UNIT 1 - Basic Structure of Computers

Functional Units

Figure 1.1. Basic functional units of a computer.

Information Handled by a Computer

- Instructions/machine instructions
- \geq Govern the transfer of information within a computer as well as between the computer and its I/O devices
- \geq Specify the arithmetic and logic operations to be performed
- \triangleright Program

Data

- \triangleright Used as operands by the instructions
- \triangleright Source program
- Encoded in binary code 0 and 1

Memory Unit

- Store programs and data
- Two classes of storage
- \triangleright Primary storage
- Fast
- * Programs must be stored in memory while they are being executed
- * Large number of semiconductor storage cells
- **↑ Processed in words**
- *❖* Address
- RAM and memory access time
- \cdot Memory hierarchy cache, main memory
- \geq Secondary storage larger and cheaper

Arithmetic and Logic Unit (ALU)

- Most computer operations are executed in ALU of the processor.
- Load the operands into memory bring them to the processor – perform operation in ALU – store the result back to memory or retain in the processor.
- Registers
- Fast control of ALU

Control Unit

- All computer operations are controlled by the control unit.
- The timing signals that govern the I/O transfers are also generated by the control unit.
- Control unit is usually distributed throughout the machine instead of standing alone.
- Operations of a computer:
- \geq Accept information in the form of programs and data through an input unit and store it in the memory
- \triangleright Fetch the information stored in the memory, under program control, into an ALU, where the information is processed
- \triangleright Output the processed information through an output unit
- Control all activities inside the machine through a control unit

Two types of functional units:

elements that operate on data values (combinational)

elements that contain state (state elements)

Five Execution Steps

Basic Operational Concepts

Review

- Activity in a computer is governed by instructions.
- To perform a task, an appropriate program consisting of a list of instructions is stored in the memory.
- Individual instructions are brought from the memory into the processor, which executes the specified operations.
- Data to be used as operands are also stored in the memory.

A Typical Instruction

- Add LOCA, R0
- Add the operand at memory location LOCA to the operand in a register R0 in the processor.
- Place the sum into register R0.
- The original contents of LOCA are preserved.
- The original contents of R0 is overwritten.
- Instruction is fetched from the memory into the processor – the operand at LOCA is fetched and added to the contents of R0 – the resulting sum is stored in register R0.

Separate Memory Access and ALU Operation

- Load LOCA, R1
- Add R1, R0
- Whose contents will be overwritten?

Connection Between the Processor and the Memory

Figure 1.2. Connections between the processor and the memory.

Registers

- Instruction register (IR)
- Program counter (PC)
- General-purpose register $(R_{0} R_{n-1})$
- Memory address register (MAR)
- Memory data register (MDR)

Typical Operating Steps

- Programs reside in the memory through input devices
- PC is set to point to the first instruction
- The contents of PC are transferred to MAR
- A Read signal is sent to the memory
- The first instruction is read out and loaded into MDR
- The contents of MDR are transferred to IR
- Decode and execute the instruction

Typical Operating Steps (Cont')

- Get operands for ALU
	- **▶ General-purpose register**
	- \triangleright Memory (address to MAR Read MDR to ALU)
- Perform operation in ALU
- Store the result back
	- \geq To general-purpose register
	- \triangleright To memory (address to MAR, result to MDR Write)
- During the execution, PC is incremented to the next instruction

Interrupt

- Normal execution of programs may be preempted if some device requires urgent servicing.
- The normal execution of the current program must be interrupted – the device raises an *interrupt* signal.
- Interrupt-service routine
- Current system information backup and restore (PC, general-purpose registers, control information, specific information)

Bus Structures

- There are many ways to connect different parts inside a computer together.
- A group of lines that serves as a connecting path for several devices is called a *bus*.
- Address/data/control

Bus Structure

• Single-bus

Speed Issue

- Different devices have different transfer/operate speed.
- If the speed of bus is bounded by the slowest device connected to it, the efficiency will be very low.
- How to solve this?
- A common approach use buffers.

Performance

- The most important measure of a computer is how quickly it can execute programs.
- Three factors affect performance:
- \triangleright Hardware design
- \triangleright Instruction set
- Compiler

 Processor time to execute a program depends on the hardware involved in the execution of individual machine instructions.

Figure 1.5. The processor cache.

- The processor and a relatively small cache memory can be fabricated on a single integrated circuit chip.
- Speed
- Cost
- Memory management

Processor Clock

- Clock, clock cycle, and clock rate
- The execution of each instruction is divided into several steps, each of which completes in one clock cycle.
- Hertz cycles per second

Basic Performance Equation

- N number of actual machine language instructions needed to complete the execution (note: loop)
- S average number of basic steps needed to execute one machine instruction. Each step completes in one clock cycle
- $R clock rate$
- Note: these are not independent to each other

$$
T = \frac{N \times S}{R}
$$

How to improve T?

Pipeline and Superscalar Operation

- Instructions are not necessarily executed one after another.
- The value of S doesn't have to be the number of clock cycles to execute one instruction.
- Pipelining overlapping the execution of successive instructions.
- Add R1, R2, R3
- Superscalar operation multiple instruction pipelines are implemented in the processor.
- Goal reduce S (could become <1!)

Clock Rate

- Increase clock rate
- \ge Improve the integrated-circuit (IC) technology to make the circuits faster
- \triangleright Reduce the amount of processing done in one basic step (however, this may increase the number of basic steps needed)
- Increases in R that are entirely caused by improvements in IC technology affect all aspects of the processor's operation equally except the time to access the main memory.

CISC and RISC

- Tradeoff between N and S
- A key consideration is the use of pipelining
- \ge S is close to 1 even though the number of basic steps per instruction may be considerably larger
- \geq It is much easier to implement efficient pipelining in processor with simple instruction sets
- Reduced Instruction Set Computers (RISC)
- Complex Instruction Set Computers (CISC)

Compiler

- A compiler translates a high-level language program into a sequence of machine instructions.
- To reduce N, we need a suitable machine instruction set and a compiler that makes good use of it.
- Goal reduce N×S
- A compiler may not be designed for a specific processor; however, a high-quality compiler is usually designed for, and with, a specific processor.

Performance Measurement

- T is difficult to compute.
- Measure computer performance using benchmark programs.
- System Performance Evaluation Corporation (SPEC) selects and publishes representative application programs for different application domains, together with test results for many commercially available computers.
- Compile and run (no simulation)
- Reference computer

= *SPEC rating* Running time on the computer under test Running time on the reference computer

$$
SPEC \; rating = (\prod_{i=1}^{n} SPEC_i)^{\frac{1}{n}}
$$

Machine Instructions and Programs

Objectives

- Machine instructions and program execution, including branching and subroutine call and return operations.
- Number representation and addition/subtraction in the 2's-complement system.
- Addressing methods for accessing register and memory operands.
- Assembly language for representing machine instructions, data, and programs.
- Program-controlled Input/Output operations.

Memory Locations, Addresses, and Operations

Memory Location, Addresses, and Operation

- Memory consists of many millions of storage cells, each of which can store 1 bit.
- Data is usually accessed in *n*-bit groups. *n* is called word length.

Figure 2.5. Memory words.

- To retrieve information from memory, either for one
- word or one byte (8-bit), addresses for each location are needed.
- A k-bit address memory has 2^k memory locations, namely $0 - 2^k - 1$, called memory space.
- 24-bit memory: $2^{24} = 16,777,216 = 16M (1M=2^{20})$
- 32-bit memory: $2^{32} = 4G (1G=2^{30})$
- $1K(kilo)=2^{10}$
- $1T(tera)=2^{40}$

- It is impractical to assign distinct addresses to individual bit locations in the memory.
- The most practical assignment is to have successive addresses refer to successive byte locations in the memory – byteaddressable memory.
- Byte locations have addresses 0, 1, 2, ... If word length is 32 bits, they successive words are located at addresses 0, 4, 8,…

Big-Endian and Little-Endian Assignments

Big-Endian: lower byte addresses are used for the most significant bytes of the word

Little-Endian: opposite ordering. lower byte addresses are used for the less significant bytes of the word

(a) Big-endian assignment (b) Little-endian assignment

Figure 2.7. Byte and word addressing.

- Address ordering of bytes
- Word alignment
	- Words are said to be aligned in memory if they begin at a byte addr. that is a multiple of the num of bytes in a word.
		- 16-bit word: word addresses: 0, 2, 4,….
		- 32-bit word: word addresses: 0, 4, 8,....
		- 64-bit word: word addresses: 0, 8,16,….
- Access numbers, characters, and character strings

Memory Operation

- Load (or Read or Fetch)
- \geq Copy the content. The memory content doesn't change.
- \triangleright Address Load
- \triangleright Registers can be used
- Store (or Write)
- \triangleright Overwrite the content in memory
- \triangleright Address and Data Store
- \triangleright Registers can be used

Instruction and Instruction Sequencing

"Must-Perform" Operations

- Data transfers between the memory and the processor registers
- Arithmetic and logic operations on data
- Program sequencing and control
- I/O transfers

Register Transfer Notation

- Identify a location by a symbolic name standing for its hardware binary address (LOC, R0,…)
- Contents of a location are denoted by placing square brackets around the name of the location $(R1 \leftarrow [LOC], R3 \leftarrow [R1] + [R2])$
- Register Transfer Notation (RTN)

Assembly Language Notation

- Represent machine instructions and programs.
- \bullet Move LOC, R1 = R1 \leftarrow [LOC]
- Add R1, R2, R3 = R3 \leftarrow [R1]+ [R2]

CPU Organization

- Single Accumulator
	- Result usually goes to the Accumulator
	- Accumulator has to be saved to memory quite often
- **General Register**
	- Registers hold operands thus reduce memory traffic
	- Register bookkeeping
- Stack
	- Operands and result are always in the stack

- Three-Address Instructions \bullet ADD R1, R2, R3 R1 ← R2 + R3 • Two-Address Instructions \bullet ADD R1, R2 R1 ← R1 + R2 • One-Address Instructions \bullet ADD M \bullet AC ← AC + M[AR] • Zero-Address Instructions \bullet ADD \bullet TOS \leftarrow TOS + (TOS – 1)
- RISC Instructions
	- Lots of registers. Memory is restricted to Load & Store

- **•** Three-Address
	-
	- l ADD R2, C, D; R2 ← M[C] + M[D]
	- ^l MUL X, R1, R2 ; M[X] ← R1 ∗ R2
- $ADD \t R1, A, B \t ; R1 \leftarrow M[A] + M[B]$
	-
	-

- Two-Address
	- $\text{1} \quad \text{MOV} \quad \text{R1, A} \quad ; \text{R1} \leftarrow \text{M[A]}$
	-
	- $1 \quad \text{MOV} \quad \text{R2, C} \quad ; \text{R2} \leftarrow \text{M}[\text{C}]$
	-
	- ^l MUL R1, R2 ; R1 ← R1 ∗ R2
	- 1 MOV X, R1 ; M[X] \leftarrow R1
-
- l ADD R1, B ; R1 ← R1 + M[B]
	-
- l ADD R2, D ; R2 ← R2 + M[D]
	-
	-

- One-Address
	-
	-
	-
	- \blacksquare LOAD C ; AC ← M[C]
	-
	-
	- $\text{1} \quad \text{STOREX} \quad ; \text{M[X]} \leftarrow \text{AC}$
- $\mathsf{LOAD} \quad \mathsf{A} \qquad \qquad ; \mathsf{AC} \leftarrow \mathsf{M}[\mathsf{A}]$
- ADD B ; $AC \leftarrow AC + M[B]$
- \perp STORET ; M[T] \leftarrow AC
	-
	- $\begin{array}{ccc} \mathsf{ADD} & \mathsf{D} & \mathsf{; AC \leftarrow AC + M[D]} \end{array}$
- \blacksquare MUL T ; AC ← AC * M[T]
	-

- Zero-Address
	-
	- ^l PUSH B ; TOS ← B
	-
	- l PUSH C ; TOS \leftarrow C
	-
	-
	- MUL ; TOS \leftarrow (C+D)∗(A+B)
	-
- $PUSH A$; TOS $\leftarrow A$ ADD ; TOS $\leftarrow (A + B)$ PUSH D ; TOS ← D ADD ; TOS \leftarrow (C + D)
	- POP X ; M[X] $\leftarrow \mathsf{TOS}$

Instruction Formats Example: Evaluate (A+B) ∗ (C+D)

- RISC
	- $\textsf{LOAD} \quad \textsf{R1}, \textsf{A} \qquad ; \textsf{R1} \leftarrow \textsf{M}[\textsf{A}]$
	- l LOAD R2, B ; R2 ← M[B]
	- l LOAD R3, C ; R3 ← M[C]
	- l LOAD R4, D ; R4 ← M[D]
	- l ADD R1, R1, R2 ; R1 ← R1 + R2
	- l ADD R3, R3, R4 ; R3 ← R3 + R4
	- ^l MUL R1, R1, R3 ; R1 ← R1 ∗ R3
	- I STOREX, R1 ; M[X] \leftarrow R1
-
-
-
-
-
-
-
-

Using Registers

- Registers are faster
- Shorter instructions
	- The number of registers is smaller (e.g. 32 registers need 5 bits)
- Potential speedup
- Minimize the frequency with which data is moved back and forth between the memory and processor registers.

Instruction Execution and Straight-Line Sequencing

Assumptions:

- One memory operand per instruction
- 32-bit word length
- Memory is byte addressable
- Full memory address can be directly specified in a single-word instruction

Two-phase procedure -Instruction fetch -Instruction execute

Page 43

Figure 2.8. A program for $C \leftarrow [A] + [B]$.

Branching

Figure 2.9. A straight-line program for adding *n* numbers.

NUM*n*

Condition Codes

- Condition code flags
- Condition code register / status register
- N (negative)
- Z (zero)
- V (overflow)
- C (carry)
- Different instructions affect different flags

Conditional Branch Instructions

- Example:
	- A: 11110000
	- **B: 00010100**

Generating Memory Addresses

-
- How to specify the address of branch target?
- Can we give the memory operand address directly in a single Add instruction in the loop?
- Use a register to hold the address of NUM1; then increment by 4 on each pass through the loop.

• Implied

- AC is implied in "ADD M[AR]" in "One-Address" instr.
- TOS is implied in "ADD" in "Zero-Address" instr.
- Immediate
	- The use of a constant in "MOV $R1, 5$ ", i.e. R1 \leftarrow 5
- Register
	- Indicate which register holds the operand

- Register Indirect
	- Indicate the register that holds the number of the register that holds the operand MOV R1, (R2) **R1**
- Autoincrement / Autodecrement
	- Access & update in 1 instr.
- Direct Address
	- Use the given address to access a memory location

- Indirect Address
	- Indicate the memory location that holds the address of the memory location that holds the data

- Indexed
	- *EA* = Index Register + Relative Addr

2/1/10

Indexing and Arrays

- Index mode the effective address of the operand is generated by adding a constant value to the contents of a register.
- Index register
- $X(R_i)$: EA = X + [R_i]
- The constant X may be given either as an explicit number or as a symbolic name representing a numerical value.
- If X is shorter than a word, sign-extension is needed.

Indexing and Arrays

• In general, the Index mode facilitates access to an operand whose location is defined relative to a reference point within the data structure in which the operand appears.

• Several variations: (R_i, R_j) : EA = [R_i] + [R_j] $X(R_i, R_j): EA = X + [R_i] + [R_j]$

Relative Addressing

- Relative mode the effective address is determined by the Index mode using the program counter in place of the general-purpose register.
- \bullet X(PC) note that X is a signed number
- Branch>0 LOOP
- This location is computed by specifying it as an offset from the current value of PC.
- Branch target may be either before or after the branch instruction, the offset is given as a singed num.
Additional Modes

- Autoincrement mode the effective address of the operand is the contents of a register specified in the instruction. After accessing the operand, the contents of this register are automatically incremented to point to the next item in a list.
- (R_i) +. The increment is 1 for byte-sized operands, 2 for 16-bit operands, and 4 for 32-bit operands.
- Autodecrement mode: -(R_i) decrement first

Figure 2.16. The Autoincrement addressing mode used in the program of Figure 2.12.

Assembly Language

Types of Instructions

• Data Transfer Instructions

Data Transfer Instructions

 \rightarrow

Conditional Branch Instructions

Basic Input/Output Operations

I/O

- The data on which the instructions operate are not necessarily already stored in memory.
- Data need to be transferred between processor and outside world (disk, keyboard, etc.)
- I/O operations are essential, the way they are performed can have a significant effect on the performance of the computer.

- Read in character input from a keyboard and produce character output on a display screen.
- \geq Rate of data transfer (keyboard, display, processor)
- Difference in speed between processor and I/O device creates the need for mechanisms to synchronize the transfer of data.
- \geq A solution: on output, the processor sends the first character and then waits for a signal from the display that the character has been received. It then sends the second character. Input is sent from the keyboard in a similar way.

- Device interface

Figure 2.19 Bus connection for processor, keyboard, and display.

• Machine instructions that can check the state of the status flags and transfer data: READWAIT Branch to READWAIT if $SIN = 0$ Input from DATAIN to R1

WRITEWAIT Branch to WRITEWAIT if SOUT = 0 Output from R1 to DATAOUT

• Memory-Mapped I/O – some memory address values are used to refer to peripheral device buffer registers. No special instructions are needed. Also use device status registers.

READWAIT Testbit #3, INSTATUS Branch=0 READWAIT MoveByte DATAIN, R1

- Assumption the initial state of SIN is 0 and the initial state of SOUT is 1.
- Any drawback of this mechanism in terms of efficiency?
	- \bullet Two wait loops \rightarrow processor execution time is wasted
- Alternate solution?
	- Interrupt

Stack Organization

