Local Prediction

Global History

P2

Global Predict
4,096 x 2 bits

Global Prediction

gshare

Choice Predict
4,096 x 2 bits

12

Final Prediction