Computer Systems

Muchang Bahng

Spring 2024

Contents

Before we do any coding, we must learn the theory behind how computer systems work, which all starts from memory management and CPU architecture. We will use C with the gcc compiler, along with MIPS and NASM assembler. It is imperative to learn these two since given that you know a high level language pretty well (Python in my case), you want to learn C to appreciate the things Python does for you, and you want to learn Assembly to appreciate the things C does for you.[1].

To start off, we want a big overall picture of high a computer works. We introduce this with the simplest model of the computer, the Von Nuemann architecture. It consists of a **central processing unit** (CPU), **memory**, and an **input/output** (I/O) system. We show a diagram of this first for conciseness in Figure 1.

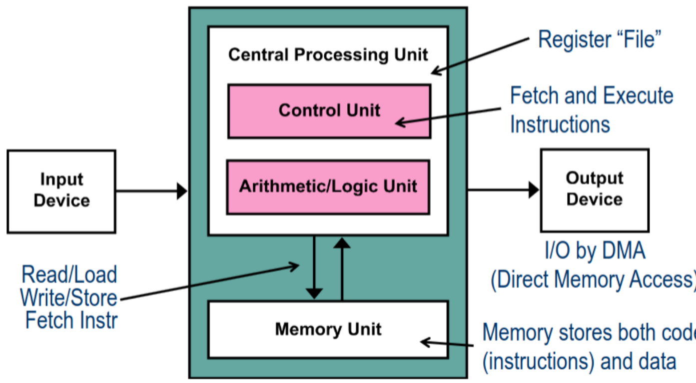


Figure 1: von Neumann Architecture

We will go through these one by one, touching on C and Assembly along the way, but the implementation of these things can differ by the **computer architecture**, so let's list some of the basic ones.

Definition 0.1 (Computer Architecture)

The **computer architecture** is the design of the computer, which includes the CPU, memory, and I/O system. There are many different architectures, but we will focus on the most common ones.

We first go over some basic theoretical properties of basic data types, focusing on C, and then we cover all the stuff about memory and then all the stuff about the CPU. This is a natural progression since to work with data, you must first know where to store the data and how it is stored (the memory), and then you want to know how data is manipulated (the CPU).

1 Encoding Schemes

In order to get into memory, it is helpful to know the theory behind how primitive types are stored in memory.

Definition 1.1 (Collections of Bits)

There are many words that are used to talk about values of different data types:

1. A **bit** (b) is either 0 or 1.
2. A **Hex** (x) is a collection of 4 bits, with a total of $2^4 = 16$ possible values, and this is used since it is easy to read for humans.
3. A **Byte** (B) is a collection of 8 bits or 2 hex, with a total of $2^8 = 256$ possible values, and most computers will work with Bytes as the smallest unit of memory.

[1] https://www.youtube.com/watch?v=XlvfHOrF26M

Definition 1.2 (Collections of Bytes)

Sometimes, we want to talk about slightly larger collections, so we group them by how many bytes they have. However, note that these may not always be the stated size, depending on what architecture or language you are using. This is more of a general term, and they may have different names in different languages. If there is a difference, we will state it explicitly.

1. A **word** (w) is 2 Bytes.
2. A **long** (l) is 4 Bytes.
3. A **quad** (q) is 8 Bytes.

Try to know which letter corresponds to which structure, since that will be useful in both C and Assembly.

1.1 Booleans and Characters

Definition 1.3 (Booleans in C)

The most basic type is the boolean, which is simply a bit. In C, it is represented as `bool`, and it is either `true` (1) or `false` (0).

We can manually check the size of the boolean type in C with the following code.

```c
#include<stdio.h>
#include<stdbool.h>

int main() {
  printf("%lu\n", sizeof(bool));
  return 0;
}
```

```
1
.
.
.
.
.
.
```

Figure 2: We can verify the size of various primitive data types in C with the `sizeof` operator.

1.2 Integer Family

The most primitive things that we can store are integers. Let us talk about how we represent some of the simplest primitive types in C: unsigned short, unsigned int, unsigned long, unsigned long long.

Definition 1.4 (Unsigned Integer Types in C)

In C, there are several integer types. We use this hierarchical method to give flexibility to the programmer on the size of the integer and whether it is signed or not.

1. An **unsigned short** is 2 bytes long and can be represented as a 4-digit hex or 16 bits, with values in $[0 : 65,535]$. Therefore, say that we have
2. An **unsigned int** is 4 bytes long and can be represented as an 8-digit hex or 32 bits, with values in $[0 : 4,294,967,295]$.
3. An **unsigned long** is 8 bytes and can be represented as an 16-digit hex or 64 bits, but they are only guaranteed to be stored in 32 bits in other systems.
4. An **unsigned long long** is 8 bytes and can be represented as an 16-digit hex or 64 bits, and they are guaranteed to be stored in 64 bits in other systems.

Theorem 1.1 (Bit Representation of Unsigned Integers in C)

To encode a signed integer in bits, we simply take the binary expansion of it.

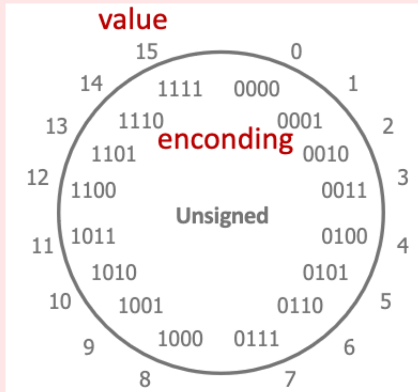


Figure 3: Unsigned encoding of 4-bit integers in C.

Example 1.1 (Bit Representation of Unsigned Integers in C)

We can see for ourselves how these numbers are represented in bits. Printing the values out in binary requires to make new functions, but we can easily convert from hex to binary.

```c
int main() {

  unsigned short x = 13;
  unsigned int y = 256;

  printf("%x\n", x);
  printf("%x\n", y);

  return 0;
}
```

```
d
100
.
.
.
.
.
.
.
.
```

So far, the process of converting unsigned numbers to bits seemed simple. Now let's introduce signed integers.

Definition 1.5 (Signed Integer Types in C)

In C, there are several signed integer types. We use this hierarchical method to give flexibility to the programmer on the size of the integer and whether it is signed or not.

1. A **signed short** is 2 bytes long and can be represented as a 4-digit hex or 16 bits, with values in $[-32,768 : 32,767]$.
2. A **signed int** is 4 bytes long and can be represented as an 8-digit hex or 32 bits, with values in $[-2,147,483,648 : 2,147,483,647]$.
3. A **signed long** is 8 bytes and can be represented as an 16-digit hex or 64 bits, but they are only guaranteed to be stored in 32 bits in other systems.
4. A **signed long long** is 8 bytes and can be represented as an 16-digit hex or 64 bits, and they are guaranteed to be stored in 64 bits in other systems.

To store signed integers, it is intuitive to simply take the first (left-most) bit and have that be the sign. Therefore, we lose one significant figure but gain information about the sign. However, this has some problems: first, there are two representations of zeros: -0 and $+0$. Second, the continuity from -1 to 0 is not natural. It is best explained through an example, which doesn't lose much insight into the general case.

Example 1.2 (Problems with the Signed Magnitude)

Say that you want to develop the signed magnitude representation for 4-bit integers in C. Then, you can imagine the following diagram to represent the numbers.

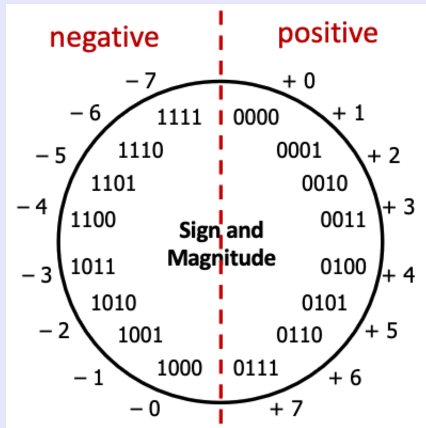


Figure 4: Signed magnitude encoding of 4-bit integers in C.

You can see that there are some problems:
1. There are two representations for 0, which is 0000 and 1000.
2. -1 (1001) plus 1 becomes -2 (1010).
3. The lowest number -7 (1111) plus 1 goes to 0 (0000) when it should go to -6 (1100).
4. The highest number 7 (0111) plus 1 goes to 0 (1000).

An alternative way is to use the two's complement representation, which solves both problems and makes it more natural.

Theorem 1.2 (Bit Representation of Signed Integers in C)

The **two's complement** representation is a way to represent signed integers in binary. It is defined as follows. Given that you want to store a decimal number p in n bits,
1. If p is positive, then take the binary expansion of that number, which should be at most $n-1$ bits (no overflow), pad it with 0s on the left.
2. If p is negative, then you can do two things: First, take the binary expansion of the positive number, flip all the bits, and add 1. Or second, represent $p = q - 2^n$, take the binary representation of q in $n-1$ bits, and add a 1 to the left.

If you have a binary number $b = b_n b_{n-1} \cdots b_1$ then to convert it to a decimal number, you simply calculate

$$q = -b_n 2^{n-1} + b_{n-1} 2^{n-2} + \cdots + b_1 \tag{1}$$

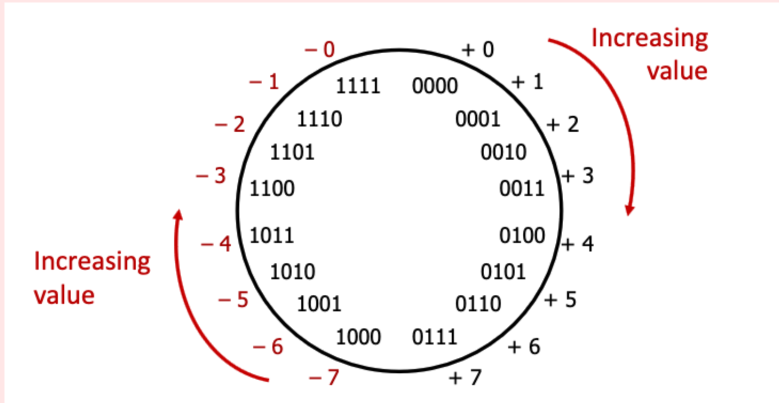


Figure 5: Two's complement encoding of 4-bit integers in C.

Example 1.3 (Bit Representation of Signed Integers in C)

We can see for ourselves how these numbers are represented in bits.

```c
int main() {

  short short_pos = 13;
  short short_neg = -25;
  int int_pos = 256;
  int int_neg = -512;

  printf("%x\n", short_pos);
  printf("%x\n", short_neg);
  printf("%x\n", int_pos);
  printf("%x\n", int_neg);

  return 0;
}
```

```
d
ffe7
100
fffffe00
.
.
.
.
.
.
.
.
.
.
```

```c
#include<stdio.h>
#include<stdbool.h>

int main() {
  printf("%lu\n", sizeof(bool));
  printf("%lu\n", sizeof(short));
  printf("%lu\n", sizeof(int));
  printf("%lu\n", sizeof(long));
  printf("%lu\n", sizeof(long long));
  return 0;
}
```

```
1
2
4
8
8
.
.
.
.
.
.
```

Figure 6: Size of various integer types in C with the `sizeof`.

1.2.1 Arithmetic Operations on Binary Numbers

Theorem 1.3 (Inversion of Binary Numbers)

Given a binary number p, to compute $-p$, simply invert the bits and add 1.

Theorem 1.4 (Addition and Subtraction of Binary Numbers)

Given two binary numbers p and q.
1. To compute $p + q$, simply add the numbers together as you would in base 10, but carry over when the sum is greater than 1.
2. To compute $p - q$, you can invert q to $-q$ and compute $p + (-q)$.

1.3 Float Family

Definition 1.6 (Floating Point Types in C)

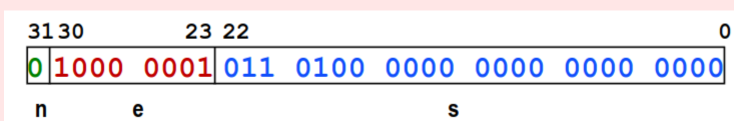
In C, there are several floating point types. We use this hierarchical method to give flexibility to the programmer on the size of the integer and whether it is signed or not.
1. A **float** is 4 bytes long and can be represented as an 8-digit hex or 32 bits, with values in $[1.2 \times 10^{-38} : 3.4 \times 10^{38}]$.
2. A **double** is 8 bytes long and can be represented as an 16-digit hex or 64 bits, with values in $[2.3 \times 10^{-308} : 1.7 \times 10^{308}]$.
3. A **long double** is 8 bytes and can be represented as an 16-digit hex or 64 bits, but they are only guaranteed to be stored in 80 bits in other systems.

Theorem 1.5 (Bit Representation of Floating Point Types in C)

Floats are actually like signed magnitude. We have

$$(-1)^n \times 2^{e-127} \times 1.s \tag{2}$$

where



Doubles encode 64 bits, so not we have exponent having 11 bits (so bias is not 1023) and 52 bits for mantissa.

2 Memory

Definition 2.1 (Memory)

The **memory** is where the computer stores data and instructions, which can be though of as a giant array of memory addresses, with each containing a byte. This data consists of graphical things or even instructions to manipulate other data. It can be visualized as a long array of boxes that each have an **address** (where it is located) and **contents** (what is stored in it).
Memory simply works as a bunch of bits in your computer with each bit having some memory address, which is also a bit. For example, the memory address `0b0010` (2) may have the bit value of `0b1` (1) stored in it.

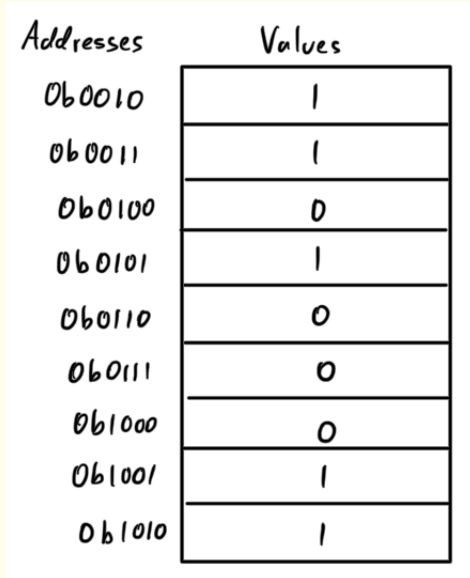


Figure 7: Visualization of memory as a long array of boxes of bits.

However, computers do not need this fine grained level of control on the memory, and they really work at the Byte level rather than the bit level. Therefore, we can visualize the memory as a long array of boxes indexed by *Bytes*, with each value being a byte as well. In short, the memory is **byte-addressable**. In certain arthitectures, some systems are **word-addressable**, meaning that the memory is addressed by words, which are 4 bytes.[a]

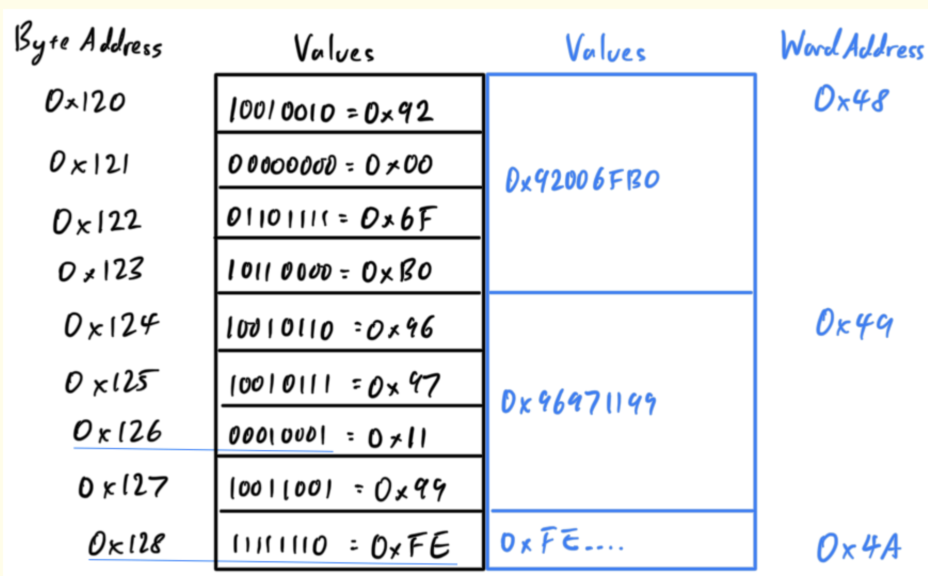


Figure 8: Visualization of memory as a long array of boxes of bytes. Every address is a byte and its corresponding value at that address is also a byte, though we represent it as a 2-digit hex.

[a]Note that in here the size of a word is 2 bytes rather than 4 as stated above. This is just how it is defined in some `x86` architectures.

In the examples above, I listed the memory addresses as a 3 hex character (1.5 bytes) for brevity. In reality, the number of bytes that a memory address takes is much longer.

Definition 2.2 (32 and 64 Bit Machines)

There are two types of machines that tend to format these boxes very differently: 32-bit and 64-bit machines.

1. 32 bit machines store addresses in 32 bits, so they can have 2^{32} addresses, which is about 4 GB of memory.
2. 64 bit machines store addresses in 64 bits, so they can have 2^{64} addresses, which is about 16 EB of memory. This does not mean that the actual RAM is 16 EB, but it means that the machine can *handle* that much memory.

```
...
0x00007FFF7FBFF860 --> 0b000000000000000000000000011111111111
                         111101111111101111111111100001100000
0x00007FFF7FBFF861 --> 0b000000000000000000000000011111111111
                         111101111111101111111111100001100001
0x00007FFF7FBFF862 --> 0b000000000000000000000000011111111111
                         111101111111101111111111100001100010
0x00007FFF7FBFF863 --> 0b000000000000000000000000011111111111
                         111101111111101111111111100001100011
0x00007FFF7FBFF864 --> 0b000000000000000000000000011111111111
                         111101111111101111111111100001100100
...
```

The numbers typically mean the size of the type that the machine works best with, so all memory addresses will be 32 or 64 bits wide. Most machines are 64-bits, and so everything in this notes will assume that we are working with a 64 bit machine. As we will later see, this is why pointers are 8 bytes long, i.e. 64 bits. This is because the memory addresses are 64 bits long, though all of them are not used.

With this structure in mind and knowing the size of some primitive types, we can now focus on how declaring them works in the backend.

Definition 2.3 (Declaration, Initialization)

Assigning a value to a variable is a two step process, which is often not distinguished in high level languages like Python.

1. You must first **initialize** the variable by setting aside the correct number of bytes in memory.
2. You must then **assign** that variable to be some actual value.

The two step process is often called declaration.

This is the reason why C is statically, or strongly, typed. In order to set aside some memory for a variable, you must know how big that variable will be, which you know by its type. This makes sense. We can first demonstrate how to both initialize and declare a variable.

```c
int main() {
  // declaring
  int x = 4;
  printf("%p\n", &x);

  // initializing and assigning
  int y;
  printf("%p\n", &y);
  y = 3;
  printf("%p\n", &y);

  return 0;
}
```

```
0x16d37ee68
0x16d37ee64
0x16d37ee64
.
.
.
.
.
.
.
.
.
.
```

Figure 9: How to declare variables in C. As you can see, by initializing `y`, the memory address is already assigned and it doesn't change when you assign it. The address is only shown to be 9 hex digits long, but it is actually 16 hex digits long and simply 0 padded on the left.

One question that may come to mind is, what is the value of the variable if you just initialize it? After all the value at that address that is initialized must be either 0s or 1s. Let's find out.

```c
int main() {
  int y;
  printf("%d\n", y);
  y = 3;
  printf("%d\n", y);

  return 0;
}
```

```
6298576
3
.
.
.
.
.
.
```

Figure 10: The value of an uninitialized variable is some random number.

It may be interesting to see how this random unititialized value is generated. It is simply the value that was stored in that memory address before, and it is not cleared when you initialize it, so you should not use this as a uniform random number generator.

2.1 Debugging and Object Dumping

Talk about gdb, lldb, objdump, etc. These are debugging tools that allow you to parse your code line by line. However, to actually see the C code, you must compile it with the debugging flag. This adds a little bit of overhead memory to the binary, but not a lot.

2.2 Endian Architecture

It is intuitive to think that given some multi-byte object like an `int` (4 bytes), the beginning of the int would be the lowest address and the end of the int would be the highest address, like how consecutive integers are stored in an array. However, this is not always the case (almost always not the case since most computers are little-endian).

Definition 2.4 (Endian Architecture)

Depending on the machine architecture, computers may store these types slightly differently in their *byte* order. Say that we have an integer of value `0xA1B2C3D4` (4 bytes). Then,

1. A **big-endian architecture** (e.g. SPARC, z/Architecture) will store it so that the least significant byte has the highest address.
2. A **little-endian architecture** (e.g. x86, x86-64, RISC-V) will store it so that the least significant byte has the lowest address.
3. A **bi-endian architecture** (e.g. ARM, PowerPC) can specify the endianness as big or little.

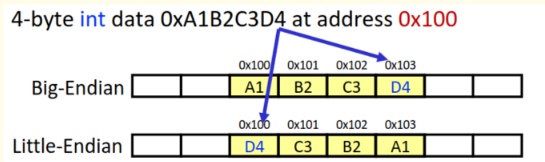


Figure 11: The big vs little endian architectures.

We can simply print out the hex values of primitive types to see how they are stored in memory, but it does not provide the level of details that we want on which bytes are stored where. At this point, we must use certain **debuggers** to directly look at the memory. For x86 architectures, we can use `gdb` and for ARM architectures, we can use `lldb`. At this point, we need to understand assembly to look through debuggers, so we will provide the example here.

Example 2.1 (Endianness of C Int in x86-64)

To do.

Example 2.2 (Endianness of C Int in ARM64)

To do.

2.3 Type Casting

2.4 Pointers

We have learned how to declare/initialize a variable, which frees up some space in the memory and possibly assigns a value to it. One great trait of C is that we can also store the memory address of a variable in another variable called a pointer. You access both the memory and the value at that memory with this pointer variable.

Definition 2.5 (Pointer Variable)

A **pointer** variable/type is a variable that stores the memory address of another variable.

1. You can declare a pointer in the same way that you declare a variable, but you must add a asterisk in front of the variable name.
2. The size of this variable is the size of the memory address, which is 8 bytes in a 64-bit architecture.
3. To get the value of the variable that the pointer points to, called **dereferencing**, you simply put a asterisk in front of the pointer. This is similar to how you put a ampersand in front of a variable to get its memory address.

```c
int main() {
  // declare an integer
  int x = 4;
  printf("x = %d\n", x);
  printf("&x = %p\n", &x);

  // declare pointer
  int *p = &x;
  printf("p = %p\n", p);
  printf("*p = %d\n", *p);

  // initialize pointer
  int *q;
  q = &x;
  printf("q = %p\n", q);
  printf("*q = %d\n", *q);
  return 0;
}
```

```
x = 4
&x = 0x16d49ae68
p = 0x16d49ae68
*p = 4
q = 0x16d49ae68
*q = 4
.
.
.
.
.
.
.
.
.
.
.
.
```

Figure 12

Since the size of addresses are predetermined by the architecture, it may not seem like we need to know the underlying data type of what it points to, so why do we need to write strongly type the underlying data type? Remember that to do pointer arithmetic, you need to know how large the underlying data type is so that you can know how many bytes to move when traversing down an array.

One of the reasons why pointers are so valuable is that they allow you to pass by reference, which is a way to change the value of a variable in a function.

2.4.1 Call by Value vs Call by Reference

Definition 2.6 (Call by Value)

Definition 2.7 (Call by Reference)

2.4.2 Pointer Errors

Just like for regular variables, you may be curious on the value of an unassigned pointer. Let's take a look.

Example 2.3 (Uninitialized Pointers)

```c
int main() {
  int x = 4;
  int *p;
  printf("p = %p\n", p);
  printf("*p = %x\n", *p);

  return 0;
}
```

```
p = 0x10249ff20
*p = d100c3ff
.
.
.
.
.
.
```

Figure 13: The value of an uninitialized pointer is some random address and at a random address it would be some random byte.

This is clearly not good, especially since the program compiles correctly and runs without any errors. This kind of pointer that hasn't been initialized is called a wild pointer.

Definition 2.8 (Wild Pointer)

A **wild pointer** is a pointer that has not been initialized to a known value.

To fix this, we must always initialize a pointer to a known value. This may come at a disadvantage, since now we can't reap the benefits of initializing first and assigning later. A nice compromise is to initialize the pointer to a null pointer.

Definition 2.9 (Null Pointer)

A **null pointer** is a pointer that has been initialized to a known value, which is the address 0x0. You can set the type of the pointer and then initialize it to `NULL`.

```c
int main() {
  int *p = NULL;
  printf("p = %p\n", p);

  // the code below gives seg fault
  /* printf("*p = %d\n", *p); */

  int x = 4;
  p = &x;
  printf("p = %p\n", p);
  printf("*p = %d\n", *p);
  return 0;
}
```

```
p = 0x0
p = 0x16da72e5c
*p = 4
.
.
.
.
.
.
.
.
.
.
```

Figure 14: Initializing a null pointer. It is a good practice to initialize a pointer to a null value.

Therefore, the null pointer allows you to set the type of the underlying data type, but the actual address will be 0x0. You cannot dereference a null pointer, and doing so will give you a segmentation fault. There may be times when you do not even know the data type of the pointer, and for this you can use the void pointer, which now doesn't know the type of the variable that it points to but it does allocate address.

Definition 2.10 (Void Pointer)

A **void pointer** is a pointer that does not know the type of the variable that it points to. We can initialize it by simply setting the underlying type to be void. This initializes the address, which should always be 8 bytes, but trying to access the value of the variable is not possible.

```c
int main() {
  void *p;
  printf("p = %p\n", p);
  int x = 4;
  p = &x;
  printf("%d", *((int*)p));
  return 0;
}
```

```
p = 0x102553f54
4
.
.
.
.
.
.
```

Figure 15: Initialize a void pointer and then use typecasting to access the value of the variable that it points to.

2.5 Pointer Arithmetic

With pointers out of the way, we can talk about how arrays are stored in memory.

Definition 2.11 (Array)

A C array is a collection of elements of the same type, which are stored in contiguous memory locations. You can initialize and declare arrays in many ways, and access their elements with the index, e.g. `arr[i]`.

1. You declare an array of some constant number of elements n with the elements themselves.

```c
int arr[5] = {1, 2, 3, 4, 5};
```

2. You declare an array with out its size n and simply assign them. Then n is automatically determined.

```c
int arr[] = {1, 2, 3, 4, 5};
```

3. You initialize an array of some constant size c, and then you assign each element of the array.

```c
int arr[5];
for (int i = 0; i < 5; i++) {
  arr[i] = i + 1;
}
```

Unfortunately, C does not provide a built-in way to get the size of the array (like `len` in Python), so we must keep track of the size of the array ourselves. Furthermore, the address of the array is the address of where it begins at, i.e. the address of the first element.

You can literally see that the elements of the array are contiguous in memory by iterating through each element and printing out its address.

```c
int main(void) {
  // initialize array
  int arr[5];
  for (int val = 1; val < 6; val++) {
    arr[val-1] = val * val;
  }

  int* p = &arr[0];
  for (int i = 0; i < 5; i++) {
    printf("Value at position %d : %d\n", i,
    arr[i]);
    printf("Address at position %d : %p\n",
    i, p + i);
  }

  return 0;
}
```

```
Value at position 0 : 1
Address at position 0 : 0x7ffd8636b0d0
Value at position 1 : 4
Address at position 1 : 0x7ffd8636b0d4
Value at position 2 : 9
Address at position 2 : 0x7ffd8636b0d8
Value at position 3 : 16
Address at position 3 : 0x7ffd8636b0dc
Value at position 4 : 25
Address at position 4 : 0x7ffd8636b0e0
.
.
.
.
.
.
.
```

Figure 16: Ints are 4 bytes long, so the address of the next element is 4 bytes away from the previous element, making this a contiguous array.

The most familiar implementation of an array is a string in C.

Definition 2.12 (String)

A string is an array of characters, which is terminated by a null character `\0`. You can initialize them in two ways:

1. You can declare a string with the characters themselves, which you must make sure to end with the null character.

```c
char str[6] = {'H', 'e', 'l', 'l', 'o', '\0'};
```

2. You can declare them with double quotes, which automatically adds the null character.

```c
char str[] = "Hello";
```

Note that for whatever string we initialize, the size of the array is the number of characters plus 1.

To access elements of an array, you simply use the index of the element, e.g. `arr[i]`, but in the backend, this is implemented with *pointer arithmetic*.

Definition 2.13 (Pointer Arithmetic)

Pointer arithmetic is the arithmetic of pointers, which is done by adding or subtracting an integer to a pointer.

1. If you add an integer n to a pointer p, e.g. `p + n`, then the new pointer will point to the nth element after the current element, with the next element being `sizeof(type)` bytes away from the pervious element.
2. If you subtract an integer n from a pointer, then the pointer will point to the nth element before the current element.

This is why you can access the elements of an array with the index, since the index is simply the number of elements away from the first element.

Example 2.4 (Pointer Arithmetic with Arrays of Ints and Chars)

Ints have a size of 4 bytes and chars 1 byte. You can see that using pointer arithmetic, the addresses of the elements of ints increment by 4 and those of the char array increment by 1.

```c
int main() {
  int integers[3] = {1, 2, 3};
  char characters[3] = {'a', 'b', 'c'};
  int *p = &integers[0];
  char *q = &characters[0];

  printf("Array of Integers\n");
  for (int i = 0; i < 3; i++) {
    printf("%p\n", integers+i); }

  printf("Array of Characters\n");
  for (int i = 0; i < 3; i++) {
    printf("%p\n", characters+i); }
  return 0;
}
```

```
Array of Integers
0x16d39ee58
0x16d39ee5c
0x16d39ee60
.
Array of Characters
0x16d39ee50
0x16d39ee51
0x16d39ee52
.
.
.
.
.
.
```

Therefore, we can think of accessing the elements of an array as simply pointer arithmetic.

Theorem 2.1 (Bracket Notation is Pointer Arithmetic)

The bracket notation is simply pointer arithmetic in the backend.

```c
int main() {
  int arr[3] = {1, 2, 3};
  int *p = &arr[0];

  for (int i = 0; i < 3; i++) {
    printf("%d\n", arr[i]);
    printf("%d\n", *(p+i));
  }
  return 0;
}
```

```
1
1
2
2
3
3
.
.
.
.
```

Figure 17: Accessing the elements of the list using both ways is indeed the same.

2.6 Global, Stack, and Heap Memory

Everything in a program is stored in memory, variables, functions, and even the code itself. However, we will find out that they are stored in different parts of the memory. When a program runs, its application memory consists of four parts, as visualized in the Figure 18.

1. The **code** is where the code text is stored.
2. The **global memory** is where all the global variables are stored.
3. The **stack** is where all of the functions and local variables are stored.
4. The **heap** is variable and can expand to as much as the RAM on the current system. We can specifically store whatever variables we want in the heap.

We provide a visual of these four parts first, and we will go into them later.

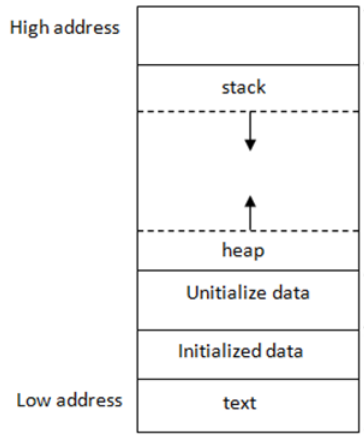


Figure 18: The four parts of memory in a C program.

Definition 2.14 (Code Memory)

This is where the code text is stored. It is read-only and is not modifiable.

In high level languages, we always talk about local and global scope. That is, variables defined within functions have a local scope in the sense that anything we modify in the local scope does not affect the global scope. We can now understand what this actually means by examining the backend. The global scope variables are stored in the global memory, and all local variables (and functions) are stored in the stack.

Definition 2.15 (Global Memory)

This is where all the global variables are stored.

Definition 2.16 (Stack Memory)

This is where all of the functions and local variables are stored. As we will see later, the compiler will always run the main function, which must exist in your file. By the main function is a function itself, and therefore it has its own local scope.
Then, when you initialize any functions or local variables within those functions (which will be the majority of your code), all these will be stored in the stack, which is an literally an implementation of the stack data structure. It is LIFO, and the first thing that goes in is the `main` function and its local variables, which is referred to as the **stack frame**. You can't free memory in the stack unless its in the top of the stack.

To see what happens in the stack, we can go through an example.

Example 2.5 (Going through the Stack)

Say that you have the following code:

```c
int total;
int Square(int x) {
  return x*x;
}
int SquareOfSum(int x, int y) {
  int z = Square(x + y);
  return z;
```

```c
}
int main() {
  int a = 4, b = 8;
  total = SquareOfSum(a, b);
  printf("output = %d", total);
  return 0;
}
```

The memory allocation of this program will run as such:

1. The `total` variable is initialized and is put into global memory.
2. `main` is called. It is put into the stack.
3. The local variables `a=4` and `b=8` are initialized and are put into the stack.
4. The `SquareOfSum` function is called and put into the stack.
5. The input local variables `x=4`, `y=8`, `z` are initialized and put into the stack.
6. `x + y=12` is computed and put into the stack.
7. The `Square` function is called and put into the stack.
8. The `x=12` local variable of `Square` is initialized and put into the stack.
9. The CPU computes `x*x=144` and returns the output. The `Square` function is removed from the stack.
10. We assign `z=144` and `SquareOfSum` returns it. Now `SquareOfSum` is removed from the stack.
11. `total=144` is assigned in the global memory still.
12. The `printf` function is called and put into the stack.
13. The `printf` function prints the output and is removed from the stack.
14. The `main` function returns `0` and is removed from the stack, ending our application.

One limitation of the stack is that its total available memory is fixed from the start, ranging from 1MB to 8MB, and so you can't initialize arrays of billions of integers in the stack. It will cause a memory overflow. In fact, the memory of the stack, along with the global and text memory, are assigned at compile time, making it a **static memory**.

Since the stack is really just a very small portion of available memory, the heap comes into rescue, which is the pool of memory available to you in RAM.

Definition 2.17 (Heap Memory)

The **heap memory** (nothing to do with the heap data structure) is a variable length (meaning it can grow at runtime) and **dynamically allocated** (meaning that we can assign memory addresses during runtime) memory that is limited to your computer's hardware. Unlike simply initializing variables to allocate memory as in the stack, we must use the **malloc** and **free** functions in C, and **new** and **delete** operations in C++.

Definition 2.18 (malloc)

Definition 2.19 (free)

The stack can store pointer variables that point to the memory address in the heap. So the only way to access variables in the heap is through pointer reference, and the stack provides you that window to access that big pool of heap memory.

One warning: if you allocate another address, the previous address does not get deallocated off the memory.

Definition 2.20 (Memory Leak)

On the other hand, if you free an address but have a pointer still pointing to that address, this is also a problem called the dangling pointer.

Definition 2.21 (Dangling Pointer)

At this point, we might be wondering why we need both a stack and a heap. Well the benefits of heaps are clearer since you can dynamically allocate memory, and you don't have the LIFO paradigm that is blocking you from deallocating memory that has been allocated in the beginning of your program. A problem with just having heap is that stacks can be orders of magnitude times faster when allocating/deallocating from it than the heap, and the sequence of function calls is naturally represented as a stack.

2.7 Dynamic Memory Allocation

Let's talk about how `malloc` and `free` are implemented in C. If you make a for loop and simply print all the addresses that you allocate to. You will find that they can be quite random. After a program makes some calls to malloc and free, the heap memory can becomes fragmented, meaning that there are chunks of free heap space interspersed with chunks of allocated heap space. The heap memory manager typically keeps lists of different ranges of sizes of heap space to enable fast searching for a free extent of a particular size. In addition, it implements one or more policies for choosing among multiple free extents that could be used to satisfy a request.

The free function may seem odd in that it only expects to receive the address of the heap space to free without needing the size of the heap space to free at that address. That's because malloc not only allocates the requested memory bytes, but it also allocates a few additional bytes right before the allocated chunk to store a header structure. The header stores metadata about the allocated chunk of heap space, such as the size. As a result, a call to free only needs to pass the address of heap memory to free. The implementation of free can get the size of the memory to free from the header information that is in memory right before the address passed to free.

3 Implementations of Memory Structures in C

3.1 Arrays

3.2 Strings

3.3 Structs

3.4 Functions

3.5 Classes (for C++)

3.6 Input Output

We have standard in, standard out, and standard error.

4 Central Processing Unit

Now let's talk about how functions work on a deeper level. When we write a command, like `int x = 4`, we are manually looking for an address (in the stack, global, or heap) and rewriting the bits that are at that address. Functions are just an automated way to do this, and all these modifications and computations are done by the CPU.

Definition 4.1 (Central Processing Unit)

The CPU is responsible for taking instructions (data) from memory and executing them.

1. The CPU is composed of **registers** (different from the cache), which are small, fast storage locations. These registers can either be **general purpose** (can be used with most instructions) or **special purpose** (can be accessed through special instructions, or have special meanings/uses, or are simply faster when used in a specific way).
2. The CPU also has an **arithmetic unit** and **logic unit**, which is responsible for performing arithmetic and logical operations.
3. The CPU also has a **control unit**, which is responsible for fetching instructions from memory through the **databus**, which is literally a wire connecting the CPU and RAM, and executing them.

It executes instructions from memory one at a time and executes them, known as the **fetch-execute cycle**. It consists of 4 main operations.

1. **Fetch**: The **program counter**, which holds the memory address of the next instruction to be executed, tells the control unit to fetch the instruction from memory through the databus.
2. **Decode**: The fetched data is passed to the **instruction decoder**, which figures out what the instruction is and what it does and stores them in the registers.
3. **Execute**: The arithmetic and logic unit then carries out these operations.
4. **Store**: Then it puts the results back on the databus, and stores them back into memory.

The CPU's **clock cycle** is the time it takes for the CPU to execute one instruction. More specifically, the clock cycle refers to a single oscillation of the clock signal that synchronizes the operations of the processor and the memory (e.g. fetch, decode, execute, store), and decent computers have clock cycles of at least 2.60GHz (2.6 billion clock cycles per second).

Therefore, in order to actually do computations on the data stored in the memory, the CPU must first fetch the data, perform the computations, and then store the results back into memory. This can be done in two ways.

1. Load and Store Operations: CPUs use load instructions to move data from memory to registers (where operations can be performed more quickly) and store instructions to move the modified data back into memory.
2. If the data is too big to fit into the registers, the CPU will use the **cache** to store the data, and in worse cases, the actual memory itself. Compilers optimize code by maximizing the use of registers for operations to minimize slow memory access. This is why you often see assembly code doing a lot in registers.

To clarify, let us compare registers and memory. Memory is addressed by an unsigned integer while registers have names like `%rsi`. Memory is much bigger at several GB, while the total register space is much smaller at around 128 bytes (may differ depending on the architecture). The memory is much slower than registers, which is usually on a sub-nanosecond timescale. The memory is dynamic and can grow as needed while the registers are static and cannot grow.

The specific structure/architecture of the CPU is determined by the instruction set architecture (ISA), which can be thought of as a subset of the general computer architecture.

Definition 4.2 (Instruction Set Architecture)

The **ISA** or just **architecture** of a CPU is a high level description of what it can do. Some differences are listed here:

1. What instructions it can execute.
2. The instruction length and decoding, along with its complexity.
3. The performance vs power efficiency.

ISAs can be classified into two types.

1. The **complex instruction set computer** (CISC) is characterized by a large set of complex instructions, which can execute a variety of low-level operations. This approach aims to reduce the number of instructions per program, attempting to achieve higher efficiency by performing more operations with fewer instructions.
2. The **reduced instruction set computer** (RISC) emphasizes simplicity and efficiency with a smaller number of instructions that are generally simpler and more uniform in size and format. This approach facilitates faster instruction execution and easier pipelining, with the philosophy that simpler instructions can provide greater performance when optimized.

Just like how memory addressing is different between 32 and 64 bit machines, CPUs also use these schemes. While 32-bit processors have 2^{32} possible addresses in their cache, it turns out that 64-bit processors have a 48-address space. This is because CPU manufacturers took a shortcut. They use an instruction set which allows a full 64-bit address space, but current CPUs just only use the last 48-bits. The alternative was wasting transistors on handling a bigger address space which wasn't going to be needed for many years (since 48-bits is about 256TB). Just a bit of history for you. Finally, just to briefly mention, the input/output device, as the name suggests, processes inputs and displays outputs, which is how you can see what the program does.

Example 4.1 (x86 Architecture)

The x86 architecture is a CISC architecture, which is the most common architecture for personal computers. Here are important properties:

1. It is a complex instruction set computer (CISC) architecture, which means that it has a large set of complex instructions[a].
2. Byte-addressing is enabled and words are stored in little-endian format.
3. In the x86_64 architecture, registers are 8 bytes long (and 4 bytes in x86_32) and there are 16 total general purpose registers, for a total of only 128 bytes (very small compared to many GB of memory). Other special purpose registers are also documented in the wikipedia page, but it is not fully documented.

[a] https://en.wikipedia.org/wiki/X86_instruction_listings

Example 4.2 (ARM Archiecture)

Mainly in phones, tablets, laptops.

Example 4.3 (MIPS Architecture)

MIPS is a RISC architecture, which is used in embedded systems such as digital home and networking equipment.

Definition 4.3 (Input/Output Device)

The input device can read/load/write/store data from the outside world. The output device, which has **direct memory address**, can display data to the outside world.

One final note to mention, there are many assembly languages out there and various syntaxes.

Example 4.4 (Assembly Syntax)

The two most popular syntaxes are AT&T and Intel.

1. **Intel Syntax**: Specifies memory operands without any special prefixes. Square brackets [] are used to denote memory addresses. For example, mov eax, [ebx] means move the contents of the

memory location pointed to by ebx into eax.

2. **AT&T Syntax**: Memory operands are denoted with parentheses () and include the % prefix for registers. An instruction moving data from a memory location into a register might look like movl (%ebx), %eax, with additional prefixes for immediate values and segment overrides.

Example 4.5 (Assembly Languages)

The various assembly languages are as follows:

1. **x86 Assembly** : The assembly language for Intel and AMD processors using the x86 architecture. Both AT&T and Intel syntax are available. Tools or environments often allow switching between the two, with AT&T being the default in GNU tools like GDB.
2. **ARM Assembly** : The assembly language for ARM processors. Has its own unique syntax, not categorized as AT&T or Intel. ARM syntax is closely tied to its instruction set architecture and is distinct from the x86 conventions.
3. **MIPS Assembly** : The assembly language for MIPS processors. MIPS uses its own assembly language syntax, which is neither AT&T nor Intel. MIPS syntax is designed around the MIPS instruction set architecture.
4. **PowerPC Assembly** : The assembly language for PowerPC processors. PowerPC has its own syntax style, tailored to its architecture and instruction set, distinct from the AT&T and Intel syntax models.
5. **6502 Assembly** : Used in many early microcomputers and gaming consoles. Utilizes a syntax unique to the 6502 processor, not following AT&T or Intel conventions.
6. **AVR Assembly** : The assembly language for Atmel's AVR microcontrollers. AVR assembly follows its own syntax style, designed specifically for AVR microcontrollers and not based on AT&T or Intel syntax.
7. **Z80 Assembly** : Associated with the Z80 microprocessor, used in numerous computing devices in the late 20th century. Z80 assembly language has its own syntax that does not adhere to AT&T or Intel syntax guidelines.

The most common one is the x86_64, which is the one that we will be focusing on, with the AT%T syntax.

4.1 Circuits

Let's go over some common logic gates since this is at the basis of how to construct arithmetic operations.

Definition 4.4 (AND, NOT, OR)

Definition 4.5 (XOR, NAND, NOR)

Definition 4.6 (NAND)

Talk about how to construct arithmetic operations with these gates such as adding two integers or multiplying them, and not just that, but other operations that we may need in a programming language.

Theorem 4.1 (Implementation of Moving Data in Circuits)

Theorem 4.2 (Implementation of Addition, Subtraction in Circuits)

Theorem 4.3 (Implementation of Multiplication in Circuits)

Theorem 4.4 (Implementation of Bitwise Operations in Circuits)

Theorem 4.5 (Implementation of Bitshift Operations)

We also want some sort of conditionals. This then can be used to implement loops by checking some conditional.

Theorem 4.6 (Implementation of Conditionals in Circuits)

As a bonus, we talk about the difference between volatile and non-volatile memory. We already learned that RAM is volatile, and this is simple to implement in a circuit since we can manually set all the bits to 0 or just deplete all power. If this is the case, then how does non-volatile memory like SSDs maintain their state?

Theorem 4.7 (Implementation of Volatile Memory)

Theorem 4.8 (Implementation of Non-Volatile Memory)

4.2 Registers

To understand anything that the CPU does, we must understand assembly language. In here, everything is done within registers, and we can see how the CPU fetches, decodes, and executes instructions. So what exactly are these registers?

Definition 4.7 (Register)

A register is a small, fast storage location within the CPU. It is used to store data that is being used immediately, and is the only place where the CPU can perform operations, which is why it must move data from memory to registers before it can perform operations on it. Everything in a register is in binary, at most 8 bytes, or 64 bits.
There are very specific types of registers that you should know. All of these registers are implemented for all assembly languages and are integral to the workflow of the CPU.

1. **parameter registers** which store the parameters of a function.
2. **Return registers** which store return values of functions.
3. **stack pointers** which point to the top of the stack (at the top of the current stack frame).
4. **frame pointers** which point to the base of the current stack frame.
5. **instruction pointers** which point to the next instruction to be executed.

4.2.1 x86 Assembly Registers

The specific type of registers that are available to a CPU depends on the computer architecture, or more specifically, the ISA, but here is a list of common ones for the x86-64. We have `%rax`, `%rbx`, `%rcx`, `%rdx`, `%rsi`, `%rdi`, `%rbp`, `%rsp`, `%r8`, `%r9`, `%r10`, `%r11`, `%r12`, `%r13`, `%r14`, `%r15`. Therefore, the x86-64 Intel CPU has a total of 16 registers for storing 64 bit data. However, it is important to know which registers are used for what.

Definition 4.8 (Parameter Registers)

Compilers typically store the first six parameters of a function in registers

$$\texttt{\%rdi}, \texttt{\%rsi}, \texttt{\%rdx}, \texttt{\%rcx}, \texttt{\%r8}, \texttt{\%r9}, \tag{3}$$

respectively.

Definition 4.9 (Return Register)

The return value of a function is stored in the

$$\texttt{\%rax} \tag{4}$$

register.

Definition 4.10 (Stack and Frame Pointers)

The `%rsp` register is the **stack pointer**, which points to the top of the stack. The `%rbp` register is the **frame pointer**, or **base pointer**, which points to the base of the current stack frame. In a typical function prologue, **%rbp** is set to the current stack pointer (**%rsp**) value, and then `%rsp` is adjusted to allocate space for the local variables of the function. This establishes a fixed point of reference (`%rbp`) for accessing those variables and parameters, even as the stack pointer (`%rbp`) moves.

Definition 4.11 (Instruction Pointer)

The `%rip` register is the **instruction pointer**, which points to the next instruction to be executed. Unlike all the registers that we have shown so far, programs cannot write directly to `%rip`.

Definition 4.12 (Notation for Accessing Lower Bytes of Registers)

Sometimes, we need a more fine grained control of these registers, and x86-64 provides a way to access the lower bits of the 64 bit registers. We can visualize them with the diagram below.

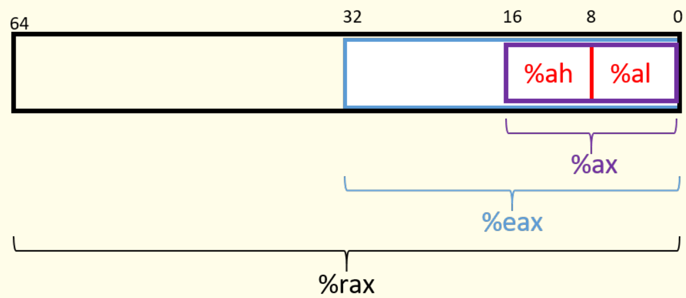


Figure 19: The names that refer to subsets of register **%rax**.

A complete list is shown below.

64-bit Register	32-bit Register	Lower 16 Bits	Lower 8 Bits
%rax	%eax	%ax	%al
%rbx	%ebx	%bx	%bl
%rcx	%ecx	%cx	%cl
%rdx	%edx	%dx	%dl
%rdi	%edi	%di	%dil
%rsi	%esi	%si	%sil
%rsp	%esp	%sp	%spl
%rbp	%ebp	%bp	%bpl
%r8	%r8d	%r8w	%r8b
%r9	%r9d	%r9w	%r9b
%r10	%r10d	%r10w	%r10b
%r11	%r11d	%r11w	%r11b
%r12	%r12d	%r12w	%r12b
%r13	%r13d	%r13w	%r13b
%r14	%r14d	%r14w	%r14b
%r15	%r15d	%r15w	%r15b

Table 1: Register mapping in x86-64 architecture

4.2.2 ARM Assembly Registers

4.3 Addressing Modes

Registers being 8 bytes mean that we can store memory addresses, and if we can store memory addresses, we can access memory, i.e. the values at those memory addresses. There are 4 ways to do this, called **addressing modes**: immediate, normal, displacement, and indexed. When we parse an instruction, its operands are either

1. Constant (literal) values
2. Registers
3. Memory forms

Definition 4.13 (Immediate Addressing)

Immediate addressing is when the operand is a constant value, used with a $ sign.

$$\texttt{\$val} \tag{5}$$

Definition 4.14 (Normal Addressing)

Normal addressing is when the operand is a register, used with a % sign and the following syntax. The parentheses are used to dereference the memory address like dereferencing a pointer in C.

$$\texttt{(R) = Mem[Reg[R]]} \tag{6}$$

where `R` is the register name, `Reg[R]` is the value in the register, and `Mem[Reg[R]]` is the value in the memory address pointed to by the register.

Definition 4.15 (Displacement Addressing)

When we have a memory address stored in a register, we can add an offset to it to access a different memory address.

$$\texttt{D(R) = Mem[Reg[R] + D]} \tag{7}$$

where `R` is the register name and `D` is a constant displacement that specifies offset.

Definition 4.16 (Indexed Addressing)

Indexed addressing gives us more flexibility, allowing us to multiply the value in the register by a constant and add it to the value in another register. The general formula is shown as the top, but there are special cases:

$$\begin{aligned}
\texttt{D(Rb, Ri, S)} &= \texttt{Mem[Reg[Rb] + S*Reg[Ri] + D]} \\
\texttt{D(Rb, Ri)} &= \texttt{Mem[Reg[Rb] + Reg[Ri] + D]} \\
\texttt{(Rb, Ri, S)} &= \texttt{Mem[Reg[Rb] + S*Reg[Ri]]} \\
\texttt{(Rb, Ri)} &= \texttt{Mem[Reg[Rb] + Reg[Ri]]} \\
\texttt{(, Ri, S)} &= \texttt{Mem[S*Reg[Ri]]}
\end{aligned}$$

where `D` is a constant displacement of 1, 2, or 4 bytes, `Rb` is the base register (can be any of 8 integer registers), `Ri` is the index register (can be any register except `rsp`), and `S` is the scale factor (1, 2, 4, or 8).

4.3.1 x86 Assembly Addressing Modes

Example 4.6 (Immediate Addressing)

```
movq $0x4, %rax
```

Example 4.7 (Normal Addressing)

The following example shows the source operand being a memory address, with normal addressing, and the destination operand being a register.

```
movq (%rax), %rbx
```

Example 4.8 (Displacement Addressing)

The following example shows the source operand being a memory address and the destination operand being a register. They are both addressed normally.

```
movq 8(%rdi), %rdx
```

Example 4.9 (Indexed Addressing)

The following shows the source operand being a memory address and the destination operand being a register. Say that `%rdx = 0xf000` and `%rcx = 0x0100`. Then

$$\texttt{0x80(,\%rdx,2) = Mem[2*0xF000 + 0x80] = Mem[0x1E080]} \quad (8)$$

We see that

```
movq 0x100(%rdi, %rsi, 8), %rdx
```

4.3.2 ARM Assembly Addressing Modes

4.4 Instructions

Now that we've gotten a sense of what these registers are and some commonalities between them, let's do some operations on them with instructions.

Definition 4.17 (Instruction)

An instruction is a single line of assembly code. It consists of some instruction followed by its (one or more) operands. The instruction is a mnemonic for a machine language operation (e.g. `mov`, `add`, `sub`, `jmp`, etc.). The **size specifier** can be appended to this instruction mnemonic to specify the size of the operands.

1. **b** (byte) for 1 byte
2. **w** (word) for 2 bytes
3. **l** (long) for 4 bytes
4. **q** (quad word) for 8 bytes

Note that due to backwards compatibility, word means 2 bytes in instruction names. Furthermore, the maximum size is 8 bytes since that is the size of each register in x86_64. An operand can be of 3 types, determined by their **mode of access**:

1. **Immediate addressing** is denoted with a `$` sign, e.g. a constant integer data `$1`.
2. **Register addressing** is denoted with a `%` sign with the following register name, e.g. `%rax`.
3. **Memory addressing** is denoted with the hexadecimal address in memory, e.g. `0x034AB`.

Like higher level programming languages, we can perform operations, do comparisons, and jump to different parts of the code. Instructions can be generally categorized into three types:

1. **Data Movement**: These instructions move data between memory and registers or between the registery and registery. Memory to memory transfer cannot be done with a single instruction.

```
%reg = Mem[address]     # load data from memory into register
Mem[address] = %reg     # store register data into memory
```

2. **Arithmetic Operation**: Perform arithmetic operation on register or memory data.

```
%reg = %reg + Mem[address]     # add memory data to register
%reg = %reg - Mem[address]     # subtract memory data from register
%reg = %reg * Mem[address]     # multiply memory data to register
%reg = %reg / Mem[address]     # divide memory data from register
```

3. **Control Flow**: What instruction to execute next both unconditional and conditional (if statements) ones. With if statements, loops can then be defined.

```
jmp label     # jump to label
je label      # jump to label if equal
jne label     # jump to label if not equal
jg label      # jump to label if greater
jl label      # jump to label if less
call label    # call a function
ret           # return from a function
```

Now unlike compiled languages, which are translated into machine code by a compiler, assembly code is translated into machine code through a two-step process. First, we **assemble** the assembly code into an **object file** by an **assembler**, and then we **link** the object file into an executable by a **linker**. Some common assemblers are **NASM** (Netwide Assembler) and **GAS/AS** (GNU Assembler), and common linkers are **ld** (GNU Linker) and **lld** (LLVM Linker), both installable with **sudo pacman -S nasm ld**.

4.4.1 Moving and Arithmetic

Again, it is more important to have a general feel of what instructions every assembly language should and get the ideas down rather than the syntax. We list them here, beginning with simply moving.

Definition 4.18 (Moving)

Next we want to have some sort of arithmetic to do calculations and to compare values.

Definition 4.19 (Arithmetic Operations)

4.4.2 Conditionals

Definition 4.20 (Conditionals)

4.4.3 Control Transfer on Stack

These are really the three basic functions needed to do anything in assembly, but let's talk about an important implementation called the **control transfer**. Say that you want to compute a function.

1. Then we must retrieve the data from the memory.

2. We must load it into our registers in the CPU and perform some computation.

3. Then we must store the data back into memory.

Let's begin with a refresher on how the call stack is managed. Recall that `%rsp` is the stack pointer and always points to the top of the stack. The register `%rbp` represents the base pointer (also known as the frame pointer) and points to the base of the current stack frame. The stack frame (also known as the activation frame or the activation record) refers to the portion of the stack allocated to a single function call. The currently executing function is always at the top of the stack, and its stack frame is referred to as the active frame. The active frame is bounded by the stack pointer (at the top of stack) and the frame pointer (at the bottom of the frame). The activation record typically holds local variables for a function.

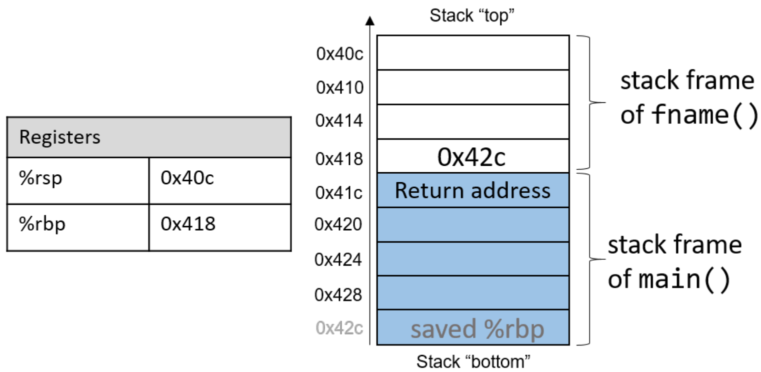


Figure 20: The current active frame belongs to the callee function (fname). The memory between the stack pointer and the frame pointer is used for local variables. The stack pointer moves as local values are pushed and popped from the stack. In contrast, the frame pointer remains relatively constant, pointing to the beginning (the bottom) of the current stack frame. As a result, compilers like GCC commonly reference values on the stack relative to the frame pointer. In Figure 1, the active frame is bounded below by the base pointer of fname, which is stack address 0x418. The value stored at address 0x418 is the "saved" `%rbp` value (0x42c), which itself is an address that indicates the bottom of the activation frame for the main function. The top of the activation frame of main is bounded by the return address, which indicates where in the main function program execution resumes once the callee function fname finishes executing.

Once we have done this we are really done. Formally, this is called Turing complete (?).

Definition 4.21 (Control Transfers)

We list some.

1. Push
2. Pop
3. Call to call a function
4. Return to return from a function
5. Continue
6. Get out of stack with leave.

Example 4.10 (Control Transfer Example)

We show this with a minimal example with psuedocode.

4.4.4 Multiple Functions

Now what happens if there are multiple functions calling each other? Take a look at the following example with two functions.

Example 4.11 (Multiple Functions Example)

There is a bit of a concern here from the previous example. The main function had two functions that returned two values. As the subfunction stack frame is removed from the stack, the return value is stored in the `%rax` register. If another function is called right after, then the return value of the second function will overwrite that of the previous one. This was not a problem in the previous example since the return value of the `assign` function was not used. However, if it was, then the return value of the `adder` function would have overwritten it. This is known as register saving.

1. For **caller-saved registers**, the caller function is responsible for saving the value of the register before calling a function and restoring it after the function returns. The caller should save values in its stack frame before calling the callee function, e.g. by pushing all the return values of each callee in the caller stack frame. Then it will restore values after the call.

   Therefore, if we have a set of registers {`%reg`}, the caller must take everything and push them in the caller stack frame. Then it will restore them after the call.

2. For **callee-saved registers**, it is the callee's responsibility to save any data in these registers before using the registers.

   Therefore, if we have a set of registers {`%reg`}, then inside the callee stack frame, the callee must take everything and push them in the callee stack frame. Once it computes the final return value, then it will restore all the saved register values from the callee stack frame back into the registers for the caller to use.

Ideally, we want *one* calling convention to simply separate implementation details between caller and callee. In general, however, neither is best. If the caller isn't using a register, then caller-save is better, and if callee doesn't need a register, then callee-save is better. If we do need to save, then callee save generally makes smaller programs, so we compromise and use a combination of both caller-save and callee-save.

4.4.5 x86-64 Instructions

Let's talk about moving instructions first.

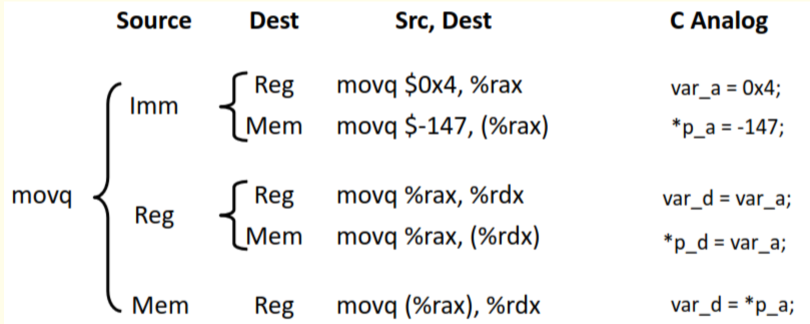
Definition 4.22 (mov)

Let's talk about the `mov` instruction which copies data from the source to the destination (the data in the source still remains!) and has the syntax

```
mov_ src, dest
```
(9)

1. The source can be a register (`%rsi`), a value (`$0x4`), or a memory address (`0x4`).
2. The destination can be a register or a memory address.
3. The `_` is defined to be one of the size operands, which determine how big the data is. For example, we can call `movq` to move 8 bytes of data (which turns about to be the maximum size of a register).

A good diagram to see is the following:



Even with just the mov instruction, we can look at a practical implementation of a C program in Assembly.

Example 4.12 (Swap Function)

Let us take a look at a function that swaps two integers. Let's see what they do.

1. In C, we dereference both `xp` and `yp` (note that they are pointers to longs, so they store 8 bytes), and assign these two values to two temporary variables. Then, we assign the value of `yp` to `xp` and the value of `xp` to `yp`.
2. In Assembly, we first take the registers `%rdi` and `%rsi`, which are the 1st and 2nd arguments of the function, dereference them with the parantheses, and store them in the temporary registers `%rax` and `%rdx`. Then, we store the value of `%rdx` into the memory address of `%rdi` and the value of `%rax` into the memory address of `%rsi`. Note that the input values (the actual of)

```c
void swap(long *xp, long *yp) {
  long t0 = *xp;
  long t1 = *yp;
  *xp = t1;
  *yp = t0;
}
```

```
swap:
  movq (%rdi), %rax
  movq (%rsi), %rdx
  movq %rdx, (%rdi)
  movq %rax, (%rsi)
  ret
```

Definition 4.23 (movz and movs)

The `movz` and `movs` instructions are used to move data from the source to the destination, but with zero and sign extension, respectively. It is used to copy from a smaller source value to a larger destination, with the syntax

```
movz__ src, dest
movs__ src, dest
```

where the first _ is the size of the source and the second _ is the size of the destination.

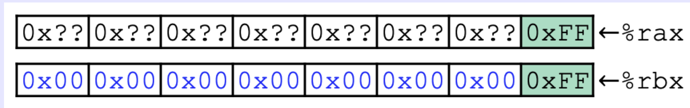
1. The source can be from a memory or register.
2. The destination must be a register.

Example 4.13 (Simple example with movz)

Take a look at the code below.

```
movzbq %al, %rbx
```

The `%al` represents the last byte of the `%rax` register. It is 1 byte long. The `%rbx` register is 8 bytes long, so we can fill in the rest of the 7 bytes with zeros.

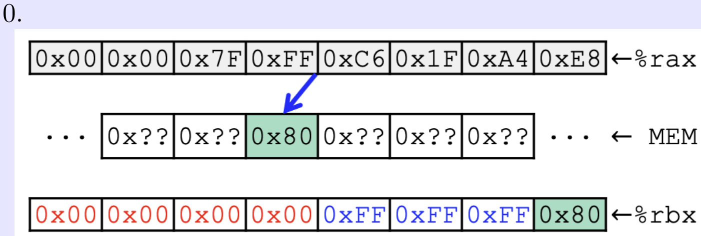


Example 4.14 (Harder example with movs)

Take a look at the code below.

```
movsbl (%rax), %ebx
```

You want to move the value at the memory address in `%rax` into `%ebx`. Since the source size is set to 1 byte, you take that byte, say it is `0x80`, from the memory, and then sign extend it (by a size of 4 bytes!) into `%ebx`. Note that therefore, the first four bytes of `%rbx` will not be affected since it's not a part of `%ebx`. An exception to this is that in x86-64, any instruction that generates a 32-bit long word value for a register also sets the high-order 32 bits of the register to 0, so this ends up clearing the first 4 bytes to 0.



Now we can talk about control transfer. Say that you have the following C and Assembly code.

```c
int add(int x) {
  return x + 2;
}

int main() {
  int a = 2;
  int b = add(a);
  return 0;
}
```

```
add:
  movq %rdi, %rax
  addq $2, %rax
  ret
main:
  movq $3, $rdi
  call add
  movq $0, %rax
  ret
```

Figure 21: A simple function.

If you go through the instructions, you see that in main, you first move `$3` into the `%rdi` register. Then, you call the `add` function, and within it you also have the `%rdi` register. This is a conflict in the register, and we don't want to simply overwrite the value of `%rdi` in the `main` function. Simply putting it to another register isn't a great idea since we can't always guarantee that it will be free. Therefore, we must use the memory itself.

Recall the stack, which we can think of as a giant array in which data gets pushed and popped in a last-in-first-out manner. The stack is used to store data and return addresses, and is used to manage function calls. Visually, we want to think of the elements getting pushed in from the bottom (upside down) towards lower memory addresses.

Definition 4.24 (Stack Pointer)

Note that every time we want to push or pop something from the stack, we must know *where* to push or pop it. This is where the **stack pointer** comes in. It is a special register that always points to the top of the stack, and is used to keep track of the stack.

Definition 4.25 (Push and Pop)

The `push` and `pop` instructions are used to push and pop data onto and off the stack, respectively.

```
push_ src        rsp = rsp - 8; Mem[rsp] = src
pop_ dest        dest = Mem[rsp]; rsp = rsp + 8
```

1. When we push the source, we fetch the value at the source and store it at the memory address pointed to by the stack pointer `%rsp`. Then, we decrement `%rsp` by 8.
2. When we pop from the stack, we fetch the value at the memory address pointed to by the stack pointer `%rsp` and store it in the destination. Then, we increment `%rsp` by 8.

Note that no matter what the size of the operand, we always subtract 8 from the stack pointer. This is because the stack grows downwards, and we want to make sure that the next element is pushed into the next available space.

Note that the register `%rsp` is the stack pointer, which points to the top of the stack. The stack is used to store data and return addresses, and is used to manage function calls.

Definition 4.26 (Push and Pop)

The `push` and `pop` instructions are used to push and pop data onto and off the stack, respectively.

```
push_ src        rsp = rsp - 8; Mem[rsp] = src
pop_ dest        dest = Mem[rsp]; rsp = rsp + 8
```

The `_` is a size operand, which determines how big the data is.

Definition 4.27 (Call and Ret)

The `call` instruction pushes the return address onto the stack and jumps to the function. The `ret` instruction pops the return address from the stack and jumps to it.

We also talked about how there is instruction code that is even below the stack that is stored. This is where all the machine code/assembly is stored, and we want to find out where we are currently at in this code. This is done with the program counter.

Definition 4.28 (Program Counter, Instruction Pointer)

The **program counter**, or **instruction pointer**, is a special register **rip** that points to the current instruction in the program. It is used to keep track of the next instruction to be executed.

Let's go through one long example to see in detail how this is calculated.

Example 4.15 (Evaluating a Function)

Say that we have the following C code.

```c
int adder2(int a) {
  return a + 2;
}

int main() {
  int x = 40;
  x = adder2(x);
  printf("x is: %d\n", x);
  return 0;
}
```

When we compile this program, we can view its full assembly code by calling `objdump -d a.out`. The output is quite long, so we will focus on the instruction for the `adder2` function.

```
0000000000400526 <adder2>:
400526:       55                      push   %rbp
400527:       48 89 e5                mov    %rsp,%rbp
40052a:       89 7d fc                mov    %edi,-0x4(%rbp)
40052d:       8b 45 fc                mov    -0x4(%rbp),%eax
400530:       83 c0 02                add    $0x2,%eax
400533:       5d                      pop    %rbp
400534:       c3                      retq
```

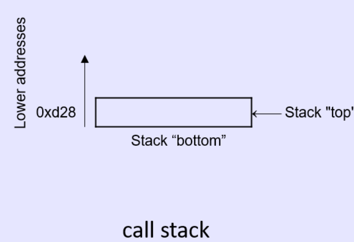
Figure 22: The output of objdump for the `adder2` function. The leftmost column represents the addresses (in hex) of where the actual instructions lie. The second column represents the machine code that is being executed. The third column represents the assembly code.

Note some things. Since `adder2` is taking in an integer input value, we want to load it into the lower 32 bits (4 bytes) of the `%rdi` register, which is the first parameter. So we use `%edi`. Likewise for the return value, we want to output an int so we use `%eax` rather than `%rax`. Let's go through some of the steps.

1. By the time we get into calling `adder2`, we can take a look at the relevant registers.

```
0x526  push  %rbp
0x527  mov   %rsp, %rbp
0x52a  mov   %edi, -0x4(%rbp)
0x52d  mov   -0x4(%rbp), %eax
0x530  add   $0x2, %eax
0x533  pop   %rbp
0x534  retq
```

Registers	
%eax	0x123
%edi	0x28
%rsp	0xd28
%rbp	0xd40
%rip	0x526



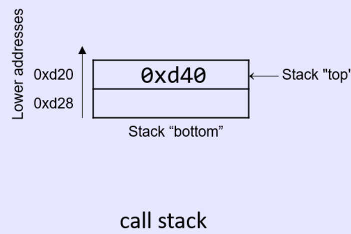
   (a) First, the `%eax` is filled with garbage, which are leftovers from previous programs that haven't been overwritten yet.
   (b) Second, the `%edi=0x28` since we have set `x=40` in `main`, before calling `adder2`, so it lingers on.
   (c) `%rsp=0xd28` since that is where the top of the stack is.
   (d) `%rbp=0xd40`
   (e) `%rip=0x526` since that is where we are currently at in our instruction (we are about to do

it, but haven't done it yet).

2. When we execute the first line of code, we simply push the value at `%rbp` into the stack. The top of the stack gets decremeneted by 8 and the value at `%rbp` is stored there. This means that the top of the stack is at `%rsp=0xd20` and the next instruction will be at `%rip=0x527`.

```
➡ 0x526  push  %rbp
  0x527  mov   %rsp, %rbp
  0x52a  mov   %edi, -0x4(%rbp)
  0x52d  mov   -0x4(%rbp), %eax
  0x530  add   $0x2, %eax
  0x533  pop   %rbp
  0x534  retq
```

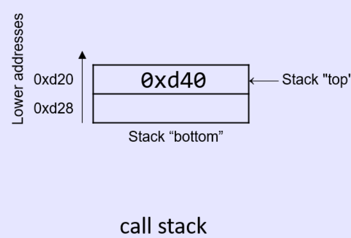
Registers	
%eax	0x123
%edi	0x28
%rsp	**0xd20**
%rbp	0xd40
%rip	**0x527**



3. The reason we have pushed `%rbp` onto the stack is that we want to save it before it gets overwritten by this next execution. We basically move the value of `%rsp` into `%rbp`, and the `%rip` advances to the next instruction. `%rip` moves to the next instruction.

```
  0x526  push  %rbp
➡ 0x527  mov   %rsp, %rbp
  0x52a  mov   %edi, -0x4(%rbp)
  0x52d  mov   -0x4(%rbp), %eax
  0x530  add   $0x2, %eax
  0x533  pop   %rbp
  0x534  retq
```

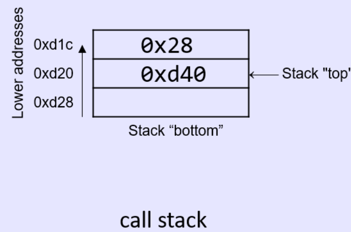
Registers	
%eax	0x123
%edi	0x28
%rsp	0xd20
%rbp	**0xd20**
%rip	**0x52a**



4. Now we want to take our first argument `%edi` and store it in memory. Note that since this is 4 bytes, we can move this value into memory that is 4 bytes below the stack (`-0x4(%rbp)`). Note that the storing the value of `%edi` into memory doesn't affect the stack pointer `%rsp`. As far as the program is concerned, the top of this stack is still address `0xd20`.

```
    0x526  push  %rbp
    0x527  mov   %rsp, %rbp
➡  0x52a  mov   %edi, -0x4(%rbp)
    0x52d  mov   -0x4(%rbp), %eax
    0x530  add   $0x2, %eax
    0x533  pop   %rbp
    0x534  retq
```

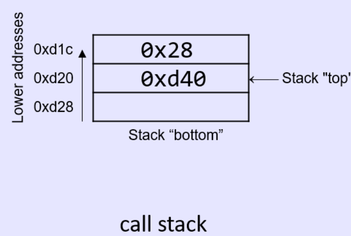
Registers	
%eax	0x123
%edi	0x28
%rsp	0xd20
%rbp	0xd20
%rip	**0x52d**



5. The next instruction simply goes into memory 4 bytes below the stack pointer, takes the value there, and stores it into `%eax`. This is the value of `%edi` that we just stored. This may seem redundant since we are making a round trip to memory and back to ultimately move the value of `%edi` into `%eax`, but compilers are not smart and just follow these instructions.

```
    0x526  push  %rbp
    0x527  mov   %rsp, %rbp
    0x52a  mov   %edi, -0x4(%rbp)
➡  0x52d  mov   -0x4(%rbp), %eax
    0x530  add   $0x2, %eax
    0x533  pop   %rbp
    0x534  retq
```

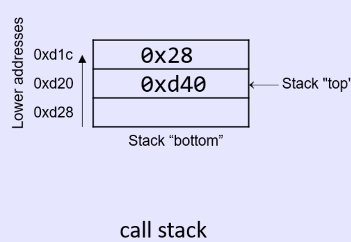
Registers	
%eax	**0x28**
%edi	0x28
%rsp	0xd20
%rbp	0xd20
%rip	**0x530**



6. Finally, we add the value `$0x2` to `%eax` and store it back into `%eax`.

```
    0x526  push  %rbp
    0x527  mov   %rsp, %rbp
    0x52a  mov   %edi, -0x4(%rbp)
    0x52d  mov   -0x4(%rbp), %eax
➡  0x530  add   $0x2, %eax
    0x533  pop   %rbp
    0x534  retq
```

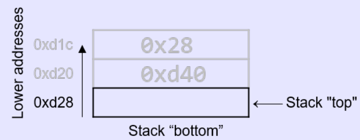
Registers	
%eax	**0x2A**
%edi	0x28
%rsp	0xd20
%rbp	0xd20
%rip	**0x533**



7. Finally, we pop the value at the top of the stack and store it into `%rbp`. Note that this is *not* the value `0x28`. It is simply the value that is stored at `%rsp=0xd20`, which is `(%rsp)=0xd40`.

```
0x526  push  %rbp
0x527  mov   %rsp, %rbp
0x52a  mov   %edi, -0x4(%rbp)
0x52d  mov   -0x4(%rbp), %eax
0x530  add   $0x2, %eax
0x533  pop   %rbp
0x534  retq
```

Registers	
%eax	0x2A
%edi	0x28
%rsp	0xd28
%rbp	0xd40
%rip	0x534



call stack

8. Finally, we return the value with `retq`.

Note that the final values in the registers `%rsp` and `%rip` are `0xd28` and `0x534`, respectively, which are the same values as when the function started executing! This is normal and expected behavior with the call stack, which just stores temporary variable sand data of each function as it executes a program. Once a function completes executing, the stack returns to the state it was in prior to the function call. Therefore, it is common to see the following two instructions at the beginning of a function:

```
push %rbp
mov %rsp, %rbp
```

and the following two at the end of a function

```
pop %rbp
retq
```

Now arithemtic operations are quite simple.

Definition 4.29 (Add, Subtract, Multiply)

The **add** and **sub** instructions are used to add and subtract data from the destination.

```
add_ src, dest          dest = dest + src
sub_ src, dest          dest = dest - src
```

The **imul** instruction is used to multiply data between the source and destination and store it in the destination.

```
imul_ src, dest         dest = dest * src
```

Again the _ is a size operand, which determines how big the data is.

Definition 4.30 (Increment, Decrement)

The **inc** and **dec** instructions are used to increment and decrement the value in the destination.

```
inc_ dest               dest = dest + 1
dec_ dest               dest = dest - 1
```

Definition 4.31 (Negative)

The **neg** instruction is used to negate the value in the destination.

`neg_ dest`	`dest = -dest`

Example 4.16 (Basic Arithmetic Function)

The following represents the same program in C and in assembly. Let's go through each one:

1. In C, we first initialize `a = 4`, then `b = 8`, add them together to get `c`, and then return `c`.
2. In Assembly, we move the value 4 to the `%rax` register, then move the value 8 to the `%rbx` register, add the two values together to store it into `%rax`, and then return the value in the `%rax` register.

```c
int main() {
  int a = 4, b = 8;
  int c = a + b;
  return c;
}
```

```
main:
  movq $4, %rax
  movq $8, %rbx
  addq %rbx, %rax
  ret
```

It is slightly different in Assembly since rather than storing 4 in some intermediate register, we immediately store it in the return register. In a way it is more optimized, and this is what the compiler does for you so that as few registers are used.

A shorthand way to do this is with `lea`, which stands for load effective address.

Definition 4.32 (Load Effective Address)

The **lea** instruction is used to load the effective address of the source into the destination. For now, we will focus on the arithmetic operations that it can do

`lea_ (src1, src2), dest`	`dest = src1 + src2`
`lea_ (src1, src2, scale), dest`	`dest = src1 + src2*scale`
`lea_ const(src1, src2), dest`	`dest = src1 + src2 + const`
`lea_ const(src1, src2, scale), dest`	`dest = src1 + src2*scale + const`

This is useful for doing arithmetic operations on the address of a variable.

Definition 4.33 (Bitwise)

The **and**, **or**, **xor**, and **not** instructions are used to perform bitwise operations on the source and destination.

`and src, dest`	`dest = dest & src`
`or src, dest`	`dest = dest | src`
`xor src, dest`	`dest = dest ŝrc`
`neg dest`	`dest = -dest`
`not dest`	`dest = ∼dest`

Definition 4.34 (Arithmetic and Logical Bit Shift)

The `sal` arithmetic instruction is used to shift the bits of the destination to the left by the number of bits specified in the source. The `shr` instruction is used to shift the bits of the destination to the right by the number of bits specified in the source.

```
sal src, dest        dest = dest « src
shr src, dest        dest = dest » src
```

The `sar` instruction is used to shift the bits of the destination to the right by the number of bits specified in the source, and fill the leftmost bits with the sign bit. The `shl` instruction is used to shift the bits of the destination to the left by the number of bits specified in the source, and fill the rightmost bits with zeros.

```
sar src, dest        dest = dest » src
shl src, dest        dest = dest « src
```

Example 4.17 (Harder Arithmetic Example)

The following two codes are equivalent.

```c
long arith(long x, long y, long z) {
  long t1 = x + y;
  long t2 = z + t1;
  long t3 = x + 4;
  long t4 = y * 48;
  long t5 = t3 + t4;
  long rval = t2 * t5;
  return rval;
}
.
.
.
.
.
```

```
arith:
  # rax/t1 = x + y
  leaq  (%rdi, %rsi), %rax
  # rax/t2 = z + t1
  addq  %rdx, %rax
  #rdx = 3 * y
  leaq  (%rsi, %rsi, 2), %rdx
  #rdx/t4 = (3*y) * 16
  salq  $4, %rdx
  #rcx/t5 = x + t4 + 4
  leaq  4(%rdi, %rdi), %rcx
  # rax/rval = t5 * t2
  imulq %rcx, %rax
  ret
```

The final thing in our list is condition codes.

Sometimes, we want to move (really copy) some value to another register if some condition is met. This is where we use conditional moves. These conditions are met by the flags register, which is a special register that stores the status of the last operation. It is the value of these flags that determine whether all future conditional statements are met in assembly.

Definition 4.35 (Condition Code Flags)

The flags register in the x86 CPU keeps 4 *condition code* flag bits internally. Think of these as status flags that are *implicitly* set by the most recent arithmetic operation (think of it as side effects). Note that condition codes are NOT set by `lea` or `mov` instructions!

1. **Zero Flag**: if the last operation resulted in a zero value.
2. **Sign Flag**: if the last operation resulted in a negative value (i.e. the most significant bit is 1).
3. **Overflow Flag**: if the last operation resulted in a signed overflow.
4. **Carry Flag**: if the last operation resulted in a carry out of the most significant bit, i.e. an unsigned overflow.

Every operation may or may not changes these flags to test for zero or nonzero, positive or negative, or overflow conditions, and combinations of these flags express the full range of conditions and cases, e.g. for signed and unsigned values.

Example 4.18 (Zero Flag)

If the code below was just run, then ZF would be set to 1.

```
movq $2, %rax
subq $2, %rax
```

Example 4.19 (Sign Flag)

If the code below was just run, then SF would be set to 1.

```
movq $2, %rax
subq $4, %rax
```

Example 4.20 (Overflow Flag)

If either code below was just run, then OF would be set to 1.

```
movq $0x7fffffffffffffff, %rax
addq $1, %rax
```

```
movq 0x8000000000000000, %rax
addq 0xffffffffffffffff, %rax
```

This is because in the left in signed arithmetic, we have a positive + positive = negative (result is `0x8000000000000000`), which is a signed overflow. Furthermore, in the right we have negative + negative = positive (result is `0x7fffffffffffffff`).

Example 4.21 (Carry Flag)

If the code below was just run, then CF would be set to 1.

```
movq $0xffffffffffffffff, %rax
addq $1, %rax
```

This is because the result is $0x0$, which is a carry out of the most significant bit and an unsigned overflow.

It would be tedious to always set these flags manually, so there are two methods that can be used to *explicitly* set these flags.

Definition 4.36 (Compare)

The **cmp** instruction is used to perform a subtraction between the source and destination, and set the flags accordingly, but it does not store the result.

`cmp_ src, dest` `dest - src`

The following flags are set if the conditions are met:
1. **ZF = 1** if `dest == src`
2. **SF = 1** if `dest < src` (MSB is 1)
3. **OF = 1** if signed overflow

4. **CF = 1** if unsigned overflow

Definition 4.37 (Test)

The **test** instruction is used to perform a bitwise AND operation between the source and destination, and set the flags accordingly.

```
test_ src, dest                    dest & src
```

The following flags are set if the conditions are met. Note that you can't have carry out (CF) or overflow (OF) if these flags are set.

1. **ZF = 1** if `dest & src == 0`
2. **SF = 1** if `dest & src < 0` (MSB is 1)

Example 4.22 (Compare)

Assuming that `%al = 0x80` and `%bl = 0x81`, which flags are set when we execute `cmpb %al, %bl`? Well we must first compute

$$\texttt{\%bl - \%al = 0x81 - 0x80 = 0x81 + } \sim \texttt{ 0x80 + 1 = 0x81 + 0x7F + 1 = 0x101 = 0x01} \quad (10)$$

1. CF=1 since the result is greater than 0xFF (i.e. larger than byte)
2. ZF=0 since the result is not 0
3. SF=0 since the MSB is 0, i.e. there is unsigned overflow
4. OF=0 since there is no signed overflow

For conditional moves and jumps later shown, it basically uses these explicit sets and always compares them to 0. We will see what this means later.

Finally, we can actually set a byte in a register to 1 or 0 based on the value of a flag.

Definition 4.38 (Set)

We can then talk about conditional moves and jumps.

Definition 4.39 (Equality with 0)

The `test` instruction is used to perform a bitwise AND operation between the source and destination, and set the flags accordingly.

```
test_ src, dest                    dest & src
```

The `sete` instruction is used to set the destination to 1 if the zero flag is set, and 0 otherwise.

```
sete_ dest                 dest = (ZF == 1) ?  1 :  0
```

The `cmovne` instruction is used to move the source to the destination if the zero flag is not set.

```
cmovne_ src, dest           dest = (ZF == 0) ?  src :  dest
```

Definition 4.40 (Jump)

There are several jump instructions, but essentially they are used to jump to another part of the code. We can use the following mnemonic to jump to a label.

Letter	Word
j	jump
n	not
e	equal
s	signed
g	greater (signed interpretation)
l	less (signed interpretation)
a	above (unsigned interpretation)
b	below (unsigned interpretation)

Table 2: Letter to Word Mapping

Figure 23: Mnemonic for Jump Instructions

For completeness, we include all the jump instructions.

Signed Comparison	Unsigned Comparison	Description
je (jz)		jump if equal (==) or jump if zero
jne (jnz)		jump if not equal (!=)
js		jump if negative
jns		jump if non-negative
jg (jnle)	ja (jnbe)	jump if greater (>)
jge (jnl)	jae (jnb)	jump if greater than or equal (>=)
jl (jnge)	jb (jnae)	jump if less (<)
jle (jng)	jbe (jna)	jump if less than or equal (<=)

Table 3: Comparison Instructions in Assembly

Figure 24: All jump instructions

Definition 4.41 (int)

The `int` instruction is used to generate a software interrupt. It is often used to invoke a system call.

Definition 4.42 (ret)

The `ret` instruction is used to return from a function. It returns the value in the `%rax` register.

Now we can have a basic idea of how if statements can be used as a sequence of conditionals and jump operators. Let's first look at the **goto** version of C.

Definition 4.43 (Goto Syntax)

The goto version processes instructions sequentially as long as there is no jump. This is useful because compilers translating code into assembly designate a jump when a condition is true. Contrast this behavior with the structure of an if statement, where a "jump" (to the else) occurs when conditions are not true. The goto form captures this difference in logic.

```c
int getSmallest(int x, int y) {
  int smallest;
  if ( x > y ) { //if (conditional)
    smallest = y; //then statement
  }
  else {
    smallest = x; //else statement
  }
  return smallest;
}
.
.
.
.
.
```

```c
int getSmallest(int x, int y) {
  int smallest;

  if (x <= y ) { //if (!conditional)
    goto else_statement;
  }
  smallest = y; //then statement
  goto done;

else_statement:
  smallest = x; //else statement

done:
  return smallest;
}
```

Figure 25: C vs GoTo code of the same function. While GoTo code allows us to view C more like assmebly, it is generally not readable and is not considered best practice.

Now let's see how if statements are implemented by taking a look at this function straight up in assembly.

```c
int getSmallest(int x, int y) {
  int smallest;
  if ( x > y ) { //if (conditional)
    smallest = y; //then statement
  }
  else {
    smallest = x; //else statement
  }
  return smallest;
}
.
```

```
Dump of assembler code for function getSmallest:
0x40059a <+4>:   mov    %edi,-0x14(%rbp)
0x40059d <+7>:   mov    %esi,-0x18(%rbp)
0x4005a0 <+10>:  mov    -0x14(%rbp),%eax
0x4005a3 <+13>:  cmp    -0x18(%rbp),%eax
0x4005a6 <+16>:  jle    0x4005b0 <getSmallest+26>
0x4005a8 <+18>:  mov    -0x18(%rbp),%eax
0x4005ae <+24>:  jmp    0x4005b9 <getSmallest+35>
0x4005b0 <+26>:  mov    -0x14(%rbp),%eax
0x4005b9 <+35>:  pop    %rbp
0x4005ba <+36>:  retq
```

Figure 26: Assembly code of a simple if statement

Again, note that since we are working with int types, the respective parameter registers are `%edi` and `%esi`, the respective lower 32-bits of the registers `%rdi` and `%rsi`. Let's walk through this again.

1. The first mov instruction copies the value located in register %edi (the first parameter, x) and places it at memory location %rbp-0x14 on the call stack. The instruction pointer (%rip) is set to the address of the next instruction, or 0x40059d.
2. The second mov instruction copies the value located in register %esi (the second parameter, y) and places it at memory location %rbp-0x18 on the call stack. The instruction pointer (%rip) updates to point to the address of the next instruction, or 0x4005a0.

3. The third mov instruction copies x to register %eax. Register %rip updates to point to the address of the next instruction in sequence.
4. The cmp instruction compares the value at location %rbp-0x18 (the second parameter, y) to x and sets appropriate condition code flag registers. Register %rip advances to the address of the next instruction, or 0x4005a6.
5. The jle instruction at address 0x4005a6 indicates that if x is less than or equal to y, the next instruction that should execute should be at location <getSmallest+26> and that %rip should be set to address 0x4005b0. Otherwise, %rip is set to the next instruction in sequence, or 0x4005a8.

With the `cmov` instruction, this can be a lot shorter. With the gcc compiler with level 1 optimizations turned on, we can see that a lot of redundancies are turned off.

```
<getSmallest>:
0x400546 <+0>: cmp    %esi,%edi      #compare x and y
0x400548 <+2>: mov    %esi,%eax      #copy y to %eax
0x40054a <+4>: cmovle %edi,%eax      #if (x<=y) copy x to %eax
0x40054d <+7>: retq                  #return %eax
```

Figure 27: Compiled with `gcc -01 -o getSmallest getSmallest.c`

Like if statements, loops in assembly can be implementing using jump functions that revisit some instruction address based on the result on an evaluated condition. Let's take a look at a basic loop function.

```
int sumUp(int n) {
  int total = 0;
  int i = 1;

  while (i <= n) {
    total += i;
    i++;
  }
  return total;
}
.
.
.
.
.
.
```

```
Dump of assembler code for function sumUp:
0x400526 <+0>:   push   %rbp
0x400527 <+1>:   mov    %rsp,%rbp
0x40052a <+4>:   mov    %edi,-0x14(%rbp)
0x40052d <+7>:   mov    $0x0,-0x8(%rbp)
0x400534 <+14>:  mov    $0x1,-0x4(%rbp)
0x40053b <+21>:  jmp    0x400547 <sumUp+33>
0x40053d <+23>:  mov    -0x4(%rbp),%eax
0x400540 <+26>:  add    %eax,-0x8(%rbp)
0x400543 <+29>:  add    $0x1,-0x4(%rbp)
0x400547 <+33>:  mov    -0x4(%rbp),%eax
0x40054a <+36>:  cmp    -0x14(%rbp),%eax
0x40054d <+39>:  jle    0x40053d <sumUp+23>
0x40054f <+41>:  mov    -0x8(%rbp),%eax
0x400552 <+44>:  pop    %rbp
0x400553 <+45>:  retq
```

Figure 28: Simple loop function in C and assembly.

Finally, we want to let the reader know the convention of calle and caller saved registers. The compiler tries to pick these registers, and by convention in x86, we have the following.

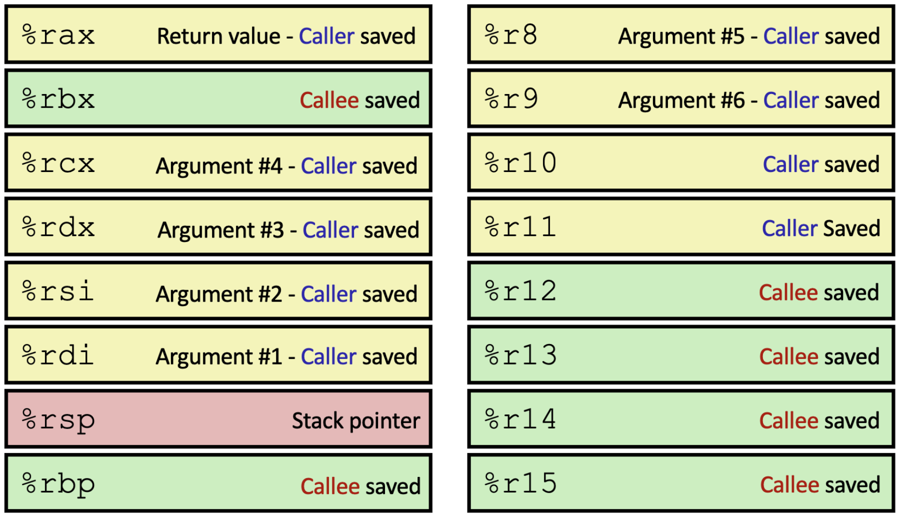


Figure 29: Caller save and callee save registers.

So far, we've traced through simple functions in assembly. In this section, we discuss the interaction between multiple functions in assembly in the context of a larger program. We also introduce some new instructions involved with function management.

Definition 4.44 (Leave)

The **leave** instruction is used to deallocate the current stack frame. For example, the leaveq instruction is a shorthand that the compiler uses to restore the stack and frame pointers as it prepares to leave a function. When the callee function finishes execution, leaveq ensures that the frame pointer is restored to its previous value. It is equivalent to the following two instructions:

```
leaveq              movq %rbp, %rsp
                         popq %rbp
```

Definition 4.45 (Call and Return)

The **call** instruction is used to call a function and the **ret** to return from a function. The callq and retq instructions play a prominent role in the process where one function calls another. Both instructions modify the instruction pointer (register %rip).

1. When the caller function executes the callq instruction, the current value of %rip is saved on the stack to represent the return address, or the program address at which the caller resumes executing once the callee function finishes. The callq instruction also replaces the value of %rip with the address of the callee function.

```
callq addr <fname>              push %rip
                           mov addr, %rip
```

2. The retq instruction restores the value of %rip to the value saved on the stack, ensuring that the program resumes execution at the program address specified in the caller function. Any value returned by the callee is stored in %rax or one of its component registers (e.g., %eax). The retq instruction is usually the last instruction that executes in any function.

```
retq                    pop %rip
```

Let's work through an example to solidify our knowledge.

Example 4.23 (Calling Functions in Assembly)

Let's take the following code and trace through main.

```c
#include <stdio.h>

int assign(void) {
    int y = 40;
    return y;
}

int adder(void) {
    int a;
    return a + 2;
}

int main(void) {
    int x;
    assign();
    x = adder();
    printf("x is:
    %d\n", x);
    return 0;
}
.
.
.
.
.
.
.
.
.
.
.
.
```

```
0000000000400526 <assign>:
  400526:       55                          push   %rbp
  400527:       48 89 e5                    mov    %rsp,%rbp
  40052a:       c7 45 fc 28 00 00 00        movl   $0x28,-0x4(%rbp)
  400531:       8b 45 fc                    mov    -0x4(%rbp),%eax
  400534:       5d                          pop    %rbp
  400535:       c3                          retq

0000000000400536 <adder>:
  400536:       55                          push   %rbp
  400537:       48 89 e5                    mov    %rsp,%rbp
  40053a:       8b 45 fc                    mov    -0x4(%rbp),%eax
  40053d:       83 c0 02                    add    $0x2,%eax
  400540:       5d                          pop    %rbp
  400541:       c3                          retq

0000000000400542 <main>:
  400542:       55                          push   %rbp
  400543:       48 89 e5                    mov    %rsp,%rbp
  400546:       48 83 ec 10                 sub    $0x10,%rsp
  40054a:       e8 e3 ff ff ff              callq  400526 <assign>
  40054f:       e8 d2 ff ff ff              callq  400536 <adder>
  400554:       89 45 fc                    mov    %eax,-0x4(%rbp)
  400557:       8b 45 fc                    mov    -0x4(%rbp),%eax
  40055a:       89 c6                       mov    %eax,%esi
  40055c:       bf 04 06 40 00              mov    $0x400604,%edi
  400561:       b8 00 00 00 00              mov    $0x0,%eax
  400566:       e8 95 fe ff ff              callq  400400
      <printf@plt>
  40056b:       b8 00 00 00 00              mov    $0x0,%eax
  400570:       c9                          leaveq
  400571:       c3                          retq
```

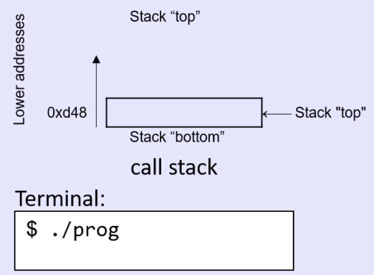
Figure 30: C code and its assembly equivalent. Main function calls two other functions.

Let's trace through what happens here in detail. This will be long.

1. `%rbp` is the base pointer that is initialized to something. Before we even begin main, say that we have the following initializations, where `%eax`, `%edi` is garbage. `%rsp` denotes where on the stack we are right before calling to main, **%rbp** is the base pointer to the current program, and **%rip** should be the address of the first instruction in main. Again since we work with integers we use the lower 32-bits of the registers. `%rip` now points to the next instruction.

```
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

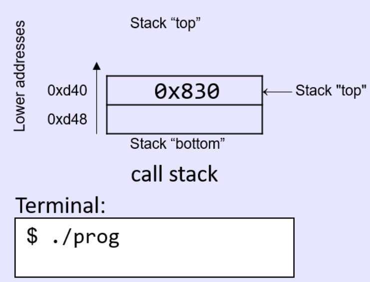
Registers	
%eax	650
%edi	1
%rsp	0xd48
%rbp	0x830
%rip	0x542



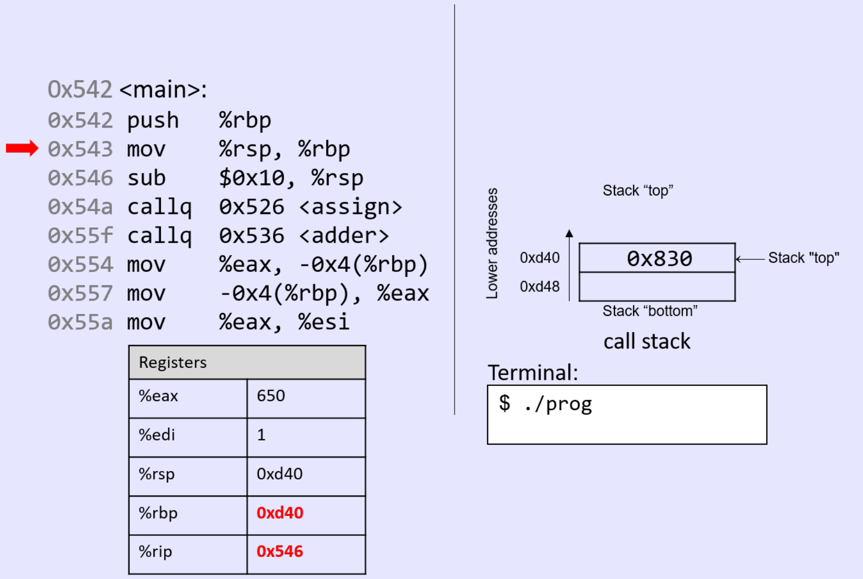
2. Now we start the main function. By calling main, the base pointer `%rbp` of the stack outside of the main frame will be overwritten by the base of the main stack frame, so we must save it for when main is done. Therefore, we push it onto the stack where `%rsp` is pointing. `%rip` now points to the next instruction.

```
0x542 <main>:
➡ 0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

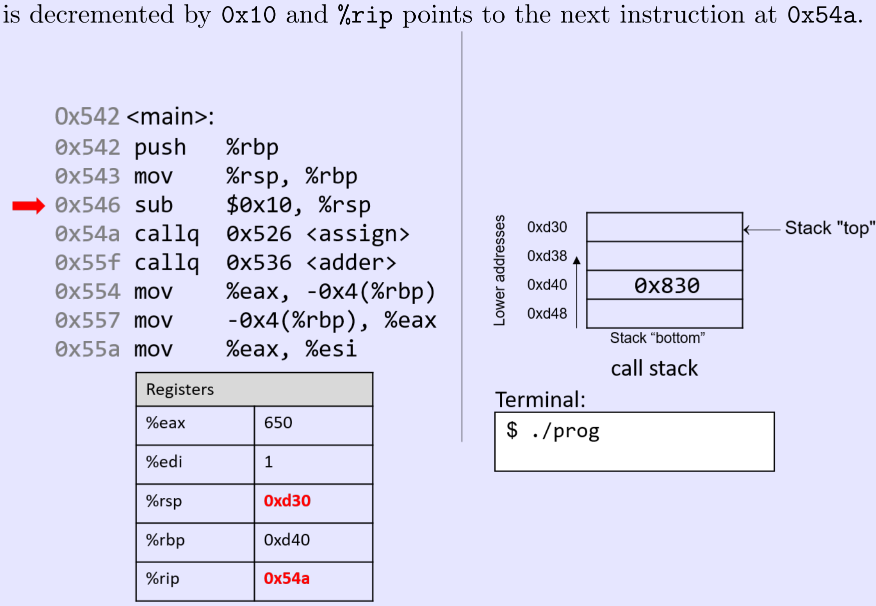
Registers	
%eax	650
%edi	1
%rsp	**0xd40**
%rbp	0x830
%rip	**0x543**



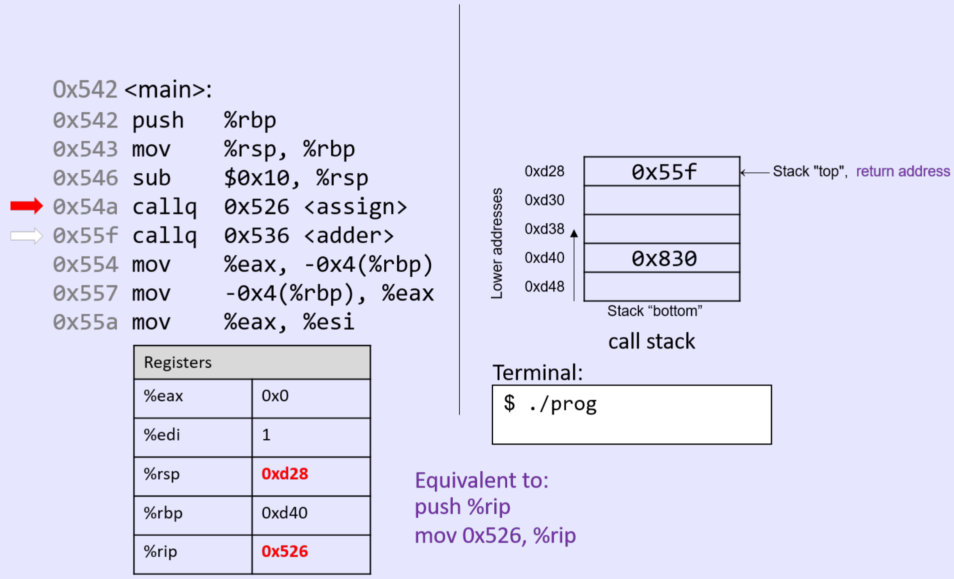
3. Then we actually change the location of the base pointer to the top of the stack, which now includes the first instruction in main.



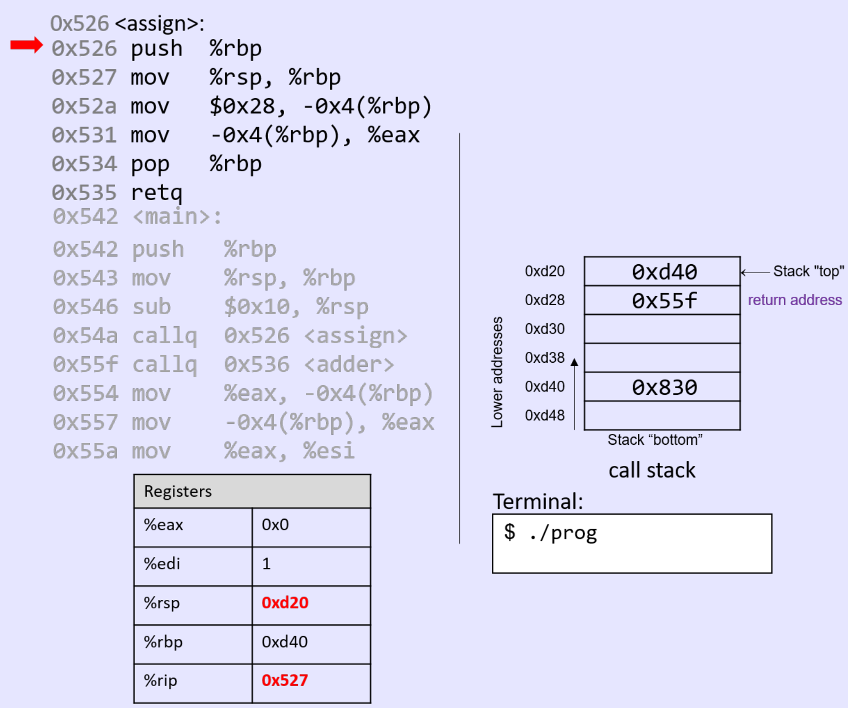
4. Now we manually change the stack pointer and have it grow by two bytes (`0x10`). Therefore, `%rsp` is decremented by `0x10` and `%rip` points to the next instruction at `0x54a`.



5. Now the next instruction pointed at by `%rip` is the `callq` instruction, which tells to go to the address of the `assign` function. We by default first update `%rip` to point to the next instruction at `0x55f`. However, this should not be the actual next instruction that we execute since we are calling another function. Rather, we want to update `%rip` to address `0x526` where `assign` is located at, but after completion we also want to know that we want to execute the instruction after it at address `0x55f`. Therefore, we should *save* address `0x55f` onto the stack and then update `%rip` to point to `0x526`. This is what we refer to as a **return address**.



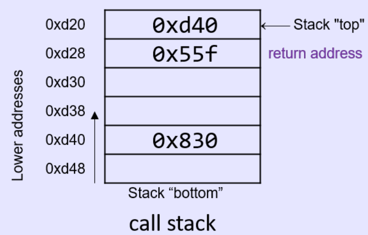
6. `%rip` is incremented to the next address. We step into the `assign` function, which is now a new stack frame, so the first thing we do is save the base pointer of the main stack frame onto the stack since we must immediately update it with the base pointer of the assign stack frame, which is where `%rsp` is pointing to.



7. `%rip` is incremented to the next address. We then update the base pointer to the top of the stack.

```
0x526 <assign>:
0x526 push  %rbp
0x527 mov   %rsp, %rbp
0x52a mov   $0x28, -0x4(%rbp)
0x531 mov   -0x4(%rbp), %eax
0x534 pop   %rbp
0x535 retq
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

Registers	
%eax	0x0
%edi	1
%rsp	0xd20
%rbp	**0xd20**
%rip	**0x52a**



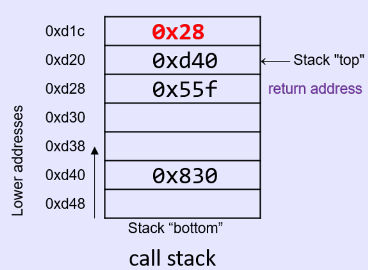
Terminal:

```
$ ./prog
```

8. Now we want to move the number 0x28 (40) into the memory location -0x4(%rbp) of the stack, which is 4 bytes above the frame pointer, which is also the stack pointer. It is common that the frame pointer is used to reference locations on the stack. Note that this does not update the stack pointer.

```
0x526 <assign>:
0x526 push  %rbp
0x527 mov   %rsp, %rbp
0x52a mov   $0x28, -0x4(%rbp)
0x531 mov   -0x4(%rbp), %eax
0x534 pop   %rbp
0x535 retq
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

Registers	
%eax	0x0
%edi	1
%rsp	0xd20
%rbp	0xd20
%rip	**0x531**



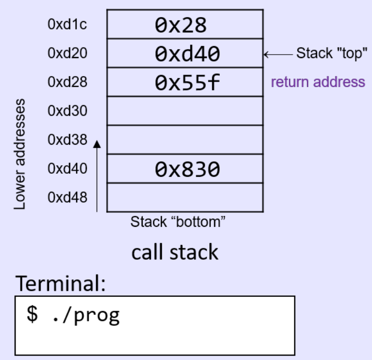
Terminal:

```
$ ./prog
```

9. Now we take the same address where we stored 0x28 to and move it into %eax, effectively loading 40 onto the return value.

```
0x526 <assign>:
0x526 push  %rbp
0x527 mov   %rsp, %rbp
0x52a mov   $0x28, -0x4(%rbp)
0x531 mov   -0x4(%rbp), %eax
0x534 pop   %rbp
0x535 retq
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

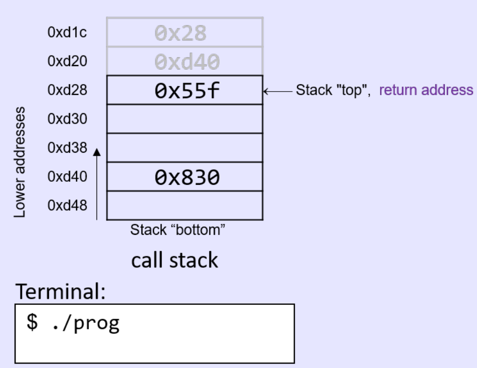
Registers	
%eax	0x28
%edi	1
%rsp	0xd20
%rbp	0xd20
%rip	0x534



10. We see that we will return this value soon, but before we do, we want to make sure that when the assign stack frame gets deleted (not really, but overwritten), we want to restore the base pointer of the main stack frame. We have already saved this before at `%rsp`, which hasn't changed since we only worked with displacements from the base pointer. We retrieve the main stack pointer data and load it back into `%rbp`. Note that this increments `%rsp` by 8 bytes, shrinking the stack, and we are technically out of the assign stack frame.

```
0x526 <assign>:
0x526 push  %rbp
0x527 mov   %rsp, %rbp
0x52a mov   $0x28, -0x4(%rbp)
0x531 mov   -0x4(%rbp), %eax
0x534 pop   %rbp
0x535 retq
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

Registers	
%eax	0x28
%edi	1
%rsp	0xd28
%rbp	0xd40
%rip	0x535



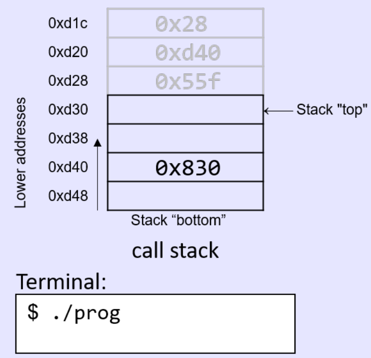
11. Note that at this point, since `%rbp` was popped off, the next value that is at the top of the stack

is the address `%rip` that we store earlier, which points to the next execution in main. When `retq` executes, this value at the top of the stack is popped into `%rip`, allowing main to continue executing within the main stack frame. Note that the return value is stored in `%eax`.

```
0x526 <assign>:
0x526 push  %rbp
0x527 mov   %rsp, %rbp
0x52a mov   $0x28, -0x4(%rbp)
0x531 mov   -0x4(%rbp), %eax
0x534 pop   %rbp
0x535 retq
0x542 <main>:
0x542 push  %rbp
0x543 mov   %rsp, %rbp
0x546 sub   $0x10, %rsp
0x54a callq 0x526 <assign>
0x55f callq 0x536 <adder>
0x554 mov   %eax, -0x4(%rbp)
0x557 mov   -0x4(%rbp), %eax
0x55a mov   %eax, %esi
```

Registers	
%eax	0x28
%edi	1
%rsp	0xd30
%rbp	0xd40
%rip	0x55f

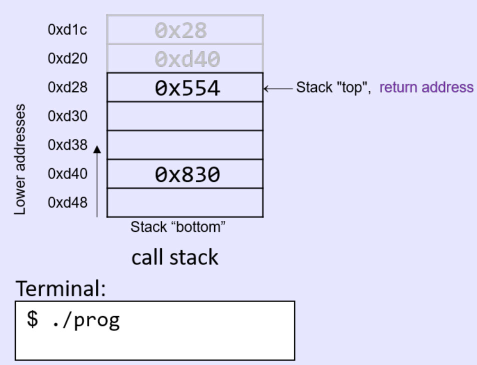
Equivalent to:
pop %rip



12. Now we execute the next instruction in `%rip` which is a call to the `adder` function. `%rip` is automatically updated to the next address at `0x554`, but since this is a `callq` instruction, we first want to store this `%rip` into the stack so we can come back to it, and then update `%rip` to the first instruction in `adder`, which is address `0x536`.

```
0x542 <main>:
0x542 push  %rbp
0x543 mov   %rsp, %rbp
0x546 sub   $0x10, %rsp
0x54a callq 0x526 <assign>
0x55f callq 0x536 <adder>
0x554 mov   %eax, -0x4(%rbp)
0x557 mov   -0x4(%rbp), %eax
0x55a mov   %eax, %esi
```

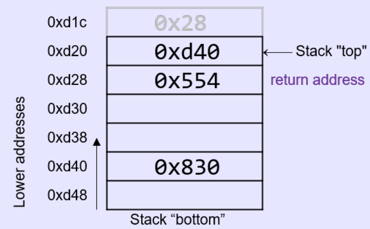
Registers	
%eax	0x0
%edi	1
%rsp	0xd28
%rbp	0xd40
%rip	0x536



13. Since we are in the adder function, this creates a new stack frame and we must update `%rbp`. Again, we don't want to overwrite the base pointer of main, so we save it onto the stack by pushing `%rbp`.

```
0x536 <adder>:
0x536 push  %rbp
0x537 mov   %rsp, %rbp
0x53a mov   $-0x4(%rbp), %eax
0x53d add   $0x2, %eax
0x540 pop %rbp
0x541 retq
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

Registers	
%eax	0x0
%edi	1
%rsp	**0xd20**
%rbp	0xd40
%rip	**0x537**



call stack

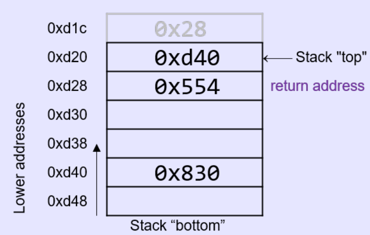
Terminal:

```
$ ./prog
```

14. Then we update %rbp to the current stack pointer.

```
0x536 <adder>:
0x536 push  %rbp
0x537 mov   %rsp, %rbp
0x53a mov   $-0x4(%rbp), %eax
0x53d add   $0x2, %eax
0x540 pop %rbp
0x541 retq
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

Registers	
%eax	0x0
%edi	1
%rsp	0xd20
%rbp	**0xd20**
%rip	**0x53a**



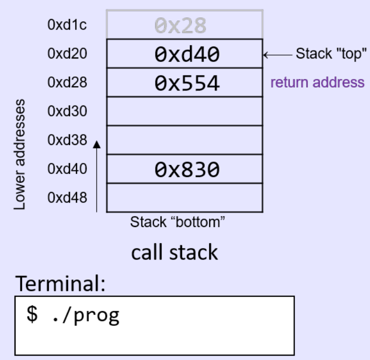
call stack

Terminal:

```
$ ./prog
```

15. This part is a bit tricky. Note that the value of 0x28 still lives at 0xd1c, which is conveniently at address -0x4(%rbp). Therefore, when we call int a; in that corresponding line in adder, we can actually add 2 to it, though it seems like there was no value assigned to it. This is just a trick though. So, we can take these remnant value and store it into %eax.

```
0x536 <adder>:
0x536 push  %rbp
0x537 mov   %rsp, %rbp
0x53a mov   $-0x4(%rbp), %eax
0x53d add   $0x2, %eax
0x540 pop %rbp
0x541 retq
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

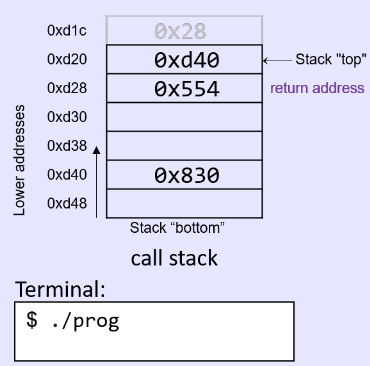


Registers	
%eax	0x28
%edi	1
%rsp	0xd20
%rbp	0xd20
%rip	0x53d

Using an old value on the stack!

16. We then add 2 to it.

```
0x536 <adder>:
0x536 push  %rbp
0x537 mov   %rsp, %rbp
0x53a mov   $-0x4(%rbp), %eax
0x53d add   $0x2, %eax
0x540 pop %rbp
0x541 retq
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
```

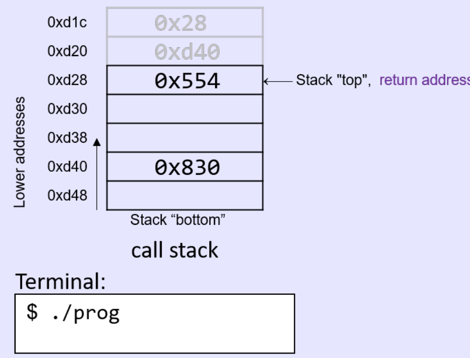


Registers	
%eax	0x2A
%edi	1
%rsp	0xd20
%rbp	0xd20
%rip	0x540

17. Now we are almost done, so we pop the base pointer of the main stack frame, at `0xd40`, back into `%rbp`.

```
  0x536 <adder>:
  0x536 push  %rbp
  0x537 mov   %rsp, %rbp
  0x53a mov   $-0x4(%rbp), %eax
  0x53d add   $0x2, %eax
➡ 0x540 pop %rbp
  0x541 retq
  0x542 <main>:
  0x542 push   %rbp
  0x543 mov    %rsp, %rbp
  0x546 sub    $0x10, %rsp
  0x54a callq  0x526 <assign>
  0x55f callq  0x536 <adder>
  0x554 mov    %eax, -0x4(%rbp)
  0x557 mov    -0x4(%rbp), %eax
  0x55a mov    %eax, %esi
```

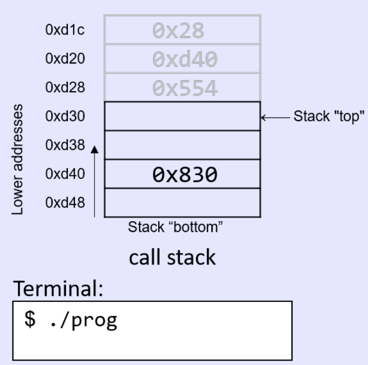
Registers	
%eax	0x2A
%edi	1
%rsp	0xd28
%rbp	0xd40
%rip	0x541



18. We now return the value in `%eax` and pop the base pointer of the adder stack frame, which simply updates the instruction pointer `%rip` back to the next instruction in main. This is equivalent to `pop %rip`, which is equivalent to moving the stack pointer `%rsp` into `%rip` and then shrinking the stack by 8 bytes `subq $8, %rsp`.

```
  0x536 <adder>:
  0x536 push  %rbp
  0x537 mov   %rsp, %rbp
  0x53a mov   $-0x4(%rbp), %eax
  0x53d add   $0x2, %eax
  0x540 pop %rbp
➡ 0x541 retq
  0x542 <main>:
  0x542 push   %rbp
  0x543 mov    %rsp, %rbp
  0x546 sub    $0x10, %rsp
  0x54a callq  0x526 <assign>
  0x55f callq  0x536 <adder>
  0x554 mov    %eax, -0x4(%rbp)
  0x557 mov    -0x4(%rbp), %eax
  0x55a mov    %eax, %esi
```

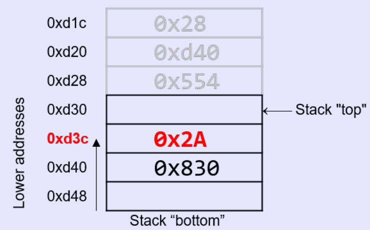
Registers	
%eax	0x2A
%edi	1
%rsp	0xd30
%rbp	0xd40
%rip	0x554



19. Now it is relatively straightforward since we do the rest in main (except for the print statement). The current value in `%eax` represents the return value of adder. We want to put this in the variable `x`, which we have already allocated some memory for right above the base pointer in the main stack frame. We move it there. Note that right after, it places this right back into

%eax.

```
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
0x55c mov    $0x400604, %edi
0x561 mov    $0x0, %eax
0x566 callq  <printf@plt>
0x56b mov    $0x0, %eax
0x570 leaveq
0x571 retq
```



call stack

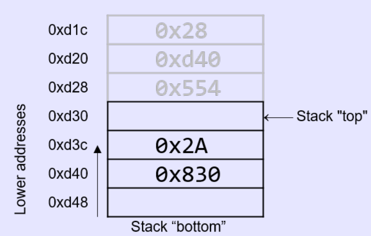
Registers	
%eax	0x2A
%edi	1
%rsp	0xd30
%rbp	0xd40
%rip	0x557

Terminal:

```
$ ./prog
```

20. the mov instruction at address 0x55a copies the value in `%eax` (or 0x2A) to register `%esi`, which is the 32-bit component register associated with `%rsi` and typically stores the second parameter to a function. We can see why since this will be put into a print statement, which is a function, and `x = %esi` is the second argument of `printf`.

```
0x542 <main>:
0x542 push   %rbp
0x543 mov    %rsp, %rbp
0x546 sub    $0x10, %rsp
0x54a callq  0x526 <assign>
0x55f callq  0x536 <adder>
0x554 mov    %eax, -0x4(%rbp)
0x557 mov    -0x4(%rbp), %eax
0x55a mov    %eax, %esi
0x55c mov    $0x400604, %edi
0x561 mov    $0x0, %eax
0x566 callq  <printf@plt>
0x56b mov    $0x0, %eax
0x570 leaveq
0x571 retq
```



call stack

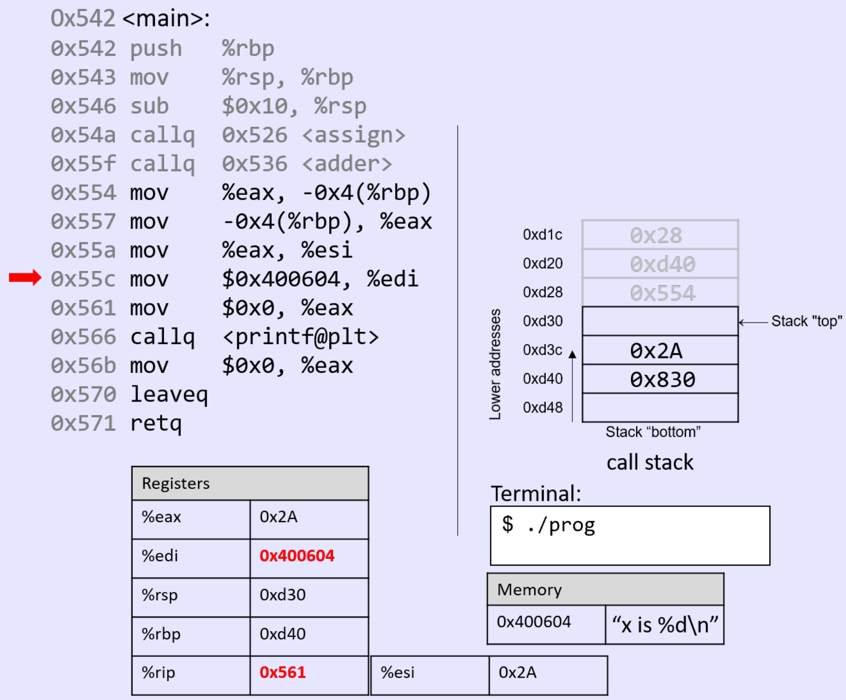
Registers	
%eax	0x2A
%edi	1
%rsp	0xd30
%rbp	0xd40
%rip	0x55c

%esi	0x2A

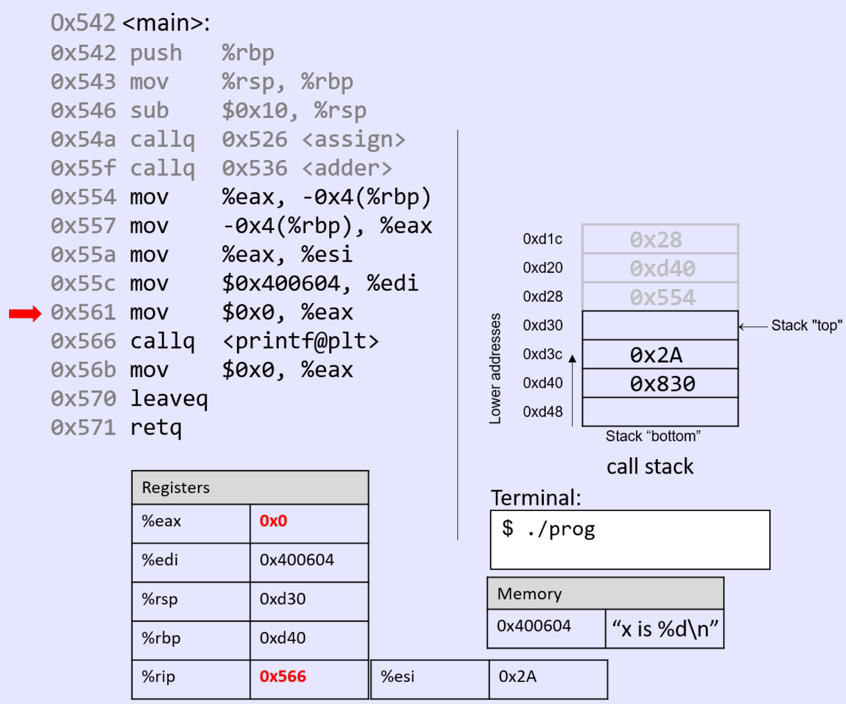
Terminal:

```
$ ./prog
```

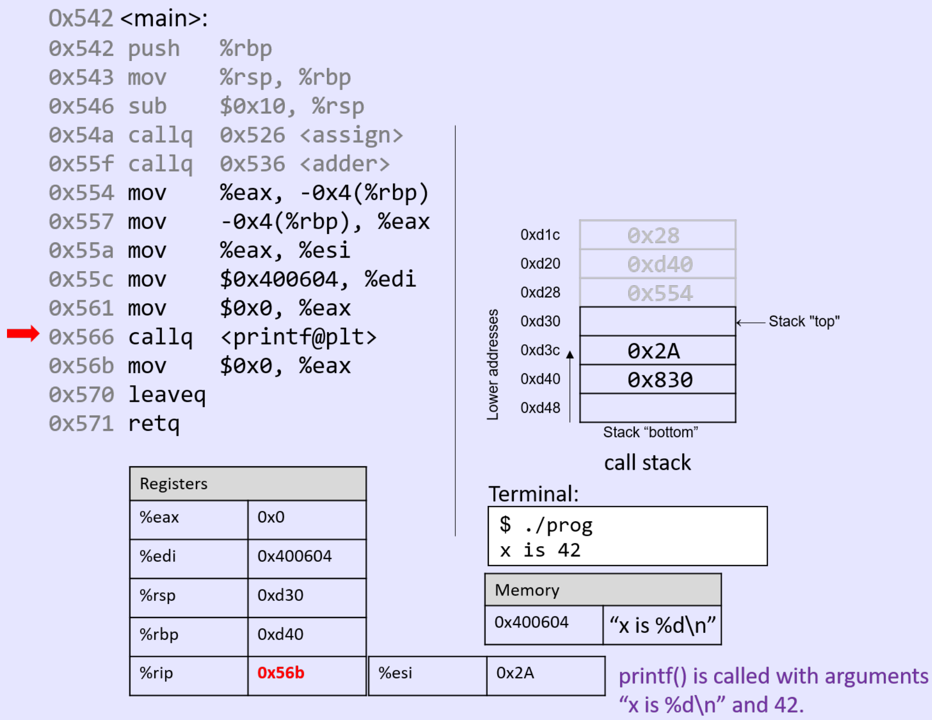
21. Now we want to retrieve the first argument of the print function. The address at `$0x400604` is some address in the code segment memory that holds the string `"x is %d\n"`.



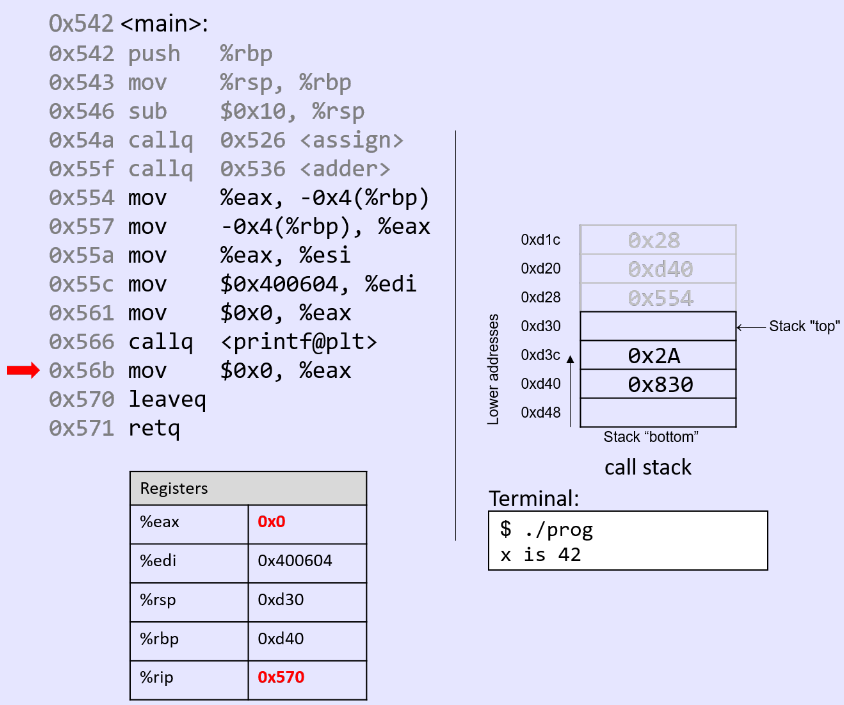
22. Then we move 0 into the `%eax` register to clear it.



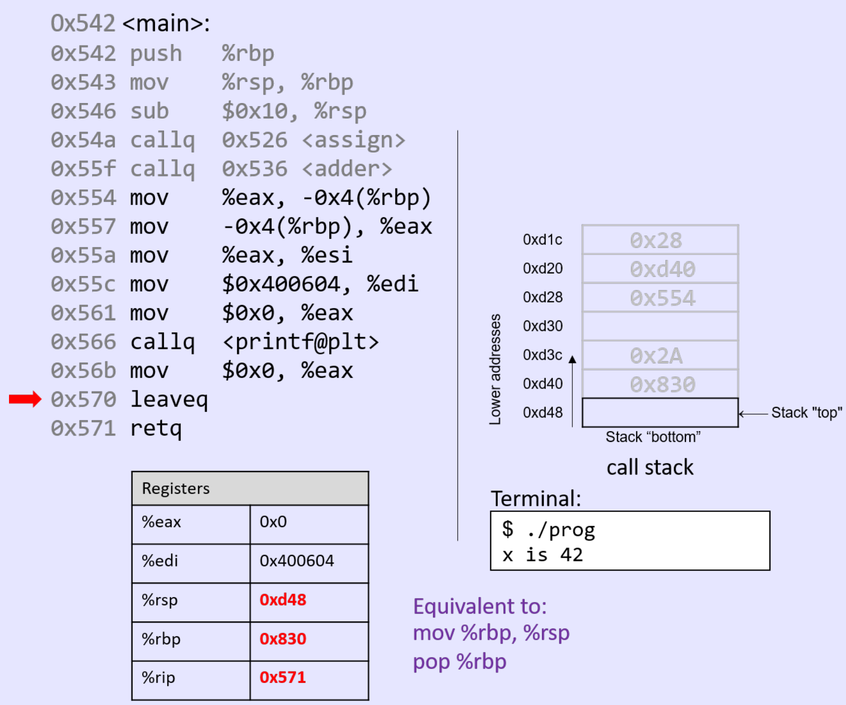
23. We then call the `printf` function, which we won't trace through but it outputs to stdout.



24. The print function might have returned something, but we don't care. We want to main function to return 0, so we move 0 into `%eax`.



25. Finally we execute `leaveq`, which prepares the stack for returning from the function call. It essentially moves the base pointer back to the stack pointer and then pops the base pointer off the stack. The new `%rbp` is the original base pointer of whatever was outside the main function, `0x830`.



26. Finally, we execute `retq`, which pops the return address off the stack and puts it into `%rip`.

We have omitted the details of caller and callee saved registers, but they do exist and are important for the general implementations.

For arrays, there's not anything new here. Let's go over some code and follow through it.

```c
int sumArray(int *array, int length) {
  int i, total = 0;
  for (i = 0; i < length; i++) {
    total += array[i];
  }
  return total;
}
```

This function takes the address of an array and the length of it and sums up all the elements in the array.

```
0x400686 <+0>: push %rbp                    # save %rbp
0x400687 <+1>: mov  %rsp,%rbp               # update %rbp (new stack frame)
0x40068a <+4>: mov  %rdi,-0x18(%rbp)        # copy array to %rbp-0x18
0x40068e <+8>: mov  %esi,-0x1c(%rbp)        # copy length to %rbp-0x1c
0x400691 <+11>:   movl $0x0,-0x4(%rbp)        # copy 0 to %rbp-0x4 (total)
0x400698 <+18>:   movl $0x0,-0x8(%rbp)        # copy 0 to %rbp-0x8 (i)
0x40069f <+25>:   jmp  0x4006be <sumArray+56> # goto <sumArray+56>
0x4006a1 <+27>:   mov  -0x8(%rbp),%eax        # copy i to %eax
0x4006a4 <+30>:   cltq                        # convert i to a 64-bit integer
0x4006a6 <+32>:   lea  0x0(,%rax,4),%rdx      # copy i*4 to %rdx
0x4006ae <+40>:   mov  -0x18(%rbp),%rax       # copy array to %rax
0x4006b2 <+44>:   add  %rdx,%rax              # compute array+i*4, store in %rax
0x4006b5 <+47>:   mov  (%rax),%eax            # copy array[i] to %eax
0x4006b7 <+49>:   add  %eax,-0x4(%rbp)        # add %eax to total
0x4006ba <+52>:   addl $0x1,-0x8(%rbp)        # add 1 to i (i+=1)
0x4006be <+56>:   mov  -0x8(%rbp),%eax        # copy i to %eax
```

```
0x4006c1 <+59>:   cmp  -0x1c(%rbp),%eax       # compare i to length
0x4006c4 <+62>:   jl   0x4006a1 <sumArray+27> # if i<length goto <sumArray+27>
0x4006c6 <+64>:   mov  -0x4(%rbp),%eax        # copy total to %eax
0x4006c9 <+67>:   pop  %rbp                   # prepare to leave the function
0x4006ca <+68>:   retq                        # return total
```

4.4.6 ARM Instructions

4.4.7 Buffer Overflows

5 Compiling and Linking

Now let's talk about how this compiling actually happens. *Compiling* is actually an umbrella term that is misused. Turning at C file into an executable file consists of multiple intermediate steps, one of which is actually compiling, but the whole series is sometimes referred to as compiling. A more accurate term would be *building*. Before we get onto it, there are two types of compilers.

Definition 5.1 (GCC, CLang)

The two mainstream compilers used is GCC (with the gdb debugger) and Clang (with lldb). For now, the difference is that

1. gcc is more established.
2. clang is newer and has more features.

A useful flag to know is that we can always specify the name of the (final or intermediary) output file with the `-o` flag.

Definition 5.2 (Complete Build Process)

To actually turn a C file into an executable file, we need to go through a series of steps. We start off with the C code, which are the `.c`, `.cpp`, or `.h` files.

1. **Preprocessing**: The precompiler step expands the *preprocessor directives* (all the `#include` and `#define` statements) and removes comments. This results in a `.i` file. The preprocessor will replace these macros with the actual code. This results in a `.i` file.

```
clang/gcc -E main.c -o main.i
```

2. **Compiling**: We take these and generate assembly code. This results in a `.asm` or `.s` file.

```
clang/gcc -S main.c -o main.s
```

3. **Assembler**: We take the assembly code and generate machine code in the form of relocatable binary object code (this is machine code, not assembly). This results in a `.o` or `.obj` file.

```
clang/gcc -c main.c -o main.o
```

4. **Linking**: We take these object files and link them together to form an executable file. This results in a `.exe` or `.out` file.

The GCC or CLang compiler automates this process for us. For example, `gcc -c hello.c` generates an object file, taking care of the preprocessing, compiling, and assembling code. Then, `gcc hello.o` links the object file to generate an executable file.

There are a lot of questions to be asked here, and we will go through them step by step.

5.1 Precompiling Stage

Just like how Python package managers like conda have specific directories that they find package in, the C library also has a certain directory.

Definition 5.3 (Standard Library Directory)

In Linux systems, there are two main directories you look at:

1. `/usr/include` contains the standard C library headers.
2. `/usr/local/include` contains the headers for libraries that you install yourself.

In Mac Silicon, these directories are a little bit more involved. You must first install the xcode command line developer tools, which will then create these directories.

1. The standard C library headers are in

   `/Library/Developer/CommandLineTools/SDKs/MacOSX*.sdk/usr/include`.

In here, we can find all the relevant import files like `stdio.h` and such. When we precompile, the output `.i` file represents a precompiled C file. It still has C code, but it has been optimized to

1. Remove comments.
2. Replace all the `#include` statements with the actual code.
3. Replace all the global variables declared in `#define` with the actual value.

Between x86 and ARM, there are no significant differences in how C files are precompiled.

Example 5.1 ()

Take a look at the following minimal example.

```c
#include "second.h"
#define a 3

int add(int x, int y) {
  return x + y;
}

int main() {
  // test comment
  int b = 5;
  int c = add(a, b);
  int d = subtract(a, b);
  return 0;
}
```

```c
int subtract(int a, int b) {
  return a - b;
}
.
.
.
.
.
.
.
.
.
.
.
```

Figure 31: I have included a `main.c` file that imports statements from a `second.h` file.

Now, I run `gcc -E main.c -o main.i` to generate the precompiled file, which gives me the following.

```
# 1 "main.c"
# 1 "<built-in>" 1
# 1 "<built-in>" 3
# 418 "<built-in>" 3
# 1 "<command line>" 1
# 1 "<built-in>" 2
# 1 "main.c" 2
# 1 "./second.h" 1
int subtract(int a, int b) {
  return a - b;
}
# 2 "main.c" 2


int add(int x, int y) {
  return x + y;
}

int main() {

  int b = 5;
  int c = add(3, b);
  int d = subtract(3, b);
  return 0;
}
```

Figure 32: The precompiled file.

Notice a few things:

1. The header file `second.h` has been replaced with the actual code.
2. The comments have indeed been removed.
3. The global variable `a` has been replaced with the actual value 3.

This leaves us with the question of what all the rest of the lines that start with a `#` are for. They are called *preprocessor directives*.

Definition 5.4 (Preprocessor Directives)

Preprocessor directives are commands that are executed before the actual compilation begins. These directives allow additional actions to be taken on the C source code before it is compiled into object code. Directives are not part of the C language itself, and they are always prefixed with a `#` symbol.

1. `#include` is used to include the contents of a file into the source file. It selects portions of the file to include based on the file name.
2. `#define` is used to define a macro, which is a way to give a name to a constant value or a piece of code.
3. `#ifdef`, `#ifndef`, `#else`, and `#endif` are used for conditional compilation.
4. `#error` is used to generate a compilation error.
5. `#pragma` is used to give the compiler specific instructions.

5.2 Compiling Stage

Once we have precompiled, we can compile the code into assembly code. For the following two examples, we will parse through the general syntax of assembly code. It is quite different between x86 and ARM, so we will use the minimal C code

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

int main() {
  int a = 3;
  int b = 5;
  int c = add(a, b);
  return 0;
}
```

for both examples.

Example 5.2 (x86 Compiled Assembly Language)

The assmebly code is shown.

```asm
.
  .file  "main.c"
  .text
  .globl add
  .type  add, @function
add:
.LFB0:
  .cfi_startproc
  endbr64
  pushq  %rbp
  .cfi_def_cfa_offset 16
  .cfi_offset 6, -16
  movq   %rsp, %rbp
  .cfi_def_cfa_register 6
  movl   %edi, -4(%rbp)
  movl   %esi, -8(%rbp)
  movl   -4(%rbp), %edx
  movl   -8(%rbp), %eax
  addl   %edx, %eax
  popq   %rbp
  .cfi_def_cfa 7, 8
  ret
  .cfi_endproc
.LFE0:
  .size  add, .-add
  .globl main
  .type  main, @function
main:
.LFB1:
  .cfi_startproc
  endbr64
  pushq  %rbp
  .cfi_def_cfa_offset 16
  .cfi_offset 6, -16
```

```
	movq	%rsp, %rbp
	.cfi_def_cfa_register 6
	subq	$16, %rsp
	movl	$3, -12(%rbp)
	movl	$5, -8(%rbp)
	movl	-8(%rbp), %edx
	movl	-12(%rbp), %eax
	movl	%edx, %esi
	movl	%eax, %edi
	call	add
	movl	%eax, -4(%rbp)
	movl	$0, %eax
	leave
	.cfi_def_cfa 7, 8
	ret
	.cfi_endproc
.LFE1:
	.size	main, .-main
	.ident	"GCC: (Ubuntu 9.4.0-1ubuntu1~20.04.2) 9.4.0"
	.section	.note.GNU-stack,"",@progbits
	.section	.note.gnu.property,"a"
	.align 8
	.long	 1f - 0f
	.long	 4f - 1f
	.long	 5
0:
	.string	 "GNU"
1:
	.align 8
	.long	 0xc0000002
	.long	 3f - 2f
2:
	.long	 0x3
3:
	.align 8
4:
```

Example 5.3 (ARM Compiled Assembly Language)

The assembly code is shown.

```
.
	.section	__TEXT,__text,regular,pure_instructions
	.build_version macos, 14, 0 sdk_version 14, 4
	.globl	_add                            ; -- Begin function add
	.p2align	2
_add:                                   ; @add
	.cfi_startproc
; %bb.0:
	sub	sp, sp, #16
	.cfi_def_cfa_offset 16
	str	w0, [sp, #12]
	str	w1, [sp, #8]
	ldr	w8, [sp, #12]
	ldr	w9, [sp, #8]
```

```
  add w0, w8, w9
  add sp, sp, #16
  ret
  .cfi_endproc
                                        ; -- End function
  .globl _main                          ; -- Begin function main
  .p2align  2
_main:                                  ; @main
  .cfi_startproc
; %bb.0:
  sub sp, sp, #48
  .cfi_def_cfa_offset 48
  stp x29, x30, [sp, #32]             ; 16-byte Folded Spill
  add x29, sp, #32
  .cfi_def_cfa w29, 16
  .cfi_offset w30, -8
  .cfi_offset w29, -16
  mov w8, #0
  str w8, [sp, #12]                   ; 4-byte Folded Spill
  stur   wzr, [x29, #-4]
  mov w8, #3
  stur   w8, [x29, #-8]
  mov w8, #5
  stur   w8, [x29, #-12]
  ldur   w0, [x29, #-8]
  ldur   w1, [x29, #-12]
  bl  _add
  mov x8, x0
  ldr w0, [sp, #12]                   ; 4-byte Folded Reload
  str w8, [sp, #16]
  ldp x29, x30, [sp, #32]             ; 16-byte Folded Reload
  add sp, sp, #48
  ret
  .cfi_endproc
                                        ; -- End function
.subsections_via_symbols
```

We can see that in both examples, there are generally two types of codes.

1. The regular CPU operations with registers and memory.
2. Some code starts off with some code that starts with a `.`. Every line that starts with a `.` are called *assembler directives*.

Let's elaborate more on what these directives are.

Definition 5.5 (Assembler Directives)

An **assembler directives** are instructions in assembly language programming that that give commands to the assembler (which then converts this to an object file) about various aspects of the assembly process, but they do not represent actual CPU instructions that execute in the program. Unlike typical assembly language instructions that directly manipulate registers and execute arithmetic or logical operations, directives are used to organize, control, and provide necessary information for the assembly and linking of binary programs. They can manage memory allocation, define symbols, control compilation settings, and much more.
There are general types of directives that are common in both x86 and ARM that we should be aware

about:
1. Section directives.
2. Data allocation directives.
3. Symbol definition directives.
4. Macro and Include directives.
5. Debugging and error handling directives.

Example 5.4 (x86 Assembly Directives)

Let us elaborate on the specific directives in the x86 assembly code, some of which are in the example above.
1. `.file "main.c"` is a directive that tells the assembler that the following code is from the file `main.c`. It is a form of metadata.
2. `.text` is a directive that tells the assembler that the following code is the text section (the text/code portion of memory) of the program. This is where the actual code is stored.
3. `.globl add` is a directive that tells the assembler that the following code is a global function called `add`.
4. `.type add, @function` is a directive that tells the assembler that the following code is a function.

Example 5.5 (ARM Assembly Directives)

You also see that there are symbols that represent memory addresses. Let's elaborate on what symbols mean.

Definition 5.6 (Symbol)

A **symbol** is a name that is used to refer to a memory location. It can be a function name, a global variable, or a local variable.
1. Global symbols are symbols that can be referenced by other object files, e.g. non-static functions and global variables.
2. Local symbols are symbols that are only visible within the object file, e.g. static functions and local variables. The linker won't know about these types.
3. External symbols are referenced by this object file but defined in another object file.

5.3 Objdump

Since we will be using the `objdump` package quite a lot, it is worth mentioning the different commands you will use and store them here as a reference. For first readers, don't expect to know what each of them do, but rather look back at this for a reference.

5.3.1 ELF and Mach-O Formats

Objdump is a command line utility that is used to display information about object files, which are often outputted in a specific format. The two main output file types are called ELF (Executable and Linkable Format) and Mach-O (Mach Object).

Definition 5.7 (ELF)

The **Executable and Linkable Format** (ELF) is a common standard file format for executables, object code, shared libraries, and core dumps. It is analogous to a book, with the following parts:

1. **Header**, which is like the cover of the book. It contains metadata about the file, such as the architecture, the entry point, and the sections.
2. **Sections**, which are like chapters. Each section contains the content for some given purpose or use wthin the program. e.g. `.binary` is just a block of bytes, `.text` contains the machine code, `.data` contains initialized data, and `.bss` contains uninitialized data.
3. **Symbol Table**, is like a detailed table of contents of all defined symbols such as functions, external (global) variables, local maps, etc.
4. **Relocation records**, which is like the index of the book that lists references to symbols.

The format is generally as such when you run `objdump -d -r hello.o` (d represents disassembly and r represents relocation entries).

```
ELF header         # file type

.text section
  - code goes here

.rodata section
  - read only data

.data section
  - initialized global variables

.bss section
  - uninitialized global variables

.symtab section
  - symbol table (symbol name, type, address)

.rel.text section
  - relocation entries for .text section
  - addresses of instructions that will need to be modified in the executable.

.rel.data section
  - relocation info for .data section
  - addresses of pointer data that will need to be modified in the merged executable.

.debug section
  - info for symbolic debugging (gcc -g)
```

Definition 5.8 (Mach-O)

5.3.2 Objdump Commands

Theorem 5.1 (File Headers with Objdump)

Given that you have an object file, the first thing you might want to do is see the file header. You do with this `objdump -f main.o`.

```
main.o:     file format elf64-x86-64
architecture: i386:x86-64, flags 0x00000011:
HAS_RELOC, HAS_SYMS
start address 0x0000000000000000
```

Theorem 5.2 (Section with Objdump)

To look at the section headers to get a closer overview, you use `objdump -h main.o`.

```
main.o:     file format elf64-x86-64

Sections:
Idx Name          Size      VMA               LMA               File off  Algn
  0 .text         0000004b  0000000000000000  0000000000000000  00000040  2**0
                  CONTENTS, ALLOC, LOAD, RELOC, READONLY, CODE
  1 .data         00000000  0000000000000000  0000000000000000  0000008b  2**0
                  CONTENTS, ALLOC, LOAD, DATA
  2 .bss          00000000  0000000000000000  0000000000000000  0000008b  2**0
                  ALLOC
  3 .comment      0000002c  0000000000000000  0000000000000000  0000008b  2**0
                  CONTENTS, READONLY
  4 .note.GNU-stack 00000000  0000000000000000  0000000000000000  000000b7  2**0
                  CONTENTS, READONLY
  5 .note.gnu.property 00000020  0000000000000000  0000000000000000  000000b8  2**3
                  CONTENTS, ALLOC, LOAD, READONLY, DATA
  6 .eh_frame     00000058  0000000000000000  0000000000000000  000000d8  2**3
                  CONTENTS, ALLOC, LOAD, RELOC, READONLY, DATA
```

Theorem 5.3 (Disassembly with Objdump)

Now you might actually want to look at the disassembly of the code, which is what we often use it for. To do this, you use `objdump -D main.o` to get the entire output.

1. The leftmost column represents the address of the instruction.
2. The next column represents the machine code of the instruction.
3. The next column represents the assembly code of the instruction.

```
main.o:     file format elf64-x86-64

Disassembly of section .text:

0000000000000000 <add>:
   0: f3 0f 1e fa             endbr64
   ...
  17: c3                      retq

0000000000000018 <main>:
  18: f3 0f 1e fa             endbr64
  ...
  4a: c3                      retq

Disassembly of section .comment:

0000000000000000 <.comment>:
   0: 00 47 43                add    %al,0x43(%rdi)
   ...
  2a: 30 00                   xor    %al,(%rax)

Disassembly of section .note.gnu.property:

0000000000000000 <.note.gnu.property>:
```

```
   0: 04 00                   add    $0x0,%al
  ...

Disassembly of section .eh_frame:

0000000000000000 <.eh_frame>:
   0: 14 00                   adc    $0x0,%al
  ...
```

If you just want to look at the contents of the executable sections, then you can use `objdump -d main.o`.

```
main.o:     file format elf64-x86-64

Disassembly of section .text:

0000000000000000 <add>:
   0: f3 0f 1e fa             endbr64
   4: 55                      push   %rbp
   5: 48 89 e5                mov    %rsp,%rbp
   8: 89 7d fc                mov    %edi,-0x4(%rbp)
   b: 89 75 f8                mov    %esi,-0x8(%rbp)
   e: 8b 55 fc                mov    -0x4(%rbp),%edx
  11: 8b 45 f8                mov    -0x8(%rbp),%eax
  14: 01 d0                   add    %edx,%eax
  16: 5d                      pop    %rbp
  17: c3                      retq

0000000000000018 <main>:
  18: f3 0f 1e fa             endbr64
  1c: 55                      push   %rbp
  1d: 48 89 e5                mov    %rsp,%rbp
  20: 48 83 ec 10             sub    $0x10,%rsp
  24: c7 45 f4 03 00 00 00    movl   $0x3,-0xc(%rbp)
  2b: c7 45 f8 05 00 00 00    movl   $0x5,-0x8(%rbp)
  32: 8b 55 f8                mov    -0x8(%rbp),%edx
  35: 8b 45 f4                mov    -0xc(%rbp),%eax
  38: 89 d6                   mov    %edx,%esi
  3a: 89 c7                   mov    %eax,%edi
  3c: e8 00 00 00 00          callq  41 <main+0x29>
  41: 89 45 fc                mov    %eax,-0x4(%rbp)
  44: b8 00 00 00 00          mov    $0x0,%eax
  49: c9                      leaveq
  4a: c3                      retq
```

If you want to see the source code intermixed with disassembly, then you can use the `-S` flag, but make sure that the object file is a generated with debugging information, i.e. use `gcc -c -g main.c -o main.o`.

```
main.o:     file format elf64-x86-64


Disassembly of section .text:

0000000000000000 <add>:
int add(int x, int y) {
   0: f3 0f 1e fa             endbr64
   4: 55                      push   %rbp
   5: 48 89 e5                mov    %rsp,%rbp
   8: 89 7d fc                mov    %edi,-0x4(%rbp)
   b: 89 75 f8                mov    %esi,-0x8(%rbp)
  return x + y;
   e: 8b 55 fc                mov    -0x4(%rbp),%edx
  11: 8b 45 f8                mov    -0x8(%rbp),%eax
  14: 01 d0                   add    %edx,%eax
}
  16: 5d                      pop    %rbp
  17: c3                      retq

0000000000000018 <main>:

int main() {
  18: f3 0f 1e fa             endbr64
  1c: 55                      push   %rbp
  1d: 48 89 e5                mov    %rsp,%rbp
  20: 48 83 ec 10             sub    $0x10,%rsp
  int a = 3;
  24: c7 45 f4 03 00 00 00    movl   $0x3,-0xc(%rbp)
  int b = 5;
  2b: c7 45 f8 05 00 00 00    movl   $0x5,-0x8(%rbp)
  int c = add(a, b);
  32: 8b 55 f8                mov    -0x8(%rbp),%edx
  35: 8b 45 f4                mov    -0xc(%rbp),%eax
  38: 89 d6                   mov    %edx,%esi
  3a: 89 c7                   mov    %eax,%edi
  3c: e8 00 00 00 00          callq  41 <main+0x29>
  41: 89 45 fc                mov    %eax,-0x4(%rbp)
  return 0;
  44: b8 00 00 00 00          mov    $0x0,%eax
}
  49: c9                      leaveq
  4a: c3                      retq
```

Figure 33: Disassembly of the object file back into assembly using `objdump -d -S main.o`.

Note that you can always see this disassembly with debuggers like gdb or lldb, but objdump generally works for all architectures.

Theorem 5.4 (Symbol Table)

If you want to look at all the symbols existing within the object file, you use `objdump -t main.o` (t for table of symbols).

1. The leftmost column represents the address of the symbol.

2. The next column represents the type of the symbol. The `g` and `l` represent global and local symbols, respectively. The `O` and `F` represent object and function symbols, while the `UND` and `ABS` represent undefined and absolute symbols.
3. The next column represents the section that the symbol is in.
4. The next column represents the size of the symbol.
5. The last column represents the name of the symbol.

```
main.o:     file format elf64-x86-64

SYMBOL TABLE:
0000000000000000 l    df *ABS*   0000000000000000 main.c
0000000000000000 l    d  .text   0000000000000000 .text
0000000000000000 l    d  .data   0000000000000000 .data
0000000000000000 l    d  .bss 0000000000000000 .bss
0000000000000000 l    d  .note.GNU-stack  0000000000000000 .note.GNU-stack
0000000000000000 l    d  .note.gnu.property  0000000000000000 .note.gnu.property
0000000000000000 l    d  .eh_frame  0000000000000000 .eh_frame
0000000000000000 l    d  .comment   0000000000000000 .comment
0000000000000000 g     F .text   0000000000000018 add
0000000000000018 g     F .text   0000000000000033 main
```

Theorem 5.5 (Relocation Table)

If you want to look then at the relocation table, then you use `objdump -r main.o`.

1. The leftmost column represents the offset of the relocation (i.e. the location within the section where this relocation needs to be applied).
2. The second column represents the type of relocation.
3. The third column represents the symbol that this relocation references.

```
main.o:     file format elf64-x86-64

RELOCATION RECORDS FOR [.text]:
OFFSET           TYPE              VALUE
000000000000003d R_X86_64_PLT32    add-0x0000000000000004


RELOCATION RECORDS FOR [.eh_frame]:
OFFSET           TYPE              VALUE
0000000000000020 R_X86_64_PC32     .text
0000000000000040 R_X86_64_PC32     .text+0x0000000000000018
```

5.4 Assembling Stage and Object Files

Now, once you have gotten the object file, you cannot simply open it up in a text edit as it is in machine code. To actually interpret anything from it, you must **disassmble** it, meaning that you convert the machine code back into assembly code. The main software that you use to do this is `objdump`. Let's take a look again at the object file.

```
Disassembly of section .text:

0000000000000000 <add>:
   0: f3 0f 1e fa              endbr64
   4: 55                       push   %rbp
   5: 48 89 e5                 mov    %rsp,%rbp
   8: 89 7d fc                 mov    %edi,-0x4(%rbp)
   b: 89 75 f8                 mov    %esi,-0x8(%rbp)
   e: 8b 55 fc                 mov    -0x4(%rbp),%edx
  11: 8b 45 f8                 mov    -0x8(%rbp),%eax
  14: 01 d0                    add    %edx,%eax
  16: 5d                       pop    %rbp
  17: c3                       retq

0000000000000018 <main>:
  18: f3 0f 1e fa              endbr64
  1c: 55                       push   %rbp
  1d: 48 89 e5                 mov    %rsp,%rbp
  20: 48 83 ec 10              sub    $0x10,%rsp
  24: c7 45 f4 03 00 00 00     movl   $0x3,-0xc(%rbp)
  2b: c7 45 f8 05 00 00 00     movl   $0x5,-0x8(%rbp)
  32: 8b 55 f8                 mov    -0x8(%rbp),%edx
  35: 8b 45 f4                 mov    -0xc(%rbp),%eax
  38: 89 d6                    mov    %edx,%esi
  3a: 89 c7                    mov    %eax,%edi
  3c: e8 00 00 00 00           callq  41 <main+0x29>
  41: 89 45 fc                 mov    %eax,-0x4(%rbp)
  44: b8 00 00 00 00           mov    $0x0,%eax
  49: c9                       leaveq
  4a: c3                       retq
```

Figure 34: Disassembly of the object file back into assembly using `objdump -d main.o`.

Let's note a couple things.

1. The functions are organized by their starting address followed by their name, e.g.

   ```
   0000000000000000 <add>:
   ```

   Within each function, each line of assembly code is shown. To find the total memory the function takes up, you can just take the address of the last line and subtract it from the address of the first line. Or you can literally count the number of bytes in each line (remember 2 hex is 1 byte).

2. The line that calls the `add` function is `0x0` (`00 00 00 00`), with is the *relative target address* intended to be filled in by the linker. The actual assembly line just says that the function continues on to the next line at address `0x41`. This is because the object file is not aware of where it will be loaded into memory, and all lines with this opcode `e8 00 00 00 00` is intended to be filled in by the linker.

3. Look at address `0x3c`. It is calling another function, but the values starting from address `0x3d` is `00 00 00 00`, which is not the actual address of the function but also a dummy address. This is because the object file is not aware of where the function is located in memory.

5.5 Linking Stage and Relocation

5.5.1 Relocation

If the object file is already in machine code, then why do we need a separate linking stage that converts `main.o` into `main` the binary? The reason is stated in the previous section: because the object files uses relative memory addressing and does not know about which memory is accessed in other object files, we need to **relocate** the symbols in the object file to their proper addresses. So how does the linker actually know how to relocate these symbols into their proper addresses? It uses the *relocation table*, which contains information about the addresses that need to be modified in the object file.

```
main.o:     file format elf64-x86-64

RELOCATION RECORDS FOR [.text]:
OFFSET           TYPE              VALUE
000000000000003d R_X86_64_PLT32    add-0x0000000000000004


RELOCATION RECORDS FOR [.eh_frame]:
OFFSET           TYPE              VALUE
0000000000000020 R_X86_64_PC32     .text
0000000000000040 R_X86_64_PC32     .text+0x0000000000000018
```

Figure 35: Relocation table for `main.o` object file.

Let's talk about how to actually read this table. We can look at the first entry, which shows an offset of `0x3d`. This represents the offset from the beginning of the `.text` section where the relocation needs to be applied. Looking back at the disassembly file, this address `0x3d` is precisely where there was a dummy address `00 00 00 00`. We want to replace this with the actual address defined in the `VALUE` column, which is `add` (with a slight offset of `0x4`, which is typically used to compensate for the PC-relative addressing mode where the CPU might be adding the length of the instruction to the program counter (PC) before the relocation value is applied). The type of relocation won't be covered in our scope. Let's go through each relocation entry:

1. The first entry is for the `add` function. If we look at the disassembly, within the `main` function, the address `0x3d` is where the `add` function is called. The linker will replace the dummy address with the actual address of the `add` function.

```
Disassembly of section .text:

0000000000000000 <add>:
   0: f3 0f 1e fa             endbr64
   4: 55                      push   %rbp
   5: 48 89 e5                mov    %rsp,%rbp
   8: 89 7d fc                mov    %edi,-0x4(%rbp)
   b: 89 75 f8                mov    %esi,-0x8(%rbp)
   e: 8b 55 fc                mov    -0x4(%rbp),%edx
  11: 8b 45 f8                mov    -0x8(%rbp),%eax
  14: 01 d0                   add    %edx,%eax
  16: 5d                      pop    %rbp
  17: c3                      retq

0000000000000018 <main>:
  18: f3 0f 1e fa             endbr64
  1c: 55                      push   %rbp
  1d: 48 89 e5                mov    %rsp,%rbp
  20: 48 83 ec 10             sub    $0x10,%rsp
```

```
  24: c7 45 f4 03 00 00 00    movl   $0x3,-0xc(%rbp)
  2b: c7 45 f8 05 00 00 00    movl   $0x5,-0x8(%rbp)
  32: 8b 55 f8                mov    -0x8(%rbp),%edx
  35: 8b 45 f4                mov    -0xc(%rbp),%eax
  38: 89 d6                   mov    %edx,%esi
  3a: 89 c7                   mov    %eax,%edi
  3c: e8 00 00 00 00          callq  41 <main+0x29>     <-- here
  41: 89 45 fc                mov    %eax,-0x4(%rbp)
  44: b8 00 00 00 00          mov    $0x0,%eax
  49: c9                      leaveq
  4a: c3                      retq
```

2. The second and third entries are for the `.eh_frame` section. We can see that the offset of `0x20` and `0x40` represents the following lines below. They also have dummy addresses that need to be replaced. They are replaced by the address `.text`, which represents the first address in the `.text` section, i.e. the address of the `add` function, and the address `.text+0x18`, which represents the address of the `main` function.

```
Disassembly of section .eh_frame:

0000000000000000 <.eh_frame>:
   0: 14 00                   adc    $0x0,%al
   2: 00 00                   add    %al,(%rax)
   4: 00 00                   add    %al,(%rax)
   6: 00 00                   add    %al,(%rax)
   8: 01 7a 52                add    %edi,0x52(%rdx)
   b: 00 01                   add    %al,(%rcx)
   d: 78 10                   js     1f <.eh_frame+0x1f>
   f: 01 1b                   add    %ebx,(%rbx)
  11: 0c 07                   or     $0x7,%al
  13: 08 90 01 00 00 1c       or     %dl,0x1c000001(%rax)
  19: 00 00                   add    %al,(%rax)
  1b: 00 1c 00                add    %bl,(%rax,%rax,1)
  1e: 00 00                   add    %al,(%rax)
  20: 00 00                   add    %al,(%rax)     <-- here for 2nd entry
  22: 00 00                   add    %al,(%rax)
  24: 18 00                   sbb    %al,(%rax)
  26: 00 00                   add    %al,(%rax)
  28: 00 45 0e                add    %al,0xe(%rbp)
  2b: 10 86 02 43 0d 06       adc    %al,0x60d4302(%rsi)
  31: 4f 0c 07                rex.WRXB or $0x7,%al
  34: 08 00                   or     %al,(%rax)
  36: 00 00                   add    %al,(%rax)
  38: 1c 00                   sbb    $0x0,%al
  3a: 00 00                   add    %al,(%rax)
  3c: 3c 00                   cmp    $0x0,%al
  3e: 00 00                   add    %al,(%rax)
  40: 00 00                   add    %al,(%rax)     <-- here for 3rd entry
  42: 00 00                   add    %al,(%rax)
  44: 33 00                   xor    (%rax),%eax
```

Therefore, we can see that the object file generates a "skeleton" code that contains all the instructions, with some dummy addresses that need to be replaced. The relocation table T tells us exactly where these dummy addresses are in the code and what they need to be replaced with. Therefore, if we want to call a function `printf` that is in the text section at address `0x30`, then we can actually look at the value at `T[30]` to see where the actual address is. At this point, note that we still do not know the actual memory address of `add`. This is determined by the linker.

5.5.2 Linking with One Object File

Now let's see what happens once we link the object file `main.o` into the final executable `main`. If we disassemble it, then we can see a few things:

1. The addresses of all the functions have been changed. `add` starts on address `0x1129` rather than `0x0` and `main` starts on address `0x1141` rather than `0x18`.
2. The dummy address `0x0` of the call to function `add` in `main` have been replaced with the actual addresses `0x1129`.

```
0000000000001129 <add>:
  1129:  f3 0f 1e fa             endbr64
  112d:  55                      push   %rbp
  112e:  48 89 e5                mov    %rsp,%rbp
  1131:  89 7d fc                mov    %edi,-0x4(%rbp)
  1134:  89 75 f8                mov    %esi,-0x8(%rbp)
  1137:  8b 55 fc                mov    -0x4(%rbp),%edx
  113a:  8b 45 f8                mov    -0x8(%rbp),%eax
  113d:  01 d0                   add    %edx,%eax
  113f:  5d                      pop    %rbp
  1140:  c3                      retq

0000000000001141 <main>:
  1141:  f3 0f 1e fa             endbr64
  1145:  55                      push   %rbp
  1146:  48 89 e5                mov    %rsp,%rbp
  1149:  48 83 ec 10             sub    $0x10,%rsp
  114d:  c7 45 f4 03 00 00 00    movl   $0x3,-0xc(%rbp)
  1154:  c7 45 f8 05 00 00 00    movl   $0x5,-0x8(%rbp)
  115b:  8b 55 f8                mov    -0x8(%rbp),%edx
  115e:  8b 45 f4                mov    -0xc(%rbp),%eax
  1161:  89 d6                   mov    %edx,%esi
  1163:  89 c7                   mov    %eax,%edi
  1165:  e8 bf ff ff ff          callq  1129 <add>     <-- replaced with actual address
  116a:  89 45 fc                mov    %eax,-0x4(%rbp)
  116d:  b8 00 00 00 00          mov    $0x0,%eax
  1172:  c9                      leaveq
  1173:  c3                      retq
  1174:  66 2e 0f 1f 84 00 00    nopw   %cs:0x0(%rax,%rax,1)
  117b:  00 00 00
  117e:  66 90                   xchg   %ax,%ax
```

5.5.3 Global vs External Symbols

So far, we have talked about using the `#include` as a precompiling command that says "put all the text from this other file right here." Take the following code for instance.

```c
// file1.c
#include "sum.h"

int array[2] = {1, 2};

int main() {
  int val = sum(array, 2);
  return val;
}
```

```c
// sum.h
int sum(int *a, int n) {
  int i, s = 0;
  for (i = 0; i < n; i++) {
    s += a[i];
  }
  return s;
}
.
```

Figure 36: Including a header file in `file1.c` to import functions and variables.

However, there is another way to do this. We can use *external symbols* to access. Rather than simply copying and pasting the code into the file, the `extern` keyword marks that the variable or function exists externally to this source file and does not allocate storage for it.

```c
// main.c
extern int sum(int *array, int n);

int array[2] = {1, 2};

int main(void) {
  int val = sum(array, 2);
  return val;
}
```

```c
// sum.c
int sum(int *array, int n) {
  int i, s = 0 ;
  for (int i = 0; i < n; i++) {
      s += array[i];
    }
  return s;
}
.
```

Figure 37: Using external symbols to access functions and variables.

One is not a replacement for the other, so what advantage does this have? Well, as we will see, if we have multiple object (source) files, say `A.c`, `B.c`, and `C.c`, that need to reference the same function or variable `var` in `ext.c`, then how would we do this? If we simply put `#include "ext.h"` in all the files, then we would have multiple copies of the same code. This means that for each source there would be its own copy of `var` created and the linker would be unable to resolve this symbol. However, if we put `extern int var;` at the top of each source file, then only one copy of `var` would be created (in `ext.c`), which creates a single instance of `var` for the linker to resolve. [2]

Therefore, there are three types of symbols (variables, functions, etc.) that we need to consider:

1. **Global symbols** that are defined in the global scope of a C file.
2. **Local symbols** that are defined in the local scope of a C file, e.g. within functions, loops, etc.
3. **External symbols** that are defined in another C file referenced by the `extern` keyword.

Linkers will only know about global and external symbols, and will have no idea that any local symbols exist. With the information of these two types of symbols and the relocation tables of each object file, the linker can then resolve the addresses of all the symbols in the final binary.

The two types of symbols that the linker will know about are the global and external symbols. We can see that external symbols can be problematic if the object files don't know about each other.

[2] https://stackoverflow.com/questions/1330114/whats-the-difference-between-using-extern-and-including-header-files

Example 5.6 (Global and Local Symbols)

Consider the following code where the left file includes the right file.

```c
// main.c
#include "sum.h"

int array[2] = {1, 2};

int main() {
  int val = sum(array, 2);
  return val;
}
```

```c
// sum.h
int sum(int *a, int n) {
  int i, s = 0;
  for (i = 0; i < n; i++) {
    s += a[i];
  }
  return s;
}
.
```

In the left file,

1. We define the global symbol `main()`.
2. Inside main, `val` is a local symbol so the linker knows nothing about it.
3. The `sum` function is an external symbol, and it references a global symbol that's defined in `sum` the right file.
4. The `array` is a global symbol that is defined in the right file.

In the right file, the linker knows nothing of the local symbols `i` or `s`.

5.5.4 Linking with Multiple Object Files

We have seen the case of linking when we simply have one object file. The relocation was simple since the `.text` section is contiguous and so we needed simple translations of addresses to relocate `add` and `main`, along with whatever other sections and files. Now let's consider the case where we have multiple object files.

```c
// main.c
extern int sum(int *array, int n);

int array[2] = {1, 2};

int main(void) {
  int val = sum(array, 2);
  return val;
}
```

```c
// sum.c
int sum(int *array, int n) {
  int i, s = 0 ;
  for (int i = 0; i < n; i++) {
      s += array[i];
    }
  return s;
}
.
```

Now they have their own object files shown below, where I also put the source code lines to make it easier to parse. Note that again, in `main.o` the call to function `sum` is a dummy address that needs to be replaced. Furthermore, in both `main.o` and `sum.o`, the `.text` section is at address `0x0`, where the addresses of the function `main` and `sum` are, respectively. This causes an overload in the address space.

To demonstrate what happens, we look at how the disassembly, symbol tables, and relocation tables are updated before (with the object files) and after (in the binary) linking.

Example 5.7 (Disassembly of Object Files)

In here, note that both the `array` and `sum` are not initialized and are therefore set to dummy addresses.

```
main.o:     file format elf64-x86-64
Disassembly of section .text:

0000000000000000 <main>:
extern int sum(int *array, int n);

```

```
int array[2] = {1, 2};

int main(void) {
   0: f3 0f 1e fa             endbr64
   4: 55                      push   %rbp
   5: 48 89 e5                mov    %rsp,%rbp
   8: 48 83 ec 10             sub    $0x10,%rsp
  int val = sum(array, 2);
   c: be 02 00 00 00          mov    $0x2,%esi
  11: 48 8d 3d 00 00 00 00    lea    0x0(%rip),%rdi        # 18 <main+0x18>  <-- dummy
    address
  18: e8 00 00 00 00          callq  1d <main+0x1d>                          <-- dummy
    address
  1d: 89 45 fc                mov    %eax,-0x4(%rbp)
  return val;
  20: 8b 45 fc                mov    -0x4(%rbp),%eax
}
  23: c9                      leaveq
  24: c3                      retq
```

```
sum.o:     file format elf64-x86-64
Disassembly of section .text:

0000000000000000 <sum>:
int sum(int *array, int n) {
   0: f3 0f 1e fa             endbr64
   4: 55                      push   %rbp
   5: 48 89 e5                mov    %rsp,%rbp
   8: 48 89 7d e8             mov    %rdi,-0x18(%rbp)
   c: 89 75 e4                mov    %esi,-0x1c(%rbp)
  int i, s = 0;
   f: c7 45 f8 00 00 00 00    movl   $0x0,-0x8(%rbp)
  for (int i = 0; i < n; i++) {
  16: c7 45 fc 00 00 00 00    movl   $0x0,-0x4(%rbp)
  1d: eb 1d                   jmp    3c <sum+0x3c>
    s += array[i];
  1f: 8b 45 fc                mov    -0x4(%rbp),%eax
  22: 48 98                   cltq
  24: 48 8d 14 85 00 00 00    lea    0x0(,%rax,4),%rdx
  2b: 00
  2c: 48 8b 45 e8             mov    -0x18(%rbp),%rax
  30: 48 01 d0                add    %rdx,%rax
  33: 8b 00                   mov    (%rax),%eax
  35: 01 45 f8                add    %eax,-0x8(%rbp)
  for (int i = 0; i < n; i++) {
  38: 83 45 fc 01             addl   $0x1,-0x4(%rbp)
  3c: 8b 45 fc                mov    -0x4(%rbp),%eax
  3f: 3b 45 e4                cmp    -0x1c(%rbp),%eax
  42: 7c db                   jl     1f <sum+0x1f>
  }
  return s;
  44: 8b 45 f8                mov    -0x8(%rbp),%eax
}
  47: 5d                      pop    %rbp
  48: c3                      retq
```

1. In `main.o` at address `0x0`, we have the `main` function and this is because everything is stored relatively to the start of main. Once we have linked, `main` shows the absolute addresses of all the instructions.
2. In instruction 11 in `main.o` we can see that `48 8d 3d` is the `lea` instruction, which is the same as that in `main`. However, the address that is was acting on is `0x0` since the array has not been initialized yet. We can see in `main` that the address is now `0x00002ecf`.
3. The comment in `main` indicates that the final relocated address used to access the `array` is `0x4010`. To see relocated addresses in general, just look for the comments and shift them accordingly.

```
main:     file format elf64-x86-64

0000000000001129 <main>:
    1129:   f3 0f 1e fa             endbr64
    112d:   55                      push   %rbp
    112e:   48 89 e5                mov    %rsp,%rbp
    1131:   48 83 ec 10             sub    $0x10,%rsp
    1135:   be 02 00 00 00          mov    $0x2,%esi
    113a:   48 8d 3d cf 2e 00 00    lea    0x2ecf(%rip),%rdi        # 4010 <array>
    1141:   e8 08 00 00 00          callq  114e <sum>
    1146:   89 45 fc                mov    %eax,-0x4(%rbp)
    1149:   8b 45 fc                mov    -0x4(%rbp),%eax
    114c:   c9                      leaveq
    114d:   c3                      retq

000000000000114e <sum>:
    114e:   f3 0f 1e fa             endbr64
    1152:   55                      push   %rbp
    1153:   48 89 e5                mov    %rsp,%rbp
    1156:   48 89 7d e8             mov    %rdi,-0x18(%rbp)
    115a:   89 75 e4                mov    %esi,-0x1c(%rbp)
    ...
```

Example 5.8 (Symbol Tables of Object Files)

Let's take a look at the symbol table of each file as well. Again, all of the addresses of each symbol are 0s since they are using relative addressing. The `array` and `main` are global symbols since they reside in the global scope, while the `sum` function is an external and undefined symbol.

```
main.o:     file format elf64-x86-64

SYMBOL TABLE:
0000000000000000 l    df *ABS*   0000000000000000 main.c
0000000000000000 l    d  .text   0000000000000000 .text
0000000000000000 l    d  .data   0000000000000000 .data
0000000000000000 l    d  .bss 0000000000000000 .bss
0000000000000000 l    d  .note.GNU-stack  0000000000000000 .note.GNU-stack
0000000000000000 l    d  .note.gnu.property  0000000000000000 .note.gnu.property
0000000000000000 l    d  .eh_frame  0000000000000000 .eh_frame
0000000000000000 l    d  .comment   0000000000000000 .comment
0000000000000000 g     O .data   0000000000000008 array
0000000000000000 g     F .text   0000000000000025 main
0000000000000000         *UND*   0000000000000000 _GLOBAL_OFFSET_TABLE_
0000000000000000         *UND*   0000000000000000 sum
```

```
sum.o:     file format elf64-x86-64

SYMBOL TABLE:
0000000000000000 l    df *ABS*   0000000000000000 sum.c
0000000000000000 l    d  .text   0000000000000000 .text
0000000000000000 l    d  .data   0000000000000000 .data
0000000000000000 l    d  .bss 0000000000000000 .bss
0000000000000000 l    d  .note.GNU-stack  0000000000000000 .note.GNU-stack
0000000000000000 l    d  .note.gnu.property  0000000000000000 .note.gnu.property
0000000000000000 l    d  .eh_frame  0000000000000000 .eh_frame
0000000000000000 l    d  .comment   0000000000000000 .comment
0000000000000000 g     F .text   0000000000000049 sum
```

When we have the linked binary, note a few things.

1. In `main.o`, the numbers on the left represents the address of the symbol (all 0s since we haven't linked yet and their final addresses aren't known), while the addresses in `a.out` are all known.
2. In `main.o`, the `sum` function is an external symbol and is undefined. The linker will need to know where this is. In `main`, note that the `sum` function is now a global symbol and is defined, along with the size. We can now see that all the final addresses of each symbol is known, along with their sizes, and the `UND` marker is now gone as well.
3. Only the size of the global variable is known in `main.o` since we have defined it within the code. However, in `main`, the linker has now assigned an address to it.
4. To see the size in bytes of the array, you can look at the address and how much size it takes up.

```
main:     file format elf64-x86-64

SYMBOL TABLE:
...
0000000000004008 g     O .data     0000000000000000              .hidden __dso_handle
000000000000114e g     F .text     0000000000000049              sum
0000000000002000 g     O .rodata   0000000000000004              _IO_stdin_used
00000000000011a0 g     F .text     0000000000000065              __libc_csu_init
0000000000004020 g       .bss      0000000000000000              _end
0000000000001040 g     F .text     000000000000002f              _start
0000000000004018 g       .bss      0000000000000000              __bss_start
0000000000001129 g     F .text     0000000000000025              main
0000000000004018 g     O .data     0000000000000000              .hidden __TMC_END__
...
```

Example 5.9 (Relocation Tables)

Ignoring the `.eh_frame`, in `main.o` the relocation table contains entries for `array` and `sum` that must be relocated.

```
main.o:     file format elf64-x86-64

RELOCATION RECORDS FOR [.text]:
OFFSET           TYPE              VALUE
0000000000000014 R_X86_64_PC32     array-0x0000000000000004
0000000000000019 R_X86_64_PLT32    sum-0x0000000000000004

RELOCATION RECORDS FOR [.eh_frame]:
OFFSET           TYPE              VALUE
0000000000000020 R_X86_64_PC32     .text
```

```
sum.o:     file format elf64-x86-64

RELOCATION RECORDS FOR [.eh_frame]:
OFFSET           TYPE              VALUE
0000000000000020 R_X86_64_PC32     .text
```

We can see a couple things. Namely, there is nothing to be relocated in `a.out` since everything has been relocated already by the linker. So let's focus on the relocation for `main.o`. In here, we can see that in the `.text` section, there are two things being relocated:

1. The reference to the global variable `array` is being relocated. In this object file, we look at the offset `0x14` from the beginning of the `.text` section, which contains the instruction that needs to access `array`. This relocation record tells the linker to calculate the 32-bit offset from the instruction (at offset `0x14`) to the start of `array`, then adjust it by subtracting 4 bytes.
2. The reference to the `sum` function is being relocated. In this object file, we look at the offset `0x19` from the beginning of the `.text` section, which contains the instruction that needs to access `sum`. This relocation record tells the linker to calculate the 32-bit offset from the instruction (at offset `0x19`) to the start of the `.plt` section, then adjust it by subtracting 4 bytes.

```
main:     file format elf64-x86-64
```

5.6 Compiler Optimization

We have learned the complete process of compilers, but compilers can be a little smarter than just translating code line by line. They also come with flags that can optimize the code.

Definition 5.9 (gcc Optimization)

The gcc compiler can optimize the code with the `-O` flag. To run level 1 optimization, we can write

```
gcc -O1 -o main main.c
```

The level of optimizations are listed:

1. Level 1 perform basic optimizations to reduce code size and execution time while attempting to keep compile time to a minimum.
2. Level 2 optimizations include most of GCC's implemented optimizations that do not involve a space-performance trade-off.
3. Level 3 performs additional optimizations (such as function inlining) and may cause the program to take significantly longer to compile.

Let's see what common implementation are.

Definition 5.10 (Constant Folding)

Constants in the code are evaluated at compile time to reduce the number of resulting instructions. For example, in the code snippet that follows, macro expansion replaces the statement `int debug = N-5` with `int debug = 5-5`. Constant folding then updates this statement to `int debug = 0`.

```c
#define N 5
int debug = N - 5; //constant folding changes this statement to debug = 0;
```

Definition 5.11 (Constant Propagation)

Constant propagation replaces variables with a constant value if such a value is known at compile time. Consider the following code segment, where the `if (debug)` statement is replaced with `if (0)`.

```c
int debug = 0;

int doubleSum(int *array, int length){
    int i, total = 0;
    for (i = 0; i < length; i++){
        total += array[i];
        if (debug) {
            printf("array[%d] is: %d\n", i, array[i]);
        }
    }
    return 2 * total;
}
```

Definition 5.12 (Dead Code Elimination)

Dead code elimination removes code that is never executed. For example, in the code snippet that follows, the `if (debug)` statement and its body is removed since the value of `debug` is known to be 0.

```c
int debug = 0;

int doubleSum(int *array, int length){
    int i, total = 0;
    for (i = 0; i < length; i++){
        total += array[i];
        if (debug) {                                        // remove
            printf("array[%d] is: %d\n", i, array[i]);    // remove
        }                                                   // remove
    }
    return 2 * total;
}
```

Definition 5.13 (Simplifying Expressions)

Some instructions are more expensive than others, so things like

1. `2 * total` may be replaced with `total + total` because addition instruction is less expensive than multiplication.
2. `total * 8` may be replaced with `total « 3`
3. `total % 8` may be replaced with `total & 7`

Note that these optimization techniques are in no way a guarantee that the code will run faster since there are many factors and always edge cases (for example, maybe some localities are lost). Furthermore, compiler optimization will never be able to improve runtime complexity (e.g. by replacing bubble sort with quicksort).

6 Storage Hierarchy

6.1 Expanding on von Neumann Architecture

So far, our model of the computer has been a simple von Neumann architecture which consists of a CPU and memory. However, there are many other intricacies that are extremely important in practice, and we'll expand on each one by one.

Definition 6.1 (Computer Architecture)

In our elaborated computer architecture, a computer consists of the components.

1. A **CPU** that consists of an arithmetic logic unit (ALU), registers, and a **bus interface** that controls the input and output.
2. The **IO bridge** that handles communication between everything.
3. The **system bus** that connects the CPU to the IO bridge.
4. The **memory bus** that connects the memory to the IO bridge.
5. The **IO bus** that connects the IO devices and disk to the IO bridge.
6. **IO devices** like mouse, keyboard, and monitor.
7. The **disk controller and disk** that stores data.

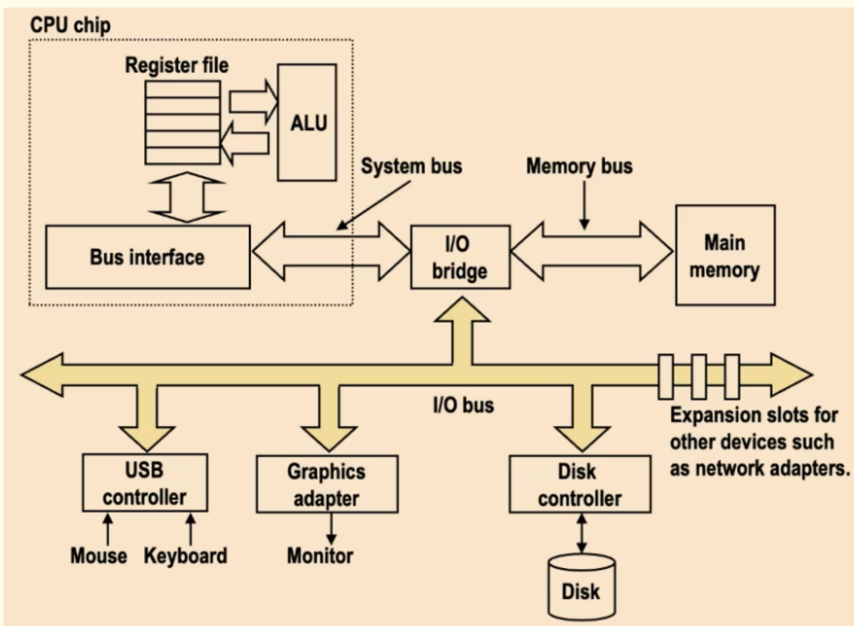


Figure 38: Diagram of the IO bus.

We can see from the diagram above that the CPU can directly access registers (since it's in the CPU itself) and the main memory (since it's connected to the memory bus). However, to access something like the disk, it must go through the disk controller. This gives us our first categorization of memory.

Definition 6.2 (Primary Storage)

Primary storage devices are directly accessible by the CPU and are used to store data that is currently being processed. This includes CPU registers, cache memory, and RAM. In memory, the basic storage unit is normally a **cell** (one bit per cell), which is the physical material that holds information. A **supercell** has address and data widths (number of bits), which is analogous to a lock number and the lock capacity, respectively. It is called random access since it takes approximately the same amount of time to access any cell in memory. There are two primary ways that this is

implemented:
1. **Static RAM (SRAM)** stores data in small electrical circuits (e.g. latches) and is typically the fastest type of memory. However, it is more expensive to build, consumers more power, and occupies more space, limiting the SRAM storage.
2. **Dynamic RAM (DRAM)** stores data using electrical components (e.g. capacitors) that hold an electrical charge. It is called *dynamic* because a DRAM system must frequently refresh the charge of its capacitors to maintain a stored value. It also requires error correction which introduces redundancy.

Device	**Capacity**	**Approx. latency**	**RAM type**
Register	4 - 8 bytes	< 1 ns	SRAM
CPU cache	1 - 32 megabytes	5 ns	SRAM
Main memory	4 - 64 gigabytes	100 ns	DRAM

Table 4: Memory hierarchy characteristics

Definition 6.3 (Secondary Storage)

Secondary storage devices are not directly accessible by the CPU and are used to store data that is not currently being processed. This includes hard drives, SSDs, and magnetic tapes. There are two primary ways:
1. **Spinning disks** store data on a magnetic surface that spins at high speeds.
2. **Solid state drives (SSDs)** store data on flash memory chips.

There are three key components of memory that we should think about:

1. The **capacity**, i.e. amount of data, it can store (how large the water tank is).
2. The **latency**, i.e. amount of time it takes for a device to respond with data after it has been instructed to perform a data retrieval operation (how fast the data flows).
3. The **transfer rate** or **thoroughput**, i.e. amount of data that can be moved between the device and main memory (how wide the pipe is). Naively, with one channel and sequential transfer the transfer rate is one over the latency.

We must provide a good balance of these three qualities, and also note that there are some physical limitations (i.e. latency cannot be faster than speed of light), and this is more effectively done through a hierarchical memory system.

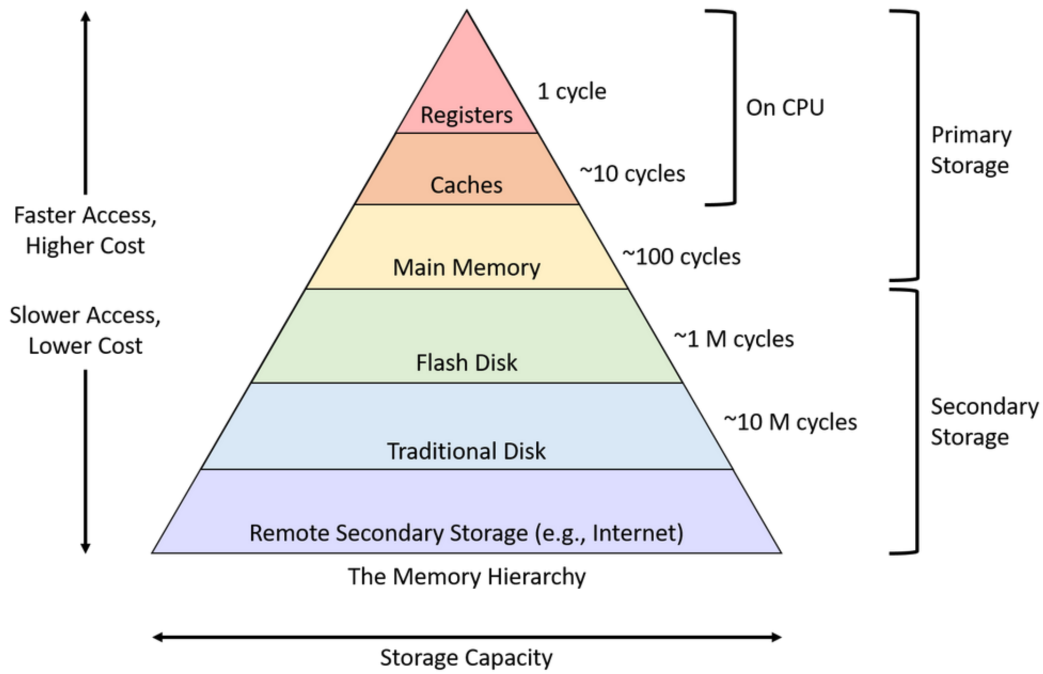


Figure 39: Memory hierarchy.

For example when we want to read from the disk, the CPU must request to the bus interface, which travels through the bus interface, I/O bridge, I/O bus, disk controller, and to the disk itself. Then the data goes back through the disk controller, I/O bus, I/O bridge, through the memory bus, and resides in the main memory. Note that disks are block addressed, so it will transfer the entire block of data into the memory. It must specify a **destination memory address (DMA)**. When the DMA completes, the disk controller notifies the CPU with an *interrupt* (i.e. asserts a special interrupt pin on the CPU), letting it know that the operation has finished. This signal goes through the disk controller to the IO bridge to the CPU. From now on, the CPU knows that there is memory that it can access to run an application loaded in memory.

6.2 Disk

Definition 6.4 (Hard Disk Drives)

Back then, there were **hard disk drives (HDDs)** that literally had a spinning wheel and a needle head that read the data.

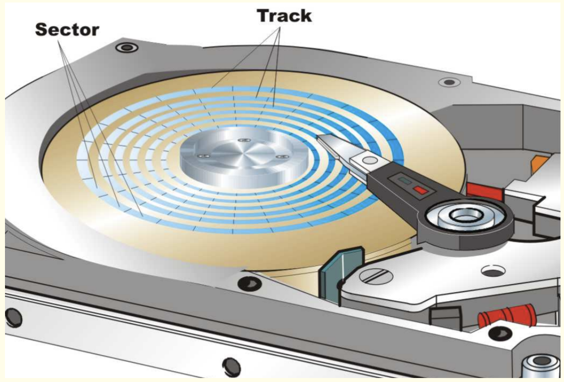


Figure 40: Visual diagram of hard disk drive with its sectors.

1. HDDs are not random access since the data must be sequentially read. This was disadvantageous since the spinning wheel had to spin to the correct location, which took time. The needle also had to move to the correct location, which also took time and therefore read and write speeds were dominated by the time it took to move the needle.
2. The smallest unit of data that can be read is a complete disk sector (not a single byte like RAM).

Definition 6.5 (Solid State Drives)

Now, we have **solid state drives (SSDs)** that store data on flash memory chips. This is advantageous since there are no moving parts, so the latency is much lower and the latency is not dominated by the time it takes to move the needle.

1. SSDs are random access.
2. The smallest unit of data is a **page**, which is usually 4KB and maybe for high scale computers 2-4 MB (but on "Big Data" applications big but computers, it can be up to 1GB).
3. A collection of pages, usually 128 pages, is called a **block**, making is 512KB.

While virtually all RAM and primary storage devices are **byte addressable** (i.e. you can access any byte in memory), secondary storage devices are **block addressable** (i.e. you can only access a block of memory at a time). Therefore, to access a single byte in secondary storage, you must first load the entire block into memory, calculate which byte from that block you want, and then access it. Therefore, you need both the block number x and the offset o to access a byte in secondary storage, which is why it is even slower than accessing RAM.

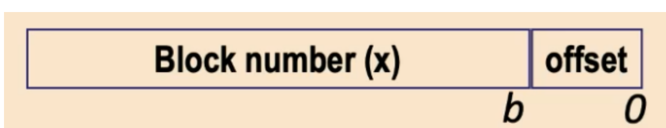


Figure 41: Block offset.

Therefore, you can think of raw data in units of blocks of size 2^b for some b bits.

1. Take the low order b bits of a byte address as an integer, which is the offset of the addressed byte in the block.
2. THe rest of the bits are the block number x, which is an unsigned long.
3. You request the block number x, receive the block contents, and then extract the requested byte at

offset in x i.e. calculate `block[x][offset]`.

6.3 Locality

So far, we have abstracted away most of these memory types as a single entity with nearly instantaneous access, but in practice this is not the case. The most simple way is to simply have RAM and our CPU registers, but by introducing more intermediate memory types, we can achieve greater efficiency.

Definition 6.6 (Locality)

Locality is a principle that generally states that a program that accesses a memory location n at time t is likely to access memory location $n + \epsilon$ at time $t + \epsilon$. This principle motivates the design of efficient caches.

1. **Temporal locality** is the idea that if you access a memory location, you are likely to access it again soon.
2. **Spatial locality** is the idea that if you access a memory location, you are likely to access nearby memory locations soon.

This generally means that if you access some sort of memory, the values around that address is also likely to be accessed and therefore it is wise to store it closer to your CPU. In CPUs, both the instructions and the data are stored in the cache, which exploits both kinds of locality (repeated operations for temporal and nearby data for spatial).

Example 6.1 (Locality)

Consider the following code.

```c
int sum_array(int *array, int len) {
  int i;
  int sum = 0;

  for (i = 0; i < len; i++) {
    sum += array[i];
  }

  return sum;
}
```

1. **Temporal Locality**
   (a) We cycle through each loop repeatedly with the same add operation, exploiting temporal locality.
   (b) The CPU accesses the same memory (stored in variables `i`, `len`, `sum`, `array`) within each iteration and therefore at similar times.
2. **Spatial Locality**
   (a) The spatial locality is exploited when the CPU accesses memory locations from each element of the array, which are contiguous in memory.
   (b) Even though the program accesses each array element only once, a modern system loads more than one `int` at a time from memory to the CPU cache. That is, accessing the first array index fills the cache with not only the first integer but also the next few integers after it too. Exactly how many additional integers get moved depends on the cache's **block size**. For example, a cache with a 16 byte block size will store `array[i]` and the elements in `i+1`, `i+2`, `i+3`.

We can see the differences in spatial locality in the following example.

Example 6.2 ()

One may find that simply changing the order of loops can cause a significant speed up in your program. Consider the following code.

```c
float averageMat_v1(int **mat, int n) {
  int i, j, total = 0;

  for (i = 0; i < n; i++) {
    for (j = 0; j < n; j++) {
      // Note indexing: [i][j]
      total += mat[i][j];
    }
  }
  return (float) total / (n * n);
}
```

```c
float averageMat_v2(int **mat, int n) {
  int i, j, total = 0;

  for (j = 0; j < n; j++) {
    for (i = 0; i < n; i++) {
      total += mat[i][j];
    }
  }
  return (float) total / (n * n);
}
.
```

Figure 42: Two implementations of taking the total sum of all elements in a matrix.

It turns out that the left hand side of the code executes about 5 times faster than the second version. Consider why. When we iterate through the `i` first and then the `j`, we access the values `array[i][j]` and then by spatial locality, the next few values in the array, which are `array[i][j+1]`, ... are stored in the cache.

1. In the left hand side of the code, these next stored values are exactly what is being accessed, and the CPU can access them in the cache rather than having to go into memory.
2. In the right hand side of the code, these next values are *not* being accessed since we want to access `array[i+1][j]`, ... Unfortunately, this is not stored in the cache and so for every n^2 loops we have to go back to the memory to retrieve it.

6.4 Caches

In theory, a cache should know which subsets of a program's memory it should hold, when it should copy a subset of a program's data from main memory to the cache (or vice versa), and how it can determine whether a program's data is present in the cache. Let's talk about the third point first. It all starts off with a CPU requesting some memory address, and we want to determine whether it is in the cache or not. To do this, we need to look a little deeper into memory addresses.

Definition 6.7 (Portions of Memory Addresses)

A memory address is a m-bit number.[a] It is divided up into three portions.

1. The **tag** field with t bits at the beginning.
2. The **index** field with i bits in the middle.
3. The **offset** field with o bits at the end.

The tag plus the index together refers to the **block number**.

Tag	Index	Offset
1010	0000011	00100

Figure 43: Portions of a 16 bit memory address with $t = 4, i = 7, o = 5$.

[a] 64 in 64-bit machines.

Before we see why we do this, we should also define the portions of a CPU.

Definition 6.8 (CPU Cache)

A **CPU cache** divides its storage space as follows. A cache is essentially an array of sets, where S is the number of sets. Each set is divided into E units called **cache lines/rows**, with each cache line independent of all others and contains two important types of information.

1. The **cache block** stores a subset of program data from main memory, of size 2^o.[a] Sometimes, the block is referred to as the cache line. Note that is the cache block size is 2^o bytes, then the block offset field has length $\log_2 2^o = o$.
2. The **metadata** stores the **valid bit** (which tells us if the actual data in memory is valid), and the **tag** of length t (the same as the tag length of the memory address) which tells us the memory address of the data in the cache.

Therefore, the **cache size** is defined to be $C = S \cdot E \cdot B$ (the metadata is not included).

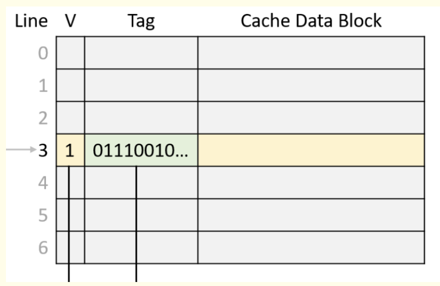


Figure 44: Diagram of a direct-mapped cache.

CPU caches are built-in fast memory (SRAM) that stores stuff. There are two types:

1. **i-cache** stores copies of instructions.
2. **d-cache** stores copies of data from commonly referenced locations.

We saw that caches come in different levels, they all just hold words retrieved from a higher level of memory.

1. CPU registers hold words retrieved from L1 cache.
2. L1 holds cache lines retrieved from L2 cache.
3. L2 cache holds cache lines retrieved from L3 cache or the main memory.
4. Main memory holds disk blocks retrieved from local disks.
5. Local disks hold blocks retrieved from remote disks or network servers.

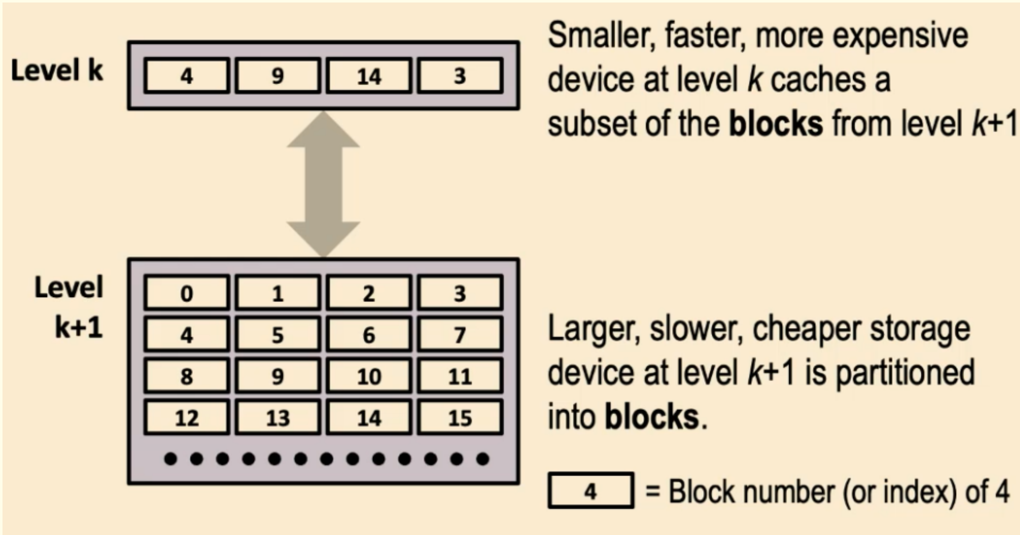


Figure 45: How caches retrieve data from higher levels of memory.

[a] In Intel computers, it is typically 64 bytes long and for Mac Silicon, it is 128 bytes.

Example 6.3 (Simple Calculations)

Given a direct-mapped cache specified by a block size of 8 bytes and a cache capacity of 4 KB,

1. the cache can hold 512 blocks.
2. the block offset field is $\log_2 8 = 3$ bits wide.
3. the address `0x1F = 0b00011111` is in block number 3 since the last three bits are the offset, and whatever is left (passed through the hashamp, which is simply modulo), is the block number.

In **I/O caches**, software keeps copies of cached items in memory, indexed by name via a hash table.

At the lowest level, registers are explicitly program-controlled, but when accessing any sort of higher memory, the CPU doesn't know whether some data is in the cache, memory, or the disk.

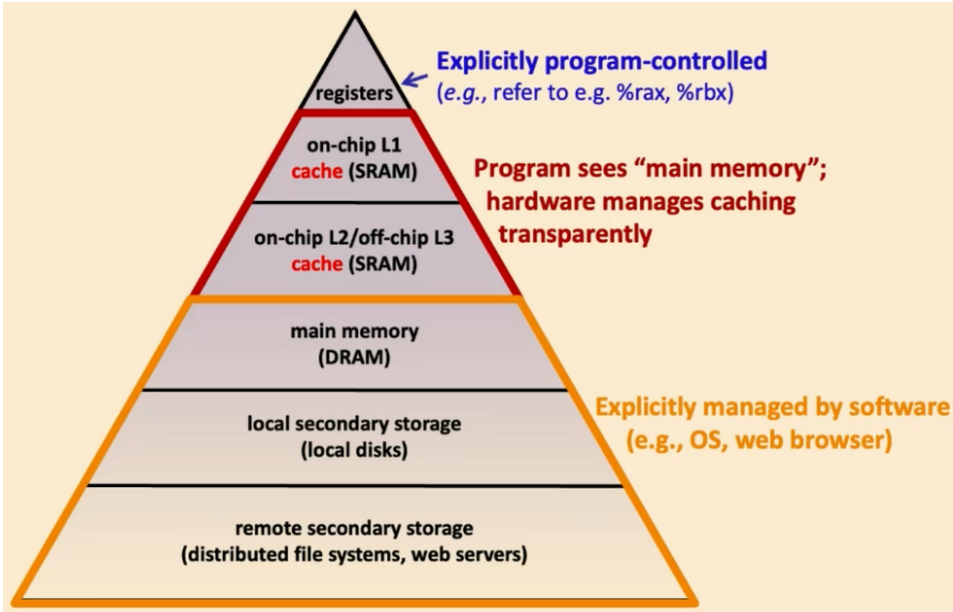


Figure 46

Finally, let's compare software vs hardware caches.

Definition 6.9 (Software Caches)

When implementing caches in software, there are large time differences (DRAM vs disk, local vs remote), and they can be tailored to specific uses cases. They also have flexible and sophisticated approaches with data structures (like trees) and can perform complex computation.

Theoretically, when implementing hash tables, you never actually have to evict something. You can have the values of the table to be a linked list where we add to the head. If there is unlimited chaining, we have a full associative cache, and if we have limited chaining (e.g. 5), it is like a 5-way set associative cache. If it goes out of bound, we can implement LRU by removing the tail of the linked list.

Definition 6.10 (Hardware Caches)

In hardware caches, there are smaller time differences, needs to be as fast as possible, and parallelization is emphasized.

There are slightly different implementations of caching, and for each implementation, we will describe

1. how to load data from memory into the cache,
2. how to retrieve data from the cache,
3. how to write data to the cache.

6.4.1 Direct Mapped Cache

A direct mapped cache is a caching implementation when we assume that $E = 1$, which means that for any given memory address, there is only one possible cache line that can store this data at that memory address. That is, the cache is really just a bunch of sets with one cache line each, and each cache line is completely isolated from the others. Whether we load data from memory into cache or try to retrieve data from the cache, it's really the same process.

Theorem 6.1 (Placement)

To load data from memory into the cache, which happens when there is a **cache miss**, we do the following.

1. The CPU requests a memory address $M = (T, I, O)$.
2. There exists a hashmap H that maps the index I to a cache line.
3. At line $H(I)$, we can get a cache miss and must load from memory into this cache.
4. We wait until the memory has retrieved the data from the portion of the memory. i.e. we wait for the 2^o bytes located at addresses $(T, I, 0 \ldots 0)$ to $(T, I, 1 \ldots 1)$. Call this data D.
5. The 2^o byte string D is stored in the cache data block at line $M(I)$,ready to be used.

Theorem 6.2 (Lookup)

To see whether a requested memory address is in the cache, we do the following.

1. The CPU requests a memory address $M = (T, I, O)$.
2. There exists a hashmap H that maps the index I to a cache line.
3. At line $H(I)$, check the cache line's valid bit. If it is not valid, then this is a cache miss and we must go to the memory to retrieve the data, leading to the above process.
4. Since there could be multiple I that maps to the same cache line, there will be overlap. But this is where the tag portion comes in. At cache line $H(I)$, the CPU checks the cache tag to see if it matches the memory tag T.
5. If it does, then we have just found a way to identify the first $t + i$ bits of the requested memory address, and we have gotten a cache hit. Now, we know that the cache's data block holds the data that the program is looking for. We use the low-order offset bits of the address to extract the program's desired data from the stored block.

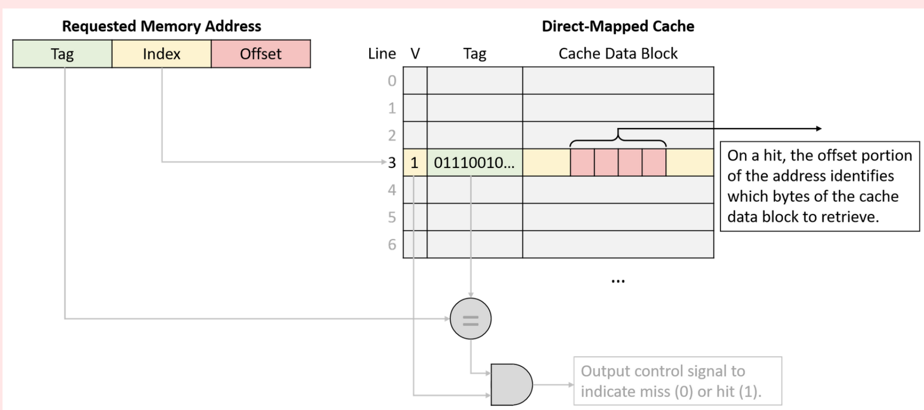


Figure 47: Diagram of a cache request. Note that since the entire data in the memory block stored in the cache, we can take advantage of spatial locality.

So far, we've talked about reading operations, but what about writing to the cache? It is generally implemented in two ways.

Definition 6.11 (Write-Through, Write-Back Cache)

Note that when we write data to cache, it does not need to be immediately written to memory, but rather it can be flushed to memory at a later time. This is efficient since if we have repeated operations on a single memory address, we don't have to go back and forth between the CPU and memory.

1. In a **write-through cache**, a memory write operation modifies the value in the cache and

simultaneously writes the value to the corresponding location in memory. It is always synchronized.

2. In a **write-back cache**, a memory write operation modifies the value stored in the cache's data block, but does *not* update main memory. Instead, the cache sets a **dirty bit** in the metadata to indicate that the cache block has been modified. The modified block is only written back to memory when the block is replaced in the cache.

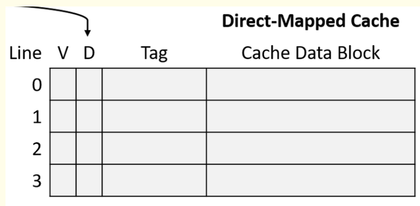


Figure 48: A dirty bit is a one bit flag that indicates whether the data stored in a cache line has been modified. When set, the data in the cache line is out o sync with main memory and must be written back (flushed) back to memory before eviction.

As usual, the difference between the designs reveals a trade-off. Write-through caches are less complex than write-back caches, and they avoid storing extra metadata in the form of a dirty bit for each line. On the other hand, write-back caches reduce the cost of repeated writes to the same location in memory.

Theorem 6.3 (Replacement)

Replacement occurs exactly the same way as if we just did a placement and is trivial. We retrieve the data block from the memory and store it in the cache. Direct-mapping conveniently determines which cache line to evict when loading new data. Given new memory $M = (T, I, O)$, you *must* evict the cache line at $H(I)$.

6.4.2 N way Set-Associative Cache

Note that for both examples, given a fixed hashmap H it is not possible to store data in two memory addresses M_1 and M_2 where both $H(I_1) = H(I_2)$. Therefore, the choice of hashing must be done so that it minimizes the number of collisions. So far, we have only considered memory read operations for which a CPU performs lookups on the cache. Caches must also allows programs to store values. However, there is a better way to do this: just construct it so that each set has more than one cache line, and so data in index portions of different memory addresses can be stored in different cache lines.

In here, we deal with $E \neq 1$, and so there are multiple set each with multiple lines. This means that the cache is more like a 2D array, and when we want to retrieve an index, we must look through the $H(I)$th line in *each* set to see if the tag matches.

Theorem 6.4 (Lookup)

To see whether a requested memory address is in the cache, we do the following.

1. The CPU requests a memory address $M = (T, I, O)$.
2. We iterate through each of the S sets in the cache, looking at cache line $M(I)$.
3. For each line, we check if it is valid and if so, whether the line tag matches the memory tag. If we get a hit, then we have found the data in the cache.

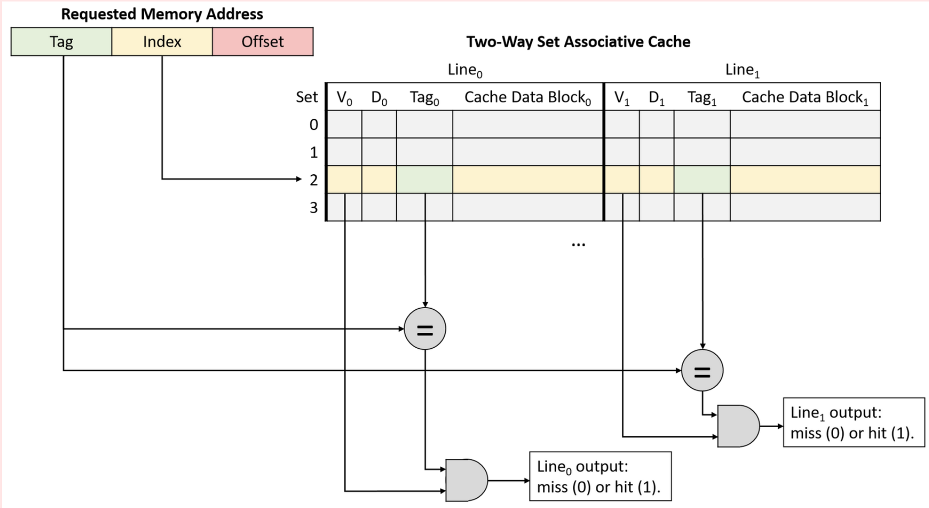


Figure 49: Diagram of a 2 set-associative cache.

If you have a **fully associative cache**, then you have one set with $E = C/B$ lines. Therefore, you can really put any memory address data in any cache line. There is a clear tradeoff here. As we increase N, we can get more flexibility in using all of our cache space, but the time complexity of retrieving and writing data scales linearly. In fact, this linear scan is too slow for a cache, which is why you need to implement some parallel tag search, but this turns out to be quite expensive to build.[3]

Though we have a more robust implementation with associative mapping, placement and replacement now face the problem of *which* set to place the data in or evict existing data.

Theorem 6.5 (Placement)

To load data from memory into the cache this is trivial since we can just go through the sets, find one where the valid bit is 0, and just place the data there.

In replacement, this is a bit trickier, but using the principle of temporal locality, we can try and replace the least recently used cache. This tries to minimize cache misses, but not slow down the lookup too much.

Theorem 6.6 (Replacement)

To replace data on the cache, we use the **least recently used (LRU)** algorithm. This matches temporal locality, but it also requires some additional state to be kept.

6.4.3 Types of Cache Misses

There are three types of cache misses.

Definition 6.12 (Cold (Compulsory) Miss)

A **cold miss** occurs when the cache is empty and the CPU requests a memory address. This is the first time the CPU is requesting this memory address, and so it must go to the memory to retrieve the data.

[3]You have to copy the request tag with a circuit and compare it to all the tags in the cache, which turns out to be a much larger circuit.

Definition 6.13 (Capacity Miss)

A **capacity miss** occurs when the cache is full and the CPU requests a memory address that is not in the cache. This is because the cache is full and so the CPU must evict some data to make space for the new data.

Definition 6.14 (Conflict Miss)

A **conflict miss** occurs from premature eviction of a warm block.

Valgrind's cachegrind mode.

7 Operating Systems

Up until now, we've seen the dynamics of how one program works in a computer system. The code, which first resides in the disk, is fetched (through blocks) into memory, and after compiling (precomiling, compiling, assembling, linking), we have a binary. The binary is then loaded into memory in the stack frame, and the CPU executes the instructions. The CPU also has a cache, which stores the most frequently accessed data during the process, taking advantage of locality for efficiency.

Our computer obvious does not just run one program. It runs several, and to run several, we need some control mechanism to manage how these programs interact with the CPU, memory, and disk. For example, one problem is that if we download application A and application B and run their binaries, how do we know whether they share memory addresses and consequently overwrite each other's data?[4] The operating system takes care of these, which manages *processes* that each have their own *virtual memory space*.

Furthermore, consider some of the components of the computer: the RAM, disk, and IO devices like your keyboard and monitor. For security reasons, it is not wise to let the user applications (e.g. Chrome or Slack) control these devices completely. Their power must be restricted in some way.

1. When you have a Chrome window and resize it, Chrome should not be able to modify the pixels outside that window.
2. When you want to print some statement using `printf`,
3. When you're editing a code file with VSCode, you want to limit the application to save to certain parts of the disk.
4. When you are running Chrome and Slack together, you don't want them to read each other's data directly.

This is also for convenience. Say that if you are creating an application that has the option to save files to disk, you don't want to write the hardware backend to write to the disk. You want to just call a function that writes to the disk, and the OS will take care of the rest.

Definition 7.1 (Operating System)

A common confusion is that people think that the **operating system** describes the computer itself, but it is really just another piece of software. What makes this piece of software so special is that it manages every other software in the computer. It provides generally three services:

1. It **multiplexes** the hardware resources. Since there are many applications/programs with finite CPU resources (number of cores) and shared access to storage devices, the OS schedules some sharing mechanism for execution time on CPU cores and manages access to storage devices.
2. It **abstracts** the hardware platform. Since each CPU core simply executes a sequence of instructions, the OS introduces processes and thread abstractions. Furthermore, it introduces

[4] This is different from linking, where we have relocation tables to ensure that *object files* do not conflict with each other.

filesystems (file/directories) on top of raw storage devices.

3. It **protects** software principals from each other. Since many applications from various users are using the CPU, the OS provides isolation between them. It enforces user access permission (read/write) for files.

The OS is booted by the system firmware (BIOS or UEFI), which lives in ROM (sRAM and therefore non-volatile) and copies the OS from a fixed part of the disk, called the **bootloader**, into the RAM, which itself then loads the OS into memory. Once the OS starts running, it loads the rest of itself from disk, discovers and initializes hardware resources, and initializes its data structures and abstractions to make the system ready for users.

Definition 7.2 (Kernel)

The **kernel** is the actual binary that is loaded into RAM that runs the OS. The kernel code and data resides in a fixed and protected range of addresses, called the **kernel space**, and user programs cannot access kernel space.

7.1 Control Flow

When we worked with jumps (conditional and unconditional), calls, and returns in assmebly, all of these operations were with respect to the **program state**, which is the isolated environment that the program is in. One program doesn't have any clue of what is going on anywhere else, such as other programs or input/output signals. This means that given what we have learned,

1. programs cannot to write files to the disk (since that is outside the program).
2. programs cannot be terminated by pressing `CTRL + C` on the keyboard.
3. programs cannot receive data that arrives from the disk.
4. programs cannot send data to the monitor to display.
5. programs cannot react accordingly when there is an instruction to divide by 0.[5]

To do all these things, we need to have access to the global **system state**, which the OS has access to.

It turns out that it is impossible for jumps and procedure calls to achieve this, and rather the system needs mechanisms for **exceptional control flow** (i.e. control flow that is not within the regular program state), or commonly referred to as **exceptions**. This requires the CPU to enter into a more powerful state than its current place in the program state, called the kernel state. The actual thing that triggers this is called an **interrupt**, which can come from both the hardware and software. In the kernel state, the CPU can access the hardware and perform operations that the program state cannot to handle these exceptions.

Example 7.1 (Interrupts)

Some examples of how the OS can be interrupted is:
1. when one's WiFi card detects a signal.
2. a hard disk drive may interrupt the OS if a read fails due to a bad sector.
3. an application may request a system call to open a file.
4. If you have 10 applications running on 1 CPU core, you may want the CPU core to run to the next application every 10 milliseconds. So, there may be a system call every 10 milliseconds in each program to the OS to switch to the next application.

[5] I guess you can use a conditional jump to check if the divisor is 0 and then jump to a different part of the code.

Definition 7.3 (Execution Modes)

The CPU helps with this by providing two execution modes, which is determined by a special bit in the CPU called the **mode bit**.

1. In **user mode**, the CPU executes only user-level instructions and accesses only the memory locations that the OS makes available to it. It also restricts which hardware components the CPU can directly access.
2. In **kernel mode**, the CPU executes any instructions and accesses any memory location (including those that store OS instructions and data). It can also directly access hardware components and execute special instructions.

Note that the execution mode is *property of the CPU*!

Example 7.2 (Monitor)

A monitor is really just some device that scans a certain portion of memory at a certain frequency that is higher than the human eye can detect. In user mode, if you try to access this memory buffer, you get an exception. No user mode can access this memory buffer.

Example 7.3 (Amazon.com)

When you are on Amazon to search up some product, you want to type in some keyword in the search bar. The web browser, say Chrome, that you are running it on, runs in user mode. When you type in the keyword, Chrome sends a system call to the OS, triggering the kernel mode which retrieves the keys that you pressed, and redirects it to Chrome. The same goes with the location of your mouse. When you move and click on a product, Chrome sends a system call to the OS, which then receives the mouse location and sends it back to Chrome. The application has no way to directly access the hardware.

Now specifically, how does one enter in this kernel mode? We've already hinted at it before, but to elaborate, there are 4 types of exceptions.

Definition 7.4 (Types of Exceptions/Interrupts)

As we have mentioned, we go into kernel mode through exceptional control flows. To go back from the kernel mode to the user mode after the exception handling is done, the kernel must explicitly give back the control to the user program, which is done with a special instruction, which changes the CPU to the user mode again. At this point, it can return back to user mode at the current instruction, next instruction, or abort it.

1. These control flows can either by **synchronous** (caused by an instruction) or **asynchronous** (caused by some other event external to the processor). Asynchronous interrupts are indicated by setting the processor's interrupt pins.
2. Furthermore, **intentional** exceptions transfer control to the OS to perform some function, and **unintentional** exceptions happen when there is a bug.

This gives us 4 categories of exceptions.

1. Intentional synchronous exceptions are **system calls**, aka **traps** (e.g. `printf`, `open`, `close`, `write`, breakpoint traps, special instructions). It returns control to the next instruction.
2. Unintentional synchronous exceptions are **faults** (possibly recoverable) or **aborts** (unrecoverable) (e.g. invalid or protected address or opcode, page fault, overflow, divide by zero). This automatically triggers the kernel mode which then uses an exception handler to kill the process.
3. Intentional asynchronous exceptions are **software interrupts**, which is when software requests an interrupt to be delivered at a later time (e.g. there's some task you want the kernel to do later).
4. Unintentional asynchronous exceptions are **hardware interrupts** caused by an external event

(e.g. IO such as `CTRL + C`, op completed, timers which may switch to another application every 10ms, power fail, keyboard, mouse click, disk, receiving a network packet). Unlike system calls, which come from executing program instructions, hardware interrupts are delivered to the CPU on an **interrupt bus**.

Once a system call or hardware interrupt is finished, the program continues to resume back in user mode.

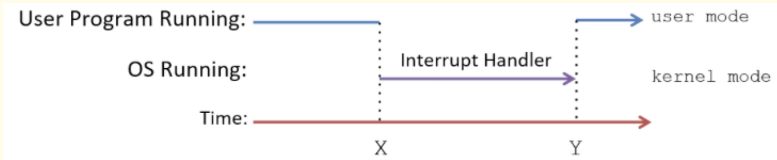


Figure 50: The CPU and interrupts. User code running on the CPU is interrupted (at time X on the time line), and OS interrupt handler code runs. After the OS is done handling the interrupt, user code execution is resumed (at time Y on the time line).

Now the question arises: how does the CPU know where to go when an system call or interrupt occurs? These are done through tables that map some unique ID number to some functionality. These tables are stored in a protected memory space reserved by the kernel.

Definition 7.5 (System Call Table)

This is done through the **system call table**, which is a table of addresses in memory that the CPU can jump to when a system call occurs. Each system call has a unique number k, and the handler function k is called each time system call k occurs.

Example 7.4 (Common System Calls)

Some common system calls, or **syscalls**, are shown below with their unique ID number (in Linux x64).

Table 5: System Call Functions

Number	Name	Description
0	read	Read file
1	write	Write file
2	open	Open file
3	close	Close file
4	stat	Get info about file
57	fork	Create process
59	execve	Execute a program
60	_exit	Terminate process
62	kill	Send signal to process

Example 7.5 (Syscalls of Open)

Look at the following objdump file below. The corresponding C code just calls `open(filename, options)` and the corresponding syscall ID is `0x2`. We are simply loading the syscall ID into the `%eax` register (only needs last 32 bits since the syscall IDs are quite small), which is then executed by the `syscall` instruction to go into the kernel mode.

```
00000000000e5d70 <__open>:
...
e5d79:  b8 02 00 00 00       mov    $0x2,%eax     # 2 is the open syscall number
e5d7e:  0f 05                syscall              # return value in %rax
e5d80:  48 3d 01 f0 ff ff   cmp    $0xfffffffffffff001,%rax
...
```

A negative number in `%eax` gives an error corresponding to negative `errorno`. It is also worth mentioning that `%eax` is used rather than `%rdi` or `%rsi` because we need these two parameter registers as arguments for the `open` function itself.

Note that whether we are in the program stack or the kernel stack, we always have stack pointers and other registers to navigate them. In fact, for every CPU core, it has its own set of registers and its own kernel stack.

Example 7.6 (Syscall of Read)

If we have read syscall, then

1. We use the syscall table to go to the trap handler for the read syscall.
2. The handler identifies the block and allocates a buffer.
3. Then it reads the block from the disk, which may take a while (in CPU time) since it is extremely slow for all IO tasks. The CPU, while waiting, can be put to sleep for other processes to run on the CPU. When the disk is done reading, it (the hardware) can send a hardware interrupt to the CPU, telling it that it is done.
4. Then it copies the block to the user buffer and returns from the syscall back into the user mode in the program state.

Definition 7.6 (Exception Table)

This is done through the **exception table**, which is a table of addresses in memory that the CPU can jump to when an exception occurs. Each type of event has a unique exception number k, and the handler function k is called each time exception k occurs.[a]

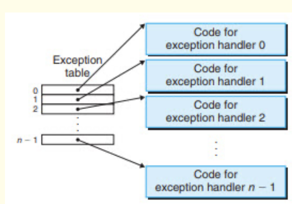


Figure 51: System call table is stored in a protected memory space reserved by the kernel.

[a] This is similar to a hardware implementation of a switch statement in C.

Example 7.7 (Common Exception Numbers)

Some common exception numbers are listed below.

Table 6: Exception Summary

Exception Number	Description	Exception Class
0	Divide Error	Fault
13	General protection fault	Fault
14	Page fault	Fault
18	Machine check	Abort
32-255	OS-defined	Interrupt or trap

From the application's point of view, even if an interrupt happens, it just thinks it is running line by line.

Definition 7.7 (Process Address Space)

Interrupts can happen at any time, and one way to efficiently support this execution context switch from user mode to kernel mode is to do the following. At boot time, the OS loads its kernel code at a fixed location in RAM. Every time you create a new program state, the OS initializes a CPU register with the starting address of the OS handler function. On an interrupt, the CPU switches to kernel mode and executes OS interrupt handler code instructions that are accessible at the top addresses in every process's address space. Because every process has the OS mapped to the same location at the top of its address space, the OS interrupt handler code is able to execute quickly in the context of any process that is running on the CPU when an interrupt occurs. This OS code can be accessed only in kernel mode, protecting the OS from user-mode accesses; during regular execution a process runs in user mode and cannot read or write to the OS addresses mapped into the top of its address space.[a]

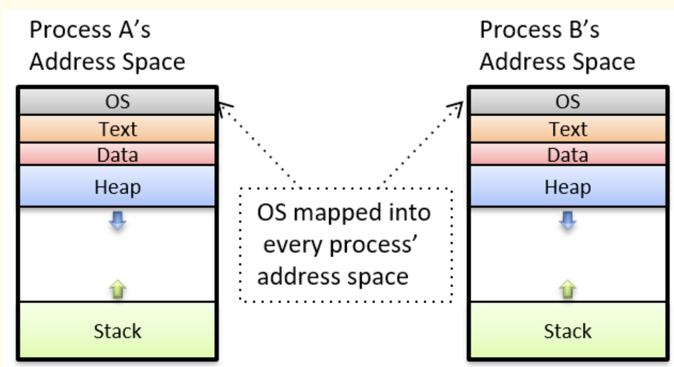


Figure 52: Process address space: the OS kernel is mapped into the top of every process's address space.

[a] However, due to security reasons where the user space can read kernel space data, this is obsolete.

In summary, a good visual is that each program runs as independent processes, with its own virtual address space (elaborated next) and the OS mediates access to shared resources.

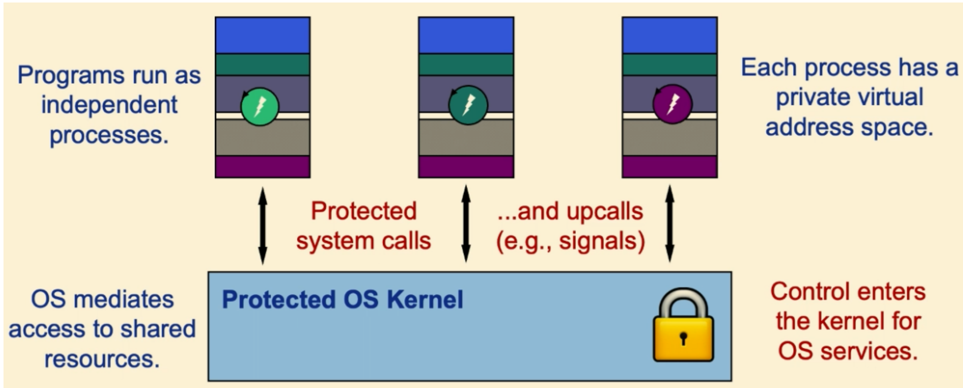


Figure 53: Multiple programs running and controlled by an operating system.

Each process can be in one of three states. It can either be currently running on the state, ready to run, or if there is a long IO operation, it can be blocked, which is then unblocked with a hardware interrupt. Usually anything that involves IO puts the state to blocked (e.g. reading data from disk, the keyboard, or the internet). The pool of processes that are concurrently running is the running and ready states.

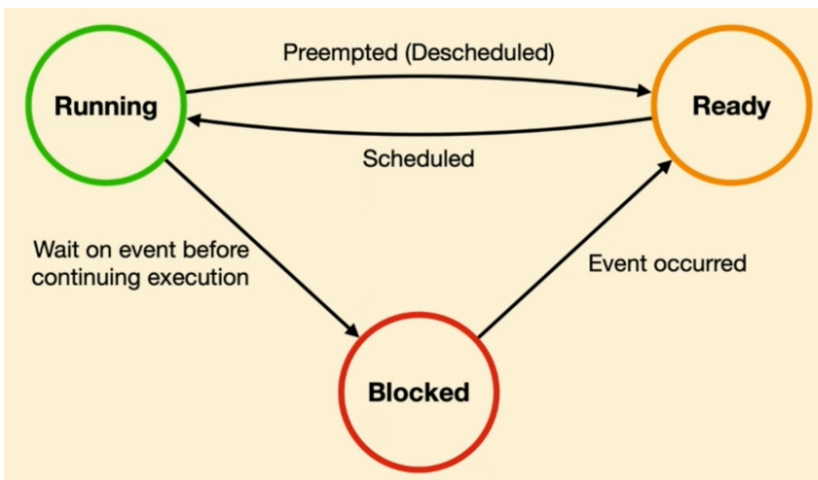


Figure 54: Three states that a single process can be in. The pool of processes that are concurrently running is the running and ready states. The blocked state is waiting to be put back into this pool by a hardware interrupt.

Example 7.8 (Running a Binary)

Therefore, to run a binary file `a.out`,

1. The kernel first loads the binary file from disk into RAM.
2. Then the OS kernel creates a new process with its own virtual memory stack and its global variables, etc.
3. Then the CPU's `%rip` register point to the address of the `main` function.
4. The kernel's virtual memory space is mapped to the top of the process's virtual memory space, where it is not visible to the user mode.

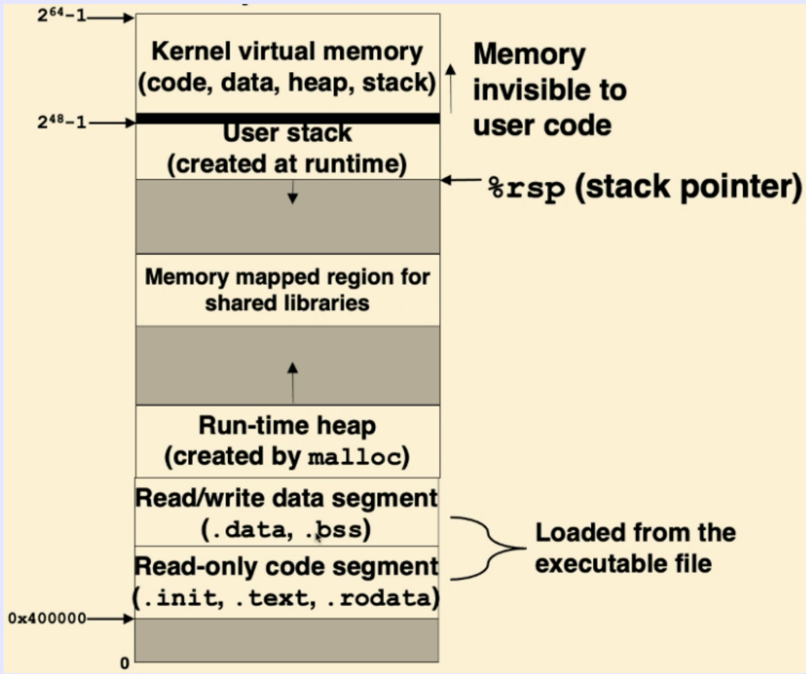


Figure 55: The kernel's virtual memory space is mapped to the top of the process's virtual memory space.

7.2 Virtual Memory

We have mentioned that there is a problem where two different application developers, who have linked their own C files to create binaries, can be installed on one computer and run at the same time. However, the linking has already been finished and the memory addresses of the symbols in each executable are fixed. This can be a problem if there are overlaps in the memory addresses.

Definition 7.8 (Virtual Memory)

The actual main memory of our system is referred to as the **physical memory**. To prevent such overlaps, the kernel and each user process has its own **virtual memory**. That is, there exists a **memory management unit (MMU)** in the CPU that translates virtual addresses to physical addresses through a hashmap.

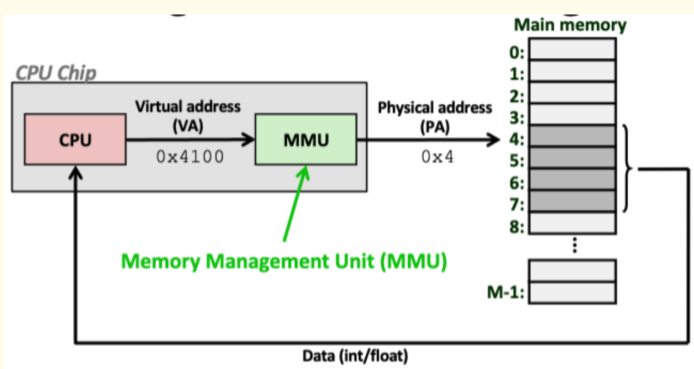


Figure 56: Memory management unit maps each virtual address to a physical address.

This allows the kernel to map the virtual memory of each process to the physical memory.

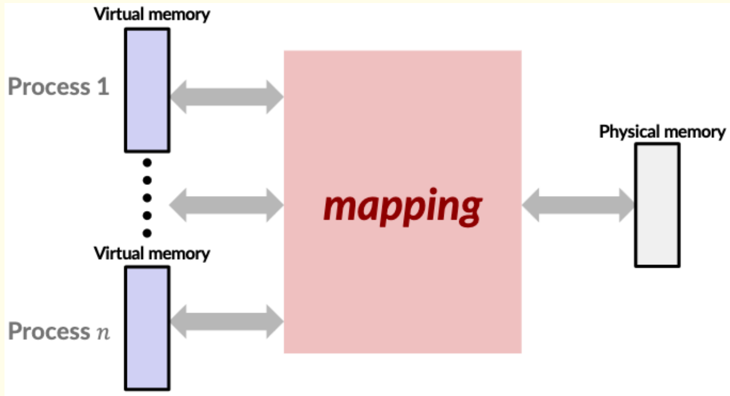


Figure 57: Each process has its own virtual memory space, which is mapped by the MMU to the physical memory space.

Example 7.9 (Virtual and Physical Memory Size)

Given a n-bit machine with 2^m-bytes of memory, $n > m$ and so there are more virtual addresses than physical addresses. If we have a 64-bit machine with 16GB of memory, then there are 2^{64} virtual addresses and $2^3 \cdot 2^{34} = 2^{37}$ bits of physical memory. If there are 8 processes running then there are $8 \cdot 2^{64} = 2^{67}$ bits of virtual memory.

There are many properties of virtual memory that solves a lot of problems and makes things more convenient. The main property is called **indirection** which means that the virtual memory is not the actual physical memory.

1. The first problem is that there are much more virtual addresses than physical addresses. Even storing a table for one process would take up more than all of your RAM. Therefore, for every byte in main memory, there exists one physical address (PA) and zero, one, or more virtual addresses (VA). We will elaborate on the specifics of this implementation later.

2. We also need to have memory management. Every process has its own stack, heap, `.text`, and `.data` sections. We must be able to allocate and deallocate memory and fit this accordingly.

3. We also need to have protection. We need to ensure that one process cannot read or write to another process's memory.

4. While we want isolation, we also want sharing between processes if needed (e.g. signing into Slack using Google on a browser). Furthermore, if there are multiple calls of the `printf` function, we can just have a single copy of the `printf` function in memory rather than having multiple copies for each process. This can be done through the concept of permissions.

Let's talk about how we should actually map these addresses. One property of this mapping is that we want contiguous addresses both in the virtual and the physical level so that we can store arrays, exploit locality, etc. Therefore, we can use larger blocks known as *pages*. Just like how we have divided memory addresses into sections that can be used to map to caches, we can divide the memory addresses into sections that can be used to map to the physical memory. Note that this also takes care of the first problem partially since now we can fit this table in the memory.

Definition 7.9 (Page)

Both in virtual and physical memory, an n-bit address can be divided into a **page number** and an **offset**. The page number is $n - 12$ and the offset is 12 bits. The page number is used to index into a **page table** that maps the page number to a physical address.

Figure 58: A page is a contiguous block of memory addresses.

While the entire page table is stored in memory (at memory stored by a protected CPU register), a portion of the page table is stored in the CPU cache.

1. The virtual page number (VPN) is equivalent to the block number.
2. The page offset is equivalent to the block offset.

Example 7.10 (Page Number)

In a 64-bit machine with 16GB of RAM, you have $2^{64}/2^{12} = 2^{52}$ virtual pages and $2^{37}/2^{12} = 2^{25}$ physical pages.

Therefore, our translation table is really a map from a virtual page number to a physical page number, rather than a virtual address to a physical address. This is created at runtime. Therefore,

1. The virtual page number VP is mapped through some map M to get the physical page number PP.
2. The virtual offset is the same as the physical offset.

Definition 7.10 (Page Table)

The **page table** is a hashmap that maps the virtual page number to the physical page number defined as the mapping

$$H : \underbrace{(VP, m)}_{\text{64 bits}} \longrightarrow PP \tag{11}$$

Each input-output pair is called a **page table entry (PTE)**, and virtual memory is **fully associative**, meaning that any virtual page can be placed in any physical page, though it requires a large mapping function (the PT), which is different from CPU caches.

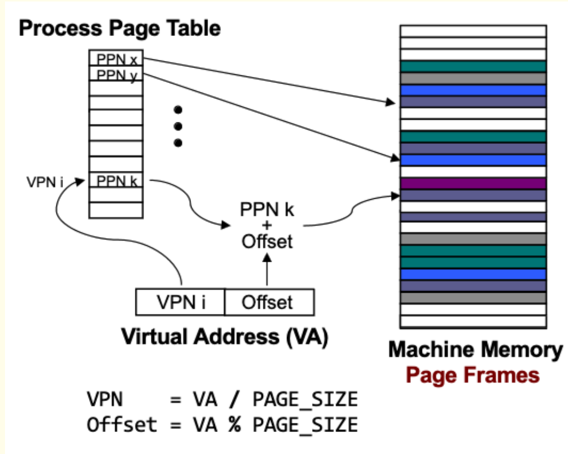


Figure 59: The page table only needs the virtual page number plus the metadata to map to the physical page number. The offset is provided by the virtual memory address itself.

Note that while we want to store the 52-bit VP in the page table, the actual input is still 64-bits, with 12 bits of metadata m. This metadata contains some information about the following

1. A bit that indicates whether the page is a read, write, or executable piece of code (3 bits).
2. A bit that indicates whether the page is valid or not.

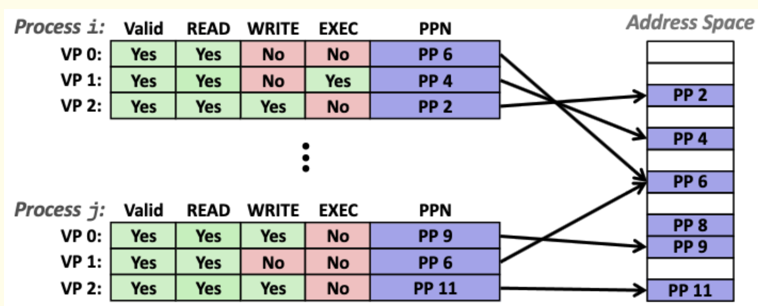


Figure 60: The page table entry contains the physical page number and some metadata.

Therefore, if you malloc, you are really just allocating some virtual memory addresses, which then get mapped to physical memory addresses in one or more pages.

Definition 7.11 (Page Fault)

It is clear that not every virtual page number can be mapped to a physical page number. If it turns out that a **page fault** happens if

1. the virtual page number maps to no physical page (i.e. is not in the page table) in the RAM
2. if some user program tries to access a physical page owned by the kernel
3. if the page number maps to some place in the disk (but it is not in physical RAM)

Page faults can be used in a lot of creative ways, but to reduce the risk of a page fault, e.g. when running out of physical memory, we can move some physical pages into disk and allocate memory by creating a new entry in our page table that maps this application's virtual page into the now empty physical page.

Note that by this construction, instructions that are contiguous in virtual memory may not be contiguous in physical memory. This may seem like it defeats the purpose of locality, but for most purposes, the 4KB page size will be enough to exploit it. We also see that malloced addresses in the heap (while we have learned that they were higher on the stack on higher addresses), are not necessarily in higher addresses in physical memory. Therefore, physical memory is scattered, and this is good since you don't need a giant contiguous block of memory to run large programs; you can divide it up into multiple physical pages.

Definition 7.12 (Swap Space)

Sometimes, the memory might not be in physical memory. Since memory is constrained (e.g. only 16GB), if we initialize a large array in the stack or global data, we may run out of memory. Therefore, the OS can flush out some physical pages in memory to disk, which is called **swapping**. The portion of the disk space that can be used in swapping is called the **swap space**.

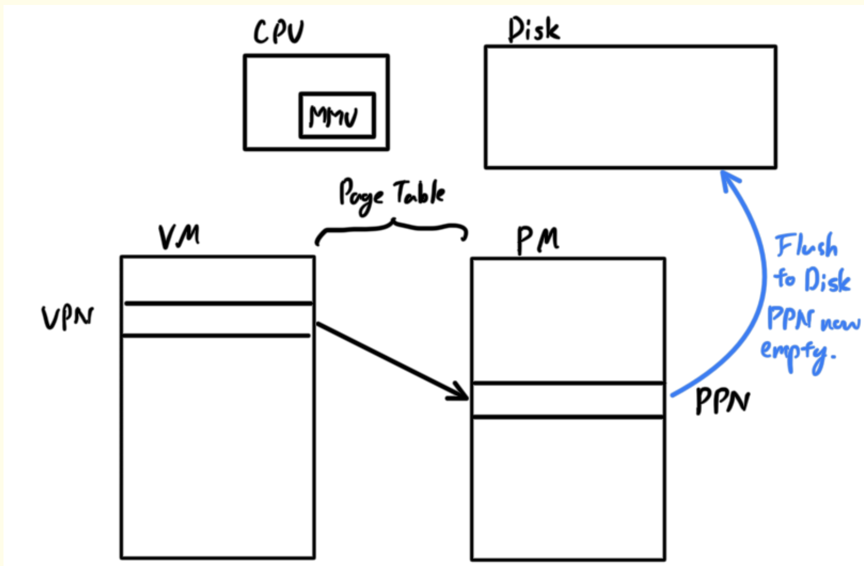


Figure 61: Swapping out physical pages to disk.

This allows us to abstract software into having almost infinite memory. Another important property is that swapping is **write-back** rather than write-through. We really don't want to write to disk every time we modify memory, so some thing may never end up on the disk (e.g. stack for short-lived processes). This is why when we open a file in C or Python, you may have to call `close()` since that will flush the memory to disk.

Example 7.11 (Page Fault)

When we swap out a physical page to disk, the physical page is now empty and accessing the virtual memory at this page table will cause a page fault. Say, when we want to write to a memory address that is swapped into the disk. The following will happen.

1. You execute code normally in user mode.
2. Then you try to write to a memory address that is swapped out, say through a `mov` operation. Say it is the following assembly code.

```
80483b7: c7 05 10 9d 04 08 0d          movl $0x0,0x8049d10
```

   This raises a page fault, an exception, and so the OS goes into kernel mode.
3. The kernel then finds the location of this physical page in the disk. The implementation is OS-specific (e.g. you can store some metadata).
4. Then it must copy the page back from disk into memory, and it may also have to swap out

some other physical page to disk to make space if needed.

5. Then the OS goes back into user mode, which now has access to the relevant memory in disk.

Ultimately, the moving operation is called twice. The first time it fails in user mode, and the second time (after the kernel mode, but now back to user mode) it succeeds. Note that this is different from a system call, which returns back to the *next* instruction. This call returns to the current instruction.

Definition 7.13 (Page Sharing)

This also makes protection and sharing to be quite nice. Given two virtual pages VP_1 and VP_2, owned by two different processes, we can have them share information by mapping to the same physical page PP.

Section	Read	Write	Execute
Stack	1	1	0
Heap	1	1	0
Static Data	1	1	0
Literals/const	1	0	0
Instructions	1	0	1

Table 7: Permissions for different sections of virtual memory.

Example 7.12 (Page Sharing Between Two Applications)

Furthermore, we can set process 1 to have only read permissions and process 2 to have read/write permissions. Therefore, say Google Chrome (process 2) can write your password into some memory, and then Slack (process 1) can read it, copy it into the CPU, and do stuff with it.

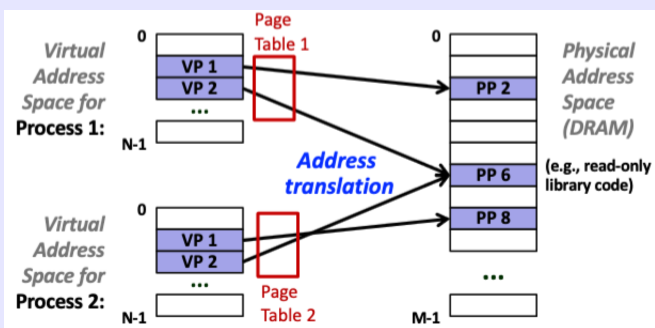


Figure 62: Sharing of data between two processes.

Now that we see how memory is swapped in the backend, we can see why larger memory can sometimes mean faster programs and why thrashing occurs.

Definition 7.14 (Thrashing)

The set of virtual pages that a program is "actively" accessing at any point in time is called its **working set**.

1. If the working set of one process is less than physical memory, then there is good performance for one process.
2. If the working set of all processes is greater than physical memory, then we have **thrashing**, which is a performance meltdown where pages are swapped between memory and disk continuously, and the CPU is always waiting or paging.

Example 7.13 (Computation Exercise)

Suppose that you have 16 KiB pages, 48-bit virtual addresses, and 16 GiB physical memory. How many bits wide are the following fields?

1. Virtual page number : $48 - 14 = 34$ bits.
2. Virtual page offset : 16 KiB is 2^{14} bytes, so we need 14 bits.
3. Physical page number : 16 GiB is 2^{34} bytes, so we need $34 - 14 = 20$ bits.
4. Physical page offset : 16 KiB is 2^{14} bytes, so we need 14 bits.

Furthermore, we have

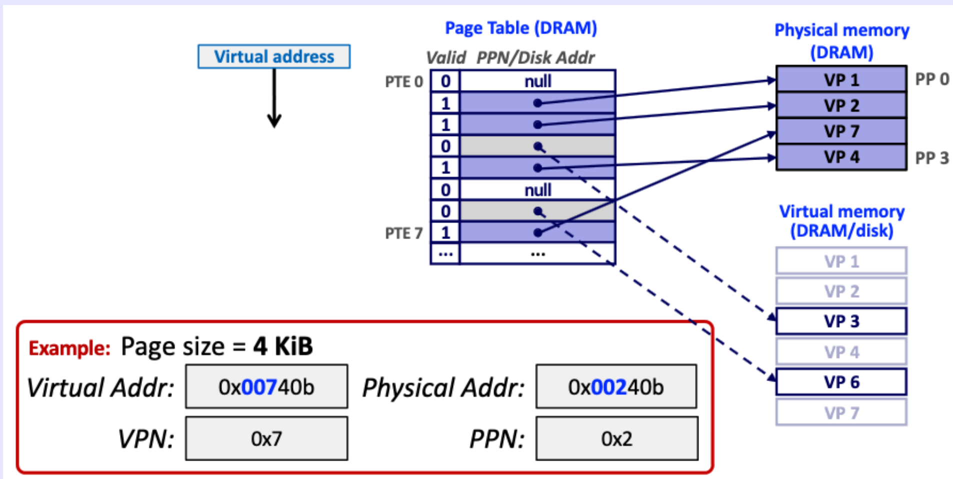


Figure 63: Given the virtual address, we can figure out the physical address, VPN, and PPN easily.

8 Shared Memory and Concurrency

So far, we've talked about everything as a sequential process of instructions. In practicality, we have improved this from the memory perspective by implementing caches, virtual memory, and swapping, but in the CPU perspective. In CPUs, we can't just simply increase the clock frequency indefinitely since there are physical limitations.[6] The current trend is to increase parallelism to compute faster, which is implemented with cores and threads. Let's clear some of these definitions up.

Definition 8.1 (Processors, Cores)

In almost every consumer computer, there exists one **processor** (CPU) in it. The CPU can have multiple **cores**. Each core has its own set of registers, L1/L2 cache, and possibly even a shared L3 cache.

Now these cores must run a certain program. Let's define what this means exactly.

Definition 8.2 (Program, Process)

A **program** can be thought of as a binary executable produced after linking. A **process** is a running instance of some program.

1. It is identified by a **process ID (PID)** number.
2. It is run on a CPU core, with its own registers.
3. It has its own virtual address space, containing the code (instructions), heap, and pagetable

[6] It turns out that power consumption increases faster than clock frequency, so it scales badly.

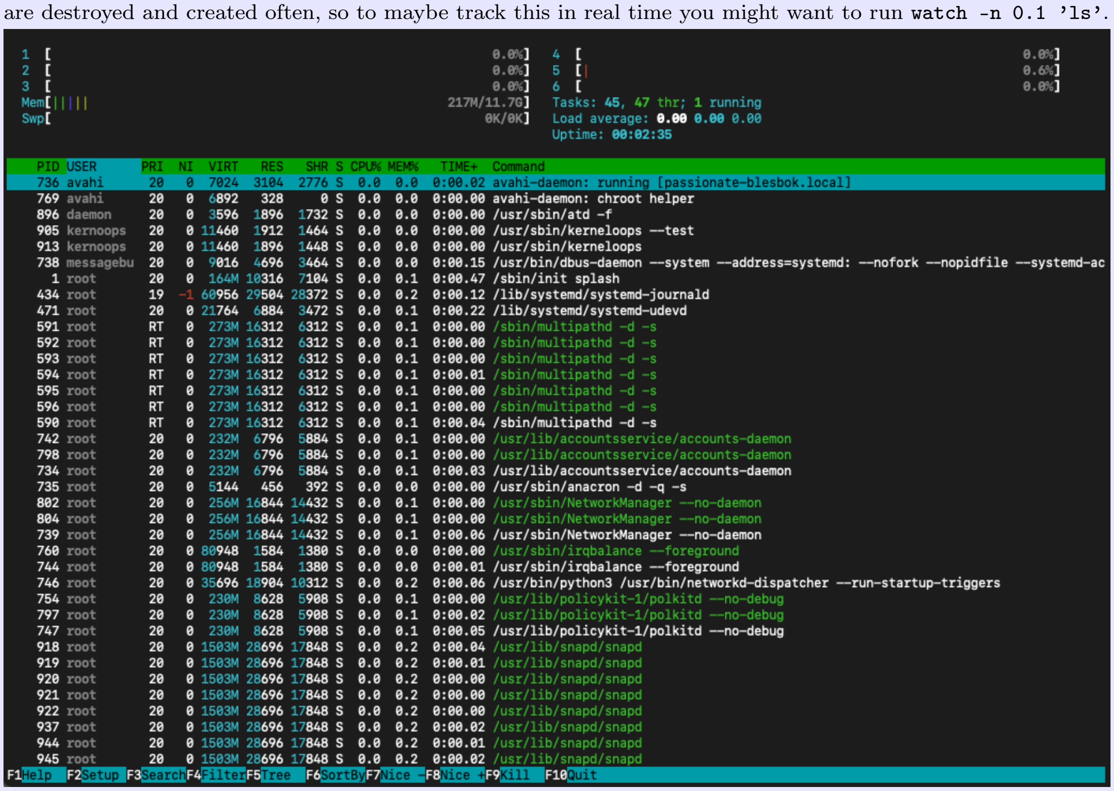
that maps it to physical memory.

Example 8.1 (Where to look for PIDs)

We can see the PIDs either by using `htop` (for UNIX systems) or by looking at the `/proc` directory in Linux systems. Each directory name represents the PID of the process.

```
ubuntu@passionate-blesbok:/proc$ ls
1     118   1368  26   44   590  762       cpuinfo      modules
10    119   1369  27   448  599  763       crypto       mounts
101   12    1370  28   45   6    764       devices      net
1012  120   1371  29   46   605  765       diskstats    pagetypeinfo
102   1232  138   3    467  608  769       driver       partitions
103   1234  139   30   468  611  796       execdomains  pressure
1031  1261  14    309  47   612  8         fb           sched_debug
1037  129   15    31   471  613  801       filesystems  schedstat
1038  13    16    32   473  629  810       fs           scsi
104   132   17    33   474  638  822       interrupts   self
1043  1342  18    34   475  651  836       iomem        slabinfo
105   1353  180   35   48   666  850       ioports      softirqs
106   1354  19    356  49   696  872       irq          stat
```

(a) You can see the PIDs of the process by looking at the `/proc` directory. This changes quite often as processes are destroyed and created often, so to maybe track this in real time you might want to run `watch -n 0.1 'ls'`.



(b) You can see the PID number of each process (binary) running on the left column when running `htop` on UNIX systems.

Figure 64: Two different ways to see the PIDs of all current processes.

There is a specific numbering to each process.

1. The process with PID 1 is always the kernel process.
2. The smaller PIDs (perhaps less than 300) are also reserved for the kernel, so don't kill it.

If you go into each process, you can see a few things.

```
ubuntu@passionate-blesbok:/proc/750$ sudo ls
attr       comm      fd       map_files   net        pagemap     sessionid    statm     uid_map
autogroup  coredump_filter  fdinfo   maps     ns      personality  setgroups     status    wchan
auxv       cpuset      gid_map   mem      numa_maps     projid_map   smaps      syscall
cgroup       cwd       io        mountinfo   oom_adj    root        smaps_rollup    task
clear_refs  environ      limits    mounts      oom_score     sched       stack       timers
cmdline     exe          loginuid  mountstats  oom_score_adj  schedstat    stat       timerslack_ns
```

Figure 65: There are many files in each PID folder that tells you about the process.

1. To get information about the status of this process, you can `cat status`.
2. The virtual address space is stored in `pagemap`. If you're on an 64-bit machine, this file will be extremely big, so just `cat pagemap` won't work. Therefore you should try `cat maps`, which shows you something like the following.

```
ubuntu@passionate-blesbok:/proc/750$ sudo cat maps
aaaac673c000-aaaac67e3000 r-xp 00000000 08:01 2576                       /usr/sbin/rsyslogd
aaaac67f3000-aaaac67f6000 r--p 000a7000 08:01 2576                       /usr/sbin/rsyslogd
aaaac67f6000-aaaac67fd000 rw-p 000aa000 08:01 2576                       /usr/sbin/rsyslogd
aaaac67fd000-aaaac67fe000 rw-p 00000000 00:00 0
aaaadfe94000-aaaadfed7000 rw-p 00000000 00:00 0                          [heap]
```

In here, you can see that the lefthand column represents the range of virtual memory address. The next column gives us the permissions (read, write, executable, shared/private).

8.1 Process Level Concurrency

Definition 8.3 (Context Switch)

Let us first start off with a single core system. At this point, everything is sequential, and to run all these processes at once[a] we want to use system calls to transition between these processes. This is called a **context switch**.

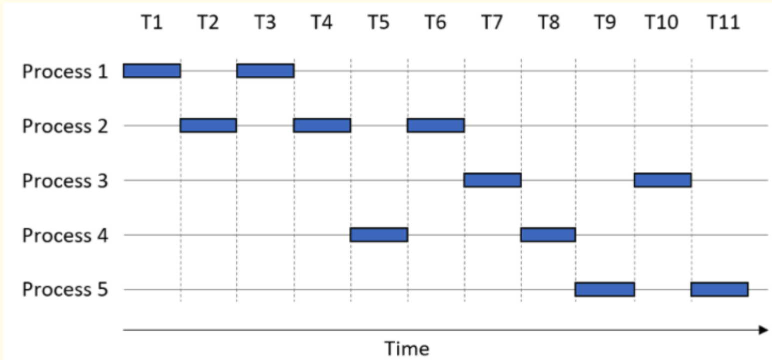


Figure 66: 5 processes may be executed as such on a single core.

Note that due to context switches, the **CPU time**, which is the time is takes to run a process on a CPU, is much shorter than the **wall-clock time**, which is the time a human perceives a process takes to complete.

[a]not programs, since there can be multiple instances of one program, like two Chrome instances. In fact, Chrome produces multiple processes to help run each part of the browser, so one program may translate to multiple processes.

Note that context switches are expensive. To do one, you must essentially replace two things.

1. First, you need to clear out all the register values. This can be done by storing them in the current stack at the VAS, which then gets mapped through the page table into the physical address space.
2. Now the register values (like the instruction and stack pointers) are stored safely in the stack in the VAS, the actual page table must be swapped out too since each process must have its own virtual address space.

Since it is quite expensive to context switch all the time, the simplest thing to do is add more cores, which gives us the double benefit of distributing the process workload *and* having to do less context switches. This is called **physical concurrency**, and given the same workload, it speeds up our computation absolutely. However, this can physically take us so far due to the limited number of cores, and we must go further and use **logical concurrency**.

8.2 Thread Level Concurrency

It turns out that it is much more expensive to reload the page table of a new process rather than clearing out the register values. So, perhaps maybe we can try to implement multiple related "processes" that *share* the same VAS, but have their own execution stream (i.e. own stack and registers). This is precisely the concept of a *thread*.

Definition 8.4 (Threads)

Threads are multiple execution streams within a single process. To summarize them, a thread is an execution context within a process that has a...

1. thread ID
2. its own stack frame
3. its own register context[a]

This is all that is really needed to execute some computation. Now, given that there are some number of threads in process K, they *share* the same virtual address space (VAS), are all under the same PID, share the same code, static data, heap, and file table. The individual stacks living within the VAS are protected from each other to avoid stack overflow.

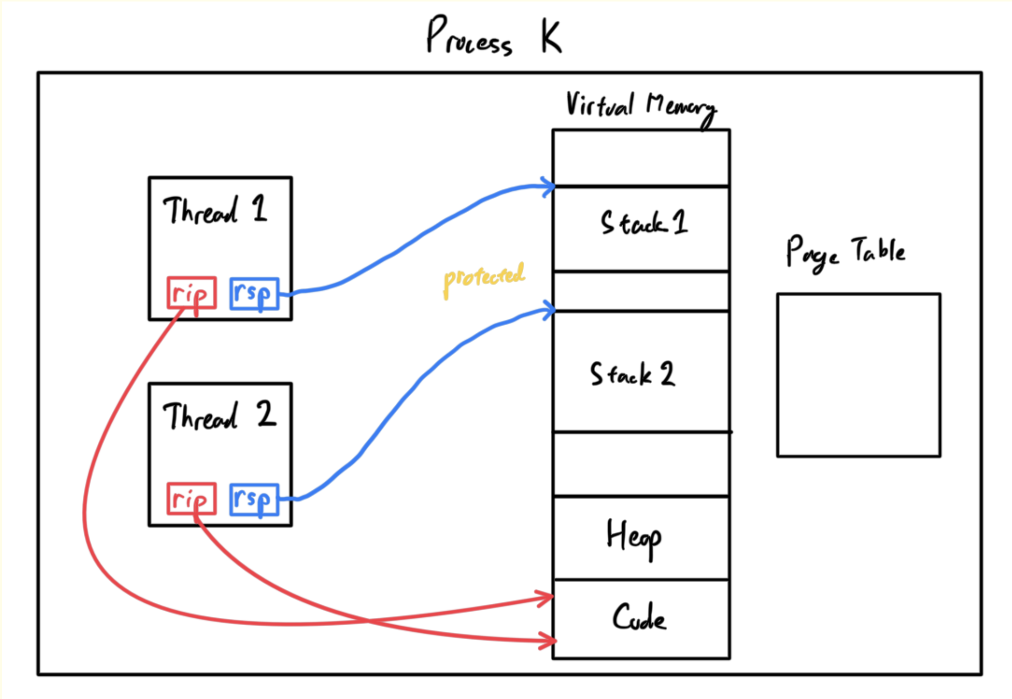


Figure 67: When there are two threads of a single process, the threads share the same virtual memory space. However, they each have their own set of registers. For example, they each have their own instruction pointer that points to the next line of code, along with their own stack pointer. Furthermore, to prevent stack overflow, there are protection mechanisms that prevent one stack from growing past a certain limit into another stack owned by a different thread.

Therefore, we can speed up our program in two ways.

1. If we have one core, we can do context switching faster between each thread (since we only have to load the register values).
2. If we have multiple cores, we can take thread 1 and have it run on one core while taking thread 2 and running it on another core. This is really analogous to having two separate processes on two cores, but these two processes simply share the same VAS, with the same code, data, and heap.

[a] This does not mean that each process has its own physical registers. It has its own *value* that is loaded into the registers. The physical number of registers is determined by the number of CPU cores.

Threads are advantageous for multiple reasons. First, by utilizing multiple cores we can speed up our program to reduce our *CPU time*. However, if we are sharing threads between one core, we're not actually speeding up anything at all but rather reducing our *wall-clock time*. The main speedup that we will feel is that latency heavy tasks will get offloaded to other threads, while more relevant programs can be run on the main thread. This is explained more in the following example.

Example 8.2 (Mobile Application)

If we have a single threaded messaging mobile app, then this is painfully slow since if we want to scroll down our messages while also sending and receiving messages, then we would have to wait for the message to receive from the server, into our disk, and into our memory, before the app responds when scrolling.

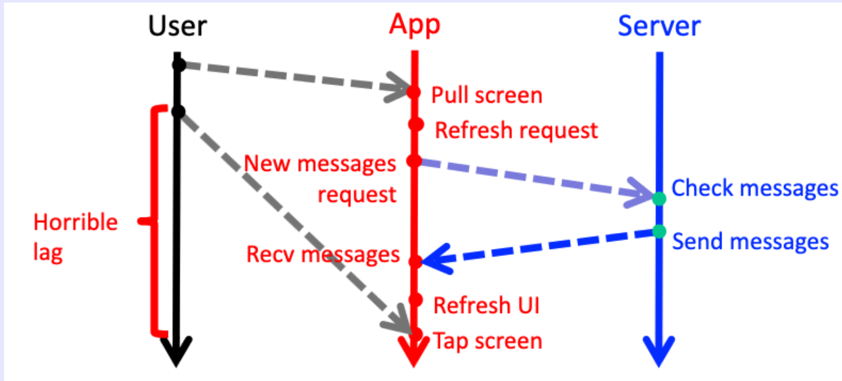


Figure 68: Single threaded app.

However, if we have a multithreaded app with one thread for the app UI and the other one for the server through a background thread, then we can have good UI response time.

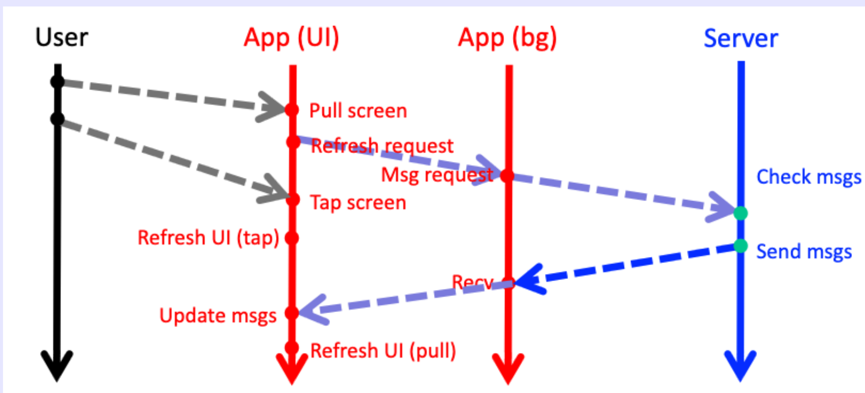


Figure 69: Multithreaded app. Methods on the UI thread must be fast to ensure user satisfaction while anything slow can run on a background thread.

The following law gives us a certain bound on how much parallelization can help us. Note that this does not talk about the responsiveness of an application due to clever thread sharing. It just says given a certain amount of computational task, how much can we reduce the CPU time with parallelization?

Theorem 8.1 (Amdahl's Law)

Say that we have code that runs in 1 second. Given that proportion f of our code can be parallelized, and the speedup for that portion is N, then the new time that our program will take is

$$T_{\text{new}} = (1 - f) + f/N \tag{12}$$

since the sequential part $1-f$ cannot be sped up, and the remaining parallel part f can be sped up by distributing over N cores. Therefore, defining the speedup as $T_{\text{new}}/T_{\text{old}}$, we get our total parallelized speedup is

$$\frac{1}{(1-f) + f/N} \tag{13}$$

Note that it is bounded by the sequential portion as $N \to \infty$.

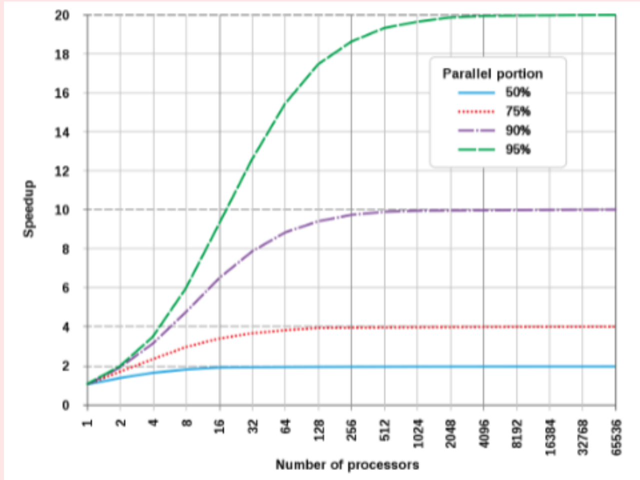


Figure 70: Amdahl's law for different f's over multiple number of processors N.

This is implemented in C with the `pthread.h` library, which is included in the standard library directory and follows the POSIX (Portable Operating System Interface) standard. Essentially, we want to do the following:

1. Define a function that will be called for each thread. It must return a void pointer `void *` and its arguments must also be a void pointer `void *`. Think of this as our new main function for each stack that will be created from each thread.[7] Since we are only restricted to a function taking in one void pointer argument, it is common to define a new struct like `arg_t` that contains all the parameters you need to run each thread. The void pointer can be typecast into the struct pointer at the beginning of each thread function.
2. We create `pthread_t` objects, which are the thread objects.
3. We call the `pthread_create` function that takes in the pointer of the thread object, some settings, the function to be called, and its arguments. At this point, the operating system will determine how these threads will be run, so you can't make any sequential assumptions about them.
4. Then we join them using `pthread_join`, which basically waits until all the threads are complete before `main` continues.

We will show two examples that go over this process. But more importantly, the concurrency of these two examples will show unpredictable behavior.

Example 8.3 (Simply Print out Thread Number)

We can make threads to print out the number. But these aren't really in the same order.

[7] Called a function pointer?

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

void* thread(void* args) {
  printf("Thread %d\n", *(int*)args);
  return NULL;
}

int main(int argc, char *argv[]) {
  int size = 10;
  pthread_t threads[size];
  int rc, i;

  // thread creation
  for (i = 0; i < size; i++) {
    rc = pthread_create(&threads[i], NULL, thread, &i);
  }

  // join waits for the threads to finish
  for (i = 0; i < size; i++) {
    rc = pthread_join(threads[i], NULL);
  }
return 0;
}
```

```
Thread 1
Thread 2
Thread 6
Thread 3
Thread 8
Thread 5
Thread 2
Thread 4
Thread 3
Thread 3
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
```

Figure 71: Threads output shows that the order in which the functions are called cannot be predicted.

8.3 Atomicity Violation Bugs and Mutex Locks

We see that there are some parts of the code that are not meant to be parallelized.

Definition 8.5 (Atomicity-Violation Bugs)

This bug happens when the desired **atomicity** (indivisibility) among multiple memory accesses is violated.

Example 8.4 (Atomicity-Violation in SQL)

For example, if we have two threads doing the following (in MySQL):

```c
Thread1::
if (thd-> proc_info) {
  ...
  fputs(thd->proc_info);
  ...
}

Thread2::
thd->proc_info = NULL;
```

If we pass the if statement but within it, `thd->proc_info` becomes `NULL`, then this would be very bad. Therefore, we should put locks around.

```c
pthread_mutex_t lock = PTHREAD_MUTEX_INITIALIZER;

  Thread1::
  pthread_mutex_lock(&lock);
  if (thd-> proc_info) {
    ...
    fputs(thd->proc_info);
    ...
  }
  pthread_mutex_unlock(&lock);

  Thread2::
  pthread_mutex_lock(&lock);
  thd->proc_info = NULL;
  pthread_mutex_unlock(&lock);
```

Example 8.5 (Incrementing Shared Counter between Two Threads)

The volatile keyword for `counter` means that it can be changed by all threads.

```c
1  #include <stdio.h>
2  #include <pthread.h>
3  #include <stdlib.h>
4
5  static volatile int counter = 0;
6
7  void* thread(void* args) {
8    printf("%s : Start \n", (char*)args);
9
10   for (int i = 0; i < 10 * 1000 * 1000; i++) {
11     counter += 1;
12   }
13
14   printf("%s : End \n", (char*)args);
15   return NULL;
16 }
17
18 int main(int argc, char *argv[]) {
19   pthread_t thread1, thread2;
20
21   int rc;
22   rc = pthread_create(&thread1, NULL, thread, "A");
23   rc = pthread_create(&thread2, NULL, thread, "B");
24
25   rc = pthread_join(thread1, NULL);
26   rc = pthread_join(thread2, NULL);
27
28   printf("Counter : %d\n", counter);
29   return 0;
30 }
```

```
A : Start
B : Start
A : End
B : End
Counter : 10229646
.
.
A : Start
B : Start
B : End
A : End
Counter : 14965289
.
.
A : Start
B : Start
A : End
B : End
Counter : 10086690
.
.
.
.
.
.
.
.
.
.
.
```

Figure 72: From looking at the behavior, we can see that the start and end times of thread A and thread B is complete unpredictable. It is only within the sequential nature within each thread (i.e. `X : start` must come before `X : End` that is predictable. More so, after all the increments, the total sum collected by counter isn't even 20 million!). The actual speed and ordering of this is determined at runtime.

To really see what's going on here, we must look into the assmebly code behind this. If we focus on line 11 when the counter is incremented and look at its assembly, we see

```
1  7a7: 8b 05 0a 0b 20 00       mov    0x200b0a(%rip),%eax        # 200c1c <counter>
2  7ad: 83 c0 01                add    $0x1,%eax
3  7b0: 89 05 01 0b 20 00       mov    %eax,0x200b01(%rip)        # 200c1c <counter>
```

What this means is that every thread consists of loading the `counter` value from the instruction pointer plus an offset into `%eax`, then adding 1 to it, and then storing it back to the same memory location (the rip changes, so the memory address between the 1st and 3rd line will be slightly different, but they are the same address). If we have threads 1 and 2, we can have the following possible interweavings:

1. (1 loads, 1 adds, 1 stores, 2 loads, 2 adds, 2 stores). This would result in a +2.
2. (1 loads, 1 adds, 2 loads, 2 adds, 2 stores, 1 stores). This would result in a +1 since thread 2 loads before 1 could store the incremented version.
3. (1 loads, 2 loads, 1 adds, 2 adds, 1 stores, 2 stores). This would also result in a +1 for the same reasons as before.

There is a big problem here: it seems that these things overwrite each other.

Definition 8.6 (Data Race)

We have just seen an example of a **data race**, which occurs if two or more threads concurrently accesses the same memory location with at least one write. The section of code where a data race can occur is called the **critical section**.

So how do we address this challenge of concurrency? This is where locks and mutexes come in.

Definition 8.7 (Lock)

A **lock** is a construct to enforce mutual exclusion in conflicting code sections (critical sections). It is implemented as a special data object in memory. We can use the API methods

1. `acquire()` or `lock()` is called when going into a critical section.
2. and `release()` or `unlock()` is called when going out of a critical section.

If the lock is already acquired by a thread and not released yet, then other threads will not be able to acquire the lock and execute the next instructions. There are two ways this is implemented. First, an incoming thread can wait if another thread holds the lock, called a **spinlock**,[a] or it can be blocked, called a **mutex** (mutual exclusion).[b][c]

[a]This is mainly implemented through kernel and used almost exclusively for OS development, not application development.

[b]This is mainly implemented in user space.

[c]This has a FIFO queue.

To implement this in C, there's a few things that we have to do.

1. First, make a `pthread_mutex_t` global variable.
2. Then, make initialize the mutex before you create the threads and destroy the mutex after you join the threads in the `main` function.
3. Finally, put the specific locks and unlocks in the locations of the functions that the thread calls.

There are few strategies to correct the counter example above, which all produce the correct output `Counter: 20000000`.

1. We can put the lock around the entire for loop of the `thread` function. However, this essentially takes us back to the sequential regime.

```c
static volatile int counter = 0;
pthread_mutex_t mutex; // global declaration of mutex

void* thread(void* args) {
  pthread_mutex_lock(&mutex); //acquire the mutex lock
  for (int i = 0; i < 10 * 1000 * 1000; i++) {
    counter += 1;
  }
  pthread_mutex_unlock(&mutex); //release the mutex lock
  return NULL;
}

int main(int argc, char *argv[]) {
  pthread_t thread1, thread2;
  int rc;
  rc = pthread_mutex_init(&mutex, NULL); //initialize the mutex
  rc = pthread_create(&thread1, NULL, thread, "A");
  rc = pthread_create(&thread2, NULL, thread, "B");
  rc = pthread_join(thread1, NULL);
  rc = pthread_join(thread2, NULL);
  pthread_mutex_destroy(&mutex); //destroy (free) the mutex
  printf("Counter : %d\n", counter);
  return 0;
}
```

Figure 73: Putting the mutex locks around the entire for loop.

2. A better idea to actually implement parallelization is to put the locks around the line that says `counter += 1;`. This is the critical code that loads the counter value from the stack into the register, increments it by 1, and sends it back to the stack. We should isolate this so that no other threads can execute during these 3 assembly lines.

```
static volatile int counter = 0;
pthread_mutex_t mutex; // global declaration of mutex

void* thread(void* args) {
  for (int i = 0; i < 10 * 1000 * 1000; i++) {
    pthread_mutex_lock(&mutex); //acquire the mutex lock
    counter += 1;
    pthread_mutex_unlock(&mutex); //release the mutex lock
  }
  return NULL;
}

int main(int argc, char *argv[]) {
  pthread_t thread1, thread2;
  int rc;
  rc = pthread_mutex_init(&mutex, NULL); //initialize the mutex
  rc = pthread_create(&thread1, NULL, thread, "A");
  rc = pthread_create(&thread2, NULL, thread, "B");
  rc = pthread_join(thread1, NULL);
  rc = pthread_join(thread2, NULL);
  pthread_mutex_destroy(&mutex); //destroy (free) the mutex
  printf("Counter : %d\n", counter);
  return 0;
}
```

Figure 74: Now we put the locks within the counter. However, this is much slower since locking and unlocking are relatively expensive. It runs in 0.274s.

3. The first two tries are not ideal, but what we can do is have each thread store its local work all within each of its stack, and then when it communicates with the shared memory on `counter`, this is where the locks should come in place. This has the double benefit of locking/unlocking very few times, along with protecting the critical section of the code.

```c
static volatile int counter = 0;
pthread_mutex_t mutex; // global declaration of mutex

void* thread(void* args) {
  int my_counter = 0;
  for (int i = 0; i < 10 * 1000 * 1000; i++) {
    my_counter += 1;
  }
  pthread_mutex_lock(&mutex); //acquire the mutex lock
  counter += my_counter;
  pthread_mutex_unlock(&mutex); //release the mutex lock
  return NULL;
}

int main(int argc, char *argv[]) {
  pthread_t thread1, thread2;
  int rc;
  rc = pthread_mutex_init(&mutex, NULL); //initialize the mutex
  rc = pthread_create(&thread1, NULL, thread, "A");
  rc = pthread_create(&thread2, NULL, thread, "B");
  rc = pthread_join(thread1, NULL);
  rc = pthread_join(thread2, NULL);
  pthread_mutex_destroy(&mutex); //destroy (free) the mutex
  printf("Counter : %d\n", counter);
  return 0;
}
```

Figure 75: In here, we have each thread increment its own local version of the counter and store it in `my_counter`. Then, we increment the global `counter` by the total `my_counter`, which will require one mutex lock. It runs in 0.049s.

Example 8.6 (Inserting into Linked List)

Inserting into a linked list is a sequential process, but what if we want to parallelize it with multiple threads? Consider the following code.

```c
typedef struct __node_t {
  int key;
  struct __node_t *next;
} node_t;

typedef struct __list_t {
  node_t *head;
} list_t;

void List_Init(list_t *L) {
  L -> head = NULL;
}

void List_Insert(list_t *L, int key) {
  // insert a new node with the key value at the beginning of the list
  node_t *new = malloc(sizeof(node_t));
  assert(new);
  new -> key = key;
  new -> next = L -> head;
  L -> head = new;
```

```c
}

int main(void) {
  list_t L;
  list_t *Lp = &L;
  List_Init(Lp);
  for (int i = 0; i < 10; i++) {
    List_Insert(Lp, i);
  }
  return 0;
}
```

Say that we want to parallelize the creation of the length 10 linked list. Simply initializing some threads and calling `List_Insert` 10 times won't properly create this linked list. This is because two threads may overwrite where `L -> head` points to. The most naive thing to do is to put locks around the whole function, but now we are back to the sequential regime! If we think about it, the `malloc` calls, asserting that there is viable memory, and assigning the input key value to `new -> key` does not overwrite anything else. It is only when we assign `new -> next` to the head value of `L` that things may get overwritten, so we put the locks shown below, while slightly modifying the `list` struct to have the `pthread_mutex_t` attribute.

```c
typedef struct __list_t {
  node_t *head;
  pthread_mutex_t lock;
} list_t;

void List_Init(list_t *L) {
  L -> head = NULL;
  pthread_mutex_init(&L -> lock, NULL);
}

void List_Insert(list_t *L, int key) {
  // insert a new node with the key value at the beginning of the list
  node_t *new = malloc(sizeof(node_t));
  assert(new);
  new -> key = key;
  pthread_mutex_lock(&L->lock)
  new -> next = L -> head;
  L -> head = new;
  pthrea_mutex_unlock(&L->lock)
}
```

8.4 Deadlock Bugs

Locks are extremely useful to segment out a portion of code that should be uninterrupted. However, there are many consequences of misusing it.

Definition 8.8 (Deadlock Bugs)

A **deadlock bug** happens when you make locks such that the program cannot run anymore. These aren't specific to threads, but also processes as well. They require the four preconditions:

1. Mutual exclusion: you must have a lock to begin with to have a deadlock, so this is trivial.
2. Hold and Wait: Threads must have the ability to hold resources (e.g. the philosophers must hold forks which prevent others from taking them) .

3. No Preemption: Preemption refers to the ability to take the fork out of someone else's hand. So if you have a thread, you can first take lock A, then look at whether lock B is available. If not, then unlock A and try again. However, this can lead to a bunch of threads just simply picking up and putting down forks, causing a *livelock*.
4. Circular Wait: That is, there exists a circular chain of threads such that each thread holds a resource needed by the next thread. A strategy is too define a fixed acquisition order for locks (e.g. lock A always before lock B).

You shouldn't try to hold multiple locks at once, but if you must, you should have a strategy to avoid deadlock. Choosing a lock order is the recommended way.

Example 8.7 (Bank Accounts)

Let's go through an example. Suppose we had the following code below.

```
struct account {
  pthread_mutex_t lock;
  int balance;
};

struct arg_t {
  struct account fromAcct;
  struct account toAcct;
  int amt;
};

pthread_mutex_t mutex; // global declaration of mutex

void* Transfer(void* args) {

  struct arg_t* data = (struct arg_t*)args;

  struct account* fromAcct = &(data -> fromAcct);
  struct account* toAcct = &(data -> toAcct);
  int amt = data -> amt;


  pthread_mutex_lock(&fromAcct->lock);
  pthread_mutex_lock(&toAcct->lock);

  fromAcct->balance -= amt;
  toAcct->balance += amt;

  pthread_mutex_unlock(&fromAcct->lock);
  pthread_mutex_unlock(&toAcct->lock);

  return NULL;
}
```

Suppose that Threads 0 and 1 are executing concurrently and represent users A and B, respectively. Now consider the situation in which A and B want to transfer money to each other: A wants to transfer 20 dollars to B, while B wants to transfer 40 to A.
Both threads concurrently execute the `Transfer` function. Thread 0 acquires the lock of acctA while Thread 1 acquires the lock of acctB. Now consider what happens. To continue executing, Thread 0 needs to acquire the lock on acctB, which Thread 1 holds. Likewise, Thread 1 needs to acquire the lock on acctA to continue executing, which Thread 0 holds. Since both threads are blocked on each

other, they are in deadlock.
This can be simply fixed by rearranging the locks so that each lock/unlock pair surrounds only the balance update statement associated with it.

```
void *Transfer(void *args){
  //argument passing removed to increase readability
  //...

  pthread_mutex_lock(&fromAcct->lock);
  fromAcct->balance -= amt;
  pthread_mutex_unlock(&fromAcct->lock);

  pthread_mutex_lock(&toAcct->lock);
  toAcct->balance += amt;
  pthread_mutex_unlock(&toAcct->lock);

  return NULL;
}
```

Example 8.8 (Dining Philosopher's Problem)

There are 5 philosophers at a roundtable with 5 plates of food with 5 forks. Each philosopher needs two forks (both on their left and on their right) to start eating their plate. However, there can be a deadlock if every philosopher takes the fork on their right and is always waiting for the left fork (this happens due to a circular dependency).

1. One way to resolve this issue is to set an ID to every fork and have the philosophers always take the lower ID fork first before trying to take the higher ID. Then this resolves and at least one philosopher can eat.

Example 8.9 (Set-Intersection Problem)

Again, like the bank account section, we have the following code that can create deadlocks if there is a code that attempts to compute $A \cap B$ and $B \cap A$ at similar times.

```
set_t *set_intersection(set_t *s1, set_t *s2) {
  set_t *result = set_create();
  pthread_mutex_lock(&s1->lock);
  pthread_mutex_lock(&s2->lock);

  for (int i = 0; i < s1->size; i++) {
    if (set_contains(s2, s1->data[i])) {
      set_add(result, s1->data[i]);
    }
  }

  pthread_mutex_unlock(&s2->lock);
  pthread_mutex_unlock(&s1->lock);
  return result;
}
```

Like using the IDs and ordering, we can resolve this by grabbing locks in a defined order, say from high to low. Therefore, we replace the locks as such.

```
if (&m1 > &m2) {
```

```
  pthread_mutex_lock(&m1);
  pthread_mutex_lock(&m2);
} else {
  pthread_mutex_lock(&m2);
  pthread_mutex_lock(&m1);
}
```

However, this still creates a deadlock if we compute $A \cap A$, so we should place a deadlock.

It turns out that we can make logically equivalent code without locks! By using atomic primitives that have implementations in C, we can create wait-free algorithms. For example, consider the `int CompAndSwap(int* addr, int expected, int new)` function. If `*addr == expected`, then we set `*addr` to `new` and return 1. Otherwise, we do nothing and return 0. Then, the two pieces of code are equivalent, but the advantage of the RHS is that there is never a deadlock.

1. In the left, we pass by pointer to increment a variable.
2. In the right, we first set `old` to the value of `val`, and if the value of `val` is equal to `old` (which may not always be true if some other thread modifies this value), then we set `old = *val` again and increment it by `amt`.

```
void add(int *val, int amt) {
  pthread_mutex_lock(&m);
  *val += amt;
  pthread_mutex_unlock(&m);
}
.
```

```
void add(int *val, int amt) {
  do {
    int old = *val;
  }
  while(!CompAndSwap(val, old, old+amt));
}
```

However, the left hand, despite the risk of deadlocks, is still recommended. The mutex API allows for better distribution of CPU.

8.5 Order Violation Bugs and Condition Variables

Definition 8.9 (Order Violation Bugs)

An **order violation bug** happens when the desired order between two memory accesses is flipped, i.e. A should always be executed before B, but the order is not enforced during execution.

1. For example, one thread that acts on a linked list may already assume that the linked list is initialized when it is not. Therefore, we should have these threads wait until initialization.
2. Another example is if there is a set of producers and consumers that manage the flow of items through a buffer. If there is no good, then consumers cannot consume and they must wait. If the supply is full then the producers must wait. This also requires some process of waiting.

We can use a regular conditional to check, but this may not be efficient due to the following example.

Example 8.10 (Condition Checks with Vanilla While Loops)

Say that you want to call the following function on a linked list, but you must wait until the list is not `NULL`.

```
int getItem(list_t *list) {
  pthread_mutex_lock(&list -> m);
  // TODO: wait until list is non-empty
  int item = list -> head -> item;
  list -> head = list -> head -> next;
```

```
  pthread_mutex_unlock(&list -> m);
  return item;
}
```

To check whether `list` is not a null pointer, we can simply use a while loop to check.

```
int tryGetItem(list_t *list, int *out) {
  int success = 0;
  ptherad_mutex_lock(&list -> m);
  if (list -> head) {
    success = 1;
    *out = list -> head -> item;
    list -> head = list -> head -> next;
  }
  pthread_mutex_unlock(&list -> m);
  return success;
}
.
.
```

```
int getItem( list_t *list) {
  pthread_mutex_lock(&list -> m);
  // TODO: wait until list is non-empty
  while (list -> head == NULL) {
    pthread_mutex_unlock(&list -> m);
    yield();        // optional
    pthread_mutex_lock(&list -> m);
  }
  int item = list -> head -> item;
  list -> head = list -> head -> next;
  pthread_mutex_unlock(&list -> m);
  return item;
}
```

Figure 76: On the LHS, this returns 0 if the retrieval is not successful, and so you must wrap this function within some while loop to check if it is actually successful. On the RHS, we can use the `yield()` function which gives control to the OS to schedule another thread.

This constant checking through a while loop leads to a waste of CPU resources. Therefore, we want to put the thread to sleep while there is no element in `list` so other processes can use the CPU core. This is the motivation behind *conditional variables*.

Definition 8.10 (Condition Variables)

Condition Variables force a thread to be blocked until a particular condition is reached. This construct is useful for scenarios in which a condition must be met before the thread does some work.

1. Every CV is bound to exactly one mutex. This is because the state of a condition, even if true on one thread, can be changed immediately by another thread, so some sort of locking is needed.
2. Condition variables have the type `pthread_cond_t`.
3. To initialize a condition variable, use the `pthread_cond_init` function.
4. To destroy it, use `pthread_cond_destroy`.
5. `pthread_cond_wait(&cond, &mutex)` takes the address of a condition variable `cond` and a mutex `mutex` as its arguments. It causes the calling thread to block on the condition variable cond until another thread signals it (or "wakes" it up).
6. The `pthread_cond_signal(&cond)` function causes the calling thread to unblock (or signal) another thread that is waiting on the condition variable `cond` (based on scheduling priority). If no threads are currently blocked on the condition, then the function has no effect. Unlike `pthread_cond_wait`, the `pthread_cond_signal` function can be called by a thread regardless of whether or not it owns the mutex in which `pthread_cond_wait` is called.

It is important to know about the states that a thread can be in. It can either be currently running/active, ready to run (perhaps through a syscall), or blocked.

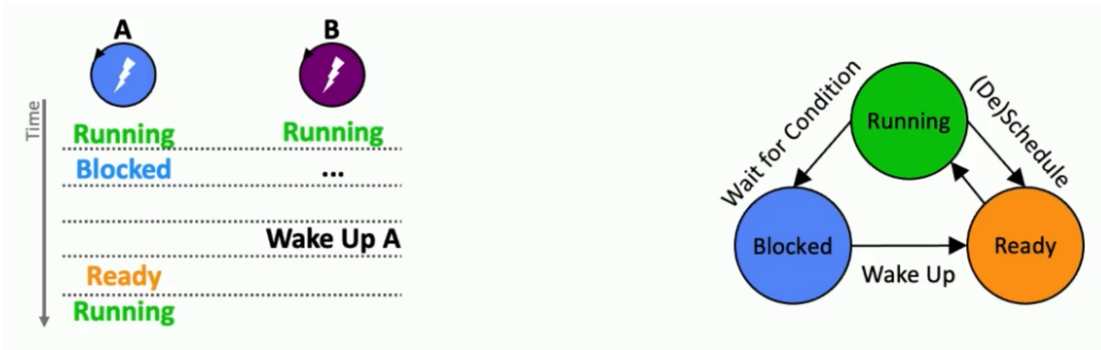


Figure 77: Note that the only way a thread can be blocked is if it is waiting for a condition to happen. If that condition happened and a signal arrives at the thread, then it "wakes up." Condition variables allow these thread to go in and out of the blocked state.

With this, the general design pattern is as such.

Theorem 8.2 (Condition Variable Design Pattern)

To implement this effectively, we must first identify a state that will be accessed by 2+ threads concurrently and add locks to protect the shared state. If we need to wait on some condition, we use condition variables.

```
methodThatWaits() {
  pthread_mutex_lock(&m);

  // Read/write shared state

  while (!checkSharedState()) {
    pthread_cond_wait(&cv, &m);
  }

  // Read/write shared state

  pthread_mutex_unlock(&m);
}
```

```
methodThatSignals() {
  ptherad_mutex_lock(&m);

  // Read/write shared state

  // If checkSharedState() is now true
  pthread_cond_signal(&cv);

  // Read/write shared state

  pthread_mutex_unlock(&m);
}
.
```

The implementation is quite complex at first, so let's go through an example.

Example 8.11 (Soda Machine)

Say that we want to model a soda machine, with consumers taking soda from the machine and producers filling the machine up.

1. We'd like to create some variables that encode the state of the soda machine. This is the **shared state**.

```
static volatile int numSodas;
#define MaxSodas 100;
```

2. We also want to implement *one* lock to protect all shared states, say

```
pthread_mutex_t sodaLock;
```

This allows us to implement mutual exclusion so that only one thread can manipulate the machine (state) at a time.

3. The **ordering constraints** are that the consumer must wait if the machine is empty (CV `hasSoda`), and the producer must wait if the machine is full (CV `hasRoom`).
4. The first thing we must do is make sure that the consumer and producer function has a lock and unlock over its body since both functions modify the vending machine.

```
consumer() {
  pthread_mutex_lock(&sodaLock);

  // take a soda from machine

  pthread_mutex_unlock(&sodaLock);
}
```

```
producer() {
  pthread_mutex_lock(&sodaLock);

  // add a soda to machine

  pthread_mutex_unlock(&sodaLock);
}
```

5. Moreover, the consumer and producer's actions of taking or adding soda is dependent on the state already. For the consumer, it should **wait** if the machine is empty for a signal that notifies that it is not empty. Once the consumer receives the signal, it takes a soda and can send a signal that the machine is not full to the producer function.

```
consumer() {
  pthread_mutex_lock(&sodaLock);
  // wait if empty
  // take a soda from machine
  // notify that it is not full
  pthread_mutex_unlock(&sodaLock);
}
```

```
producer() {
  pthread_mutex_lock(&sodaLock);
  // wait if full
  // add a soda to machine
  // notify that it is not empty
  pthread_mutex_unlock(&sodaLock);
}
```

6. To put this into code, we finally have

```
consumer() {
  pthread_mutex_lock(&sodaLock);

  while (numSodas == 0) {
    // while empty
    wait(sodaLock, hasSoda);
  }
  numSodas -= 1;    // take soda
  signal(hasRoom);
  pthread_mutex_unlock(&sodaLock);
}
```

```
producer() {
  pthread_mutex_lock(&sodaLock);

  while (numSodas == MaxSodas) {
    wait(sodaLock, hasRoom);
  }

  numSodas += 1;    // add soda
  signal(hasSoda);
  pthread_mutex_unlock(&sodaLock);
}
```

Let's go through this. From the consumer function, we have:

1. The consumer function starts by acquiring a lock on the `sodaLock` mutex using `pthread_mutex_lock()`. This ensures that only one consumer thread can access the shared resources (e.g., `numSodas`) at a time.
2. It then enters a while loop[a] that checks if `numSodas` is zero. If `numSodas` is zero, it means there are no sodas available for consumption. In this case, the consumer thread calls the `wait()` function, which unlocks the `sodaLock` mutex and waits for a signal on the `hasSoda` condition variable. This allows other threads to acquire the lock and proceed.
3. When the consumer thread receives a signal indicating that sodas are available (`hasSoda` condition variable), it wakes up and reacquires the lock on `sodaLock`.
4. The consumer thread then decrements `numSodas` by 1, indicating the consumption of a soda.
5. After consuming a soda, the consumer thread signals the `hasRoom` condition variable using the `signal()` function. This notifies any waiting producer threads that there is now room available in the soda buffer.
6. Finally, the consumer thread unlocks the `sodaLock` mutex using `pthread_mutex_unlock()`,

allowing other threads to access the shared resources.

From the producer function, we have:

1. The producer function starts by acquiring a lock on the `sodaLock` mutex using `pthread_mutex_lock()`. This ensures that only one producer thread can access the shared resources at a time.
2. It then enters a while loop that checks if `numSodas` is equal to `MaxSodas`. If `numSodas` is equal to `MaxSodas`, it means the soda buffer is full, and the producer cannot add more sodas. In this case, the producer thread calls the `wait()` function, which unlocks the `sodaLock` mutex and waits for a signal on the `hasRoom` condition variable. This allows other threads to acquire the lock and proceed.
3. When the producer thread receives a signal indicating that there is room available in the soda buffer (`hasRoom` condition variable), it wakes up and reacquires the lock on `sodaLock`.
4. The producer thread then increments `numSodas` by 1, indicating the production of a new soda.
5. After producing a soda, the producer thread signals the `hasSoda` condition variable using the `signal()` function. This notifies any waiting consumer threads that a soda is now available for consumption.
6. Finally, the producer thread unlocks the `sodaLock` mutex using `pthread_mutex_unlock()`, allowing other threads to access the shared resources.

Note that we must have a while loop since if the producer ended up broadcasting the `hasSoda` condition to say 10 threads when there are 5 sodas, then 5 of those threads will get a soda while 5 may not, and this possibility should be detected by the while loop.

[a] We want this to be a while loop since we want to reevaluate the condition even after reacquiring the lock.