Lecture 9: Multiple Issue (Superscalar and VLIW)

Iakovos Mavroidis

Computer Science Department
University of Crete
Example: Dynamic Scheduling in PowerPC 604 and Pentium Pro

- In-order Issue, Out-of-order execution, In-order Commit
Multiple Issue

\[ \text{CPI} = \text{CPI}_{\text{ideal}} + \text{Stalls}_{\text{structural}} + \text{Stalls}_{\text{RAW}} + \text{Stalls}_{\text{WAR}} + \text{Stalls}_{\text{WAW}} + \text{Stalls}_{\text{control}} \]

Προσοχή να διατηρούνται
1. Data flow
2. Exception Behavior

Έχουμε μελετήσει
θα μελετήσουμε σήμερα
Θα μελετήσουμε σε επόμενα μαθήματα

Δυναμικές δρομολόγησης
\section*{εντολών (hardware)}
• Scoreboard (ελάττωση RAW stalls)
• Register Renaming
  α) Tomasulo
  (ελάττωση WAR και WAW stalls)
  β) Reorder Buffer
• Branch prediction
  (ελάττωση Control stalls)
• Multiple Issue (CPI < 1)
• Multithreading (CPI < 1)

Στατικές (shoftware/compiler)
• Loop Unrolling
• Software Pipelining
• Trace Scheduling

\section*{Data flow behavior}
• CPI = \text{CPI}_{\text{ideal}} + \text{Stalls}_{\text{structural}} + \text{Stalls}_{\text{RAW}} + \text{Stalls}_{\text{WAR}} + \text{Stalls}_{\text{WAW}} + \text{Stalls}_{\text{control}}
Beyond CPI = 1

- Initial goal to achieve CPI = 1
- Can we improve beyond this?
- Two approaches
  - **Superscalar**:
    - varying no. instructions/cycle (1 to 8), i.e. 1-way, 2-way, ..., 8-way superscalar
    - scheduled by compiler (**statically scheduled**) or by HW (**dynamically scheduled**)  
    - e.g. IBM PowerPC, Sun UltraSparc, DEC Alpha, HP 8000
    - The successful approach (to date) for general purpose computing
  - Anticipated success lead to use of **Instructions Per Clock** cycle (**IPC**) vs. CPI
Beyond CPI = 1

• Alternative approach
• **(Very) Long Instruction Words (V)L IW:**
  – fixed number of instructions (4-16)
  – scheduled by the compiler; put ops into wide templates
  – Currently found more success in DSP, Multimedia applications
  – Intel Architecture-64 (Merced/A-64) 64-bit address
  – Style: “Explicitly Parallel Instruction Computer (EPIC)”
Getting CPI < 1: Issuing Multiple Instructions/Cycle

- Superscalar DLX: 2 instructions, 1 FP & 1 anything else
  - Fetch 64-bits/clock cycle; Int on left, FP on right
  - Can only issue 2nd instruction if 1st instruction issues
  - More ports for FP registers to do FP load & FP op in a pair

<table>
<thead>
<tr>
<th>Type</th>
<th>Pipe Stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Int. instruction</td>
<td>IF ID EX MEM WB</td>
</tr>
<tr>
<td>FP instruction</td>
<td>IF ID EX MEM WB</td>
</tr>
<tr>
<td>Int. instruction</td>
<td>IF ID EX MEM WB</td>
</tr>
<tr>
<td>FP instruction</td>
<td>IF ID EX MEM WB</td>
</tr>
<tr>
<td>Int. instruction</td>
<td>IF ID EX MEM WB</td>
</tr>
<tr>
<td>FP instruction</td>
<td>IF ID EX MEM WB</td>
</tr>
</tbody>
</table>

- 1 cycle load delay expands to **3 instructions** in SS
  - instruction in right half can’t use it, nor instructions in next slot
In-Order Superscalar Pipeline

- Fetch two instructions per cycle; issue both simultaneously if one is integer/memory and other is floating point.
- Same idea can be extended to wider issue by duplicating functional units (e.g. 4-issue UltraSPARC) but regfile ports and bypassing costs grow quickly.
Instructions in the instruction window are **free from control dependencies** due to branch prediction, and **free from name dependences** due to register renaming.

So, only **(true) data dependences and structural conflicts remain** to be solved.
Similar Technique: Superpipelined Machines

- **Machine issues instructions faster than they are executed**
- **Advantage**: increase in the number of instructions which can be in the pipeline at one time and hence the level of parallelism.
- **Disadvantage**: The larger number of instructions "in flight" (ie in some part of the pipeline) at any time, increases the potential for data dependencies to introduce stalls.
Sequential ISA Bottleneck

Sequential source code

Superscalar compiler

Find independent operations

Schedule operations

Superscalar processor

Check instruction dependencies

Schedule execution

Sequential machine code

a = foo(b);
for (i=0, i<
Review: Unrolled Loop that Minimizes Stalls for Scalar

1 Loop:  
1. LD  F0,0 (R1)  
2. LD  F6,-8 (R1)  
3. LD  F10,-16 (R1)  
4. LD  F14,-24 (R1)  
5. ADDD  F4,F0,F2  
6. ADDD  F8,F6,F2  
7. ADDD  F12,F10,F2  
8. ADDD  F16,F14,F2  
9. SD  0 (R1),F4  
10. SD  -8 (R1),F8  
11. SD  -16 (R1),F12  
12. SUBI  R1,R1,#32  
13. BNEZ  R1,LOOP  
14. SD  8 (R1),F16 ; 8-32 = -24  

14 clock cycles, or 3.5 per iteration
### Loop Unrolling in Superscalar

<table>
<thead>
<tr>
<th>Integer instruction</th>
<th>FP instruction</th>
<th>Clock cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LD</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>F0,0(R1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LD</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>F6,-8(R1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LD</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>F10,-16(R1)</td>
<td>ADDD F4,F0,F2</td>
<td></td>
</tr>
<tr>
<td>LD</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>F14,-24(R1)</td>
<td>ADDD F8,F6,F2</td>
<td></td>
</tr>
<tr>
<td>LD</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>F18,-32(R1)</td>
<td>ADDD F12,F10,F2</td>
<td></td>
</tr>
<tr>
<td>SD 0(R1),F4</td>
<td>ADDD F16,F14,F2</td>
<td></td>
</tr>
<tr>
<td>SD -8(R1),F8</td>
<td>ADDD F20,F18,F2</td>
<td></td>
</tr>
<tr>
<td>SD -16(R1),F12</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>SD -24(R1),F16</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>SUBI R1,R1,#40</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>BNEZ R1,LOOP</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>SD -32(R1),F20</td>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

- Unrolled 5 times to avoid delays (+1 due to SS)
- 12 clocks, or 2.4 clocks per iteration (1.5X)
SS Advantages and Challenges

- The potential advantages of a SS processor versus a vector or VLIW processor are their ability to extract some parallelism from less structured code (i.e. no loops) and their ability to easily cache all forms of data.

- While Integer/FP split is simple for the HW, get CPI of 0.5 only for programs with:
  - Exactly 50% FP operations
  - No hazards

- If more instructions issue at same time, greater difficulty of decode and issue
  - Even 2 way-scalar => examine 2 opcodes, 6 register specifiers, & decide if 1 or 2 instructions can issue
Example Processor: Intel Core2

Superpipelined & Superscalar (4-way)
All in one: 2-way SS + OoO + Branch Prediction + Reorder Buffer (Speculation)

<table>
<thead>
<tr>
<th>Iteration number</th>
<th>Instructions</th>
<th>Issues at clock number</th>
<th>Executes at clock number</th>
<th>Read access at clock number</th>
<th>Write CDB at clock number</th>
<th>Commits at clock number</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LD R2,0(R1)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>First issue</td>
</tr>
<tr>
<td>1</td>
<td>DADDIU R2,R2,#1</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>Wait for LW</td>
</tr>
<tr>
<td>1</td>
<td>SD R2,0(R1)</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>Wait for DADDIU</td>
</tr>
<tr>
<td>1</td>
<td>DADDIU R1,R1,#8</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>Commit in order</td>
</tr>
<tr>
<td>1</td>
<td>BNE R2,R3,LOOP</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>Wait for DADDIU</td>
</tr>
<tr>
<td>2</td>
<td>LD R2,0(R1)</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>No execute delay</td>
</tr>
<tr>
<td>2</td>
<td>DADDIU R2,R2,#1</td>
<td>4</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>Wait for LW</td>
</tr>
<tr>
<td>2</td>
<td>SD R2,0(R1)</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>11</td>
<td>Wait for DADDIU</td>
</tr>
<tr>
<td>2</td>
<td>DADDIU R1,R1,#8</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>11</td>
<td>11</td>
<td>Commit in order</td>
</tr>
<tr>
<td>2</td>
<td>BNE R2,R3,LOOP</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>11</td>
<td>11</td>
<td>Wait for DADDIU</td>
</tr>
<tr>
<td>3</td>
<td>LD R2,0(R1)</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>12</td>
<td>Earliest possible</td>
</tr>
<tr>
<td>3</td>
<td>DADDIU R2,R2,#1</td>
<td>7</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>13</td>
<td>Wait for LW</td>
</tr>
<tr>
<td>3</td>
<td>SD R2,0(R1)</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>Wait for DADDIU</td>
</tr>
<tr>
<td>3</td>
<td>DADDIU R1,R1,#8</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>14</td>
<td>14</td>
<td>Executes earlier</td>
</tr>
<tr>
<td>3</td>
<td>BNE R2,R3,LOOP</td>
<td>9</td>
<td>13</td>
<td></td>
<td></td>
<td>14</td>
<td>Wait for DADDIU</td>
</tr>
</tbody>
</table>
Alternative Solutions

• Very Long Instruction Word (VLIW)
• Explicitly Parallel Instruction Computing (EPIC)
• Simultaneous Multithreading (SMT), next lecture
• Multi-core processors, ~last lecture

• VLIW: tradeoff instruction space for simple decoding
  – The long instruction word has room for many operations
  – By definition, all the operations the compiler puts in the
    long instruction word are independent => execute in
    parallel
  – E.g., 2 integer operations, 2 FP ops, 2 Memory refs, 1
    branch
    » 16 to 24 bits per field => 7*16 or 112 bits to 7*24 or 168
    bits wide
    » Intel Itanium 1 and 2 contain 6 operations per instruction
      packet
  – Need compiling technique that schedules across
    several branches
VLIW: Very Long Instruction Word

- Multiple operations packed into one instruction
- Each operation slot is for a fixed function
- Constant operation latencies are specified
- Architecture requires guarantee of:
  - Parallelism within an instruction => no cross-operation RAW check
  - No data use before data ready => no data interlocks
VLIW Compiler Responsibilities

• Schedule operations to maximize parallel execution
• Guarantees intra-instruction parallelism
• Schedule to avoid data hazards (no interlocks)
  – Typically separates operations with explicit NOPs
Typical VLIW processor

(a) A typical VLIW processor and instruction format

(b) VLIW execution with degree $m = 3$

Figure 4.14 The architecture of a very long instruction word (VLIW) processor and its pipeline operations. (Courtesy of Multiflow Computer, Inc., 1987)
Loop Unrolling in VLIW

<table>
<thead>
<tr>
<th>Memory reference 1</th>
<th>Memory reference 2</th>
<th>FP operation 1</th>
<th>FP op. 2</th>
<th>Int. op/branch</th>
<th>Clock</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD F0,0(R1)</td>
<td>LD F6,-8(R1)</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>LD F10,-16(R1)</td>
<td>LD F14, 24(R1)</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>LD F18,-32(R1)</td>
<td>LD F22,-40(R1)</td>
<td>ADDD F4,F0,F2</td>
<td>ADDD F8,F6,F2</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>LD F26,-48(R1)</td>
<td></td>
<td>ADDD F12,F10,F2</td>
<td>ADDD F16,F14,F2</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ADDD F20,F18,F2</td>
<td>ADDD F24,F22,F2</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>SD 0(R1),F4</td>
<td>SD -8(R1),F8</td>
<td>ADDD F28,F26,F2</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>SD -16(R1),F12</td>
<td>SD -24(R1),F16</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>SD -32(R1),F20</td>
<td>SD -40(R1),F24</td>
<td></td>
<td></td>
<td>SUBI R1,R1,#48</td>
<td>8</td>
</tr>
<tr>
<td>SD -0(R1),F28</td>
<td></td>
<td></td>
<td></td>
<td>BNEZ R1,LOOP</td>
<td>9</td>
</tr>
</tbody>
</table>

Unrolled 7 times to avoid delays
7 results in 9 clocks, or 1.3 clocks per iteration (1.8X vs SS)
Average: 2.5 ops per clock, 50% efficiency
Note: Need more registers in VLIW (15 vs. 6 in SS)
Advantages of VLIW

Compiler prepares fixed packets of multiple operations that give the full "plan of execution"

- dependencies are determined by compiler and used to schedule according to function unit latencies
- function units are assigned by compiler and correspond to the position within the instruction packet ("slotting")
- compiler produces fully-scheduled, hazard-free code => hardware doesn't have to "rediscover" dependencies or schedule
Disadvantages of VLIW

• Object-code compatibility
  – have to recompile all code for every machine, even for two machines in same generation

• Object code size
  – instruction padding wastes instruction memory/cache
  – loop unrolling/software pipelining replicates code

• Scheduling variable latency memory operations
  – caches and/or memory bank conflicts impose statically unpredictable variability
  – As the issue rate and number of memory references becomes large, this synchronization restriction becomes unacceptable

• Knowing branch probabilities
  – Profiling requires an significant extra step in build process

• Scheduling for statically unpredictable branches
  – optimal schedule varies with branch path
What if there are no loops?

- Branches limit basic block size in control-flow intensive irregular code
- Difficult to find ILP in individual basic blocks
Trace Scheduling [Fisher, Ellis]

• **Trace selection:** Pick string of basic blocks, a *trace*, that represents most frequent branch path
• Use profiling feedback or compiler heuristics to find common branch paths
• **Trace Compaction:** Schedule whole “trace” at once. Packing operations to few wide instructions.
• Add fixup code to cope with branches jumping out of trace
• Effective to certain classes of programs
• Key assumption is that the trace is much more probable than the alternatives
Intel Itanium, EPIC IA-64

- EPIC is the style of architecture (cf. CISC, RISC)
  - Explicitly Parallel Instruction Computing (really just VLIW)
- IA-64 is Intel’s chosen ISA (cf. x86, MIPS)
  - IA-64 = Intel Architecture 64-bit
  - An object-code-compatible VLIW
- Merced was first Itanium implementation (cf. 8086)
  - First customer shipment expected 1997 (actually 2001)
  - McKinley, second implementation shipped in 2002
  - Recent version, Poulson, eight cores, 32nm, announced 2011

- Different instruction format than VLIW architectures using with indicators
- Support for SW speculation
Eight Core Itanium “Poulson” [Intel 2011]

- 8 cores
- 1-cycle 16KB L1 I&D caches
- 9-cycle 512KB L2 I-cache
- 8-cycle 256KB L2 D-cache
- 32 MB shared L3 cache
- 544mm² in 32nm CMOS
- Over 3 billion transistors
- Cores are 2-way multithreaded
- 6 instruction/cycle fetch
  - Two 128-bit bundles
- Up to 12 insts/cycle execute
IA-64 Registers

- 128 General Purpose 64-bit Integer Registers
- 128 General Purpose 64/80-bit Floating Point Registers
- 64 1-bit Predicate Registers
- 8 64-bit Branch Registers

Register stack mechanism: GPRs “rotate” to reduce code size for software pipelined loops
  - Rotation is a simple form of register renaming allowing one instruction to address different physical registers on each procedure call
### IA-64 Instruction Format

<table>
<thead>
<tr>
<th>Instruction 2</th>
<th>Instruction 1</th>
<th>Instruction 0</th>
<th>Template</th>
</tr>
</thead>
</table>

128-bit instruction bundle (41*3+5)

- Template bits describe grouping of these instructions with others in adjacent bundles
- Each group contains instructions that can execute in parallel

```
bundle j-1  bundle j  bundle j+1  bundle j+2

group i-1                group i                group i+1                group i+2
```
## IA-64 Template

<table>
<thead>
<tr>
<th>Template</th>
<th>Slot 0</th>
<th>Slot 1</th>
<th>Slot 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>M</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>1</td>
<td>M</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>L</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>M</td>
<td>L</td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td>M</td>
<td>M</td>
<td>I</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>M</td>
<td>I</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>M</td>
<td>I</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>M</td>
<td>I</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
<td>F</td>
<td>I</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>F</td>
<td>I</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>15</td>
<td>M</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>16</td>
<td>M</td>
<td>I</td>
<td>B</td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>I</td>
<td>B</td>
</tr>
<tr>
<td>18</td>
<td>M</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>19</td>
<td>M</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>22</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>23</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>24</td>
<td>M</td>
<td>M</td>
<td>B</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>M</td>
<td>B</td>
</tr>
<tr>
<td>28</td>
<td>M</td>
<td>F</td>
<td>B</td>
</tr>
<tr>
<td>29</td>
<td>M</td>
<td>F</td>
<td>B</td>
</tr>
</tbody>
</table>
IA-64 Basic Architecture

- Registers (both integer and floating point) are 64-bit.
- Predicate registers are 1-bit.
- 8 or more functional units.
Problem: Mispredicted branches limit ILP
Solution: Eliminate hard to predict branches with predicated execution
  – Almost all IA-64 instructions can be executed conditionally under predicate
  – Instruction becomes NOP if predicate register false

Four basic blocks

Predicated Execution

Inst 1
Inst 2
br a==b, b2 if

Inst 3
Inst 4
br b3

Inst 5
Inst 6

Inst 7
Inst 8

Inst 1
Inst 2
p1 = a!=b,p2 = a==b
(p1) Inst 3      ||  (p2) Inst 5
(p1) Inst 4      ||  (p2) Inst 6
Inst 7
Inst 8

One basic block

Mahlke et al, ISCA95: On average >50% branches removed
Branch Predication

- **Branch predication** is an aggressive compilation technique to generate code with a higher degree of instruction level parallelism.
- It lets operations from both branches of a conditional branch to be executed in parallel, to increase the amount of parallel operations.
- In this way, branches are eliminated and replaced by conditional execution.
  - Hardware support is needed, as implemented in the IA-64 architecture.

The idea is: let instructions from both branches go on in parallel, before the branch condition has been evaluated. The hardware takes care that only those corresponding to the right branch will be finally committed.
Branch Predication Example

For a branch instruction, the compiler assigns a predicate to each of the two following instruction paths.

CPU can execute instructions from different paths concurrently, but only the correct path will finally be committed.

For a VLIW machine, the instructions may be arranged as follows:

<table>
<thead>
<tr>
<th>Instruction 1</th>
<th>Instruction 2</th>
<th>Instruction 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;P₁&gt; Instruction 4</td>
<td>&lt;P₂&gt; Instruction 7</td>
<td>&lt;P₁&gt; Instruction 5</td>
</tr>
<tr>
<td>&lt;P₁&gt; Instruction 5</td>
<td>&lt;P₂&gt; Instruction 8</td>
<td>&lt;P₁&gt; Instruction 6</td>
</tr>
<tr>
<td>&lt;P₂&gt; Instruction 8</td>
<td>&lt;P₁&gt; Instruction 6</td>
<td>&lt;P₂&gt; Instruction 9</td>
</tr>
</tbody>
</table>