

18-447

Computer Architecture

Lecture 4: More ISA Tradeoffs and Single-Cycle Microarchitectures

Prof. Onur Mutlu

Carnegie Mellon University

Spring 2012, 1/25/2012

Reminder: Homeworks for Next Two Weeks

- Homework 1
 - Due tonight, 11:59pm, on Blackboard
 - MIPS warmup, ISA concepts, basic performance evaluation

- Homework 2
 - Will be out today or early tomorrow
 - Due February 13

Reminder: Lab Assignment 1

- Due this Friday (Feb 3), at the end of Friday lab

A Note on the Primary Textbook

- 4th Edition
- Revised 4th Edition

- They are very similar → you are fine if you have either
- Problem statements are different
 - We will specify the full problems in the future

Review of Last Lecture

- ISA Principles and Tradeoffs
- Elements of the ISA
 - Sequencing model, instruction processing style
 - Instructions, data types, memory organization, registers, addressing modes, orthogonality, I/O device interfacing ...
- What is the benefit of autoincrement addressing mode?
- What is the downside of having an autoincrement addressing mode?
- Is the LC-3b ISA orthogonal?
 - Can all addressing modes be used with all instructions?

Is the LC-3b ISA Orthogonal?

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
ADD ⁺	0001			DR			SR1			A	op.spec					
AND ⁺	0101			DR			SR1			A	op.spec					
BR	0000			n	z	p	PCoffset9									
JMP	1100			000			BaseR			000000						
JSR(R)	0100			A	operand.specifier											
LDB ⁺	0010			DR			BaseR			boffset6						
LDW ⁺	0110			DR			BaseR			offset6						
LEA ⁺	1110			DR			PCoffset9									
RTI	1000			000000000000												
SHF ⁺	1101			DR			SR			A	D	amount4				
STB	0011			SR			BaseR			boffset6						
STW	0111			SR			BaseR			offset6						
TRAP	1111			0000			trapvect8									
XOR ⁺	1001			DR			SR1			A	op.spec					
not used	1010															
not used	1011															

Complex vs. Simple Instructions

- Complex instruction: An instruction does a lot of work, e.g. many operations
 - Insert in a doubly linked list
 - Compute FFT
 - String copy

- Simple instruction: An instruction does small amount of work, it is a primitive
 - Add
 - XOR
 - Multiply

Complex vs. Simple Instructions

- Advantages of Complex instructions
 - + Denser encoding → smaller code size → better memory utilization, saves off-chip bandwidth, better cache hit rate (better packing of instructions)
 - + Simpler compiler: no need to optimize small instructions as much
- Disadvantages of Complex Instructions
 - Larger chunks of work → compiler has less opportunity to optimize
 - More complex hardware → translation from a high level to control signals and optimization needs to be done by hardware
 - Compiler is limited in fine-grained optimizations it can do

ISA-level Tradeoffs: Semantic Gap

- **Where to place the ISA?** Semantic gap
 - Closer to high-level language (HLL) → Small semantic gap, complex instructions
 - Closer to hardware control signals? → Large semantic gap, simple instructions
- RISC vs. CISC machines
 - FFT, QUICKSORT, POLY, FP instructions?
 - VAX INDEX instruction (array access with bounds checking)

ISA-level Tradeoffs: Semantic Gap

- Some tradeoffs (for you to think about)
- Simple compiler, complex hardware vs. complex compiler, simple hardware
 - Caveat: Translation (indirection) can change the tradeoff!
- Burden of backward compatibility
- Performance?
 - Optimization opportunity: Example of VAX INDEX instruction: who (compiler vs. hardware) puts more effort into optimization?
 - Instruction size, code size

X86: Small Semantic Gap: String Operations

- An instruction operates on a string
 - Move one string of arbitrary length to another location
 - Compare two strings
- Enabled by the ability to specify repeated execution of an instruction (in the ISA)
 - Using a “prefix” called REP prefix
- Example: REP MOVS instruction
 - Only two bytes: REP prefix byte and MOVS opcode byte (F2 A4)
 - Implicit source and destination registers pointing to the two strings (ESI, EDI)
 - Implicit count register (ECX) specifies how long the string is

X86: Small Semantic Gap: String Operations

REP MOVS (DEST SRC)

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

```
DEST ← SRC;
IF (Byte move)
  THEN IF DF = 0
    THEN
      (R)ESI ← (R)ESI + 1;
      (R)EDI ← (R)EDI + 1;
    ELSE
      (R)ESI ← (R)ESI - 1;
      (R)EDI ← (R)EDI - 1;
    FI;
  ELSE IF (Word move)
    THEN IF DF = 0
      (R)ESI ← (R)ESI + 2;
      (R)EDI ← (R)EDI + 2;
      FI;
    ELSE
      (R)ESI ← (R)ESI - 2;
      (R)EDI ← (R)EDI - 2;
    FI;
  ELSE IF (Doubleword move)
    THEN IF DF = 0
      (R)ESI ← (R)ESI + 4;
      (R)EDI ← (R)EDI + 4;
      FI;
    ELSE
      (R)ESI ← (R)ESI - 4;
      (R)EDI ← (R)EDI - 4;
    FI;
  ELSE IF (Quadword move)
    THEN IF DF = 0
      (R)ESI ← (R)ESI + 8;
      (R)EDI ← (R)EDI + 8;
      FI;
    ELSE
      (R)ESI ← (R)ESI - 8;
      (R)EDI ← (R)EDI - 8;
    FI;
  FI;
```

How many instructions does this take in 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
- Digital Equipment Corp., “[VAX11 780 Architecture Handbook](#),” 1977-78.

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 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 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 dynamic/static 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

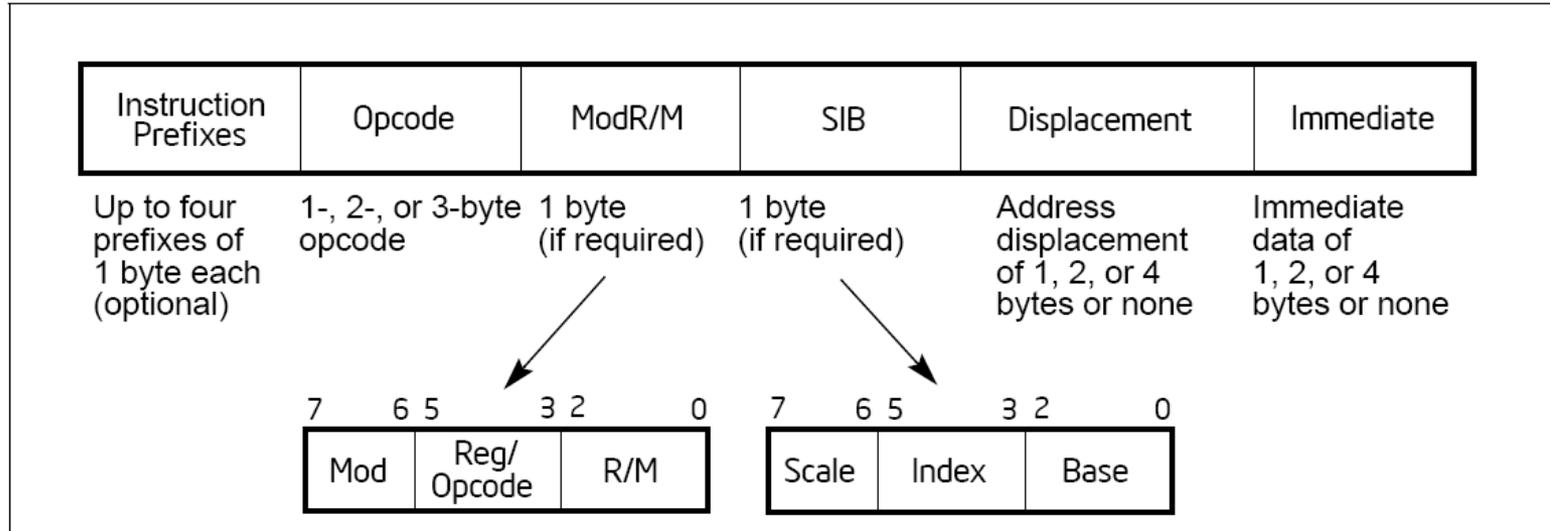
- **Fixed length:** Length of all instructions the same
 - + Easier to decode single instruction in hardware
 - + Easier to decode multiple instructions concurrently
 - Wasted bits in instructions (**Why is this bad?**)
 - Harder-to-extend ISA (how to add new instructions?)
- **Variable length:** Length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

ISA-level Tradeoffs: Uniform Decode

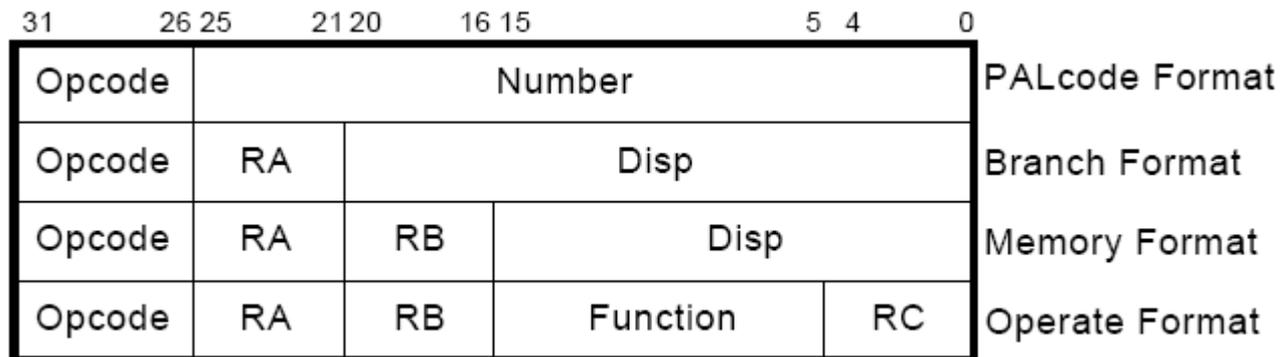
- **Uniform decode:** Same bits in each instruction correspond to the same meaning
 - Opcode is always in the same location
 - Ditto operand specifiers, immediate values, ...
 - Many “RISC” ISAs: Alpha, MIPS, SPARC
 - + Easier decode, simpler hardware
 - + Enables parallelism: generate target address 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:



■ Alpha:



MIPS Instruction Format

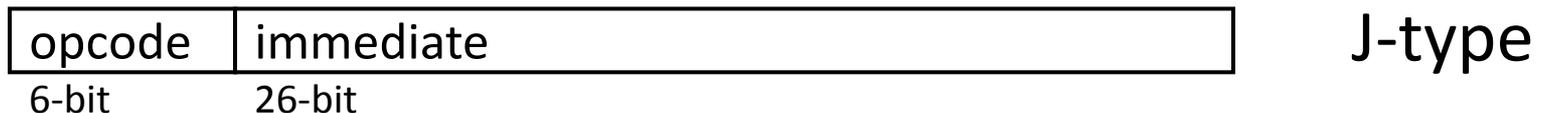
- R-type, 3 register operands



- I-type, 2 register operands and 16-bit immediate operand



- J-type, 26-bit immediate operand



- 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

ISA-level Tradeoffs: Number of Registers

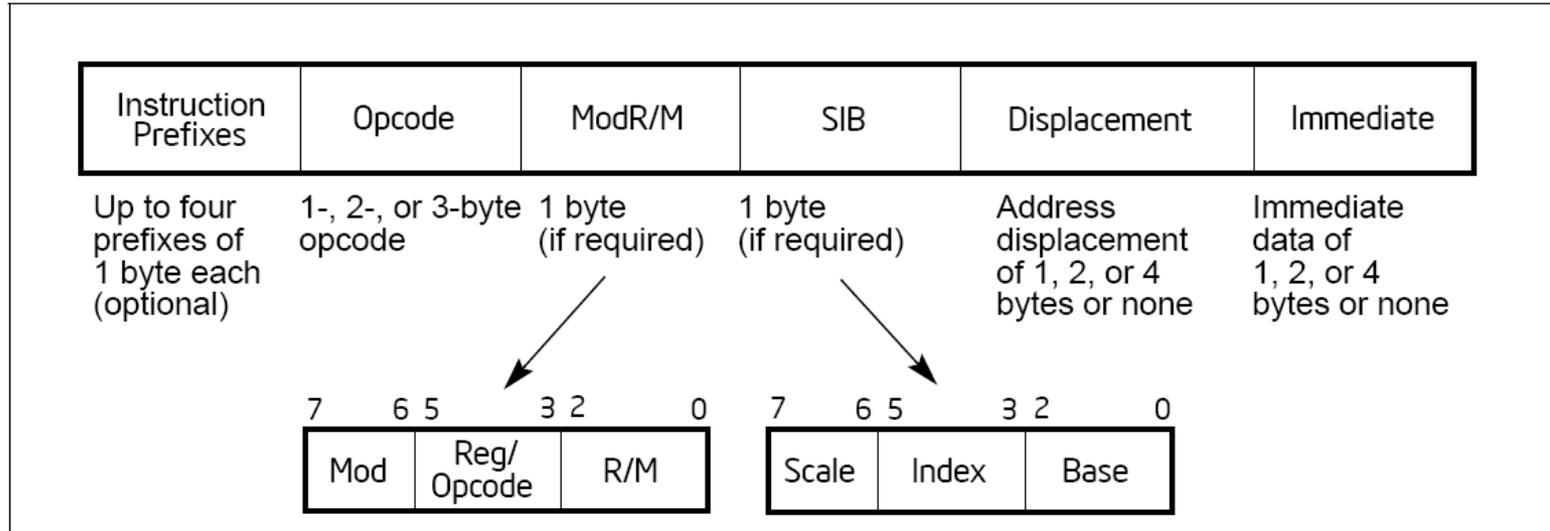
- Affects:
 - Number of bits used for encoding register address
 - Number of values kept in fast storage (register file)
 - (uarch) Size, access time, power consumption of register file
- Large number of registers:
 - + Enables better register allocation (and optimizations) by compiler → fewer saves/restores
 - Larger instruction size
 - Larger register file size

Review: 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
 - Read *Wulf*, “*Compilers and Computer Architecture*”

x86 vs. Alpha Instruction Formats

■ x86:



■ Alpha:

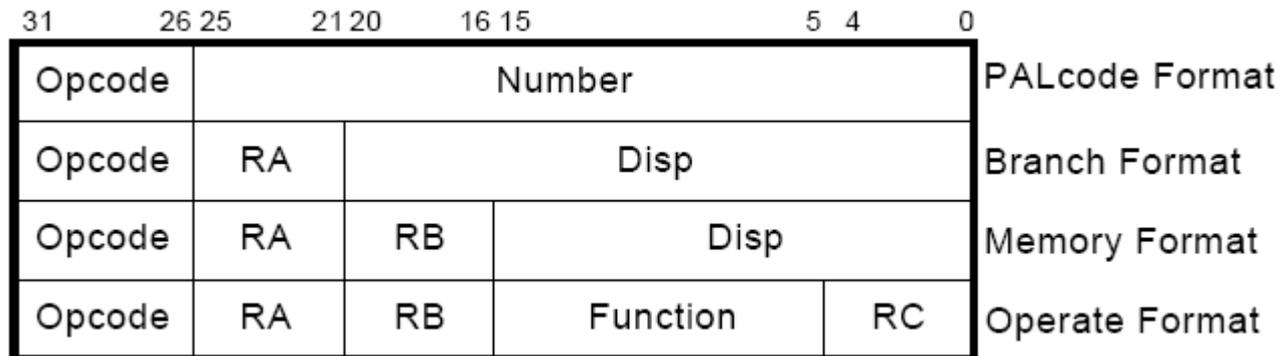


Table 2-2. 32-Bit Addressing Forms with the ModR/M Byte

x86

			AL AX	CL CX	DL DX	BL BX	AH SP	CH BP	DH SI	BH DI
			EAX	ECX	EDX	EBX	ESP	EBP	ESI	EDI
			MM0	MM1	MM2	MM3	MM4	MM5	MM6	MM7
			XMM0	XMM1	XMM2	XMM3	XMM4	XMM5	XMM6	XMM7
			0	1	2	3	4	5	6	7
			000	001	010	011	100	101	110	111
			Value of ModR/M Byte (in Hexadecimal)							
Effective Address	Mod	R/M								
[EAX]	00	000	00	08	10	18	20	28	30	38
[ECX]		001	01	09	11	19	21	29	31	39
[EDX]		010	02	0A	12	1A	22	2A	32	3A
[EBX]		011	03	0B	13	1B	23	2B	33	3B
[--][--] ¹		100	04	0C	14	1C	24	2C	34	3C
disp32 ²		101	05	0D	15	1D	25	2D	35	3D
[ESI]		110	06	0E	16	1E	26	2E	36	3E
[EDI]		111	07	0F	17	1F	27	2F	37	3F
[EAX]+disp8 ³	01	000	40	48	50	58	60	68	70	78
[ECX]+disp8		001	41	49	51	59	61	69	71	79
[EDX]+disp8		010	42	4A	52	5A	62	6A	72	7A
[EBX]+disp8		011	43	4B	53	5B	63	6B	73	7B
[--][--]+disp8		100	44	4C	54	5C	64	6C	74	7C
[EBP]+disp8		101	45	4D	55	5D	65	6D	75	7D
[ESI]+disp8		110	46	4E	56	5E	66	6E	76	7E
[EDI]+disp8		111	47	4F	57	5F	67	6F	77	7F
[EAX]+disp32	10	000	80	88	90	98	A0	A8	B0	B8
[ECX]+disp32		001	81	89	91	99	A1	A9	B1	B9
[EDX]+disp32		010	82	8A	92	9A	A2	AA	B2	BA
[EBX]+disp32		011	83	8B	93	9B	A3	AB	B3	BB
[--][--]+disp32		100	84	8C	94	9C	A4	AC	B4	BC
[EBP]+disp32		101	85	8D	95	9D	A5	AD	B5	BD
[ESI]+disp32		110	86	8E	96	9E	A6	AE	B6	BE
[EDI]+disp32		111	87	8F	97	9F	A7	AF	B7	BF
EAX/AX/AL/MM0/XMM0	11	000	C0	C8	D0	D8	E0	E8	F0	F8
ECX/CX/CL/MM1/XMM1		001	C1	C9	D1	D9	E1	E9	F1	F9
EDX/DX/DL/MM2/XMM2		010	C2	CA	D2	DA	E2	EA	F2	FA
EBX/BX/BL/MM3/XMM3		011	C3	CB	D3	DB	E3	EB	F3	FB
ESP/SP/AH/MM4/XMM4		100	C4	CC	D4	DC	E4	EC	F4	FC
EBP/BP/CH/MM5/XMM5		101	C5	CD	D5	DD	E5	ED	F5	FD
ESI/SI/DH/MM6/XMM6		110	C6	CE	D6	DE	E6	EE	F6	FE
EDI/DI/BH/MM7/XMM7		111	C7	CF	D7	DF	E7	EF	F7	FF

register
indirect

absolute

register +
displacement

register

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

r32 (In decimal) Base – (In binary) Base –			EAX 0 000	ECX 1 001	EDX 2 010	EBX 3 011	ESP 4 100	[*] 5 101	ESI 6 110	EDI 7 111
Scaled Index	SS	Index	Value of SIB Byte (in Hexadecimal)							
[EAX] [ECX] [EDX] [EBX] none [EBP] [ESI] [EDI]	00	000 001 010 011 100 101 110 111	00 08 10 18 20 28 30 38	01 09 11 19 21 29 31 39	02 0A 12 1A 22 2A 32 3A	03 0B 13 1B 23 2B 33 3B	04 0C 14 1C 24 2C 34 3C	05 0D 15 1D 25 2D 35 3D	06 0E 16 1E 26 2E 36 3E	07 0F 17 1F 27 2F 37 3F
[EAX+2] [ECX+2] [EDX+2] [EBX+2] none [EBP+2] [ESI+2] [EDI+2]	01	000 001 010 011 100 101 110 111	40 48 50 58 60 68 70 78	41 49 51 59 61 69 71 79	42 4A 52 5A 62 6A 72 7A	43 4B 53 5B 63 6B 73 7B	44 4C 54 5C 64 6C 74 7C	45 4D 55 5D 65 6D 75 7D	46 4E 56 5E 66 6E 76 7E	47 4F 57 5F 67 6F 77 7F
[EAX+4] [ECX+4] [EDX+4] [EBX+4] none [EBP+4] [ESI+4] [EDI+4]	10	000 001 010 011 100 101 110 111	80 88 90 98 A0 A8 B0 B8	81 89 91 99 A1 A9 B1 B9	82 8A 92 9A A2 AA B2 BA	83 8B 93 9B A3 AB B3 BB	84 8C 94 9C A4 AC B4 BC	85 8D 95 9D A5 AD B5 BD	86 8E 96 9E A6 AE B6 BE	87 8F 97 9F A7 AF B7 BF
[EAX+8] [ECX+8] [EDX+8] [EBX+8] none [EBP+8] [ESI+8] [EDI+8]	11	000 001 010 011 100 101 110 111	C0 C8 D0 D8 E0 E8 F0 F8	C1 C9 D1 D9 E1 E9 F1 F9	C2 CA D2 DA E2 EA F2 FA	C3 CB D3 DB E3 EB F3 FB	C4 CC D4 DC E4 EC F4 FC	C5 CD D5 DD E5 ED F5 FD	C6 CE D6 DE E6 EE F6 FE	C7 CF D7 DF E7 EF F7 FF

indexed
(base +
index)

scaled
(base +
index*4)

NOTES:

- 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:

MOD bits	Effective Address
00	[scaled index] + disp32
01	[scaled index] + disp8 + [EBP]
10	[scaled index] + disp32 + [EBP]

X86 SIB-D Addressing Mode

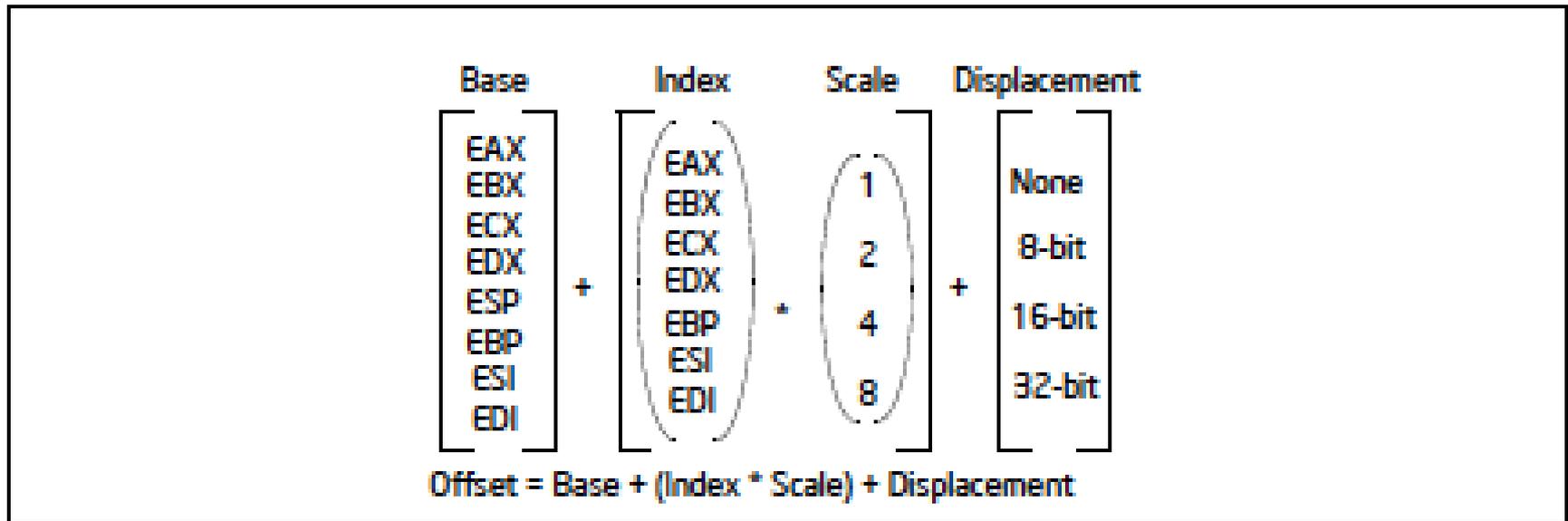


Figure 3-11. Offset (or Effective Address) Computation

X86 Manual: Suggested Uses of Addressing Modes

- **(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.

Other Example ISA-level Tradeoffs

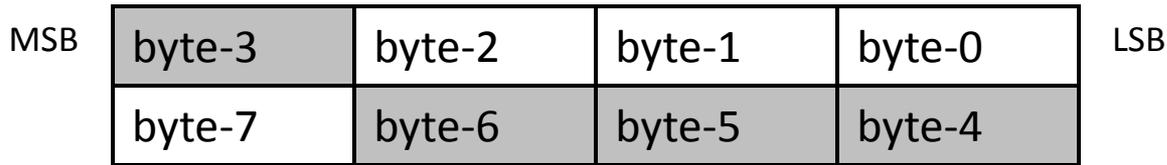
- Condition codes vs. not
- VLIW vs. single instruction
- Precise vs. imprecise exceptions
- Virtual memory vs. not
- Unaligned access vs. not
- Hardware interlocks vs. software-guaranteed interlocking
- Software vs. hardware managed page fault handling
- Cache coherence (hardware vs. software)
- ...

Back to Programmer vs. (Micro)architect

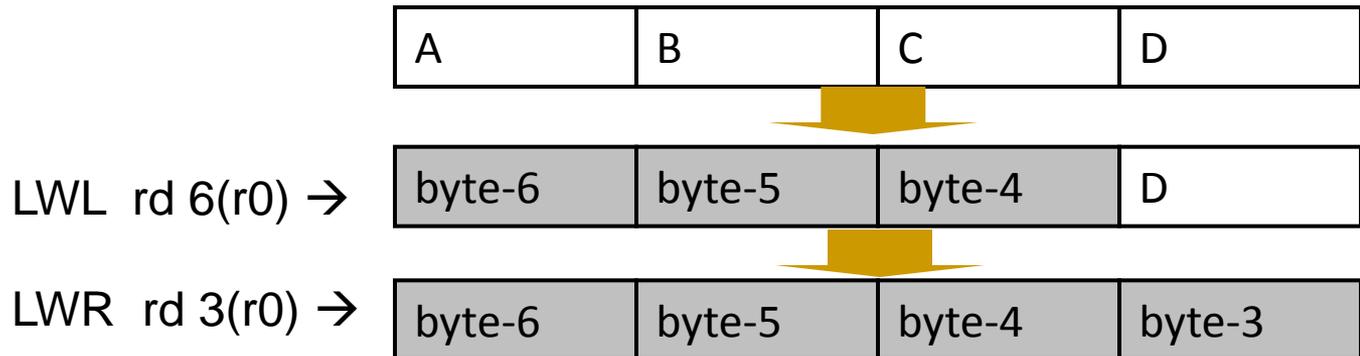
- Many ISA features designed to aid programmers
- But, complicate the hardware designer's job

- 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

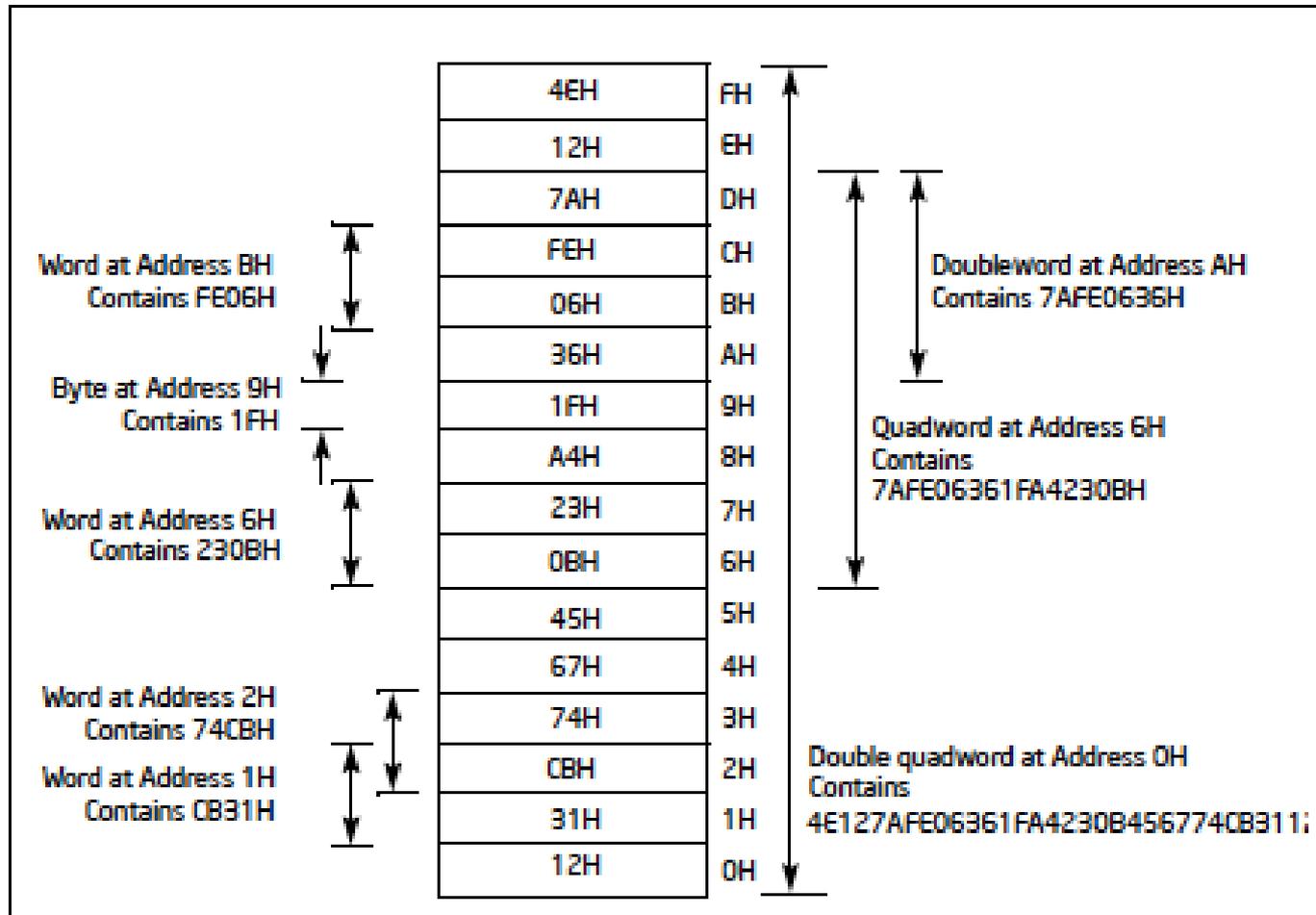


Figure 4-2. Bytes, Words, Doublewords, Quadwords, and Double Quadwords in Memory

Aligned vs. Unaligned Access

- Pros of having no restrictions on alignment

- Cons of having no restrictions on alignment

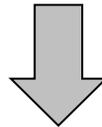
- Filling in the above: an exercise for you...

Implementing the ISA: Microarchitecture Basics

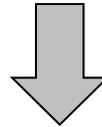
How Does a Machine Process Instructions?

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

AS = Architectural (programmer visible) state before an instruction is processed



Process instruction



AS' = Architectural (programmer visible) state after an instruction is processed

- Processing an instruction: Transforming A to A' according to the ISA specification of the instruction

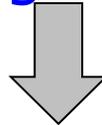
The “Process instruction” Step

- 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: $AS \rightarrow AS'$ (transform A to A' in a single clock cycle)
 - Choice 2: $AS \rightarrow AS+MS1 \rightarrow AS+MS2 \rightarrow AS+MS3 \rightarrow AS'$ (take multiple clock cycles to transform AS to AS')

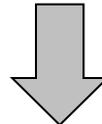
A Very Basic Instruction Processing Engine

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

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



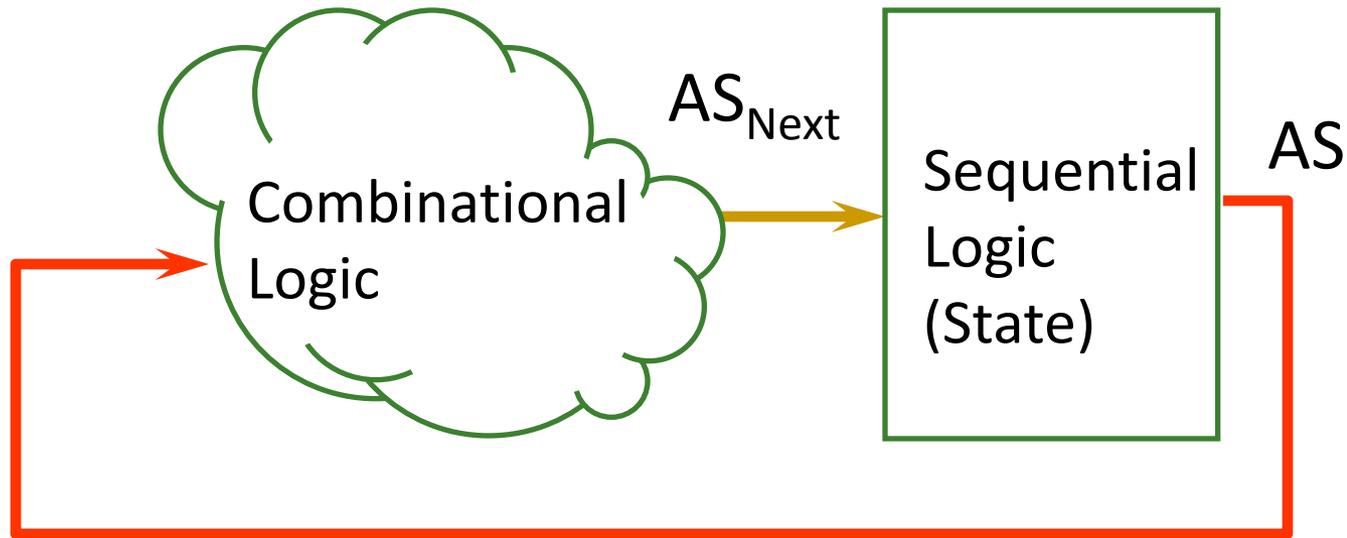
Process instruction in one clock cycle



AS' = 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 determined by?