# Think OS: A Brief Introduction to Operating Systems

Version 0.1

# Think OS

## A Brief Introduction to Operating Systems

Version 0.1

Allen B. Downey

Green Tea Press

Needham, Massachusetts

Green Tea Press
9 Washburn Ave
Needham MA 02492



The original form of this book is LaTeX source code. Compiling this code has the effect of generating a device-independent representation of a textbook, which can be converted to other formats and printed.

The LaTeX source for this book is available from http://thinkstats2.com.

The cover for this book is based on a photo by Paul Friel (http://flickr.com/people/frielp/), who made it available under the Creative Commons Attribution license. The original photo is at http://flickr.com/photos/frielp/11999738/.

# Preface

In many computer science programs, Operating Systems is an advanced topic. By the time students take it, they usually know how to program in C, and they have probably taken a class in Computer Architecture. Usually the goal of the class is to expose students to the design and implementation of operating systems, with the implied assumption that some of them will do research in this area, or write part of an OS.

This book is intended for a different audience, and it has different goals. I developed it for a class at Olin College called Software Systems.

Most students taking this class learned to program in Python, so one of the goals is to help them learn C. For that part of the class, I use Griffiths and Griffiths, *Head First C*, from O'Reilly Media. This book is meant to complement that one.

Few of my students will ever write an operating system, but many of them will write low-level applications in C, and some of them will work on embedded systems. My class includes material from operating systems, networks, databases, and embedded systems, but it emphasizes the topics programmers need to know.

This book does not assume that you have studied Computer Architecture. As we go along, I will explain what we need.

If this book is successful, it should give you a better understanding of what is happening when programs run, and what you can do to make them run better and faster.

Chapter 1 explains some of the differences between compiled and interpreted languages, with some insight into how compilers work. Recommended reading: *Head First C* Chapter 1.

Chapter 2 explains how the operating system uses processes to protect running programs from interfering with each other.

Chapter 3 explains virtual memory and address translation. Recommended reading: *Head First C* Chapter 2.

Chapter 4 is about file systems and data streams. Recommended reading: *Head First C* Chapter 3.

Chapter 5 describes how numbers, letters, and other values are encoded, and presents the bitwise operators.

Chapter 6 explains how to use dynamic memory management, and how it works. Recommended reading: *Head First C* Chapter 6.

Chapter 7 is about caching and the memory hierarchy.

Chapter 8 is about multitasking and scheduling.

Chapter 9 will be about threads.

Chapter 10 will be about synchronization with POSIX mutexes and condition variables.

## A note on this draft

The current version of this book is an early draft. While I am working on the text, I have not yet included the figures. So there are a few places where, I'm sure, the explanation will be greatly improved when the figures are ready.

## Where to get the code

The source material for this book and the supporting code are in a Git repository at .

## Contributor List

If you have a suggestion or correction, please send email to `downey@allendowney.com`. If I make a change based on your feedback, I will add you to the contributor list (unless you ask to be omitted).

If you include at least part of the sentence the error appears in, that makes it easy for me to search. Page and section numbers are fine, too, but not quite as easy to work with. Thanks!

- I am grateful to the student in Software Systems at Olin College, who tested an early draft of this book in Spring 2014. They corrected many errors and made many helpful suggestions. I appreciate their pioneering spirit!

# Contents

# Chapter 1

# Compilation

## 1.1 Compiled and interpreted languages

People often describe programming languages as either compiled, which means that programs are translated into machine language and then executed by hardware, or interpreted, which means that programs are read and executed by a software interpreter. For example, C is considered a compiled language and Python is considered an interpreted language. But the distinction is not always clear-cut.

First, many languages can be either compiled or interpreted. For example, there are C interpreters and Python compilers. Second, there are languages like Java that use a hybrid approach, compiling programs into an intermediate language and then running the translated program in an interpreter. Java uses an intermediate language called Java bytecode, which is similar to machine language, but it is executed by a software interpreter, the Java virtual machine (JVM).

So being compiled or interpreted is not an intrinsic characteristic of a language; nevertheless, there are some general differences between compiled and interpreted languages.

## 1.2 Static types

Compiled languages usually have static types, which means that you can tell by looking at the program what type each variable refers to. Interpreted languages often have dynamic types, which means you don't always know

the type of a variable until the program is running. In general, "static" refers to things that happen at compile time, and "dynamic" refers to things that happen at run time.

For example, in Python you can write a function like this:

```
def foo(x, y):
    return x + y
```

Looking at this code, you can't tell what type `x` and `y` will refer to. At run time, this function could be called several times with different types. Any types that support the addition operator will work; any other types will cause an exception or "run time error."

In C you would write the same function like this:

```
int foo(int x, int y) {
    return x + y;
}
```

The first line of the function includes "type declarations" for the parameters and the return value. `x` and `y` are declared to be integers, which means that we can check at compile time whether the addition operator is legal for this type (it is). The return value is also declared to be an integer.

Because of these declarations, when this function is called elsewhere in the program, the compiler can check whether the arguments provided have the right type, and whether the return value is used correctly.

These checks happen before the program starts executing, so errors can be found more quickly. More importantly, errors can be found in parts of the program that have never run. Furthermore, these checks don't have to happen at run time, which is one of the reasons compiled languages generally run faster than interpreted languages.

Declaring types at compile time also saves space. In dynamic languages, variable names are stored in memory while the program runs, and they are often accessible by the program. For example, in Python the built-in function `locals` returns a dictionary that contains variable names and their values. Here's an example in a Python interpreter:

```
>>> x = 5
>>> print locals()
{'x': 5, '__builtins__': <module '__builtin__' (built-in)>,
'__name__': '__main__', '__doc__': None, '__package__': None}
```

This shows that the name of the variable is stored in memory while the program is running (along with some other values that are part of the default run-time environment).

In compiled languages, variable names exist at compile-time but not at run time. The compiler chooses a location for each variable and records these locations as part of the compiled program.[1] The location of a variable is called its "address". At run time, the value of each variable is stored at its address, but the names of the variables are not stored at all (unless they are added by the compiler for purposes of debugging).

## 1.3 The compilation process

As a programmer, you should have a mental model of what happens during compilation. If you understand the process, it will help you interpret error messages, debug your code, and avoid common pitfalls.

The steps of compilation are:

1. Preprocessing: C is one of several languages that include "preprocessing directives" that take effect before the program is compiled. For example, the `#include` directive causes the source code from another file to be inserted at the location of the directive.

2. Parsing: During parsing, the compiler reads the source code and builds an internal representation of the program, called an "abstract syntax tree." Errors detected during this step are generally syntax errors.

3. Static checking: The compiler checks whether variables and values have the right type, whether functions are called with the right number and type of arguments, etc. Errors detected during this step are sometimes called "static semantic" errors.

4. Code generation: The compiler reads the internal representation of the program and generates machine code or byte code.

5. Linking: If the program uses values and functions defined in a library, the compiler has to find the appropriate library and include the required code.

[1] This is a simplification; we will go into more detail later.

6. Optimization: At several points in the process, the compiler can transform the program to generate code that runs faster or uses less space. Most optimizations are simple changes that eliminate obvious waste, but some compilers perform sophisticated analyses and transformations.

Normally when you run `gcc`, it runs all of these steps and generates an executable file. For example, here is a minimal C program:

```
#include <stdio.h>
int main()
{
    printf("Hello World\n");
    return 0;
}
```

If you save this code in a file called `hello.c`, you can compile and run it like this:

```
$ gcc hello.c
$ ./a.out
```

By default, `gcc` stores the executable code in a file called `a.out` (which originally stood for "assembler output"). The second line runs the executable. The prefix `./` tells the shell to look for it in the current directory.

It is usually a good idea to use the `-o` flag to provide a better name for the executable:

```
$ gcc hello.c -o hello
$ ./hello
```

## 1.4 Object code

The `-c` flag tells `gcc` to compile the program and generate machine code, but not to link it or generate an executable:

```
$ gcc hello.c -c
```

The result is a file named `hello.o`, where the `o` stands for "object code," which is the compiled program. Object code is not executable, but it can be linked into an executable.

The Unix command `nm` reads an object file and generates information about the names it defines and uses. For example:

```
$ nm hello.o
0000000000000000 T main
                 U puts
```

This output indicates that `hello.o` defines the name `main` and uses a function named `puts`, which stands for "put string." In this example, `gcc` performs an optimization by replacing `printf`, which is a large and complicated function, with `puts`, which is relatively simple.

In general you can control how much optimization `gcc` does with the `-O` flag. By default, it does very little optimization, which can help with debugging. The option `-O1` turns on the most common and safe optimizations. Higher numbers turn on additional optimizations that require longer compilation time.

In theory, optimization should not change the behavior of the program, other than to speed it up. But if your program has a subtle bug, you might find that optimization makes the bug appear or disappear. It is usually a good idea to turn off optimization while you are developing new code. Once the program is working and passing appropriate tests, you can turn on optimization and confirm that the tests still pass.

## 1.5 Assembly code

Similar to the `-c` flag, the `-S` flag tells `gcc` to compile the program and generate assembly code, which is basically a human-readable form of machine code.

```
$ gcc hello.c -S
```

The result is a file named `hello.s`, which might look something like this:

```
        .file        "hello.c"
        .section        .rodata
.LC0:
```

```
        .string        "Hello World"
        .text
        .globl        main
        .type        main, @function
main:
.LFB0:
        .cfi_startproc
        pushq        %rbp
        .cfi_def_cfa_offset 16
        .cfi_offset 6, -16
        movq        %rsp, %rbp
        .cfi_def_cfa_register 6
        movl        $.LC0, %edi
        call        puts
        movl        $0, %eax
        popq        %rbp
        .cfi_def_cfa 7, 8
        ret
        .cfi_endproc
.LFE0:
        .size        main, .-main
        .ident        "GCC: (Ubuntu/Linaro 4.7.3-1ubuntu1) 4.7.3"
        .section        .note.GNU-stack,"",@progbits
```

`gcc` is usually configured to generate code for the machine you are running on, so for me it generates x86 assembly language, which runs on a wide variety of processors from Intel, AMD, and others. If you are running on a different architecture, you might see different code.

## 1.6 Preprocessing

Taking another step backward through the compilation process, you can use the `-E` flag to run the preprocessor only:

```
$ gcc hello.c -E
```

The result is the output from the preprocessor. In this example, it contains the included code from `stdio.h`, and all the files included from `stdio.h`, and all the files included from those files, and so on. On my machine, the total is more than 800 lines of code. Since almost every C program includes `stdio.h`, those 800 lines of code get compiled a lot. If, like many C programs, you also include `stdlib.h`, the result is more than 1800 lines of code.

## 1.7 Understanding errors

Now that we know the steps in the compilation process, it is easier to understand errors messages. For example, if there is an error in a `#include` directive, you'll get a message from the preprocessor:

```
hello.c:1:20: fatal error: stdioo.h: No such file or directory
compilation terminated.
```

If there's a syntax error, you get a message from the compiler:

```
hello.c: In function 'main':
hello.c:6:1: error: expected ';' before '}' token
```

If you use a function that's not defined in any of the standard libraries, you get a message from the linker:

```
/tmp/cc7iAUbN.o: In function `main':
hello.c:(.text+0xf): undefined reference to `printff'
collect2: error: ld returned 1 exit status
```

`ld` is the name of the Unix linker, so named because "loading" is another step in the compilation process that is closely related to linking.

Once the program starts, C does very little run-time checking, so there are only a few run-time errors you are likely to see. If you divide by zero, or perform another illegal floating-point operation, you will get a "Floating point exception." And if you try to read or write an incorrect location in memory, you will get a "Segmentation fault."

# Chapter 2

# Processes

## 2.1 Abstraction and virtualization

Before we talk about processes, I want to define a few words:

- Abstraction: An abstraction is a simplified representation of something complicated. For example, if you drive a car, you understand that when you turn the wheel left, the car goes left, and vice versa. Of course, the steering wheel is connected to a sequence of mechanical and (often) hydraulic systems that turn the wheels, and the wheels interact with the road in ways that can be complex, but as a driver, you normally don't have to think about any of those details. You can get along very well with a simple mental model of steering. Your mental model is an abstraction.

  Similarly, when you use a web browser, you understand that when you click on a link, the browser displays the page the link refers to. The software and network communication that make that possible are complex, but as a user, you don't have to know the details.

  A large part of software engineering is designing abstractions, like these, that allow users and other programmers to use powerful and complicated systems without having to know about the details of their implementation.

- Virtualization: An important kind of abstraction is virtualization, which is the process of creating a desirable illusion.

  For example, many public libraries participate in inter-library collaborations that allow them to borrow books from each other. When I

request a book, sometimes the book is on the shelf at my local library, but other times is has to be transferred from another collection. Either way, I get a notification when it is available for pickup. I don't need to know where it came from, and I don't need to know which books my library has. As a whole, the system creates the illusion that my library has every book in the world.

The collection physically located at my local library might be small, but the collection available to me virtually includes every book in the inter-library collaboration.

As another example, most computers are only connected to one network, but that network is connected to others, and so on. What we call the Internet is a collection of networks and a set of protocols that forward packets from one network to the next. From the point of view of a user or programmer, the system behaves as if every computer on the Internet is connected to every other computer. The number of physical connections is small, but the number of virtual connections is very large.

The word "virtual" is often used in the context of a virtual machine, which is software that creates the illusion of a dedicated computer running a particular operating system, when in reality the virtual machine might be running, along with many other virtual machines, on a computer running a different operating system.

In the context of virtualization, we sometimes call what is really happening "physical", and what is virtually happening either "logical" or "abstract."

## 2.2 Isolation

One of the most important principles of engineering is isolation: when you are designing a system with multiple components, it is usually a good idea to isolate them from each other so that a change in one component doesn't have undesired effects on other components.

One of the most important goals of an operating system is to isolate each running program from the others so that programmers don't have to think about every possible interaction. The software object that provides this isolation is a **process**.

A process is a software object that represents a running program. I mean "software object" in the sense of object-oriented programming; in general,

an object contains data and provides methods that operate on the data. A process is an object that contains the following data:

- The text of the program, usually a sequence of machine language instructions,
- Data associated with the program, including static data (allocated at compile time) and dynamic data including the run-time stack and the heap,
- The state of any pending input/output operations. For example, if the process is waiting for data to be read from disk or for a packet to arrive on a network, the status of these operations is part of the process, and
- The hardware state of the program, which includes data stored in registers, status information, and the program counter, which indicates which instruction is currently executing.

Usually one process runs one program, but it is also possible for a process to load and run a new program.

It is also possible, and common, to run the same program in more than one process. In that case, the processes share the same program text, but generally have different data and hardware states.

Most operating systems provide a fundamental set of capabilities to isolate processes from each other:

- Multitasking: Most operating systems have the ability to interrupt a running process at almost any time, save its hardware state, and then resume the process later. In general, programmers don't have to think about these interruptions. The program behaves as if it is running continuously on a dedicated processor, except that the time between instructions is unpredictable.
- Virtual memory: Most operating systems create the illusion that each process has its own chunk of memory, isolated from all other processes. Again, programmers generally don't have to think about how virtual memory works; they can proceed as if every program has a dedicated chunk of memory.
- Device abstraction: Processes running on the same computer share the disk drive, the network interface, the graphics card, and other hardware. If processes interacted with this hardware directly, without coordination, chaos would ensue. For example, network data intended

for one process might be read by another. Or multiple processes might try to store data in the same location on a hard drive. It is up to the operating system to maintain order by providing appropriate abstractions.

As a programmer, you don't need to know much about how these capabilities are implemented. But if you are curious, you will find a lot of interesting things going on under the metaphorical hood. And if you know what's going on, it can make you a better programmer.

## 2.3 Unix processes

While I write this book, the process I am most aware of is my text editor, emacs. Every once in a while I switch to a terminal window, which is a window running a Unix shell that provides a command-line interface.

When I move the mouse, the window manager wakes up, sees that the mouse is over the terminal window, and wakes up the terminal. The terminal wakes up the shell. If I type `make` in the shell, it creates a new process to run Make, which creates another process to run LaTeX and then another process to display the results.

If I need to look something up, I might switch to another desktop, which wakes up the window manager again. If I click on the icon for a web browser, the window manager creates a process to run the web browser. Some browsers, like Chrome, create a new process for each window and each tab.

And those are just the processes I am aware of. At the same time there are many other processes running "in the background." Many of them are performing operations related to the operating system.

The Unix command `ps` prints information about running processes. If you run it in a terminal, you might see something like this:

```
  PID TTY          TIME CMD
 2687 pts/1    00:00:00 bash
 2801 pts/1    00:01:24 emacs
24762 pts/1    00:00:00 ps
```

The first column is the unique numerical process ID. The second column is the terminal that created the process; "TTY" stands for teletypewriter, which was the original mechanical terminal.

The third column is the total processor time used by the process, and the last column is the name of the running program. In this example, `bash` is the name of the shell that interprets the commands I type in the terminal, emacs is my text editor, and ps is the process generating this output.

By default, `ps` lists only the processes associated with the current terminal. If you use the `-e` flag, you get every process (including processes belonging to other users, which is a security flaw, in my opinion).

On my system there are currently 233 processes. Here are some of them:

```
  PID TTY          TIME CMD
    1 ?        00:00:17 init
    2 ?        00:00:00 kthreadd
    3 ?        00:00:02 ksoftirqd/0
    4 ?        00:00:00 kworker/0:0
    8 ?        00:00:00 migration/0
    9 ?        00:00:00 rcu_bh
   10 ?        00:00:16 rcu_sched
   47 ?        00:00:00 cpuset
   48 ?        00:00:00 khelper
   49 ?        00:00:00 kdevtmpfs
   50 ?        00:00:00 netns
   51 ?        00:00:00 bdi-default
   52 ?        00:00:00 kintegrityd
   53 ?        00:00:00 kblockd
   54 ?        00:00:00 ata_sff
   55 ?        00:00:00 khubd
   56 ?        00:00:00 md
   57 ?        00:00:00 devfreq_wq
```

`init` is the first process created when the operating system starts. It creates many of the other processes, and then sits idle until the processes it created are done.

`kthreadd` is a process the operating system uses to create new "threads." We'll talk more about threads soon, but for now you can think of a thread as kind of a process. The `k` at the beginning stands for "kernel" which is the part of the operating system responsible for core capabilities like creating threads. The extra `d` at the end stands for "daemon", which is another name for processes like this that run in the background and provide operating system services. In this context, "daemon" is used in the classical sense of a helpful spirit, with no connotation of evil.

Based on the name, you can infer that `ksoftirqd` is also a kernel daemon; specifically, it handles software interrupt requests, or "soft IRQ".

`kworker` is a worker process created by the kernel to do some kind of processing for the kernel.

There are often multiple processes running these kernel services. On my system at the moment, there are 8 `ksoftirqd` processes and 35 `kworker` processes.

I won't go into more details about the other processes, but you might be interested to search for more information about some of them. Also, you should run `ps` on your system and compare your results to mine.

# Chapter 3

# Virtual memory

## 3.1 A bit of information theory

A bit is a binary digit; it is also a unit of information. If you have one bit, you can specify one of two possibilities, usually written 0 and 1. If you have two bits, there are 4 possible combinations, 00, 01, 10, and 11. In general, if you have $b$ bits, you can indicate one of $2^b$ values. A byte is 8 bits, so it can hold one of 256 values.

Going in the other direction, suppose you want to store a letter of the alphabet. There are 26 letters, so how many bits do you need? With 4 bits, you can specify one of 16 values, so that's not enough. With 5 bits, you can specify up to 32 values, so that's enough for all the letters, with a few values left over.

In general, if you want to specify one of $N$ values, you should choose the smallest value of $b$ so that $2^b \geq N$. Taking the log base 2 of both sides yields $b \geq log_2 N$.

Suppose I flip a coin and tell you the outcome. I have given you one bit of information. If I roll a six-sided die and tell you the outcome, I have given you $log_2 6$ bits of information. And in general, if the probability of the outcome is 1 in $N$, then the outcome contains $log_2 N$ bits of information.

Equivalently, if the probability of the outcome is $p$, then the information content is $-log_2 p$. This quantity is called the "self-information" of the outcome. It measures how surprising the outcome is, which is why it is also called "surprisal." If your horse has only one chance in 16 of winning, and he wins, you get 4 bits of information (along with the payout). But if the favorite wins 75% of the time, the news of the win contains only 0.42 bits.

Intuitively, unexpected news carries a lot of information; conversely, if there is something you were already confident of, confirming it contributes only a small amount of information.

For several topics in this book, we will need to be comfortable converting back and forth between the number of bits, $b$, and the number of values they can encode, $N = 2^b$.

## 3.2 Memory and storage

While a process is running, most of its data is held in "main memory", which is usually some kind of random access memory (RAM). On most current computers, main memory is volatile, which means that when the computer shuts down, the contents of main memory are lost. A current typical desktop computer has 1–8 GiB of memory. GiB stands for "gibibyte," which is $2^{30}$ bytes.

If the process reads and writes files, those files are usually stored on a hard disk drive (HDD) or solid state drive (SSD). These storage devices are non-volatile, so they are used for long-term storage. Currently a typical desktop computer has a HDD with a capacity of 500 GB to 2 TB. GB stands for "gigabyte," which is $10^9$ bytes. TB stands for "terabyte," which is $10^{12}$ bytes.

You might have noticed that I used the binary unit GiB for the size of main memory and the decimal units GB and TB for the size of the HDD. For historical and technical reasons, memory is measured in binary units, and disk drives are measured in decimal units. In this book I will be careful to distinguish binary and decimal units, but you should be aware that the word "gigabyte" and the abbreviation GB are often used ambiguously.

In casual use, the term "memory" is sometimes used for HDDs and SDDs as well as RAM, but the properties of these devices are very different, so we will need to distinguish them. I will use "storage" to refer to HDDs and SDDs.

## 3.3 Address spaces

Each byte in main memory is specified by an integer "physical address." The set of valid physical addresses is called the physical "address space." It usually runs from 0 to $N - 1$, where $N$ is the size of main memory. On a

system with 1 GB of physical memory, the highest valid address is $2^{30} - 1$, which is 1,073,741,823 in decimal, or 0x03ff ffff in hexadecimal (the prefix 0x indicates a hexadecimal number).

However, most operating systems provide "virtual memory," which means that programs never deal with physical addresses, and don't have to know how much physical memory is available.

Instead, programs work with virtual addresses, which are numbered from 0 to $M - 1$, where $M$ is the number of valid virtual address. The size of the virtual address space is determined by the operating system and the hardware it runs on.

You have probably heard people talk about 32-bit and 64-bit systems. These terms indicate the size of the registers, which is usually also the size of a virtual address. On a 32-bit system, virtual addresses are 32 bits, which means that the virtual address space runs from 0 to 0xffff ffff. The size of this address space is $2^{32}$ bytes, or 4 GiB.

On a 64-bit system, the size of the virtual address space is $2^{64}$ bytes, or $4 \cdot 1024^6$ bytes. That's 16 exbibytes, which is about a billion times bigger than current physical memories. It might seem strange that a virtual address space can be so much bigger than physical memory, but we will see soon how that works.

When a program reads and writes values in memory, it generates virtual addresses. The hardware, with help from the operating system, translates to physical addresses before accessing main memory. This translation is done on a per-process basis, so even if two processes generate the same virtual address, they would map to different locations in physical memory.

Thus, virtual memory is one important way the operating system isolates processes from each other. In general, a process cannot access data belonging to another process, because there is no virtual address it can generate that maps to physical memory allocated to another process.

## 3.4 Memory segments

The data of a running process is organized into 4 segments:

- The text segment contains the program text; that is, the machine language instructions that make up the program.

- The static segment contains variables that are allocated by the compiler, including global variables and local variables that are declared `static`.
- The stack segment contains the run-time stack, which is made up of stack frames. Each stack frame contains the parameters and local variables of a function.
- The heap segment contains chunks of memory allocated at run time, usually by calling the C library function `malloc`.

The arrangement of these segments is determined partly by the compiler and partly by the operating system. The details vary from one system to another, but in the most common arrangement:

- The text segment is near the "bottom" of memory; that is at addresses near 0.
- The static segment is often just above the text segment.
- The stack is near the top of memory; that is, near the highest addresses in the virtual address space. As the stack expands, it grows down toward smaller addresses.
- The heap is often above the static segment. As it expands, it grows up toward larger addresses.

To determine the layout of these segments on your system, try running this program (you can download it from `http://todo`):

```
#include <stdio.h>
#include <stdlib.h>

int global;

int main ()
{
    int local = 5;
    void *p = malloc(128);

    printf ("Address of main is %p\n", main);
    printf ("Address of global is %p\n", &global);
    printf ("Address of local is %p\n", &local);
```

```
    printf ("Address of p is %p\n", p);

    return 0;
}
```

`main` is the name of a function; when it is used as a variable, it refers to the address of the first machine language instruction in `main`, which we expect to be in the text segment.

`global` is a global variable, so we expect it to be in the static segment. `local` is a local variable, so we expect it to be on the stack.

And `p` contains an address returned by `malloc`, which allocates space in the heap. "malloc" stands for "memory allocate."

When I run this program, the output looks like this (I added spaces to make it easier to read):

```
Address of main is   0x      40057c
Address of global is 0x      60104c
Address of local is  0x7fffd26139c4
Address of p is      0x     1c3b010
```

As expected, the address of `main` is the lowest, followed by `global` and `p`. The address of `local` is much bigger. It has 12 hexadecimal digits. Each hex digit corresponds to 4 bits, so it is a 48-bit address. That suggests that the usable part of the virtual address space is $2^{48}$ bytes.

**Exercise 3.1** Run this program on your computer and compare your results to mine.

Add a second call to `malloc` and check whether the heap on your system grows up (toward larger addresses). Add a function that prints the address of a local variable, and check whether the stack grows down.

## 3.5 Address translation

How does a virtual address (VA) get translated to a physical address (PA)? The basic mechanism is simple, but a simple implementation would be too slow and take too much space. So actual implementations are a bit more complicated.

Most processors provide a memory management unit (MMU) that sits between the CPU and main memory. The MMU performs fast translation between VAs and PAs.

1. When a program reads or writes a variable, the CPU generates a VA.
2. The MMU splits the VA into two parts, called the page number and the offset. A "page" is a chunk of memory; the size of a page depends on the operating system and the hardware, but common sizes are 1–4 KiB.
3. The MMU looks up the page number in the "page table" and gets the corresponding physical page number. Then it combines the physical page number with the offset to produce a PA.
4. The PA is passed to main memory, which reads or writes the given location.

As an example, suppose that the VA is 32 bits and the physical memory is 1 GiB, divided into 1 KiB pages.

- Since 1 GiB is $2^{30}$ bytes and 1 KiB is $2^{10}$ bytes, there are $2^{20}$ physical pages, sometimes called "frames."
- The size of the virtual address space is $2^{32}$ B and the size of a page is $2^{10}$ B, so there are $2^{22}$ virtual pages.
- The size of the offset is determined by the page size. In this example the page size is $2^{10}$ B, so it takes 10 bits to specify a byte on a page.
- If a VA is 32 bits and the offset is 10 bits, the remaining 22 bits make up the virtual page number.
- Since there are $2^{20}$ physical pages, each physical page number is 20 bits. Adding in the 10 bit offset, the resulting PAs are 30 bits.

So far this all seems feasible. But let's think about how big a page table might have to be. The simplest implementation of a page table is an array with one entry for each virtual page. Each entry would contain a physical page number, which is 20 bits in this example, plus some additional information about each frame. So we expect 3–4 bytes per entry. But with $2^{22}$ virtual pages, the page table would require $2^{24}$ bytes, or 16 MiB.

And since we need a page table for each process, a system running 256 processes would need $2^{32}$ bytes, or 4 GiB, just for page tables! And that's just with 32-bit virtual addresses. With 48- or 64-bit VAs, the numbers are ridiculous.

Fortunately, nothing like that much space is actually needed, because most processes don't use even a small fraction of their virtual address space. And if a process doesn't use a virtual page, we don't need an entry in the page table for it.

Another way to say the same thing is that page tables are "sparse," which implies that the simple implementation, an array of page table entries, is a bad idea. Fortunately, there are several good implementations for sparse arrays.

One option is a multilevel page table, which is what many operating systems, including Linux, use. Another option is an associative table, where each entry includes both the virtual page number and the physical page number. Searching an associative table can be slow in software, but in hardware we can search the entire table in parallel, so associative arrays are often used to represent pages tables in the MMU.

You can read more about these implementations at `http://en.wikipedia.org/wiki/Page_table`; you might find the details interesting. But the fundamental idea is that page tables are sparse, so we have to choose a good implementation for sparse arrays.

I mentioned earlier that the operating system can interrupt a running process, save its state, and then run another process. This mechanism is called a "context switch." Since each process has its own page table, the operating system has to work with the MMU to make sure that each process gets the right page table. In older machines, the page table information in the MMU had to be replaced during every context switch, which was expensive. In newer systems, each page table entry in the MMU includes the process ID, so page tables from multiple processes can be in the MMU at the same time.

# Chapter 4

# Files and file systems

When a process completes (or crashes), any data stored in main memory is lost. But data stored on a hard disk drive (HDD) or solid state drive (SSD) is "persistent;" that is, it survives after the process completes, even if the computer shuts down.

Hard disk drives are complicated. Data is stored in blocks, which are laid out in sectors, which make up tracks, which are arranged in concentric circles on platters.

Solid state drives are simpler in one sense, because blocks are numbered sequentially, but they raise a different complication: each block can be written a limited number of times before it becomes unreliable.

As a programmer, you don't want to deal with these complications. What you want is an appropriate abstraction of persistent storage hardware. The most common abstraction is called a "file system."

Abstractly:

- A "file system" is a mapping from a file name to a file. If you think of file names as keys, and the file contents as values, you can think of a file system as a simple key-value store, which is one category of NoSQL database. (see `http://en.wikipedia.org/wiki/NoSQL`).
- A "file" is a sequence of bytes.

File names are usually strings, and they are usually "hierarchical;" that is, the string specifies a path from a top-level directory (or folder), through a series of subdirectories, to a specific file.

The primary difference between the abstraction and the underlying mechanism is that files are byte-based and persistent storage is block-based. The operating system translates byte-based file operations in the C library into block-based operations on storage devices. Typical block sizes are 1–8 KiB.

For example, the following code opens a file and reads the first byte:

```
FILE *fp = fopen("/home/downey/file.txt", "r");
char c = fgetc(fp);
fclose(fp);
```

When this code runs:

1. `fopen` uses the filename to find the top-level directory, called `/`, the subdirectory `home`, and the sub-subdirectory `downey`.

2. It finds the file named `file.txt` and "opens" it for reading, which means it creates a data structure that represents the file being read. Among other things, this data structure keeps track of how much of the file has been read, called the "file position."

   In DOS, this data structure is called a File Control Block, but I want to avoid that term because in UNIX it means something else. In UNIX, there seems to be no good name for it. It is an entry in the open file table, but "open file table entry" is hard to parse, so I will call it an OpenFileTableEntry.

3. When we call `fgetc`, the operating system checks whether the next character of the file is already in memory. If so, it reads the next character, advances the file position, and returns the result.

4. If the next character is not in memory, the operating system issues an I/O request to get the next block. Disk drives are slow, so a process waiting for a block from disk is usually interrupted so another process can run until the data arrives.

5. When the I/O operation is complete, the new block of data is stored in memory, and the process resumes.

6. When the process closes the file, the operating system completes any pending operations, removes any data stored in memory, and frees the OpenFileTableEntry.

The process for writing a file is similar, but there are some additional steps. Here is an example that opens a file for writing and changes the first character.

```
FILE *fp = fopen("/home/downey/file.txt", "w");
fputc('b', fp);
fclose(fp);
```

When this code runs:

1. Again, `fopen` uses the filename to find the file. If it does not already exist, it creates a new file and adds an entry in the parent directory, `/home/downey`.

2. The operating system creates an OpenFileTableEntry that indicates that the file is open for writing, and sets the file position to 0.

3. `fputc` attempts to write (or re-write) the first byte of the file. If the file already exists, the operating system has to load the first block into memory. Otherwise it allocates a new block in memory and requests a new block on disk.

4. After the block in memory is modified, it might not be copied back to the disk right away. In general, data written to a file is "buffered," which means it is stored in memory and only written to disk when there is at least one block to write.

5. When the file is closed, any buffered data is written to disk and the OpenFileTableEntry is freed.

To summarize, the C library provides the abstraction of a file system that maps from file names to streams of bytes. This abstraction is built on top of storage devices that are actually organized in blocks.

## 4.1 Disk performance

I mentioned earlier that disk drives are slow. On current HDDs, the time to read a block from disk to memory is typically 2–6 ms. SSDs are faster, taking 25 $\mu$s to read a 4 KiB block and 250 $\mu$s to write one (see `http://en.wikipedia.org/wiki/Ssd#Controller`).

To put these numbers in perspective, let's compare them to the clock cycle of the CPU. A processor with clock rate 2 GHz completes one clock cycle every 0.5 ns. The time to get a byte from memory to the CPU is typically around 100 ns. If the processor completes one instruction per clock cycle, it would complete 200 instructions while waiting for a byte from memory.

In one microsecond, it would complete 2000 instructions, so while waiting for a byte from an SSD, it would complete 50,000.

In one millisecond, it would complete 2,000,000 instructions, so while waiting for a byte from a HDD, it might complete 10 million. If there's nothing for the CPU to do while it waits, it would be idle. That's why the operating system generally switches to another process while one is waiting for disk.

The gap in performance between main memory and persistent storage is one of the major challenges of computer system design. Operating systems and hardware provide several features intended to "fill in" this gap:

- Block transfers: The time it takes to load a single byte from disk is about 5 ms. By comparison, the additional time to load an 8 KiB block is negligible. If the processor does 5 ms of work on each block, it might be possible to keep the processor busy.
- Prefetching: Sometimes the operating system can predict that a process will read a block and start loading it before it is requested. For example, if you open a file and read the first block, there is a good chance you will go on to read the second block. The operating system might start loading additional blocks before they are requested.
- Buffering: As I mentioned, when you write a file, the operating system stores the data in memory and only writes it to disk later. If you modify the block several times while it is in memory, the system only has to write it to disk once.
- Caching: If a process has used a block recently, it is likely to use it again soon. If the operating system keeps a copy of the block in memory, it can handle future requests at memory speed.

Some of these features are also implemented in hardware. For example, some disk drives provide a cache that stores recently-used blocks, and many disk drives read more than one block at a time, even if only one is requested.

These mechanisms generally improve the performance of programs, but they don't change the behavior. Usually programmers don't have to think about them, with two exceptions: (1) if the performance of a program is unexpectedly bad, you might have to know something about these mechanisms to diagnose the problem, and (2) when data is buffered, it can be harder to debug a program. For example, if a program prints a value and then crashes, the value might not appear, because it might be in a buffer. Similarly, if a program writes data to disk and then the computer loses power, the data might be lost if it is in a cache and not yet on disk.

## 4.2 Disk metadata

The blocks that make up a file might be arranged contiguously on disk, and file system performance is generally better if they are, but most operating systems don't require contiguous allocation. They are free to place a block anywhere on disk, and they use various data structures to keep track of them.

In many UNIX file systems, that data structure is called an "inode," which stands for "index node." More generally, information about files, including the location of their blocks, is called "metadata." (The content of the file is data, so information about the file is data about data, hence "meta").

Since inodes reside on disk along with the rest of the data, they are designed to fit neatly into disk blocks. An UNIX inode contains IDs for the owner of the file and the group associated with it, permission flags indicating who is allowed to read, write, or execute it, and timestamps that indicate when it was last modified and accessed. In addition, it contains block numbers for the first 12 blocks that make up the file.

If the block size is 8 KiB, the first 12 blocks make up 96 KiB. On most systems, that's big enough for a large majority of files, but it's obviously not big enough for all of them. That's why the inode also contains a pointer to an "indirection block," which contains nothing but pointers to other blocks.

The number of pointers in an indirection block depends on the sizes of the blocks and the block numbers, but it is often 1024. With 1024 block numbers and 8 KiB blocks, an indirection block can address 8 MiB. That's big enough for all but the largest files, but still not big enough for all.

That's why the inode also contains a pointer to a "double indirection block," which contains pointers to indirection blocks. With 1024 indirection blocks, we can address 8 GiB.

And if that's not big enough, there is (finally) a triple indirection block, which contains pointers to double indirection blocks, yielding a maximum file size of 8 TiB. When UNIX inodes were designed, that seemed big enough to serve for a long time. But that was a long time ago.

As an alternative to indirection blocks, some files systems, like FAT, use a File Allocation Table that contains one entry for each block (called a "cluster" in this context). A root directory contains a pointer to the first cluster in each file. The FAT entry for each cluster points to the next cluster in the file, similar to a linked list. See `http://en.wikipedia.org/wiki/File_Allocation_Table`.

## 4.3 Block allocation

File systems have to keep track of which blocks belong to each file; they also have to keep track of which blocks are available for use. When a new file is created, the file system finds an available block and allocates it. When a file is deleted, the file system makes its blocks available for re-allocation.

The goals of the block allocation system are:

- Speed: Allocating and freeing blocks should be fast.
- Minimal space use: The data structures used by the allocator should be small, leaving as much space as possible for data.
- Minimal fragmentation: If some blocks are left unused, or some are only partially used, the unused space is called "fragmentation."
- Maximum contiguity: Data that is likely to be used at the same time should be physically contiguous, if possible, to improve performance.

It is hard to design a file system that achieves all of these goals, especially since file system performance depends on "workload characteristics," that include file sizes, access patterns, etc. A file system that is well tuned for one workload might not perform as well for a different workload.

For this reason, most operating systems support several kinds of file systems, and file system design is an active area of research and development. In the last decade, Linux systems have migrated from ext2, which was a conventional UNIX file system, to ext3, a "journaling" file system intended to improve speed and contiguity, and more recently to ext4, which can handle larger files and file systems. Within the next few years, there might be another migration to the B-tree file system, Btrfs.

## 4.4 Everything is a file?

The file abstraction is really a "stream of bytes" abstraction, which turns out to be useful for many things, not just file systems.

One example is the UNIX pipe, which is a simple form of inter-process communication. Processes can be set up so that output from one process is taken as input into another process. For the first process, the pipe behaves like a file open for writing, so it can use C library functions like `fputs`

and `fprintf`. For the second process, the pipe behaves like a file open for reading, so it uses `fgets` and `fscanf`.

Network communication also uses the stream of bytes abstraction. A UNIX socket is a data structure that represents a communication channel between processes on different computers (usually). Again, processes can read data from and write data to a socket using "file" handling functions.

Reusing the file abstraction makes life easier for programmers, since they only have to learn one API (application program interface). It also makes programs more versatile, since a program intended to work with files can also work with data coming from pipes and other sources.

# Chapter 5

# More bits and bytes

## 5.1 Representing integers

You probably know that computers represent numbers in base 2, also known as binary. For positive numbers, the binary representation is straightforward; for example, the representation for $5_{10}$ is $b101$.

For negative numbers, the most obvious representation uses a sign bit to indicate whether a number is positive or negative. But there is another representation, called "two's complement" that is much more common because it is easier to work with in hardware.

To find the two's complement of a negative number, $-x$, find the binary representation of $x$, flip all the bits, and add 1. For example, to represent $-5_{10}$, start with the representation of $5_{10}$, which is $b00000101$ if we write the 8-bit version. Flipping all the bits and adding 1 yields $b11111011$.

In two's complement, the leftmost bit acts like a sign bit; it is 0 for positive numbers and 1 for negative numbers.

To convert from an 8-bit number to 16-bits, we have to add more 0's for a positive number and add 1's for a negative number. In effect, we have to copy the sign bit into the new bits. This process is called "sign extension."

In C all integer types are signed (able to represent positive and negative numbers) unless you declare them `unsigned`. Operations on unsigned integers don't use sign extension.

## 5.2 Bitwise operators

People learning C are sometimes confused about the bitwise operators `&` and `|`. These operators treat integers as bit vectors and compute logical operations on corresponding bits.

For example, `&` computes the AND operation, which yields 1 if both operands are 1, and 0 otherwise. Here is an example of `&` applied to two 4-bit numbers:

```
  1100
& 1010
  ----
  1000
```

In C, this means that the expression `12 & 10` has the value 8.

Similarly, `|` computes the OR operation, which yields 1 if either operand is 1, and 0 otherwise.

```
  1100
| 1010
  ----
  1110
```

So the expression `12 | 10` has the value 14.

Finally, `^` computes the XOR operation, which yields 1 if either operand is 1, but not both.

```
  1100
^ 1010
  ----
  0110
```

So the expression `12 ^ 10` has the value 6.

Most commonly, `&` is used to clear a set of bits from a bit vector, `|` is used to set bits, and `^` is used to flip, or "toggle" bits. Here are the details:

**Clearing bits**: For any value $x$, $x$&0 is 0, and $x$&1 is $x$. So if you AND a vector with 3, it selects only the two rightmost bits, and sets the rest to 0.

```
  xxxx
& 0011
  ----
  00xx
```

In this context, the value 3 is called a "mask" because it selects some bits and masks the rest.

**Setting bits**: Similarly, for any $x$, $x|0$ is x, and $x|1$ is 1. So if you OR a vector with 3, it sets the rightmost bits, and leaves the rest alone:

```
  xxxx
| 0011
  ----
  xx11
```

**Toggling bits**: Finally, if you XOR a vector with 3, it flips the rightmost bits and leaves the rest alone. As an exercise, see if you can compute the two's complement of 12 using `~`. Hint: what's the two's complement representation of -1?

C also provides shift operators, `<<` and `>>`, which shift bits left and right. Each left shift doubles a number, so `5 << 1` is 10, and `5 << 2` is 20. Each right shift divides by two (rounding down), so `5 >> 1` is 2 and `5 >> 1` is 1.

## 5.3 Representing floating-point numbers

Floating-point numbers are represented using the binary version of scientific notation. In decimal notation, large numbers are written as the product of a coefficient and 10 raised to an exponent. For example, the speed of light in m/s is approximately $2.998 \cdot 10^8$.

Most computers use the IEEE standard for floating-point arithmetic. The C type `float` usually corresponds to the 32-bit IEEE standard; `double` usually corresponds to the 64-bit standard.

In the 32-bit standard, the leftmost bit is the sign bit, $s$. The next 8 bits are the exponent, $q$, and the last 23 bits are the coefficient, $c$. The value of a floating-point number is

$$(-1)^s c \cdot 2^q$$

Well, that's almost correct, but there is one more wrinkle. Floating-point numbers are usually normalized so that there is one digit before the point. For example, in base 10, we prefer $2.998 \cdot 10^8$ rather than $2998 \cdot 10^5$ or any other equivalent expression. In base 2, a normalized number always has the digit 1 before the binary point. Since the digit in this location is always 1, we can save space by leaving it out of the representation.

For example, the integer representation of $13_{10}$ is $b1101$. In floating point, that's $1.101 \cdot 2^3$, so the exponent is 3 and the part of the coefficient that would be stored is 101 (followed by 20 zeros).

Well, that's almost correct. But there's one more wrinkle. The exponent is stored with a "bias". In the 32-bit standard, the bias is 127, so the exponent 3 would be stored as 130.

To pack and unpack floating-point numbers in C, we can use a union and bitwise operations. Here's an example:

```
union {
    float f;
    unsigned int u;
} p;

p.f = -13.0;
unsigned int sign = (p.u >> 31) & 1;
unsigned int exp = (p.u >> 23) & 0xff;

unsigned int coef_mask = (1 << 23) - 1;
unsigned int coef = p.u & coef_mask;

printf("%d\n", sign);
printf("%d\n", exp);
printf("0x%x\n", coef);
```

The union allows us to store a floating-point value using `p.f` and then read it as an unsigned integer using `p.u`.

To get the sign bit, we shift the bits to the right 31 places and then use a 1-bit mask to select only the rightmost bit.

To get the exponent, we shift the bits 23 places, then select the rightmost 8 bits (the hexadecimal value `0xff` has eight 1's).

To get the coefficient, we need to extract the 23 rightmost bits and ignore the rest. We do that by making a mask with 1s in the 23 rightmost places and 0s on the left. The easiest way to do that is by shifting 1 to the left by 23 places and then subtracting 1.

The output of this program is:

```
1
130
0x500000
```

As expected, the sign bit for a negative number is 1. The exponent is 130, including the bias. And the coefficient, which I printed in hexadecimal, is 101 followed by 20 zeros.

As an exercise, try assembling and disassembling a `double`, which uses the 64-bit standard. See `http://en.wikipedia.org/wiki/IEEE_floating_point`.

## 5.4 Unions and memory errors

There are two common uses of C unions. One, which we saw in the previous section, is to access the binary representation of data. Another is to store heterogeneous data. For example, you could use a union to represent a number that might be an integer, float, complex, or rational number.

However, unions are error-prone. It is up to you, as the programmer, to keep track of what type of data is in the union; if you write a floating-point value and then interpret it as an integer, the result is usually nonsense.

Actually, the same thing can happen if you read a location in memory incorrectly. One way that can happen is if you read past the end of an array.

To see what happens, I'll start with a function that allocates an array on the stack and fills it with the numbers from 0 to 99.

```
void f1() {
    int i;
    int array[100];

    for (i=0; i<100; i++) {
        array[i] = i;
    }
}
```

Next I'll define a function that creates a smaller array and deliberately accesses elements before the beginning and after the end:

```
void f2() {
    int x = 17;
    int array[10];
    int y = 123;

    printf("%d\n", array[-2]);
    printf("%d\n", array[-1]);
    printf("%d\n", array[10]);
    printf("%d\n", array[11]);
}
```

If I call `f1` and then `f2`, I get these results:

```
17
123
98
99
```

The details here depend on the compiler, which arranges variables on the stack. From these results, we can infer that the compile put `x` and `y` next to each other, "below" the array (at a lower address). And when we read past the array, it looks like we are getting values that were left on the stack by the previous function call.

In this example, all of the variables are integers, so it is relatively easy to figure out what is going on. But in general when you read beyond the bounds of an array, the values you read might have any type. For example, if I change `f1` to make an array of floats, the results are:

```
17
123
1120141312
1120272384
```

The latter two values are what you get if you interpret a floating-point value as an integer. If you encountered this output while debugging, you would have a hard time figuring out what's going on.

## 5.5 Representing strings

Related issues sometimes come up with strings. First, remember that C strings are null-terminated. When you allocate space for a string, don't forget the extra byte at the end.

Also, remember that the letters *and numbers* in C strings are encoded in ASCII. The ASCII codes for the digits "0" through "9" are 48 through 57, *not* 0 through 9. The ASCII code 0 is the NUL character that marks the end of a string. And the ASCII codes 1 through 9 are special characters used in some communication protocols. ASCII code 7 is a bell; on some terminals, printing it makes a sound.

The ASCII code for the letter "A" is 65; the code for "a" is 97. here are those codes in binary:

```
65 = b0100 0001
97 = b0110 0001
```

A careful observer will notice that they differ by a single bit. And this pattern holds for the rest of the letters; the sixth bit (counting from the right) acts as a "case bit," 0 for upper-case letters and 1 for lower case letters.

As an exercise, write a function that takes a string and converts from lower-case to upper-case by flipping the sixth bit. As a challenge, you can make a faster version by reading the string 32 or 64 bits at a time, rather than one character at a time. This optimization is made easier if the length of the string is a multiple of 4 or 8 bytes.

If you read past the end of a string, you are likely to see strange characters. Conversely, if you write a string and then accidentally read it as an int or float, the results will be hard to interpret.

For example, if you run:

```
char array[] = "allen";
float *p = array;
printf("%f\n", *p);
```

You will find that the ASCII representation of the first 8 characters of my name, interpreted as a double-precision floating point number, is 6977971387880058545766 4.

# Chapter 6

# Memory management

C provides 4 functions for dynamic memory allocation:

- `malloc`, which takes an integer size, in bytes, and returns a pointer to a newly-allocated chunk of memory with (at least) the given size. If it can't satisfy the request, it returns the special pointer value NULL.
- `calloc`, which is the same as `malloc` except that it also clears the newly allocated chunk; that is, sets all bytes in the chunk to 0.
- `realloc`, which takes a pointer to a previously allocated chunk and a new size. It allocates a chunk of memory with the new size, copies data from the old chunk to the new, and returns a pointer to the new chunk.
- `free`, which takes a pointer to a previously allocated chunk and deallocates it; that is, makes the space available for future allocation.

This API is notoriously error-prone and unforgiving. Memory management is one of the most challenging parts of designing large software systems, which is why most modern language provide higher-level memory management features like garbage collection.

## 6.1 Memory errors

The C memory management API is a bit like Jasper Beardly, a minor character on the animated television program *The Simpson*, who appears as a strict

substitute teacher who imposes corporal punishment, a "paddlin,"' for all infractions.

Here are some of things a program can do that deserve a paddling:

- If you access any chunk that has not been allocated, that's a paddling.
- If you free an allocated chunk and then access it, that's a paddling.
- If you try to free a chunk that has not been allocated, that's a paddling.
- If you free the same chunk more than once, that's a paddling.
- If you call `realloc` with a chunk that was not allocated, or was allocated and then freed, that's a paddling.

It might not sound difficult to follow these rules, but in a large program a chunk of memory might be allocated in one part of the program, used in several other parts, and freed in yet another part. So changes in one part of the program can require changes in many other parts.

Also, there might be many aliases, or references to the same allocated chunk, in different parts of the program. The chunk should not be freed until all references to the chunk are no longer in use. Getting this right often requires careful analysis across all parts of the program, which is difficult and contrary to fundamental principles of good software engineering.

Ideally, every function that allocates memory should include, as part of the documented interface, information about how that memory should be freed. Mature libraries often do this well, but in the real world, software engineering practice often falls short of this ideal.

To make matters worse, memory errors can be extremely difficult to find because the symptoms are unpredictable. For example:

- If you read a value from an unallocated chunk, the system *might* detect the error, trigger a Segmentation Fault, and stop the program. This outcome is desirable, because it indicates the location in the program that caused the error. But, sadly, this outcome is *rare*. More often, the program reads unallocated memory without detecting the error, and the value is whatever happened to be stored at a particular location. If the value is not interpreted as the right type, it will often cause program behavior that is unexpected and hard to interpret. For example, if you read bytes from a string and interpret them as a floating-point

value, the result might be an invalid number, or a number that is extremely large or small. If you pass that value to a function that isn't prepared to handle it, the results can be bizarre.

- If you write a value to an unallocated chunk, things are even worse, because after the bad value is written, a long time might pass before it is read and causes problems. At that point it will be very difficult to find the source of the problem.

  And things can be even worse than that! One of the most common problems with C-style memory management is that the data structures used to implement `malloc` and `free` (which we will see soon) are often stored along with the allocated chunks. So if you accidentally write past the end of a dynamically-allocated chunk, you are likely to mangle these data structures. The system usually won't detect the problem until much later, when you call `malloc` or `free`, and those functions fail in some inscrutable way.

One conclusion you should draw from this is that safe memory management requires design and discipline. If you write a library or module that allocates memory, you should also provide an interface to free it, and memory management should be part of the API design from the beginning.

If you use a library that allocates memory, you should be disciplined in your use of the API. For example, if the library provides functions to allocate and deallocate storage, you should use those functions and not, for example, call `free` on a chunk you did not `malloc`. And you should avoid keeping multiple references to the same chunk in different parts of your program.

Often there is a tradeoff between safe memory management and performance. For example, the most common source of memory errors is writing beyond the bounds of an array. The obvious remedy for this problem is bounds checking; that is, every access to the array should check whether the index is out of bounds. High-level libraries that provide array-like structures usually perform bounds checking. But C arrays and most low-level libraries do not.

## 6.2 Memory leaks

There is one more memory error that may or may not deserve a paddling. If you allocate a chunk of memory and never free it, that's a "memory leak."

For some programs, memory leaks are ok. For example, if your program allocates memory, performs computations on it, and then exits, it is probably not necessary to free the allocated memory. When the program exits, all of its memory is deallocated by the operating system. Freeing memory immediately before exiting might feel more responsible, but it is mostly a waste of time.

But if a program runs for a long time and leaks memory, its total memory use will increase indefinitely. Eventually the program will run out of memory and, probably, crash. But even before that, a memory hog might slow down other processes (for reasons we'll see soon) or cause them to run out of memory and fail.

Many large complex programs, like web browsers, leak memory, causing their performance to degrade over time. Users who have observed this pattern are often in the habit of restarting these programs periodically.

To see which programs on your system are using the most memory, you can use the UNIX utilities `ps` and `top`.

## 6.3 Implementation

When a process starts, the system allocates space for the text segment and statically allocated data, space for the stack, and space for the heap, which contains dynamically allocated data.

Not all programs allocate data dynamically, so the initial size of the heap might be small or zero. Initially the heap contains only one free chunk.

When `malloc` is called, it checks whether it can find a free chunk that's big enough. If not, it has to request more memory from the system. The function that does that is `sbrk`, which sets the "program break," which you can think of as a pointer to the end of the heap.

When `sbrk` is called, it allocates new pages of physical memory, updates the process's page table, and updates the program break.

In theory, a program could call `sbrk` directly (without using `malloc`) and manage the heap itself. But `malloc` is easier to use and, for most memory-use patterns, it runs fast and uses memory efficiently.

Most Linux systems use an implementation of `ptmalloc`, which is based on `dlmalloc`, written by Doug Lea. A short paper that describes key elements

of the implementation is available at `http://gee.cs.oswego.edu/dl/html/malloc.html`.

For programmers, the most important elements to be aware of are:

- The run time of `malloc` does not usually depend on the size of the chunk, but does depend on how many free chunks there are. `free` is usually fast, regardless of the number of free chunks. Because `calloc` clears every byte in the chunk, the run time depends on chunk size (as well as the number of free chunks).

  `realloc` is sometimes fast, if the new size is smaller or if space is available to expand the chunk. If not, it has to copy data from the old chunk to the new; in that case, the run time depends on the size of the old chunk.

- Boundary tags: When `malloc` allocates a chunk, it adds space at the beginning and end to store information about the chunk, including its size and the state (allocated or free). These bits of data are called "boundary tags." Using these tags, `malloc` can get from any chunk to the previous chunk and the next chunk in memory. In addition, free chunks are chained into a doubly-linked list, so each chunk contains pointers to the next and previous chunks in the "free list."

  The boundary tags and free list pointers make up `malloc`'s internal data structures. These data structures are interspersed with program data, so it is easy for a program error to damage them.

- Space overhead: Boundary tags and free list pointers take up space. The minimum chunk size on most systems is 16 bytes. So for very small chunks, `malloc` is not space efficient. If your program requires large numbers of small structures, it might be more efficient to allocate them in arrays.

- Fragmentation: If you allocate and free chunks with varied sizes, the heap will tend to become fragmented. That is, the free space might be broken into many small pieces. Fragmentation wastes space; it also slows the program down by making memory caches less effective.

- Binning and caching: The free list is sorted by size into bins, so when `malloc` searches for a chunk with a particular size, it knows what bin to search in. Also, if you free a chunk and then immediately allocate a chunk with the same size, `malloc` will usually be very fast.

# Chapter 7

# Caching

In order to understand caching, you have to understand how computers execute programs. For a deep understanding of this topic, you should study computer architecture. My goal in this chapter is to provide a simple model of program execution.

When a program starts, the code (or text) is usually on a hard disk or solid state drive. The operating system creates a new process to run the program, then the **loader** copies the text from disk into main memory and starts the program by calling `main`.

While the program is running, most of its data is stored in main memory, but some of the data is in **registers**, which are small units of memory on the CPU. These registers include:

- The program counter, or PC, which contains the address (in memory) of the next instruction in the program.
- The instruction register, or IR, which contains the instruction currently executing.
- The stack pointer, or SP, which contains the address of the stack frame for the current function, which contains its parameters and local variables.
- General-purpose registers that hold the data the program is currently working with.
- A status register, or flag register, that contains information about the current computation. For example, the flag register usually contains a bit that is set if the result of the previous operation was zero.

When a program is running, the CPU executes the following steps, called the **instruction cycle**:

- Fetch: The next instruction is fetched from memory and stored in the instruction register.
- Decode: Part of the CPU, called the **control unit**, decodes the instruction and send signals to the other parts of the CPU.
- Execute: Signals from the control unit cause the appropriate computation to occur.

Most computers can execute a few hundred different instructions, called the **instruction set**. But most instructions fall into a few general categories:

- Load: Transfers a value from memory to a register.
- Arithmetic/logic: Loads operands from registers, performs a mathematical operation, and stores the result in a register.
- Store: Transfers a value from a register to memory.
- Jump/branch: Changes the program counter, causing the flow of execution to jump to another location in the program. Branches are usually conditional, which means that they check a flag in the flag register and jump only if it is set.

Some instructions sets, including the ubiquitous x86, provide instructions that combine a load and an arithmetic operation.

During each instruction cycle, one instruction is read from the program text. In addition, about half of the instructions in a typical program load or store data. And therein lies one of the fundamental problems of computer architecture: the **memory bottleneck**.

In current desktop computers, a typical CPU runs at 2 GHz, which means that it initiates a new instruction every 0.5 ns. But the time it takes to transfer data to and from memory is about 10 ns. If the CPU has to wait 10 ns to fetch the next instruction, and another 10 ns to load data, it would take 40 clock cycles to complete one instruction.

## 7.1 Cache performance

The solution to this problem, or at least a partial solution, is caching. A **cache** is a small, fast memory on the same chip as the CPU. In current computers, a cache might be 1–2 MiB, and the access time might be 1–2 ns.

When the CPU loads a value from memory, it stores a copy in the cache. If the same value is loaded again, the CPU gets the cached copy and doesn't have to wait for memory.

Eventually the cache gets full. Then, in order to bring something new in, we have to kick something out. So if the CPU loads a value and then loads it again much later, it might not be in cache any more.

The performance of many programs is limited by the effectiveness of the cache. If the data the CPU needs is usually in cache, the program can run at the full speed of the CPU. If the CPU frequently needs data that is not in cache, the program is limited by the speed of memory.

The **cache hit rate**, $h$, is the fraction of memory accesses that find data in cache; the **miss rate**, $m$, is the fraction of memory accesses that have to go to memory. If the time to process a cache hit is $T_h$ and the time for a cache miss is $T_m$, the average time for each memory access is

$$hT_h + mT_m$$

Equivalently, we could define the **miss penalty** as the extra time to process a cache miss, $T_p = T_m - T_h$. Then the average access time is

$$T_h + mT_p$$

When the miss rate is low, the average access time can be close to $T_h$. That is, the program can perform as if memory ran at cache speeds.

## 7.2 Locality

When a program reads a byte for the first time, the cache usually loads a **block** or **line** of data that includes the requested byte and some of its neighbors. If the program goes on to read one of the neighbors, it will find it in cache.

As an example, suppose that the block size is 64 B. And suppose you read a string with length 64, and the first byte of the string happens to fall at

the beginning of a block. When you load the first byte, you would incur a miss penalty, but after that the rest of the string would be in cache. After reading the whole string, the hit rate would be 63/64. If the string spans two blocks, you would incur 2 miss penalties. But even then the hit rate would be 62/64, or almost 97%.

On the other hand, if the program jumps around unpredictably, reading data from scattered locations in memory, and seldom accessing the same location twice, cache performance would be poor.

The tendency of a program to use the same data more than once is called **temporal locality**. The tendency to use data in nearby locations is called **spatial locality**. Fortunately, many programs naturally display both kinds of locality:

- Most programs contain blocks of code with no jumps or branches. Within these blocks, instructions run sequentially, so the access pattern has spatial locality.
- In a loop, programs execute the same instructions many times, so the access pattern has temporal locality.
- The result of one instruction is often used immediately as an operand of the next instruction, so the data access pattern has temporal locality.
- When a program executes a function, its parameters and local variables are stored together on the stack; accessing these values has spatial locality.
- One of the most common processing patterns is to read or write the elements of an array sequentially; this pattern also has spatial locality.

The next section explores the relationship between a program's access pattern and cache performance.

## 7.3 Measuring cache performance

When I was a graduate student at U.C. Berkeley I was a teaching assistant for Computer Architecture with Brian Harvey. One of my favorite exercises involved a program that iterates through an array and measures the average time to read and write an element. By varying the size of the array, it is possible to infer the size of the cache, the block size, and some other attributes.

You can download my modified version of this program from TBA.

The kernel of the program is this loop:

```
    iters = 0;
    do {
        sec0 = get_seconds();

        for (index = 0; index < limit; index += stride)
            array[index] = array[index] + 1;

        iters = iters + 1;
        sec = sec + (get_seconds() - sec0);

    } while (sec < 0.1);
```

The inner `for` loop traverses the array. `limit` determines how much of the array it traverses; `stride` determines how many elements it skips over. For example, if `limit` is 16 and `stride` is 4, the loop would increment elements 0, 4, 8, and 12.

`sec` keeps track of the total CPU time used by the inner loop. The outer loop runs until `sec` exceeds 0.1 seconds, which is long enough that we can compute the average time with sufficient precision.

`get_seconds` uses the system call `clock_gettime`, converts to seconds, and returns the result as a `double`:

```
double get_seconds(){
    struct timespec ts;
    clock_gettime(CLOCK_PROCESS_CPUTIME_ID, &ts);
    return ts.tv_sec + ts.tv_nsec / 1e9;
}
```

To isolate the time to access the elements of the array, the program runs a second loop that is almost identical except that the inner loop doesn't touch the array; it always increments the same variable:

```
    iters2 = 0;
    do {
        sec0 = get_seconds();

        for (index = 0; index < limit; index += stride)
```

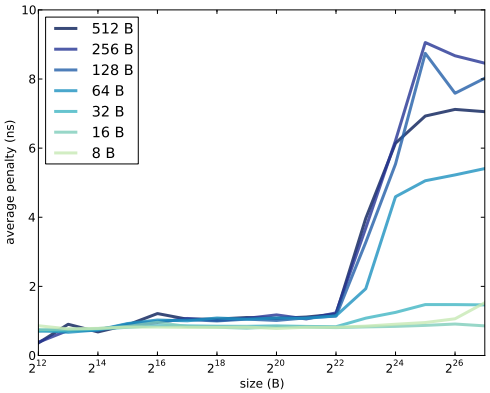


Figure 7.1: Average miss penalty as a function of array size and stride.

```
            temp = temp + index;

        iters2 = iters2 + 1;
        sec = sec - (get_seconds() - sec0);

    } while (iters2 < iters);
```

The second loop runs the same number of iterations as the first. After each iteration, it *subtracts* the elapsed time from `sec`. When the loop completes, `sec` contains the total time for all array accesses, minus the total time it took to increment `temp`. This difference is the total miss penalty incurred by all accesses. Finally, we divide by the number of accesses to get the average miss penalty per access, in ns:

```
sec * 1e9 / iters / limit * stride
```

If you compile and run `cache.c` you should see output like this:

```
Size:     4096 Stride:        8 read+write: 1.1699 ns
Size:     4096 Stride:       16 read+write: 0.7023 ns
Size:     4096 Stride:       32 read+write: 0.7105 ns
Size:     4096 Stride:       64 read+write: 0.6858 ns
```

If you have Python and `matplotlib` installed, you can use `graph_cache_data.py` to graph the results. Figure 7.1 shows the results when I ran it on a Dell Optiplex 7010. Notice that the array size and stride are reported in bytes, not number of array elements.

Take a minute to consider this graph, and see what you can infer about the cache. Here are some things to think about:

- The program reads through the array many times, so it has plenty of temporal locality. If the entire array fits in cache, we expect the average miss penalty to be near 0.
- When the stride is 4 bytes, we read every element of the array, so the program has plenty of spatial locality. For example, if the block size is big enough to contain 64 elements, the hit rate would be 63/64, even if the array does not fit in cache.
- If the stride is equal to the block size (or greater), the spatial locality is effectively zero, because each time we read a block, we only access one element. In that case we expect to see the maximum miss penalty.

In summary, we expect good cache performance if the array is smaller than the case size *or* if the stride is smaller than the block size. Performance only degrades if the array is bigger than the cache *and* the stride is large.

In Figure 7.1, cache performance is good, for all strides, as long as the array is less than $2^{22}$ B. We can infer that the cache size is near 4 MiB; in fact, according to the specs, it is 3 MiB.

When the stride is 8, 16, or 32 B, cache performance is good. At 64 B it starts to degrade, and for larger block sizes the average miss penalty is about 9 ns. We can infer that the block size is 128 B.

Many processors use **multi-level caches** that include a small, fast cache and a bigger, slower cache. In this example, it looks like the miss penalty increases a little when the array size is bigger than $2^{14}$ B, so it's possible that this processor also has a 16 KB cache with an access time less than 1 ns.

## 7.4 Programming for cache performance

Memory caching is implemented in hardware, so most of the time programmers don't need to know much about it. But if you know how caches work, you can write programs that use them more effectively.

For example, if you are working with a large array, it might be faster to traverse the array once, performing several operations with each element, rather than traversing the array several times.

If you are working with a 2-D array, it might be stored as an array of rows. If you traverse through the elements, it would be faster to go row-wise, so

the effective stride is 1 element, rather than column-size, with a stride equal to the row length.

Linked data structures don't always exhibit spatial locality, because the nodes aren't necessarily contiguous in memory. But you allocate many nodes at the same time, they are usually co-located in the heap. Or, even better, if you can allocate an array of nodes, you know they will be contiguous.

Recursive strategies like mergesort often have good cache behavior because they break big arrays into smaller pieces and then work with the pieces. Sometimes these algorithms can be tuned to take advantage of cache behavior.

In extreme cases where performance is very important, it is possible to design algorithms that depend on the size of the cache, the block size, and other hardware characterstics. Algorithms like that are called **cache-aware**. The obvious drawback of cache-aware algorithms is that they are hardware-specific.

## 7.5 The memory hierarchy

At some point during this chapter, a question like the following might have occurred to you: "If caches are so much faster than main memory, why not make a really big cache and forget about memory?"

Without going too far into computer architecture, there are two reasons: electronics and economics. Caches are fast because they are small and close to the CPU, which minimizes delays due to capacitance and signal propagation. If you make a cache big, it will be slower.

Also, caches take up space on the processor chip, and bigger chips are more expensive. Main memory is usually dynamic random-access memory (DRAM), which uses only one transistor and one capacitor per bit, so it is possible to pack more memory into the same amount of space. But this way of implementing memory is slower than the way caches are implemented.

Also main memory is usally packaged in a dual in-line memory module (DIMM) that includes 16 or more chips. Several small chips are cheaper than one big one.

The tradeoff between speed, size, and cost is the fundamental reason for caching. If there were one memory technology that was fast, big, and cheap, we wouldn't need anything else.

The same principle applies to storage as well as memory. Flash drives are fast, but they are more expensive than hard drives, so they tend to be smaller. Tape drives are even slower than hard drives, but they can store very large amounts of data relatively cheaply.

The following table shows the access time, typical size, and cost for each of these technologies:

| Device | Access time | Typical size | Cost |
|---|---|---|---|
| Register | 0.5 ns | 256 B | ? |
| Cache | 1 ns | 2 MiB | ? |
| DRAM | 10 ns | 4 GiB | $10 / GiB |
| SSD | 10 $\mu$s | 100 GiB | $1 / GiB |
| HDD | 5 ms | 500 GiB | $0.25 / GiB |
| Tape | 0.5 ns | 1–2 TiB | $20 / TiB |

The number and size of registers depends on details of the architecture. Current computers have about 32 general-purpose registers, each storing one **word**. On a 32-bit computer, a word is 32 bits or 4 B. On a 64-bit computer, a word is 64 bits or 8 B. So the total size of the register file is 100–300 B.

The cost of registers and caches is hard to quantify. They contribute to the cost of the chips they are on, but consumers don't see that cost directly.

For the other numbers in the table, I looked at the specifications for typical hardware for sale from online computer hardware stores. By the time you read this, these numbers will be obsolete, but they give you an idea of what the performance and cost gaps look like at one point in time.

These technologies make up the **memory hierarchy** (note that this use of "memory" also includes storage). Each level of the hierarchy is bigger and slower than the one above it. And in some sense, each level acts as a cache for the one below it. You can think of main memory as a cache for programs and data that are stored permanently on SSDs and HHDs. And if you are working with very large datasets stored on tape, you could use hard drives to cache one subset of the data at a time.

## 7.6 Caching policy

The memory hierarchy suggests a framework for thinking about caching. At every level of the hierarchy, we have to address four fundamental ques-

tions of caching:

- Who moves data up and down the hierarchy? In enterprise systems, administrators move data explicitly between disk and tape. Users implicitly move data from disk to memory when they execute programs and open files. Hardware on the CPU handles the memory cache. And register allocation is usually done by the compiler.
- What gets moved? In general, block sizes are small at the top of the hierarchy, and bigger at the bottom. In a memory cache, a typical block size is 128 B. Pages in memory might be 4 KiB, but when the operating system reads a file from disk, it might read 10 or 100 blocks at a time.
- When does data get moved? In the most basic cache, data gets moved into cache when it is used for the first time. But many caches use some kind of **prefetching**, meaning that data is loaded before it is explicitly requested. We have already seen a simple form of preloading: loading an entire block when only part of it is requested.
- Where in the cache does the data go? When the cache is full, we can't bring anything in without kicking something out. Ideally, we want to keep data that will be used again soon and replace data that will not be used again.

The answers to these questions make up the **cache policy**. Near the top of the hierarchy, cache policies tend to be simple because they have to be fast and they are implemented in hardware. Near the bottom of the hierarchy, there is more time to make decisions, and well-designed policies can make a big difference.

Most cache policies are based on the principle that history repeats itself; if we have information about the recent past, we can use it to predict the immediate future. For example, if a block of data has been used recently, we expect it to be used again soon. This principle suggests a replacement policy called "least recently used," or LRU, which removes from the cache a block of data that has not been used recently. For more on this topic, see `http://en.wikipedia.org/wiki/Cache_algorithms`.

# Chapter 8

# Multitasking

In many current systems, the CPU contains multiple cores, which means it can run several processes at the same time. In addition, each core is capable of **multitasking**, which means it can switch from one process to another quickly, creating the illusion that many processes are running at the same time.

The part of the operating system that implements multitasking is the **kernel**. In a nut or seed, the kernel is the innermost part, surrounded by a shell. In an operating system, the kernel is the lowest level of software, surrounded by several other layers, including an interface called a "shell." Computer scientists love extended metaphors.

It's hard to define precisely which parts of the operating system are part of the kernel. But at its most basic, the kernel's job is to handle interrupts. An **interrupt** is an event that stops the normal instruction cycle and causes the flow of execution to jump to a special section of code called an **interrupt handler**.

A hardware interrupt is caused when a device sends a signal to the CPU. For example, a network interface might cause an interrupt when a packet of data arrives, or a disk drive might cause an interrupt when a data transfer is complete. Many systems also have timers that cause interrupts at regular intevals.

A software interrupt is caused by a running program. For example, if a computation cannot execute in hardware, it might trigger an interrupt so the condition can be handled in software. Some floating-point errors, like division by zero, can be handled using interrupts.

Also, when a program needs to access a hardware device, it makes a **system call**, which is similar to a function call, except that instead of jumping to the beginning of the function, it executes a special instruction that triggers an interrupt, causing the flow of execution to jump to the kernel. The kernel reads the parameters of the system call, performs the requested operation, and then resumes the interrupted process.

## 8.1 Hardware state

Handling interrupts requires cooperation between hardware and software. When an interrupt occurs, there might be several instructions running on the CPU, and data stored in registers.

Usually the hardware is responsible for bringing the CPU to a consistent state; for example, every instruction should either complete or behave as if it never started. No instruction should be left half complete. Also, the hardware is responsible for saving the program counter (PC), so the kernel knows where to resume.

Then, usually, it is the responsibility of the interrupt handler to save the contents of the registers. In order to do its job, the kernel needs to run code, which modifies data registers and the flag register. So this **hardware state** needs to be saved.

Here is an outline of this sequence of events:

1. When the interrupt occurs, the hardware saves the program counter in a special register and jumps to the appropriate interrupt handlers.

2. The interrupt handlers stores the program counter and the flag register in memory, along with the contents of any data registers it plans to use.

3. The interrupt handler runs whatever code is needed to handle the interrupt.

4. Then it restores the contents of the saved registers. Finally, it restores the program counter of the interrupted process, which has the effect of jumping back to the interrupted instruction.

If this mechanism works correctly, there is generally no way for the interrupted process to know there was an interruption, unless it can detect small changes in the time between instructions.

## 8.2 Context switching

Interrupt handlers can be fast because they don't have to save the entire hardware state; they only have to save registers they are planning to use.

But when an interrupt occurs, the kernel does not always resume the interrupted process. It has the option of switching to another process. This mechanism is called a **context switch**.

In general, the kernel does not know which registers a process will use, so it has to save all of them. Also, when it switches to a new process, it might have to clear data stored in the MMU (see Section **??**). And after the context switch, it might take some time to reload data into the cache. For these reasons, context switches are relatively slow, on the order of thousands of cycles, or a few milliseconds.

In a multi-tasking system, each process is allowed to run for a short period of time called a **time slice** or **quantum**. During the context switch, the kernel sets a hardware timer that causes an interrupt at the end of the time slice. When the interrupt occurs, the kernel can switch to another process or allow the interrupted process to resume. The part of the operating system that makes this decision is the **scheduler**.

## 8.3 The process life cycle

When a process is created, the operating system allocates a data structure that contains information about the process, called a **process control block** or PCB. Among other things, the PCB keeps track of the process state, which is one of:

- Running, if the process is currently running on a core.
- Ready, if the process could be running, but isn't, usually because there are more ready processes than cores.
- Blocked, if the process cannot run because it is waiting for a future event like network communication or a disk read.
- Done, if the process has completed, but has exit status information that has not been read yet.

Here are the events that cause a process to transition from one state to another:

- A process is created when the running program executes a system call. At the end of the system call, the new process is usually ready. Then the scheduler might resume the original process (the "parent") or start the new process (the "child").
- When the scheduler starts a process, it switches the state from ready to running.
- When the scheduler chooses not to resume a running process, it switches to ready.
- If a process executes a system call that cannot complete immediately, like a disk request, it becomes blocked and the scheduler chooses another process.
- When an operation like a disk request completes, it causes an interrupt. The interrupt handler figures out which process was waiting for the operating to complete and switches the state from blocked to ready, or possibly running.
- When a process completes, its exit code is stored in the PCB and its state switches to done.

## 8.4 Scheduling

As we saw in Section **??** there might be hundreds of processes on a computer, but usually most of them are blocked. Most of the time, there are only a few processes that are ready or running. When an interrupt occurs, the scheduler decides which process to start or resume.

On a workstation or laptop, the primary goal of the scheduler is to minimize response time; that is, the computer should respond quickly to user actions. Response time is also important on a server, but in addition the scheduler might try to maximize throughput, which is the number of requests that complete per unit of time.

In general the scheduler doesn't have much information about what processes are doing, so its decisions are usually based on a few heuristics:

- Processes might be limited by different resources. A process that does a lot of computation is probably CPU-bound, which means that its run time depends on how much CPU time it gets. A process that reads data from a network or disk might be I/O-bound, which means that

it would run faster if data input and output went faster, but would not run faster with more CPU time. Finally, a process that interacts with the user is probably blocked, most of the time, waiting for user actions.

The operating system can sometimes classify processes based on their past behavior, and schedule them accordingly. For example, when an interactive process is unblocked, it should probably run immediately, because a user is probably waiting for a reply. On the other hand, a CPU-bound process that has been running for a long time might be less time-sensitive.

- If a process is likely to run for a short time and then make a blocking request, it should probably run immediately, for two reasons: (1) if the request takes some time to complete, we should start it as soon as possible, and (2) it is better for a long-running process to wait for a short one, rather than the other way around.

  As an analogy, suppose you are making an apple pie. The crust takes 5 minutes to prepare, but then it has to chill for half an hour. It takes 20 minutes to prepare the filling. If you prepare the crust first, you can prepare the filling while the crust is chilling, and you can finish the pie in 35 minutes. If you prepare the filling first, the process takes 55 minutes.

Most schedulers use some form of priority-based scheduling, where each process has a priority that can be adjusted up or down over time. When the scheduler runs, it chooses the runnable process with the highest priority, or at least it is more likely to choose a process with high priority.

Here are some of the factors that determine a process's priority:

- A process usually starts with a relatively high priority so it starts running quickly.
- If a process makes a request and blocks before its time slice is complete, it is more likely to be interactive or I/O-bound, so its priority should go up.
- If a process runs for an entire time slice, it is more likely to be long-running and CPU-bound, so its priority should go down.
- If a task blocks for a long time and then becomes ready, it should get a priority boost so it can respond to whatever it was waiting for.

- If process A is blocked waiting for process B, for example if they are connected by a pipe, the priority of process B should go up.
- The system call `nice` allows a process to decrease (but not increase) its own priority, allowing programmers to pass explicit information to the scheduler.

For most systems running normal workloads, scheduling algorithms don't have a substantial effect on performance. Simple scheduling policies are usually good enough.

## 8.5 Real-time scheduling

However, for programs that interact with the real world, scheduling can be very important. For example, a program that reads data from sensors and controls motors might have to complete recurring tasks at some minimum frequency and react to external events with some maximum response time. These requirements are often expressed in terms of **tasks** that must be completed before **deadlines**.

Scheduling tasks to meet deadlines is called **real-time scheduling**. For some applications, a general-purpose operating system like Linux can be modified to handle real-time scheduling. These modifications might include:

- Providing richer APIs for controlling task priorities.
- Modifying the scheduler to guarantee that the process with highest priority runs within a fixed amount of time.
- Reorganizing interrupt handlers to guarantee a maximum completion time.
- Modifying locks and other synchronization mechanisms to allow a high-priority task to preempt a lower-priority task.
- Choosing an implementation of dynamic memory allocation that guarantees a maximum completion time.

For more demanding applications, especially in domains where real-time response is a matter of life and death, **real-time operating systems** provide specialized capabilities, often with much simpler designs than general purpose operating systems.