Advanced Processor Architecture

Jin-Soo Kim (jinsookim@skku.edu)
Computer Systems Laboratory
Sungkyunkwan University
http://csl.skku.edu
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

- 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

---

**Clock cycles**

**Inst's**
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
Skylake Microarchitecture

Front End
- Branch Prediction (BFU)
- L1 Instruction Cache
  - 32KB 8-Way
- Instruction Fetch & Prefetching
  - 8-Way Prefetch
- Instruction Queue (4K, 256 entries)
  - 4-Way Instruction
- Microcode
  - RON (RMPCM)
- Decoded Diving Buffer (50KB)
  - 50KB (8-Way)
  - 50KB Output
- Allocation Group (512, 512KB)
  - 512KB Output
  - 512KB Output

Memory
- L1 Cache
  - 32KB 4-Way
- L2 Cache
  - 256KB 4-Way
- L3 Cache
  - 512KB 4-Way

Execution Engine
- Move Elimination
- Rename / Allocate / Retire
- ReOrder Buffer (224 entries)
- Scheduling
  - Unified Reservation Station (UR)
    - 97 entries

Scheduler
- Integer-Physical Register File (158)
- Vector-Physical Register File (158)

ALU
- Add
- Subtract
- Multiply
- Divide
- Branch
- Shift
- Compare

Data TLB
- 64KB/Cycle

Support
- 2x32B/Cycle Read
- 2x32B/Cycle Write

32B/Cycle Store
- L1 Data Cache
  - 32KB 8-Way
- L1 Instruction Cache
  - 32KB 8-Way
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 Opteron: dual-core (Apr. 2005)
  - 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

Stuttering

- Transistors per chip, ’000
- Clock speed (max), MHz
- Thermal design power*, w

Chip introduction dates, selected

Transistors bought per $, m

Sources: Intel; press reports; Bob Colwell; Linley Group; IB Consulting; *The Economist

*Maximum safe power consumption
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

- **Performance**
- **Power**

**Single-Core**
- Raise Clock (20%)
- \(1.73x\) PERFORMANCE
- \(1.00x\) POWER

**Dual-Core**
- Lower Clock (20%)
- \(0.87x\) PERFORMANCE
- \(0.51x\) POWER

**More MIPS/watt**
- \(1.73x\)
- \(1.02x\)

Source: Intel