Dynamic Branch Prediction

Readings: 4.8

Branches introduce control hazards
    Determine the right next instruction in time for instruction fetch

Previous solutions:
    Stall
    Statically predict not taken
    Branch Delay slot

Better:
    Branch-prediction buffers (caches)

Problems With Static Branch Predictors

Are all conditional branches created equal?
Branch Prediction Buffer

Direct-mapped cache w/1-bit history
Predict taken/not taken by previous execution
If incorrect prediction, annul instructions incorrectly started

Tags?
Valid bits?

Thought Experiment

Consider the following code segment:

```
while (1) {
    <code 1>
    for(int i=0; i<9; i++) {
        <code 2>
    }
    <code 3>
}
```

1-bit prediction accuracy?

```
ADDI C, X31, #9 // Const. 9
WHILE_TOP:
    <code 1>
    ADD i, X31, X31 // Init i
    B FOR_TEST
FOR_TOP:
    <code 2>
    ADDI i, i, #1 // i++
FOR_TEST:
    CMP i, C // i < 9?
    B.LT FOR_TOP
    <code 3>
    B WHILE_TOP // Endwhile
```
2-bit Predictor

while (1) {
    if (normal_condition) {
        <code1>
    }
    for(int i=0; i<9; i++) {
        if (exception) {
            return (FALSE);
        }
        <code 2>
    }
    if (random 50/50 chance) {
        <code 3>
    }
}

# of predictors

Branch Predictor

2

# of predictors
Instruction-Level Parallelism & Advanced Architectures

Readings: 4.10

Key to pipelining was dealing with hazards

Advanced processors require significant hazard avoidance/flexibility

ILP = Instruction-Level Parallelism

Why ILP

Advanced processors optimize two factors:

- Reduce clock period by heavy pipelining
  - Greater pipelining means more hazards, delay slots

- Reduce CPI
  - What if we want a CPI < 1.0?
ILP Example

Source code:

```
ADDI X0, X0, #15
SUB  X2, X1, X0
EORI X3, X0, #15
ORR  X4, X3, X0
```

Constraint graph:

---

Complex ILP Example

Note: Assume no delay slots.

```
1: LDUR X0, [X3, #0]
2: ADD  X1, X0, X2
3: SUB  X2, X3, X4
4: ANDI X2, X5, #57
5: ORR  X7, X5, X1
6: STUR X5, [X9, #0]
7: CBZ  X7, LOOP
8: EOR  X6, X8, X6
```
Types of Hazards

Data Hazards (multiple uses of the same location)
- RAW – Read after write
- WAW – Write after write
- WAR – Write after read
- RAR – Read after read

Memory

Control Hazards
- Branches

Hazard Minimization

Hardware and software techniques for improving performance
- Loop unrolling
- Register Renaming
- Predicated Instructions

Responsibility
- Hardware
- Software (Compiler)
Loop-Level Parallelism

LLPex(char *src) {
    char *end = src + 1000;
    while(src<end) {
        *src = (*src)+3; src++;
    }
}

Loop Unrolling

Better loop structure?
Compiler Register Renaming

Compiler reduces hazards by removing name dependences

Hardware Register Renaming

CPU dynamically associates each register read with the most recent instruction issue that writes that register.

0: ADDI X0, X0, #4
1: LDUR X1, [X0, #0]
2: STUR X1, [X0, #100]
3: ADDI X0, X0, #4
4: LDUR X1, [X0, #0]
5: STUR X1, [X0, #100]
6: ADDI X0, X0, #4
7: LDUR X1, [X0, #0]
8: STUR X1, [X0, #100]
Control Hazards

if (c == 0) { t = s; }
if (a[0] == 0)
    a[0] = b[0];
else
    a[0] = a[0] + 4;

Branches introduce hazards that limit ILP
Branch prediction

Conditional/Predicated instructions
Predicated Instructions

if (c == 0) { t = s; }

Normal:

W/Conditional move (instructions with internal if-like operation – no branches)
CMOVZ <dest>, <src>, <cond> // move src to dest if cond == 0
CMOVNZ <dest>, <src>, <cond> // move src to dest if cond != 0

Note: ARM actually uses CSEL, a Mux-like instruction instead. But, predication important enough we’ll pretend there’s a CMOV instruction…

Predicated Instructions (cont.)

d = a - b;
if (d<0)
    sum -= d;
else
    sum += d;

Normal:

Predicated Instructions:
ARM CSEL Instruction: the MUX

\[
\text{CSEL } \langle \text{dest} \rangle, \langle \text{src1} \rangle, \langle \text{src2} \rangle, \langle \text{cond} \rangle \\
\langle \text{dest} \rangle = \text{if } \langle \text{cond} \rangle \text{ then } \langle \text{src1} \rangle \text{ else } \langle \text{src2} \rangle \\
\langle \text{cond} \rangle: \text{EQ, NE, LT, etc.}
\]

\[
d = a - b; \\
\text{if } (d<0) \\
\quad \text{sum }-= d; \\
\text{else} \\
\quad \text{sum }+= d;
\]

Why ILP

Advanced processors optimize two factors:

- Reduce clock period by heavy pipelining
  Greater pipelining means more hazards, delay slots

- Reduce CPI
  What if we want a CPI < 1.0?
**Superpipelining**

Divide datapath into **multiple** pipeline stages

Pentium 4 Processor ("NetBurst"): 20 stages

---

**Multiple Issue**

Replicate Execution Units

ALU
Shifter
Floating Point
Register File
Memory
Ifetch/Branch
VLIW

Very Long Instruction Word

<table>
<thead>
<tr>
<th>ALU</th>
<th>ALU</th>
<th>Load/Store</th>
<th>Branch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Max CPI?

Concerns

VLIW Scheduling

Schedule the code for a 4-way VLIW. Assume no delay slots, and all instructions in parallel with a branch still execute.

1: LDUR X6, [X0, #4]
2: ADD X7, X6, X0
3: LDUR X8, [X1, #8]
4: SUB X7, X8, X2
5: STUR X10, [X3, #0]
6: CBZ X4, FOO
7: AND X12, X4, X5

<table>
<thead>
<tr>
<th>ALU1</th>
<th>ALU2</th>
<th>Load/Store</th>
<th>Branch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ALU1</th>
<th>ALU2</th>
<th>Load/Store</th>
<th>Branch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ALU1</th>
<th>ALU2</th>
<th>Load/Store</th>
<th>Branch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ALU1</th>
<th>ALU2</th>
<th>Load/Store</th>
<th>Branch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Superscalar

Dynamically schedule multiple instructions, based on hazard detection

Example Execution on Modern Processors

Bringing everything together, how would this code be run on an advanced CPU? No delay slots, instructions in parallel with a branch still execute

0: LDUR X0, [X1, #8]
1: ADDI X2, X0, #4
2: STUR X2, [X1, #8]
3: CBZ X1, SKIP
4: ADDI X3, X4, #0
SKIP:
5: ADDI X5, X5, #1
6: EORI X0, X1, #5
7: LSL X0, X0, #2
8: STUR X0, [X1, #16]
Example on VLIW

How would this be scheduled on a 2-way VLIW? Assume no delay slots.

<table>
<thead>
<tr>
<th>ALU/Load/Store</th>
<th>ALU/Branch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example on Superscalar

How would this execute on a 2-way superscalar? Assume it schedules the earliest ready instruction first, and cache miss only stalls dependent instructions.

<table>
<thead>
<tr>
<th>ALU/Load/Store</th>
<th>ALU/Branch</th>
</tr>
</thead>
<tbody>
<tr>
<td>No miss</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>2-cycle miss</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Symmetric Multi-Threading (SMT)

~Superscalar drawing from multiple programs or program threads

<table>
<thead>
<tr>
<th>Instr L + 7</th>
<th>Instr L + 6</th>
<th>Instr L + 5</th>
<th>Instr L + 4</th>
<th>Instr L + 3</th>
<th>Instr L + 2</th>
<th>Instr L + 1</th>
<th>Instr L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instr K + 7</td>
<td>Instr K + 6</td>
<td>Instr K + 5</td>
<td>Instr K + 4</td>
<td>Instr K + 3</td>
<td>Instr K + 2</td>
<td>Instr K + 1</td>
<td>Instr K</td>
</tr>
<tr>
<td>Instr J + 7</td>
<td>Instr J + 6</td>
<td>Instr J + 5</td>
<td>Instr J + 4</td>
<td>Instr J + 3</td>
<td>Instr J + 2</td>
<td>Instr J + 1</td>
<td>Instr J</td>
</tr>
<tr>
<td>Instr I + 7</td>
<td>Instr I + 6</td>
<td>Instr I + 5</td>
<td>Instr I + 4</td>
<td>Instr I + 3</td>
<td>Instr I + 2</td>
<td>Instr I + 1</td>
<td>Instr I</td>
</tr>
</tbody>
</table>

Schedule/Dispatch
Hazard Detection

Mem
PC

Issues

Multicore

Add multiple processors (separate control & datapath)

<table>
<thead>
<tr>
<th>Computer</th>
<th>Memory</th>
<th>Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processor 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processor 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processor 4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Input
Output

Issues
Multicore Example

Using a multicore is significantly more complex.

Example: Uniprocessor MAX

```c
int max(int vals[], int len) {
    int result = -infinity;
    for (int i=0; i<len; i++) {
        if (vals[i] > result)
            result = vals[i];
    }
    return result;
}
```

Multicore MAX

```c
int max(int vals[], int len) {
    int global_result = -infinity;
    int lenT = len/num_procs;
    for (int i=0; i<num_procs; i++)
        process maxT(&vals[i*lenT], lenT);
    }
}
```

```c
void maxT(int vals[], int len) {
    .
    .
    .
}
```