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

- 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 (1)

- 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 runtime
    » Register renaming reduces the dependency further.
    » MIPS R10000, Alpha 21264, POWER/PowerPC, Pentium Pro, Pentium 4, Core 2 Duo
## Intel Architectures

<table>
<thead>
<tr>
<th>Architecture Brand Name</th>
<th>XScale (IA-64)</th>
<th>IA-32</th>
<th>Intel 64 (IA-32e, EM64T, x86-64)</th>
<th>Itanium (IA-64)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Microarchitecture Brand Name</strong></td>
<td><strong>P5</strong></td>
<td><strong>P6</strong></td>
<td><strong>NetBurst</strong></td>
<td><strong>Mobile</strong></td>
</tr>
<tr>
<td><strong>Processor Brand Name</strong></td>
<td>Pentium Pentium MMX</td>
<td>Pentium Pro Pentium II Pentium III</td>
<td>Pentium 4 Pentium D</td>
<td>Pentium M</td>
</tr>
</tbody>
</table>

### Core
- Core 2 Quad/Duo/Solo
  - Conroe, Allendale, Wolfdale – Merom, Penryn – Kentsfield, Yorkfield
Pentium Integer Pipeline

- 2-way superscalar with 5 stages

PF

D1

D2

D2

E

E

WB

WB

- Fetch and align instruction
- Decode instruction & generate control word
- Check for instruction pairability
- Perform branch prediction
- Decode control word
- Compute memory address
- Access data cache or calculate ALU result
- Write result
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
Intel Pentium 4

- **Netburst microarchitecture**
  - Hyper pipelined for clock rates > 1 GHz
  - Need deeper pipelines
    - Diminishing performance returns
    - Requires deeper buffering to cover the longer pipelines
  - Clock skew, jitter, and latch/wire delay matter
    - Reduced percentage of the clock cycle usable by actual logic
    - Heavily depends on circuit design techniques, process technology, power and thermal constraints, etc.
  - Increases complexity
    - Harder to balance
    - More challenges to architect around
    - Greater validation effort
NetBurst Pipeline

- **Hyper pipelined technology**
  - 20 (Willamette) ~ 31 (Prescott) stages
  - Minimum misprediction penalty: 20 cycles
  - Max clock rate: 3.80 GHz (Prescott)

### P6 misprediction pipeline (introduced at 733MHz @ 0.18μ)

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetch</td>
<td>Fetch</td>
<td>Decode</td>
<td>Decode</td>
<td>Decode</td>
<td>Rename</td>
<td>ROB Rd</td>
<td>Rdy/Sch</td>
<td>Dispatch</td>
<td>Exec</td>
</tr>
</tbody>
</table>

### NetBurst misprediction pipeline (introduced at ≥ 1.4GHz @ 0.18μ)

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC Nxt IP</td>
<td>TC Fetch</td>
<td>Drive Alloc</td>
<td>Rename</td>
<td>Que</td>
<td>Sch</td>
<td>Sch</td>
<td>Disp</td>
<td>Disp</td>
<td>RF</td>
</tr>
<tr>
<td>15 RF</td>
<td>16 RF</td>
<td>17 Ex</td>
<td>18 Flgs</td>
<td>19 Br CkDrive</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Core Microarchitecture

- 128 Entry ITLB
- 32 KB I-cache (8 way)
- x86 Instruction Pre-Decode, Fetch Buffer
- Complex Decoder
- Simple Decoder
- Simple Decoder
- Simple Decoder
- μcode Sequencer
- 7+ Entry μop Buffer
- Register Alias Table and Allocator
- 96 Entry Reorder Buffer (ROB)
- 32 Entry Reservation Station
- Port 0
- Port 1
- Port 2
- Port 3
- Port 4
- Port 5
- 64 bit ALU
- 128 bit SSE
- ALU Shift Rotate
- 128 bit SSE
- 64 bit ALU Branch
- 128 bit SSE
- 128 bit FADD
- 128 bit FMUL
- 128 bit FDIV
- Internal Results Bus
- Memory Reorder Buffer (MOB)
- 32 KB dual ported D-cache (8 way)
- 256 Entry DTLB

Shared Bus Interface Unit
4MB Shared L2 Cache (16 way)
Retirement Register File (Program Visible State)
Multi-core (1)

- **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, Xeon: dual-core
    - Sun UltraSparc T1: eight-core, 32 threads (Nov. 2005)
    - Intel Core 2 Extreme Quad-core processor (Sep. 2006)
    - Intel Xeon X7460 Six-core processor (Sep. 2008)
    - Microsoft’s Xbox 360: triple-core PowerPC microprocessor
Multi-core (2)

Intel CPU Trends
(sources: Intel, Wikipedia, K. Olukotun)

Montecito

Moore’s Law

Pentium

386

Transistors (000)
Clock Speed (MHz)
Power (W)
Perf/Clock (ILP)
Multi-core (3)

- **Memory wall**
  - CPU 55%/year, Memory 10%/year (1986 ~ 2000)
  - Caches show diminishing returns.

- **ILP wall**
  - Control dependency
  - Data dependency

- **Power wall**
  - Dynamic power $\propto$ Frequency$^3$
  - Static power $\propto$ Frequency
  - Total power $\propto$ The number of cores
Multi-core (4)

Performance

Power

1.00x

Raise Clock (20%)

0.87x

Lower Clock (20%)

1.13x

1.73x

More MIPS/watt

1.73x

1.02x

Source: Intel