Simultaneous Multithreading: Multiplying Alpha Performance

Dr. Joel Emer
Principal Member Technical Staff
Alpha Development Group
Compaq Computer Corporation
Outline

- Alpha Processor Roadmap
- Motivation for Introducing SMT
- Implementation of an SMT CPU
- Performance Estimates
- Architectural Abstraction
Alpha Microprocessor Overview

Higher Performance

Lower Cost

1998  1999  2000  2001  2002  2003

First System Ship

21264 EV6

0.35 \( \mu \text{m} \)

21264 EV67

0.28 \( \mu \text{m} \)

21264 EV68

0.18 \( \mu \text{m} \)

EV7

0.18 \( \mu \text{m} \)

EV78

0.125 \( \mu \text{m} \)

EV8

0.125 \( \mu \text{m} \)

First System Ship
EV8 Technology Overview

- Leading edge process technology – 1.2-2.0GHz
  - 0.125µm CMOS
  - SOI-compatible
  - Cu interconnect
  - low-k dielectrics

- Chip characteristics
  - ~1.2V Vdd
  - ~250 Million transistors
  - ~1100 signal pins in flip chip packaging

www.compaq.com
EV8 Architecture Overview

- Enhanced out-of-order execution
- 8-wide superscalar
- Large on-chip L2 cache
- Direct RAMBUS interface
- On-chip router for system interconnect
- Glueless, directory-based, ccNUMA for up to 512-way SMP
- 4-way simultaneous multithreading (SMT)
Goals

- Leadership single stream performance

- Extra multistream performance with multithreading
  - Without major architectural changes
  - Without significant additional cost
Instruction Issue

Reduced function unit utilization due to dependencies
Superscalar leads to more performance, but lower utilization
Predicated Issue

Time

Adds to function unit utilization, but results are thrown away
Chip Multiprocessor

Limited utilization when only running one thread
Fine Grained Multithreading

Intra-thread dependencies still limit performance
Simultaneous Multithreading

Maximum utilization of function units by independent operations
Basic Out-of-order Pipeline

Fetch Decode / Map Queue Reg Read Execute Dcache / Store Buffer Reg Write Retire

Thread-blind
## SMT Pipeline

<table>
<thead>
<tr>
<th>fetch</th>
<th>decode / map</th>
<th>queue</th>
<th>reg read</th>
<th>execute</th>
<th>dcache / store buffer</th>
<th>reg write</th>
<th>retire</th>
</tr>
</thead>
<tbody>
<tr>
<td>icache</td>
<td>register map</td>
<td>pc</td>
<td>reg map</td>
<td>regs</td>
<td>dcache</td>
<td>regs</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><a href="http://www.compaq.com">www.compaq.com</a></td>
</tr>
</tbody>
</table>
Changes for SMT

- Basic pipeline – unchanged
- Replicated resources
  - Program counters
  - Register maps
- Shared resources
  - Register file (size increased)
  - Instruction queue
  - First and second level caches
  - Translation buffers
  - Branch predictor
Multiprogrammed workload
Decomposed SPEC95 Applications

![Decomposed SPEC95 Applications](chart.png)
Multithreaded Applications

- Barnes
- Chess
- Sort
- TP

- 1T
- 2T
- 4T
Architectural Abstraction

- 1 CPU with 4 Thread Processing Units (TPUs)
- Shared hardware resources
Quiescing Idle Threads

Problem:
Spin looping thread consumes resources

Solution:
Provide quiescing operation that allows a TPU to sleep until a memory location changes
Summary

- Alpha will maintain single stream performance leadership

- SMT will significantly enhance multistream performance
  - Across a wide range of applications,
  - Without significant hardware cost, and
  - Without major architectural changes
References

- "Simultaneous Multithreading: Maximizing On-Chip Parallelism" by Tullsen, Eggers and Levy in ISCA95.

- "Exploiting Choice: Instruction Fetch and Issue on an Implementable Simultaneous Multithreaded Processor" by Tullsen, Eggers, Emer, Levy, Lo and Stamm in ISCA96.
