Chapter 1
Introduction

Unlike most assembly language books, this one does not emphasize writing programs in assembly language. Higher-level languages, e.g., C, C++, Java, are much better for that. You should avoid writing in assembly language whenever possible.

You may wonder why you should study assembly language at all. The usual reasons given are:

1. *Assembly language is more efficient.* This does not always hold. Modern compilers are excellent at optimizing the machine code that is generated. Only a very good assembly language programmer can do better, and only in some situations. Assembly language programming is very tedious, even for the best programmers. Hence, it is very expensive. The possible gains in efficiency are seldom worth the added expense.
2. *There are situations where it must be used.* This is more difficult to evaluate. How do you know whether assembly language is required or not?

Both these reasons presuppose that you know the assembly language equivalent of the translation that your compiler does. Otherwise, you would have no way of deciding whether you can write a more efficient program in assembly language, and you would not know the machine level limitations of your higher-level language. So this book begins with the fundamental high-level language concepts and "looks under the hood" to see how they are implemented at the assembly language level.

There is a more important reason for reading this book. The interface to the hardware from a programmer's view is the *instruction set architecture* (ISA). This book is a description of the ISA of the x86 architecture as it is used by the C/C++ programming languages. Higher-level languages tend to hide the ISA from the programmer, but good programmers need to understand it. This understanding is bound to make you a better programmer, even if you never write a single assembly language statement after reading this book.

Some of you will enjoy assembly language programming and wish to carry on. If your interests take you into systems programming, e.g., writing parts of an operating system, writing a compiler, or even designing another higher-level language, an understanding of assembly language is required. There are many challenging opportunities in programming embedded systems, and much of the work in this area demands at least an understanding of the ISA. This book serves as an introduction to assembly language programming and prepares you to move on to the intermediate and advanced levels.

In his book *The Design and Evolution of C++*[32] Bjarne Stroustrup nicely lists the purposes of a programming language:

- a tool for instructing machines
- a means of communicating between programmers
- a vehicle for expressing high-level designs
- a notation for algorithms
- a way of expressing relationships between concepts
- a tool for experimentation
- a means of controlling computerized devices.

It is assumed that you have had at least an introduction to programming that covered the first five items on the list. This book focuses on the first item — instructing machines — by studying assembly language programming of a 64-bit x86 architecture computer. We will use C as an example higher-level language and study how it instructs the computer at the assembly language level. Since there is a one-to-one correspondence between

assembly language and machine language, this amounts to a study of how C is used to instruct a machine (computer).

You have already learned that a compiler (or interpreter) translates a program written in a higher-level language into machine language, which the computer can execute. But what does this mean? For example, you might wonder:

- How is an integer stored in memory?
- How is a computer instructed to implement an if-else construct?
- What happens when one function calls another function?
- How does the computer know how to return to the statement following the function call statement?
- How is a computer instructed to display a simple character string — for example, "Hello, world" — on the screen?

It is the goal of this book to answer these and many other questions. The specific higher-level programming language concepts that are addressed in this book include:

General concept	***C/C++ implementation***
Program organization	Functions, variables, literals
Allocation of variables for storage of primitive data types — integers, characters	`int`, `char`
Program flow control constructs — loops, two-way decision	while and for; if-else
Simple arithmetic and logical operations	`+, -, *, /, %, &, |`
Boolean operators	`!, &&, ||`
Data organization constructs — arrays, records, objects	Arrays, structs, classes (C++ only)
Passing data to/from named procedures	Function parameter lists; return values
Object operations	Invoking a member function (C++ only)

This book assumes that you are familiar with these programming concepts in C, C++, and/or Java.

1.1 Computer Subsystems

We begin with a very brief overview of computer hardware. The presentation here is intended to provide you with a rough context of how things fit together. In subsequent chapters we will delve into more details of the hardware and how it is controlled by software.

We can think of computer hardware as consisting of three separate subsystems as shown in Fig. 1.1.

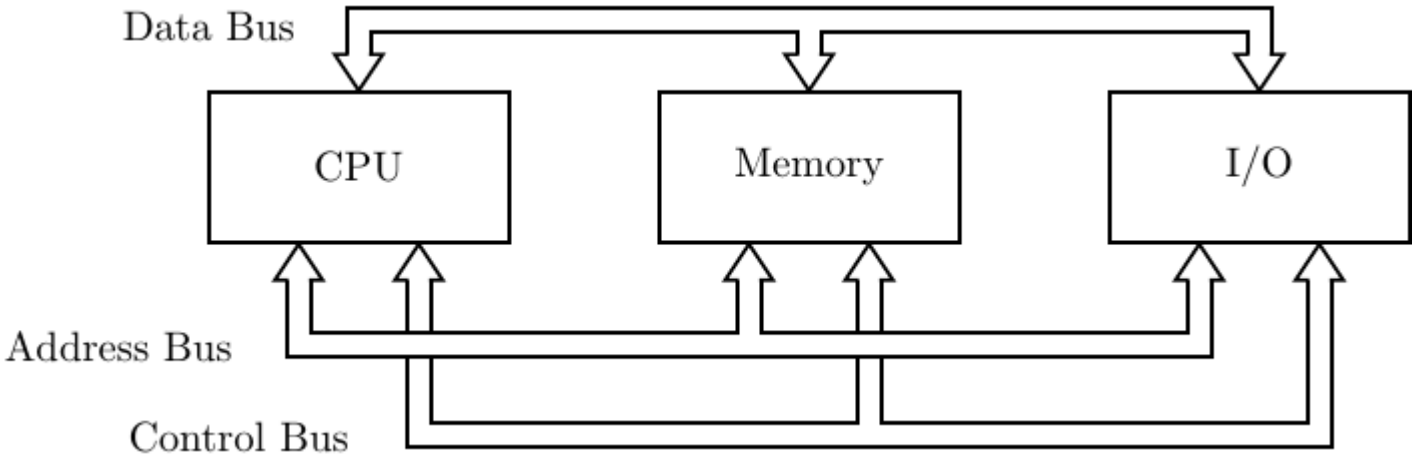

Figure 1.1: Subsystems of a computer. The CPU, Memory, and I/O subsystems communicate with one another via the three buses.

Central Processing Unit (CPU)
: controls most of the activities of the computer, performs the arithmetic and logical operations, and contains a small amount of very fast memory.

Memory
: provides storage for the instructions for the CPU and the data they manipulate.

Input/Output (I/O)
: communicates with the outside world and with mass storage devices (e.g., disks).

When you create a new program, you use an editor program to write your new program in a high-level language, for example, C, C++, or Java. The editor program sees the source code for your new program as data, which is typically stored in a file on the disk. Then you use a compiler program to translate the high-level language statements into machine instructions that are stored in a disk file. Just as with the editor program, the compiler program sees both your source code and the resulting machine code as data.

When it comes time to execute the program, the instructions are read from the machine code disk file into memory. At this point, the program is a sequence of instructions stored in memory. Most programs include some constant data that are also stored in memory. The CPU executes the program by fetching each instruction from memory and executing it. The data are also fetched as needed by the program.

This computer model — both the program instructions and data are stored in a memory unit that is separate from the processing unit — is referred to as the *von Neumann architecture*. It was described in 1945 by John von Neumann [35], although other computer science pioneers of the day were working with the same concepts. This is in contrast to a fixed-program computer, e.g., a calculator. A compiler illustrates one of the benefits of the von Neumann architecture. It is a program that treats the source file as data, which it translates into an executable binary file that is also treated as data. But the executable binary file can also be run as a program.

A downside of the von Neumann architecture is that a program can be written to view itself as data, thus enabling a self-modifying program. GNU/Linux, like most modern, general purpose operating systems, prohibits applications from modifying themselves.

Most programs also access I/O devices, and each access must also be programmed. I/O devices vary widely. Some are meant to interact with humans, for example, a keyboard, a mouse, a screen. Others are meant for machine readable I/O. For example, a program can store a file on a disk or read a file from a network. These devices all have very different behavior, and their timing characteristics differ drastically from one another. Since I/O device programming is difficult, and every program makes use of them, the software to handle I/O devices is included in the operating system. GNU/Linux provides a rich set of functions that an applications programmer can use to perform I/O actions, and we will call upon these services of GNU/Linux to perform our I/O

operations. Before tackling I/O programming, you need to gain a thorough understanding of how the CPU executes programs and interacts with memory.

The goal of this book is study how programs are executed by the computer. We will focus on how the program and data are stored in memory and how the CPU executes instructions. We leave I/O programming to more advanced books.

1.2 How the Subsystems Interact

The subsystems in Figure 1.1 communicate with one another via buses. You can think of a *bus* as a communication pathway with a protocol specifying exactly how the pathway is used. The buses shown here are logical groupings of the signals that must pass between the three subsystems. A given bus implementation may not have physically separate paths for each of the three types of signals. For example, the PCI bus standard uses the same physical pathway for the address and the data, but at different times. Control signals indicate whether there is an address or data on the lines at any given time.

A program consists of a sequence of instructions that is stored in memory. When the CPU is ready to execute the next instruction in the program, the location of that instruction in memory is placed on the *address bus*. The CPU also places a "read" signal on the *control bus*. The memory subsystem responds by placing the instruction on the *data bus*, where the CPU can then read it. If the CPU is instructed to read data from memory, the same sequence of events takes place.

If the CPU is instructed to store data in memory, it places the data on the data bus, places the location in memory where the data is to be stored on the address bus, and places a "write" signal on the control bus. The memory subsystem responds by copying the data on the data bus into the specified memory location.

If an instruction calls for reading or writing data from memory or to memory, the next instruction in the program sequence cannot be read from memory over the same bus until the current instruction has completed the data transfer. This conflict has given rise to another stored-program architecture. In the *Harvard architecture* the program and data are stored in different memories, each with its own bus connected to the CPU. This makes it possible for the CPU to access both program instructions and data simultaneously. The issues should become clearer to you in Chapter 6.

In modern computers the bus connecting the CPU to external memory modules cannot keep up with the execution speed of the CPU. The slowdown of the bus is called the *von Neumann bottleneck*. Almost all modern CPU chips include some cache memory, which is connected to the other CPU components with much faster internal buses. The cache memory closest to the CPU commonly has a Harvard architecture configuration to achieve higher throughput of data processing.

CPU interaction with I/O devices is essentially the same as with memory. If the CPU is instructed to read a piece of data from an input device, the particular device is specified on the address bus and a "read" signal is placed on the control bus. The device responds by placing the data item on the data bus. And the CPU can send data to an output device by placing the data item on the data bus, specifying the device on the address bus, and placing a "write" signal on the control bus. Since the timing of various I/O devices varies drastically from CPU and memory timing, special programming techniques must be used. Chapter 16 provides an introduction to I/O programming techniques.

These few paragraphs are intended to provide you a very general overall view of how computer hardware works. The rest of the book will explore many of these concepts in more depth. Most of the discussion is at the ISA level, but we will also take a peek at the hardware implementation. In Chapter 4 we will even look at some transistor circuits. The goal of the book is to provide you with an introduction to computer architecture as seen from a software point of view.