18-447
Computer Architecture
Lecture 4: ISA Tradeoffs (Continued) and Single-Cycle Microarchitectures

Prof. Onur Mutlu
Carnegie Mellon University
Spring 2014, 1/22/2014
X86: Small Semantic Gap: String Operations

**REP MOVVS (DEST SRC)**

IF AddressSize = 16
    THEN
        Use CX for CountReg;
    ELSE IF AddressSize = 64 and REX.W used
        THEN Use RCX for CountReg; Fl;
    ELSE
        Use ECX for CountReg;
    Fl;
WHILE CountReg ≠ 0
    DO
        Service pending interrupts (if any);
        Execute associated string instruction;
        CountReg ← (CountReg − 1);
        IF CountReg = 0
            THEN exit WHILE loop; Fl;
        IF (Repeat prefix is REPZ or REPE) and (ZF = 0)
            or (Repeat prefix is REPNZ or REPNE) and (ZF = 1)
                THEN exit WHILE loop; Fl;
    OD;

How many instructions does this take in ARM and MIPS?
Small Semantic Gap Examples in VAX

- FIND FIRST
  - Find the first set bit in a bit field
  - Helps OS resource allocation operations
- SAVE CONTEXT, LOAD CONTEXT
  - Special context switching instructions
- INSQUEUE, REMQUEUE
  - Operations on doubly linked list
- INDEX
  - Array access with bounds checking
- STRING Operations
  - Compare strings, find substrings, ...
- Cyclic Redundancy Check Instruction
- EDITPC
  - Implements editing functions to display fixed format output

Small versus Large Semantic Gap

- **CISC vs. RISC**
  - Complex instruction set computer → complex instructions
    - Initially motivated by “not good enough” code generation
  - Reduced instruction set computer → simple instructions
    - John Cocke, mid 1970s, IBM 801
      - Goal: enable better compiler control and optimization

- **RISC motivated by**
  - Memory stalls (no work done in a complex instruction when there is a memory stall?)
    - When is this correct?
  - Simplifying the hardware → lower cost, higher frequency
  - Enabling the compiler to optimize the code better
    - Find fine-grained parallelism to reduce stalls
How High or Low Can You Go?

- **Very large semantic gap**
  - Each instruction specifies the complete set of control signals in the machine
  - Compiler generates control signals
  - Open microcode (John Cocke, circa 1970s)
    - Gave way to optimizing compilers

- **Very small semantic gap**
  - ISA is (almost) the same as high-level language
  - Java machines, LISP machines, object-oriented machines, capability-based machines
A Note on ISA Evolution

- ISAs have evolved to reflect/satisfy the concerns of the day

- Examples:
  - Limited on-chip and off-chip memory size
  - Limited compiler optimization technology
  - Limited memory bandwidth
  - Need for specialization in important applications (e.g., MMX)

- Use of translation (in HW and SW) enabled underlying implementations to be similar, regardless of the ISA
  - Concept of translation/interpretation interface
  - Contrast it with hardware/software interface
Effect of Translation

- One can translate from one ISA to another ISA to change the semantic gap tradeoffs

- Examples
  - Intel's and AMD's x86 implementations translate x86 instructions into programmer-invisible microoperations (simple instructions) in hardware
  - Transmeta's x86 implementations translated x86 instructions into "secret" VLIW instructions in software (code morphing software)

- Think about the tradeoffs
ISA-level Tradeoffs: Instruction Length

- **Fixed length**: Length of all instructions the same
  - + Easier to decode single instruction in hardware
  - + Easier to decode multiple instructions concurrently
  - -- Wasted bits in instructions *(Why is this bad?)*
  - -- Harder-to-extend ISA (how to add new instructions?)

- **Variable length**: Length of instructions different (determined by opcode and sub-opcode)
  - + Compact encoding *(Why is this good?)*
    - Intel 432: Huffman encoding (sort of). 6 to 321 bit instructions. How?
  - -- More logic to decode a single instruction
  - -- Harder to decode multiple instructions concurrently

- **Tradeoffs**
  - Code size (memory space, bandwidth, latency) vs. hardware complexity
  - ISA extensibility and expressiveness vs. hardware complexity
  - Performance? Smaller code vs. ease of decode
**ISA-level Tradeoffs: Uniform Decode**

- **Uniform decode:** Same bits in each instruction correspond to the same meaning
  - Opcode is always in the same location
  - Ditto operand specifiers, immediate values, ...
  - Many “RISC” ISAs: Alpha, MIPS, SPARC
  + Easier decode, simpler hardware
  + Enables parallelism: generate target address before knowing the instruction is a branch
  -- Restricts instruction format (fewer instructions?) or wastes space

- **Non-uniform decode**
  - E.g., opcode can be the 1st-7th byte in x86
  + More compact and powerful instruction format
  -- More complex decode logic
x86 vs. Alpha Instruction Formats

- **x86:**

  ![x86 Instruction Format Diagram]

- **Alpha:**

  ![Alpha Instruction Format Diagram]
MIPS Instruction Format

- **R-type**, 3 register operands
  - 
    |    | rs | rt | rd | shamt | funct |
    |----|----|----|----|-------|-------|
    | 0  | 5-bit | 5-bit | 5-bit | 5-bit | 6-bit |
  - R-type

- **I-type**, 2 register operands and 16-bit immediate operand
  - 
    | opcode | rs | rt | immediate |
    |-------|----|----|------------|
    | 6-bit | 5-bit | 5-bit | 16-bit |
  - I-type

- **J-type**, 26-bit immediate operand
  - 
    | opcode | immediate |
    |-------|------------|
    | 6-bit | 26-bit |
  - J-type

- **Simple Decoding**
  - 4 bytes per instruction, regardless of format
  - must be 4-byte aligned (2 lsb of PC must be 2b’00)
  - format and fields easy to extract in hardware
### ARM Instruction Set Formats

**Figure 4-1: ARM instruction set formats**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Opcode</th>
<th>S</th>
<th>Rn</th>
<th>Rd</th>
<th>Operand 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 0 0 0 0 0 A S</td>
<td>Rd</td>
<td>Rn</td>
<td>Rs</td>
<td>1 0 0 1</td>
<td>Rm</td>
</tr>
<tr>
<td>0 0 0 0 1 U A S</td>
<td>RdHi</td>
<td>RdLo</td>
<td>Rn</td>
<td>1 0 0 1</td>
<td>Rm</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0 B 0 0</td>
<td>Rn</td>
<td>Rd</td>
<td>0 0 0 0 1 0 0 1</td>
<td>Rm</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1</td>
<td>Rn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 0 0 P U 0 W L</td>
<td>Rn</td>
<td>Rd</td>
<td>0 0 0 1 S H 1</td>
<td>Rm</td>
<td></td>
</tr>
<tr>
<td>0 0 0 P U 1 W L</td>
<td>Rn</td>
<td>Rd</td>
<td>Offset 1 S H 1 Offset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 1 1 P U B W L</td>
<td>Rn</td>
<td>Rd</td>
<td>Offset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 1 1</td>
<td>Offset 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 0 0 P U S W L</td>
<td>Rn</td>
<td>Register List</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 0 1 L</td>
<td>Offset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 1 0 P U N W L</td>
<td>Rn</td>
<td>CRd</td>
<td>CP#</td>
<td>Offset</td>
<td></td>
</tr>
<tr>
<td>1 1 1 0</td>
<td>CP Opc</td>
<td>CRn</td>
<td>CRd</td>
<td>CP#</td>
<td>CP</td>
</tr>
<tr>
<td>1 1 1 0</td>
<td>CP Opc</td>
<td>CRn</td>
<td>Rd</td>
<td>CP#</td>
<td>CP</td>
</tr>
<tr>
<td>1 1 1 1</td>
<td>Ignored by processor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Data Processing / PSR Transfer**
- **Multiply**
- **Multiply Long**
- **Single Data Swap**
- **Branch and Exchange**
- **Halfword Data Transfer: register offset**
- **Halfword Data Transfer: immediate offset**
- **Single Data Transfer**
- **Undefined**
- **Block Data Transfer**
- **Branch**
- **Coprocessor Data Transfer**
- **Coprocessor Data Operation**
- **Coprocessor Register Transfer**
- **Software Interrupt**
A Note on Length and Uniformity

- Uniform decode usually goes with fixed length

- In a variable length ISA, uniform decode can be a property of instructions of the same length
  - It is hard to think of it as a property of instructions of different lengths
A Note on RISC vs. CISC

- Usually, ...

- **RISC**
  - Simple instructions
  - Fixed length
  - Uniform decode
  - Few addressing modes

- **CISC**
  - Complex instructions
  - Variable length
  - Non-uniform decode
  - Many addressing modes
ISA-level Tradeoffs: Number of Registers

- **Affects:**
  - Number of bits used for encoding register address
  - Number of values kept in fast storage (register file)
  - (uarch) Size, access time, power consumption of register file

- **Large number of registers:**
  + Enables better register allocation (and optimizations) by compiler → fewer saves/restores
  -- Larger instruction size
  -- Larger register file size
ISA-level Tradeoffs: Addressing Modes

- Addressing mode specifies how to obtain an operand of an instruction
  - Register
  - Immediate
  - Memory (displacement, register indirect, indexed, absolute, memory indirect, autoincrement, autodecrement, ...)

- More modes:
  + help better support programming constructs (arrays, pointer-based accesses)
  -- make it harder for the architect to design
  -- too many choices for the compiler?
    - Many ways to do the same thing complicates compiler design
x86 vs. Alpha Instruction Formats

- **x86:**

  - Instruction Prefixes: Up to four prefixes of 1 byte each (optional)
  - Opcode: 1-, 2-, or 3-byte opcode
  - ModR/M: 1 byte (if required)
  - SIB: 1 byte (if required)
  - Displacement: Address displacement of 1, 2, or 4 bytes or none
  - Immediate: Immediate data of 1, 2, or 4 bytes or none

- **Alpha:**

  - Opcode: RA, RB, Function, RC
  - Number: Disp, Disp, RC
  - PALcode Format
  - Branch Format
  - Memory Format
  - Operate Format
### Table 2-2. 32-Bit Addressing Forms with the ModR/M Byte

<table>
<thead>
<tr>
<th>Effective Address</th>
<th>Mod</th>
<th>R/M</th>
<th>Value of ModR/M Byte (in Hexadecimal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[EAX]</td>
<td>00</td>
<td>00</td>
<td>08 10 18 20 28 30 38</td>
</tr>
<tr>
<td>[ECX]</td>
<td>00</td>
<td>01</td>
<td>09 11 19 21 29 31 39</td>
</tr>
<tr>
<td>[EDX]</td>
<td>01</td>
<td>02</td>
<td>0A 12 1A 22 2A 32 3A</td>
</tr>
<tr>
<td>[EBX]</td>
<td>01</td>
<td>03</td>
<td>0B 13 1B 23 2B 33 3B</td>
</tr>
<tr>
<td>[ESP]</td>
<td>100</td>
<td>04</td>
<td>0C 14 1C 24 2C 34 3C</td>
</tr>
<tr>
<td>[EBP]</td>
<td>101</td>
<td>05</td>
<td>0D 15 1D 25 2D 35 3D</td>
</tr>
<tr>
<td>[EDI]</td>
<td>110</td>
<td>06</td>
<td>0E 16 1E 26 2E 36 3E</td>
</tr>
<tr>
<td>[ESI]</td>
<td>111</td>
<td>07</td>
<td>0F 17 1F 27 2F 37 3F</td>
</tr>
<tr>
<td>[EAX]+disp32</td>
<td>01</td>
<td>00</td>
<td>40 48 50 58 60 68 70 78</td>
</tr>
<tr>
<td>[ECX]+disp32</td>
<td>01</td>
<td>01</td>
<td>41 49 51 59 61 69 71 79</td>
</tr>
<tr>
<td>[EDX]+disp32</td>
<td>01</td>
<td>02</td>
<td>42 50 52 60 68 76 74 7C</td>
</tr>
<tr>
<td>[EBX]+disp32</td>
<td>01</td>
<td>03</td>
<td>43 4B 53 63 6B 73 7B</td>
</tr>
<tr>
<td>[ESP]+disp32</td>
<td>100</td>
<td>04</td>
<td>44 4C 54 64 74 7C</td>
</tr>
<tr>
<td>[EBP]+disp32</td>
<td>101</td>
<td>05</td>
<td>45 4D 55 65 75 7D</td>
</tr>
<tr>
<td>[ESI]+disp32</td>
<td>110</td>
<td>06</td>
<td>46 4E 56 66 76 7E</td>
</tr>
<tr>
<td>[EDI]+disp32</td>
<td>111</td>
<td>07</td>
<td>47 4F 57 67 77 7F</td>
</tr>
<tr>
<td>[EAX]+disp8</td>
<td>10</td>
<td>00</td>
<td>80 88 90 98 A0 A8 B0 B8</td>
</tr>
<tr>
<td>[ECX]+disp8</td>
<td>10</td>
<td>01</td>
<td>81 89 91 99 A1 A9 B1 B9</td>
</tr>
<tr>
<td>[EDX]+disp8</td>
<td>10</td>
<td>02</td>
<td>82 8A 92 9A A2 AA B2 BA</td>
</tr>
<tr>
<td>[EBX]+disp8</td>
<td>10</td>
<td>03</td>
<td>83 8B 93 9B A3 AB B3 BB</td>
</tr>
<tr>
<td>[ESP]+disp8</td>
<td>10</td>
<td>04</td>
<td>84 8C 94 9C A4 AC B4 BC</td>
</tr>
<tr>
<td>[EBP]+disp8</td>
<td>10</td>
<td>05</td>
<td>85 8D 95 9D A5 AD B5 BD</td>
</tr>
<tr>
<td>[ESI]+disp8</td>
<td>11</td>
<td>06</td>
<td>86 8E 96 9E A6 AE B6 BE</td>
</tr>
<tr>
<td>[EDI]+disp8</td>
<td>11</td>
<td>07</td>
<td>87 8F 97 9F A7 AF B7 BF</td>
</tr>
<tr>
<td>[EAX/AX/AL/MMO/XMM0]</td>
<td>11</td>
<td>00</td>
<td>C0 C8 D0 D8 E0 E8 F0 F8</td>
</tr>
<tr>
<td>[ECX/EC/CL/MMO/XMM1]</td>
<td>11</td>
<td>01</td>
<td>C1 C9 D1 D9 E1 E9 F1 F9</td>
</tr>
<tr>
<td>[EDX/DX/CL/MM2/XMM2]</td>
<td>11</td>
<td>02</td>
<td>C2 CA D2 DA E2 EA F2 FA</td>
</tr>
<tr>
<td>[EBX/BL/MM3/XMM3]</td>
<td>11</td>
<td>03</td>
<td>C3 CB D3 DB E3 EB F3 FB</td>
</tr>
<tr>
<td>[EBP/BP/MM5/XMM5]</td>
<td>11</td>
<td>05</td>
<td>C5 CD D5 DD E5 ED F5 FD</td>
</tr>
<tr>
<td>[EDI/DI/MM7/XMM7]</td>
<td>11</td>
<td>07</td>
<td>C7 CF D7 DF E7 EF FF</td>
</tr>
</tbody>
</table>

**NOTES:**

1. The [-][-] nomenclature means a SIB follows the ModR/M byte.
2. The disp32 nomenclature denotes a 32-bit displacement that follows the ModR/M byte (or the SIB byte if one is present) and that is added to the index.
3. The disp8 nomenclature denotes an 8-bit displacement that follows the ModR/M byte (or the SIB byte if one is present) and that is sign-extended and added to the index.

Table 2-3 is organized to give 256 possible values of the SIB byte (in hexadecimal). General purpose registers used as a base are indicated across the top of the table.
### Table 2-3. 32-Bit Addressing Forms with the SIB Byte

<table>
<thead>
<tr>
<th>Indexed Addressing Form</th>
<th>Scaled Index</th>
<th>SS Index</th>
<th>Value of SIB Byte (in Hexadecimal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[EAX]</td>
<td>0000</td>
<td>00</td>
<td>00 01 02 03 04 05 06 07</td>
</tr>
<tr>
<td>[ECX]</td>
<td>0001</td>
<td>08 09 0A 0B 0C 0D 0E 0F</td>
<td></td>
</tr>
<tr>
<td>[EDX]</td>
<td>0010</td>
<td>10 11 12 13 14 15 16 17</td>
<td></td>
</tr>
<tr>
<td>[EBX]</td>
<td>0011</td>
<td>18 19 1A 1B 1C 1D 1E 1F</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>1000</td>
<td>20 21 22 23 24 25 26 27</td>
<td></td>
</tr>
<tr>
<td>[EBP]</td>
<td>1001</td>
<td>28 29 2A 2B 2C 2D 2E 2F</td>
<td></td>
</tr>
<tr>
<td>[ESI]</td>
<td>1100</td>
<td>30 31 32 33 34 35 36 37</td>
<td></td>
</tr>
<tr>
<td>[EDI]</td>
<td>1101</td>
<td>38 39 3A 3B 3C 3D 3E 3F</td>
<td></td>
</tr>
<tr>
<td>[EAX*2]</td>
<td>0000</td>
<td>40 41 42 43 44 45 46 47</td>
<td></td>
</tr>
<tr>
<td>[ECX*2]</td>
<td>0001</td>
<td>48 49 4A 4B 4C 4D 4E 4F</td>
<td></td>
</tr>
<tr>
<td>[EDX*2]</td>
<td>0010</td>
<td>50 51 52 53 54 55 56 57</td>
<td></td>
</tr>
<tr>
<td>[EBX*2]</td>
<td>0011</td>
<td>58 59 5A 5B 5C 5D 5E 5F</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>1000</td>
<td>60 61 62 63 64 65 66 67</td>
<td></td>
</tr>
<tr>
<td>[EBP*2]</td>
<td>1001</td>
<td>68 69 6A 6B 6C 6D 6E 6F</td>
<td></td>
</tr>
<tr>
<td>[ESI*2]</td>
<td>1100</td>
<td>70 71 72 73 74 75 76 77</td>
<td></td>
</tr>
<tr>
<td>[EDI*2]</td>
<td>1101</td>
<td>78 79 7A 7B 7C 7D 7E 7F</td>
<td></td>
</tr>
<tr>
<td>[EAX*4]</td>
<td>0000</td>
<td>80 81 82 83 84 85 86 87</td>
<td></td>
</tr>
<tr>
<td>[ECX*4]</td>
<td>0001</td>
<td>88 89 8A 8B 8C 8D 8E 8F</td>
<td></td>
</tr>
<tr>
<td>[EDX*4]</td>
<td>0010</td>
<td>90 91 92 93 94 95 96 97</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>1000</td>
<td>A0 A1 A2 A3 A4 A5 A6 A7</td>
<td></td>
</tr>
<tr>
<td>[EBP*4]</td>
<td>1001</td>
<td>A8 A9 AA AB AC AD AE AF</td>
<td></td>
</tr>
<tr>
<td>[ESI*4]</td>
<td>1100</td>
<td>B0 B1 B2 B3 B4 B5 B6 B7</td>
<td></td>
</tr>
<tr>
<td>[EDI*4]</td>
<td>1101</td>
<td>B8 B9 BA BB BC BD BE BF</td>
<td></td>
</tr>
<tr>
<td>[EAX*8]</td>
<td>0000</td>
<td>C0 C1 C2 C3 C4 C5 C6 C7</td>
<td></td>
</tr>
<tr>
<td>[ECX*8]</td>
<td>0001</td>
<td>C8 C9 CA CB CC CD CE CF</td>
<td></td>
</tr>
<tr>
<td>[EDX*8]</td>
<td>0010</td>
<td>D0 D1 D2 D3 D4 D5 D6 D7</td>
<td></td>
</tr>
<tr>
<td>[EBX*8]</td>
<td>0011</td>
<td>DB DA DC DD DE DF</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>1000</td>
<td>E0 E1 E2 E3 E4 E5 E6 E7</td>
<td></td>
</tr>
<tr>
<td>[EBP*8]</td>
<td>1001</td>
<td>E8 EA EB EC ED EE EF</td>
<td></td>
</tr>
<tr>
<td>[ESI*8]</td>
<td>1100</td>
<td>F0 F1 F2 F3 F4 F5 F6 F7</td>
<td></td>
</tr>
<tr>
<td>[EDI*8]</td>
<td>1101</td>
<td>F8 F9 FA FB FC FD FE FF</td>
<td></td>
</tr>
</tbody>
</table>

**NOTES:**

1. The [*] nomenclature means a disp32 with no base if the MOD is 00B. Otherwise, [*] means disp8 or disp32 + [EBP]. This provides the following address modes:

<table>
<thead>
<tr>
<th>MOD bits</th>
<th>Effective Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>[scaled index] + disp32</td>
</tr>
<tr>
<td>01</td>
<td>[scaled index] + disp8 + [EBP]</td>
</tr>
<tr>
<td>10</td>
<td>[scaled index] + disp32 + [EBP]</td>
</tr>
</tbody>
</table>
X86 SIB-D Addressing Mode

Figure 3-11. Offset (or Effective Address) Computation

Offset = Base + (Index * Scale) + Displacement

x86 Manual Vol. 1, page 3-22 -- see course resources on website
Also, see Section 3.7.3 and 3.7.5
The following addressing modes suggest uses for common combinations of address components.

- **Displacement** — A displacement alone represents a direct (uncomputed) offset to the operand. Because the displacement is encoded in the instruction, this form of an address is sometimes called an absolute or static address. It is commonly used to access a statically allocated scalar operand.

- **Base** — A base alone represents an indirect offset to the operand. Since the value in the base register can change, it can be used for dynamic storage of variables and data structures.

- **Base + Displacement** — A base register and a displacement can be used together for two distinct purposes:
  - As an index into an array when the element size is not 2, 4, or 8 bytes—The displacement component encodes the static offset to the beginning of the array. The base register holds the results of a calculation to determine the offset to a specific element within the array.
  - To access a field of a record: the base register holds the address of the beginning of the record, while the displacement is a static offset to the field.

An important special case of this combination is access to parameters in a procedure activation record. A procedure activation record is the stack frame created when a procedure is entered. Here, the EBP register is the best choice for the base register, because it automatically selects the stack segment. This is a compact encoding for this common function.

---

x86 Manual Vol. 1, page 3-22 -- see course resources on website
Also, see Section 3.7.3 and 3.7.5
(Index * Scale) + Displacement — This address mode offers an efficient way to index into a static array when the element size is 2, 4, or 8 bytes. The displacement locates the beginning of the array, the index register holds the subscript of the desired array element, and the processor automatically converts the subscript into an index by applying the scaling factor.

Base + Index + Displacement — Using two registers together supports either a two-dimensional array (the displacement holds the address of the beginning of the array) or one of several instances of an array of records (the displacement is an offset to a field within the record).

Base + (Index * Scale) + Displacement — Using all the addressing components together allows efficient indexing of a two-dimensional array when the elements of the array are 2, 4, or 8 bytes in size.

x86 Manual Vol. 1, page 3-22  -- see course resources on website
Also, see Section 3.7.3 and 3.7.5
Other Example ISA-level Tradeoffs

- Condition codes vs. not
- VLIW vs. single instruction
- Precise vs. imprecise exceptions
- Virtual memory vs. not
- Unaligned access vs. not
- Hardware interlocks vs. software-guaranteed interlocking
- Software vs. hardware managed page fault handling
- Cache coherence (hardware vs. software)
- ...

23
Many ISA features designed to aid programmers
But, complicate the hardware designer’s job

Virtual memory
- vs. overlay programming
- Should the programmer be concerned about the size of code blocks fitting physical memory?

Addressing modes

Unaligned memory access
- Compile/programmer needs to align data
MIPS: Aligned Access

- LW/SW alignment restriction: 4-byte word-alignment
  - not designed to fetch memory bytes not within a word boundary
  - not designed to rotate unaligned bytes into registers
- Provide separate opcodes for the “infrequent” case

- LWL/LWR is slower
- Note LWL and LWR still fetch within word boundary
X86: Unaligned Access

- LD/ST instructions automatically align data that spans a “word” boundary
- Programmer/compiler does not need to worry about where data is stored (whether or not in a word-aligned location)

4.1.1 Alignment of Words, Doublewords, Quadwords, and Double Quadwords

Words, doublewords, and quadwords do not need to be aligned in memory on natural boundaries. The natural boundaries for words, double words, and quadwords are even-numbered addresses, addresses evenly divisible by four, and addresses evenly divisible by eight, respectively. However, to improve the performance of programs, data structures (especially stacks) should be aligned on natural boundaries whenever possible. The reason for this is that the processor requires two memory accesses to make an unaligned memory access; aligned accesses require only one memory access. A word or doubleword operand that crosses a 4-byte boundary or a quadword operand that crosses an 8-byte boundary is considered unaligned and requires two separate memory bus cycles for access.
X86: Unaligned Access

Figure 4-2. Bytes, Words, Doublewords, Quadwords, and Double Quadwords in Memory
What About ARM?

- https://www.scss.tcd.ie/~waldroj/3d1/arm_arm.pdf
  - Section A2.8
Aligned vs. Unaligned Access

- Pros of having no restrictions on alignment

- Cons of having no restrictions on alignment

- Filling in the above: an exercise for you...
Implementing the ISA: Microarchitecture Basics
How Does a Machine Process Instructions?

- What does processing an instruction mean?
- Remember the von Neumann model

\[ A = \text{Architectural (programmer visible) state before an instruction is processed} \]

\[ \text{Process instruction} \]

\[ A' = \text{Architectural (programmer visible) state after an instruction is processed} \]

- Processing an instruction: Transforming A to A’ according to the ISA specification of the instruction
The “Process instruction” Step

- ISA specifies abstractly what A’ should be, given an instruction and A
  - It defines an abstract finite state machine where
    - State = programmer-visible state
    - Next-state logic = instruction execution specification
  - From ISA point of view, there are no “intermediate states” between A and A’ during instruction execution
    - One state transition per instruction

- Microarchitecture implements how A is transformed to A’
  - There are many choices in implementation
  - We can have programmer-invisible state to optimize the speed of instruction execution: multiple state transitions per instruction
    - Choice 1: $A \rightarrow A'$ (transform A to A’ in a single clock cycle)
    - Choice 2: $A \rightarrow A+MS1 \rightarrow A+MS2 \rightarrow A+MS3 \rightarrow A'$ (take multiple clock cycles to transform A to A’)

32
A Very Basic Instruction Processing Engine

- Each instruction takes a single clock cycle to execute
- Only combinational logic is used to implement instruction execution
  - No intermediate, programmer-invisible state updates

A = Architectural (programmer visible) state at the beginning of a clock cycle

Process instruction in one clock cycle

A’ = Architectural (programmer visible) state at the end of a clock cycle
A Very Basic Instruction Processing Engine

- Single-cycle machine

What is the *clock cycle time* determined by?

What is the *critical path* of the combinational logic determined by?
Remember: Programmer Visible (Architectural) State

Memory
array of storage locations
indexed by an address

Program Counter
memory address
of the current instruction

Instructions (and programs) specify how to transform
the values of programmer visible state

Registers
- given special names in the ISA
  (as opposed to addresses)
- general vs. special purpose
Single-cycle vs. Multi-cycle Machines

- **Single-cycle machines**
  - Each instruction takes a single clock cycle
  - All state updates made at the end of an instruction’s execution
  - **Big disadvantage:** The slowest instruction determines cycle time → long clock cycle time

- **Multi-cycle machines**
  - Instruction processing broken into multiple cycles/stages
  - State updates can be made during an instruction’s execution
  - Architectural state updates made only at the end of an instruction’s execution
  - **Advantage over single-cycle:** The slowest “stage” determines cycle time

- Both single-cycle and multi-cycle machines literally follow the von Neumann model at the microarchitecture level
Instruction Processing “Cycle”

- Instructions are processed under the direction of a “control unit” step by step.
- Instruction cycle: Sequence of steps to process an instruction
- Fundamentally, there are six phases:
  - Fetch
  - Decode
  - Evaluate Address
  - Fetch Operands
  - Execute
  - Store Result
- Not all instructions require all six stages (see P&P Ch. 4)
Instruction Processing “Cycle” vs. Machine Clock Cycle

- Single-cycle machine:
  - All six phases of the instruction processing cycle take a *single machine clock cycle* to complete

- Multi-cycle machine:
  - All six phases of the instruction processing cycle can take *multiple machine clock cycles* to complete
  - In fact, *each phase can take multiple clock cycles to complete*
Instruction Processing Viewed Another Way

- Instructions transform Data (AS) to Data’ (AS’)
- This transformation is done by functional units
  - Units that “operate” on data
- These units need to be told what to do to the data

- An instruction processing engine consists of two components
  - Datapath: Consists of hardware elements that deal with and transform data signals
    - functional units that operate on data
    - hardware structures (e.g. wires and muxes) that enable the flow of data into the functional units and registers
    - storage units that store data (e.g., registers)
  - Control logic: Consists of hardware elements that determine control signals, i.e., signals that specify what the datapath elements should do to the data
Single-cycle vs. Multi-cycle: Control & Data

- **Single-cycle machine:**
  - Control signals are generated in the same clock cycle as data signals are operated on.
  - Everything related to an instruction happens in one clock cycle.

- **Multi-cycle machine:**
  - Control signals needed in the next cycle can be generated in the previous cycle.
  - Latency of control processing can be overlapped with latency of datapath operation.

- We will see the difference clearly in *microprogrammed multi-cycle microarchitecture*.
Many Ways of Datapath and Control Design

- There are many ways of designing the data path and control logic

- Single-cycle, multi-cycle, pipelined datapath and control

- Single-bus vs. multi-bus datapaths
  - See your homework 2 question

- Hardwired/combinational vs. microcoded/microprogrammed control
  - Control signals generated by combinational logic versus
  - Control signals stored in a memory structure

- Control signals and structure depend on the datapath design
Flash-Forward: Performance Analysis

- Execution time of an instruction
  - \{\text{CPI}\} \times \{\text{clock cycle time}\}

- Execution time of a program
  - Sum over all instructions [\{\text{CPI}\} \times \{\text{clock cycle time}\}]
  - \{\# \text{ of instructions}\} \times \{\text{Average CPI}\} \times \{\text{clock cycle time}\}

- Single cycle microarchitecture performance
  - CPI = 1
  - Clock cycle time = long

- Multi-cycle microarchitecture performance
  - CPI = different for each instruction
    - Average CPI \rightarrow \text{hopefully small}
  - Clock cycle time = short

Now, we have two degrees of freedom to optimize independently