Speculative Execution with Late Recovery

2 Speculative Execution with Early Recovery
   2.1. Adding Speculative Bits
   2.2. Adding Rename-Table Snapshots

3 Complete Out-of-Order Superscalar PARCv2 Processor
1. Speculative Execution with Late Recovery

- Every instruction is actually speculative because an older in-flight instruction might cause an exception

- We recover from exceptions at the commit point (C-stage) which is late in the pipeline

- With out-of-order load/store issue, loads (and dependent instructions) are also speculative

- We recover from incorrect speculation in the C stage which is late in the pipeline
1. Speculative Execution with Late Recovery

- Branches also require speculative execution
- Recover mispredictions late in the pipeline?

<p>| | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>lw</td>
<td>r1, 0(r2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>mul</td>
<td>r3, r1, r4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>sw</td>
<td>r3, 0(r5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>addiu</td>
<td>r2, r2, 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>addiu</td>
<td>r5, r5, 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>addiu</td>
<td>r6, r6, -1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>bne</td>
<td>r6, r0, loop</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Branches are far more common than exceptions and memory-dependence violations
- Accurate branch prediction helps, but some branches are just inherently difficult to predict
- **Key Idea:** Recover from branch mispredictions as soon as possible
2. Speculative Execution with Early Recovery

We will explore early recovery in two steps:

- Adding speculative bits
- Adding rename-table snapshots

2.1. Adding Speculative Bits

- Add a speculative bit to the IQ, ROB, FSB, FLB, and functional units
- Add a speculative mode bit in the D stage

- In D stage for a branch
  - Set speculative mode bit
  - All inst after branch carry speculative bit into IQ, ROB, FSB, LB, func units

- In X stage for a correctly predicted branch
  - Broadcast clear speculative bit from X stage to all data structures

- In X stage for an incorrectly predicted branch
  - Broadcast squash signal from X stage to all of these data structures
  - Each data structure invalidates entry/inst for which speculative bit is set
  - Start fetching from correct address

- Multiple speculative enable multiple speculative branches in flight
  - Given instruction can be squashed by multiple branches
  - Treat multiple speculative bits as “branch mask”
Do not copy ARF into PRF on branch misprediction recovery

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>addiu r1, r2, 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>branch L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>addiu r1, r3, 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>opA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>opB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>opC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>opD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>L1: addiu r4, r1, 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Copy ARF into PRF on branch misprediction recovery

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>addiu r1, r2, 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>addiu r1, r3, 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>addiu r4, r1, 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>branch L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>opA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>opB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>opC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>opD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td>L1: addiu r5, r6, 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Need to make copy of “precise” ARF in D on every branch ...
- ... but ARF is not precise in D
- Need “view” of what precise ARF would be in D on every branch ...
- ... this is the rename table!
2.2. Adding Rename-Table Snapshots

- Add a speculative bit to the IQ, ROB, FSB, FLB, and functional units
- Add a speculative mode bit in the D stage
- Add a rename table snapshot in the D stage

- In D stage for a branch
  - Set speculative mode bit
  - All inst after branch carry speculative bit into IQ, ROB, FSB, LB, func units
  - Create a RT snapshot to save “view” of precise ARF for branch

- In X stage for a correctly predicted branch
  - Broadcast clear speculative bit from X stage to all data structures

- In X stage for an incorrectly predicted branch
  - Broadcast squash signal from X stage to all of these data structures
  - Each data structure invalidates entry/inst for which speculative bit is set
  - Restore RT from snapshot
  - Start fetching from correct address

- Need multiple speculative bits and multiple snapshots to support multiple speculative branches in flight
RT snapshots squash speculative state

```
a: addiu r1, r2, 1
b: branch L1
c: addiu r1, r3, 1
d: opA
e: opB
f: opC
g: opD
h: L1: addiu r4, r1, 1
```

RT snapshots prevent overwriting non-speculative state

```
a: addiu r1, r2, 1
b: addiu r1, r3, 1
c: addiu r4, r1, 1
d: branch L1
e: opA
f: opB
g: opC
h: opD
i: L1: addiu r5, r6, 1
```
3. Complete Out-of-Order Superscalar PARCv2 Processor

- **Superscalar execution**: two-way every stage, aligned fetch blocks
- **Out-of-order execution**: IO2L with IQ and ROB
- **Register renaming**: pointer-based scheme with URF and ART
- **Memory disambiguation**: OOO load/store issue with FSB and FLB
- **Branch prediction**: BTB with generalized two-level BHT
- **Speculative execution**: speculative bits with rename table snapshots

### Vector-Vector Add Microbenchmark

<table>
<thead>
<tr>
<th>Microarchitecture</th>
<th>cycles/itr</th>
<th>actual CPI</th>
<th>actual IPC</th>
<th>peak IPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-Order Single-Issue PARCv1</td>
<td>12</td>
<td>1.33</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>In-Order Dual-Issue PARCv1</td>
<td>10</td>
<td>1.11</td>
<td>0.90</td>
<td>2</td>
</tr>
<tr>
<td>Out-of-Order Dual-Issue PARCv1</td>
<td>5</td>
<td>0.55</td>
<td>1.80</td>
<td>2</td>
</tr>
</tbody>
</table>