Advanced Processor Architecture

Jinkyu Jeong (jinkyu@skku.edu)
Computer Systems Laboratory
Sungkyunkwan University
http://csl.skku.edu

Advacned Processor Architecture
Modern Microprocessors

- More than just GHz

<table>
<thead>
<tr>
<th>CPU</th>
<th>Clock Speed</th>
<th>SPECint2000</th>
<th>SPECfp2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athlon 64 FX-55</td>
<td>2.6GHz</td>
<td>1854</td>
<td>1782</td>
</tr>
<tr>
<td>Pentium 4 Extreme Edition</td>
<td>3.46GHz</td>
<td>1772</td>
<td>1724</td>
</tr>
<tr>
<td>Pentium 4 Prescott</td>
<td>3.8GHz</td>
<td>1671</td>
<td>1842</td>
</tr>
<tr>
<td>Opteron 150</td>
<td>2.4GHz</td>
<td>1655</td>
<td>1644</td>
</tr>
<tr>
<td>Itanium 2 9MB</td>
<td>1.6GHz</td>
<td>1590</td>
<td>2712</td>
</tr>
<tr>
<td>Pentium M 755</td>
<td>2.0GHz</td>
<td>1541</td>
<td>1088</td>
</tr>
<tr>
<td>POWER5</td>
<td>1.9GHz</td>
<td>1452</td>
<td>2702</td>
</tr>
<tr>
<td>SPARC64 V</td>
<td>1.89GHz</td>
<td>1345</td>
<td>1803</td>
</tr>
<tr>
<td>Athlon 64 3200+</td>
<td>2.2GHz</td>
<td>1080</td>
<td>1250</td>
</tr>
<tr>
<td>Alpha 21264C</td>
<td>1.25GHz</td>
<td>928</td>
<td>1019</td>
</tr>
</tbody>
</table>
Pipelining

• Sequential execution

• Pipelining (RISC)
Superpipelining

- Superpipelining
  - Subdivide each pipeline stage
  - Higher clock speed
Superscalar

- Superscalar
  - The execution stage has a bunch of different functional units
  - Execute multiple instructions in parallel
  - Pentium: 2-way superscalar
Superpipelined Superscalar

- Superpipelining + Superscalar
  - 2-way: MIPS R5000
  - 3-way: PowerPC G3/G4, Pentium Pro/II/III/M/4, Athlon
  - 4-way: UltraSparc, MIPS R10000, PowerPC G4e, Alpha 21164 & 21264, Core 2 Duo
  - 5-issue: PowerPC G5
Tackling Instruction Dependencies

• Branch prediction + speculative execution
  – Mispredict penalty: 10 – 15 cycles in Pentium Pro/II/III

• Instruction scheduling
  – In-order execution + compiler optimization
    • Rearrange the instructions at compile time
    • Compiler can see further down the program than the hardware
    • SuperSparc, HyperSparc, UltraSparc, Alpha 21064 & 21164
  – Out-of-order execution
    • Reorder instruction execution sequence in hardware at run time
    • Register renaming reduces the dependency further
    • MIPS R10000, Alpha 21264, POWER/PowerPC, Pentium Pro, Pentium 4, Core 2 Duo, Core i7, …
Intel Pentium Pro

- **In-order front-end**
  - Multiple branch prediction
  - Micro-operations
  - Register renaming

- **Out-of-order execution core**
  - 3-way superscalar
  - Multiple execution units
  - Dataflow analysis
  - Speculative execution

- **In-order retirement**
  - Precise faulting semantics
P6 Microarchitecture

```
Instr TLB (32 entry)
Branch Target Buffer
Instruction Fetch Unit

IN-ORDER SECTION

8K Instruction Cache
Simple Decoder
Simple Decoder
General Decoder
Uop Sequencer
Reorder Buffer (40 entries)
RAT | RRF

Reservation Station (20 entries)
Store Data
Store Addr Unit
Load Addr Unit
Integer ALU
FP Unit
Integer Unit
Memory Reorder Buffer (MOB)

OUT-OF-ORDER EXECUTION ENGINE

8K Dual-Ported Data Cache
System Bus Interface
L2 Cache Interface

Data TLB (64 entry)

load data
1 store
1 load
32
8
64
64 data
64 data
36 addr
```
Hyper-Threading

- Simultaneous multithreading technology (SMT)
  - Utilizes thread-level parallelism
  - Fill pipelines with the instructions from multiple threads running at the same time
  - An SMT processor appears as if it were multiple independent processors
  - Uses processor resources more effectively
  - Cost: <5% in added die area
Multi-core

• Put two or more processor cores onto a single chip
  – Previously called CMP (Chip Multiprocessor)

• Examples
  – AMD dual-core Athlon 64 X2: dual-core (May 2005)
  – Intel Core Duo, Core 2 Duo: dual-core
  – Sun UltraSparc T1: eight-core, 32 threads (Nov. 2005)
  – Intel Xeon X7460: six-core (Sep. 2008)
  – Intel Xeon E7-8890 v4: 24-core (Jun. 2016)
CPU Trends
Why Multi-core?

• Memory wall
  – CPU 55%/year, Memory 10%/year (1986 – 2000)
  – Caches show diminishing returns

• ILP (Instruction Level Parallelism) wall
  – Control dependency
  – Data dependency

• Power wall
  – Dynamic power $\propto$ Frequency$^3$
  – Static power $\propto$ Frequency
  – Total power $\propto$ The number of cores
Single-core vs. Multi-core

- **Single-Core**
  - Performance: 1.00x
  - Power: 1.02x
  - Source: Intel

- **Multi-Core**
  - Performance: 1.73x
  - Power: 0.87x

**More MIPS/watt**

- **Raise Clock (20%)**
  - Performance: 1.73x
  - Power: 1.13x

- **Lower Clock (20%)**
  - Performance: 0.87x
  - Power: 0.51x

Source: Intel