Techniques for Mitigating Memory Latency Effects in the PA-8500 Processor

David Johnson
Systems Technology Division
Hewlett-Packard Company
Presentation Overview

• PA-8500 Overview
• Instruction Fetch Capabilities
• Reorder Buffers ("The Queue")
• Data Cache
• System Bus
PA-8500 Processor Core

Instruction Fetch Unit
BHT, BTAC

System Bus Interface

Inst. Cache

Dual 64-bit Integer ALUs

Dual Shift/ Merge Units

Dual FP Multiply/ Accumulate Units

Dual FP Divide/ SQRT Units

Rename Registers

ALU Buffer 28 entries

Memory Buffer 28 entries

Sort

Retire

Architected Registers

Rename Registers

Address Reorder Buffer 28 entries

Dual Load/Store Address Adders

TLB

Data Cache

Runway bus
Memory Latency

Year 1990 1995 2000
Speed (MHz) 10 100 1000

Latency Problems
Instruction Fetches & Loads

Techniques for Hiding Latency
High hit-rate caches
Prefetching
Overlapping cache misses

Hewlett Packard
Instruction Fetch Features

- Instruction Cache
  - 0.5 MB on-chip cache
  - 4-way set associative
  - Pipelined 2-cycle access
  - Provides 4 instructions per cycle to CPU core
  - Supports 32-byte and 64-byte line sizes
- Instruction Prefetching
PA-8500 I-Cache Composition

4 Instructions per cycle to Queue from a 0.5 MB cache
PA-8500 Instruction Prefetching

1. I-Miss from cache
2. I-Miss issued to Runway Bus
3. I-Prefetch issued to Runway Bus
4. I-Miss Return inserted into cache
5. I-Prefetch Return held in Prefetch Buffer
6. I-Miss from Cache causes Prefetch Buffer Hit
7. I-Miss moved from Prefetch Buffer to Cache
8. I-Prefetch issued to Runway Bus (next line)
Reorder Buffers

Cycle by cycle progression of a load instruction

<table>
<thead>
<tr>
<th>Insert</th>
<th>Launch</th>
<th>Address</th>
<th>Cache</th>
<th>Cache</th>
<th>RR</th>
<th>Retire</th>
</tr>
</thead>
</table>
LOAD-MISS Overlapping

The Problem

PA-8500 Solution
Address Reorder Buffer: High-Speed Custom Circuitry

- 28 Entries

- Front-end Address Reorder Buffer

- Matches grantL requests

- 28 Entry Cache port arbitration circuits

- Insert Addresses from ALU to Cache

- Launch Addresses to Runway

- Miss Addresses

- 0-catcher
Data Prefetching

The Problem
Avoid the LOAD-MISS latency

Solution
Compiler inserts Prefetch instruction (LOAD to GR0)
Independent instructions executed ("Instr")
Data is resident in cache for LOAD (LOAD-HIT)
Data Cache Features

- 1.0 MB on-chip cache
- 4-way set associative
- 2-cycle pipelined access
- Two accesses per cycle
- Supports 32-byte and 64-byte line sizes
- Sophisticated Store Queue
Data Cache
Single-Level vs. Multi-Level Cache Designs

1.5 MB @ 2 cycles

Core

1 MB Data Cache

0.5 MB Instr Cache

System Bus

32 KB (each) @ 1 cycle

L1 Data

L1 Instr

Core

Off-chip L2

4 MB @ ~15 cycles

System Bus
System Bus Interface

- Split-transaction bus with out-of-order returns
- Multiple transactions in flight simultaneously
- Priority given to latency-sensitive transactions
- Asynchronous Interface
- Turbo Mode
Turbo Mode

High-Speed Data Transfer between Memory and CPU
Mitigating Memory Latency Effects

- Large Caches
- Out-of-Order Queue
- Flexible System Interface
- Custom Circuit Design

The PA-8500 Achieves Superb Performance!