An Overview of the Alpha AXP™ 21164 Micro-Architecture

The World’s Highest Performance Microprocessor

John Edmondson and Paul Rubinfeld
Digital Equipment Corporation
Semiconductor Engineering Group
Hudson, MA

digital

Alpha AXP 21164 Overview

Key Attributes

- 4-way issue superscalar
- Large on-chip L2 cache
- 7-stage integer pipeline
- 9-stage floating point pipeline
- Emphasis on low latency at high clock rate
- High-throughput memory subsystem
Instruction Issue Pipeline

Instruction Prefetching

- Aggressive prefetching from L2 cache using high-bandwidth capability
  - At least three 32-byte blocks ahead of the current issue point
  - Continuous integer instruction issue possible out of L2 cache (2 per cycle)
  - 60% of peak issue rate possible out of L2 cache (2.4 per cycle)
### Execution Pipeline

- Int mul
- Integer Pipeline 0: arith, logical, ld/st, shift
- Integer Pipeline 1: arith, logical, ld, br/jmp
- FP div
- FP Pipeline 0: add, subtract, compare, FP br
- FP Pipeline 1: multiply

### Instruction Latency

<table>
<thead>
<tr>
<th>Instruction Type</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most integer ops</td>
<td>1</td>
</tr>
<tr>
<td>CMOV</td>
<td>2</td>
</tr>
<tr>
<td>Integer multiply</td>
<td>8-16</td>
</tr>
<tr>
<td>Floating point ops</td>
<td>4</td>
</tr>
<tr>
<td>Loads (L1 cache hit)</td>
<td>2</td>
</tr>
<tr>
<td><strong>Special Case Bypass</strong></td>
<td></td>
</tr>
<tr>
<td>CMOV or conditional BR dependent on a compare or logical operation</td>
<td>0</td>
</tr>
<tr>
<td><strong>Example:</strong> CMP R1, R2, R3</td>
<td></td>
</tr>
<tr>
<td>BEQ R3, LABEL</td>
<td></td>
</tr>
</tbody>
</table>
High-Throughput Load Execution

Addr from integer pipeline 0
Dual-Ported L1 Data Cache (8Kbyte, write thru)

Addr from integer pipeline 1
Miss Address File (MAF) 6 entry

Bus Address File (BAF) 2 entry
On-Chip L2 Cache (96Kbyte, 3-way set assoc., writeback, pipelined)

Miss Address File Details

4

6

<table>
<thead>
<tr>
<th>Address</th>
<th>Rn</th>
<th>Rn</th>
<th>Rn</th>
<th>Rn</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

- MAF merges loads to the same cache block
- Up to 21 loads
- Multiple loads merge, regardless of order
- Up to two register file fills per cycle
L3 Cache (off-chip)

- L3 cache is a direct-mapped writeback superset of on-chip L2 cache
- Up to 2 reads (or outstanding read commands) in L3 cache
- Programmable wave pipelining for L3 cache
- L3 cache is optional

Latency & Bandwidth of Memory Operations

<table>
<thead>
<tr>
<th></th>
<th>Latency (cycles)</th>
<th>Bandwidth (bytes/cycle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Data Cache</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>L2 Cache</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>L3 Cache</td>
<td>≥12</td>
<td>≤4</td>
</tr>
</tbody>
</table>

- L1 cache block size is 32 bytes
- L2 and L3 cache block sizes are each 64 bytes (with a 32-byte block size option)
Improvements Over the Previous Generation

- Reduced key latencies

<table>
<thead>
<tr>
<th></th>
<th>21164</th>
<th>21064/21064A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shift/byte ops</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Integer multiply</td>
<td>8-16</td>
<td>19-23</td>
</tr>
<tr>
<td>CMP → BR</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>FP latency</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>L1 data cache</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

- Wider issue rate
  - 4 vs. 2

- Cycle time improvement
  - Greater than simple technology scaling

Estimated Performance Results

- Better than 1 SPECint92 per MHz
- Better than 1.5 SPECfp92 per MHz
- Better than 2 TPS per MHz
Alpha AXP Processor Road Map

Summary

- The Alpha AXP 21164 is totally new design
  - Quad instruction issue
  - On-chip secondary cache
  - Achieves short latency at a high speed clock

- It contains significant micro-architecture and circuit advances over the first implementation

- This chip is the world's highest performance microprocessor