Introduction to Compilers and Language Design

Hardcover ISBN: 978-0-359-13804-3
Paperback ISBN: 978-0-359-14283-5
First edition.

You can find the latest version of the PDF edition, and purchase inexpensive hardcover copies at `http://compilerbook.org`

Revision Date: October 9, 2018

Chapter 10 – Assembly Language

10.1 Introduction

In order to build a compiler, you must have a working knowledge of at least one kind of assembly language. And, it helps to see two or more variations of assembly, so as to fully appreciate the distinctions between architectures. Some of these differences, such as register structure, are quite fundamental, while some of the differences are merely superficial.

We have observed that many students seem to think that assembly language is rather obscure and complicated. Well, it is true that the complete manual for a CPU is extraordinarily thick, and may document hundreds of instructions and obscure addressing modes. However, it's been our experience that it is really only necessary to learn a small subset of a given assembly language (perhaps 30 instructions) in order to write a functional compiler. Many of the additional instructions and features exist to handle special cases for operating systems, floating point math, and multi-media computing. You can do almost everything needed with the basic subset.

We will look at two different CPU architectures that are in wide use today: X86 and ARM. The Intel X86 is a CISC architecture that has evolved since the 1970s from 8-bit to 64-bit and is now the dominant chip in personal computers, laptops, and high performance servers. The ARM processor is a RISC architecture began life as a 32-bit chip for the personal computer market, and is now the dominant chip for low-power and embedded devices such as mobile phones and tablets.

This chapter will give you a working knowledge of the basics of each architecture, but you will need a good reference to look up more details. We recommend that you consult the Intel Software Developer Manual [1] and the ARM Architecture Reference Manual [3] for the complete details. (Note that each section is meant to be parallel and self-contained, so some explanatory material is repeated for both X86 and ARM.)

10.2 Open Source Assembler Tools

A given assembly language can have multiple dialects for the same CPU, depending on whether one uses the assembler provided by the chip vendor, or other open source tools. For consistency, we will give examples in the assembly dialect supported by the GNU compiler and assembler, which are known as `gcc` and `as` (or sometimes `gas`.)

A good way to get started is to view the assembler output of the compiler for a C program. To do this, run `gcc` with the `-S` flag, and the compiler will produce assembly output rather than a binary program. On Unix-like systems, assembly code is stored in files ending with `.s`, which indicates "source" file.

If you run `gcc -S hello.c -o hello.s` on this C program:

```
#include <stdio.h>

int main( int argc, char *argv[] )
{
    printf("hello %s\n","world");
    return 0;
}
```

then you should see output similar to this in `hello.s`

```
.file   "test.c"
.data
.LC0:
        .string "hello %s\n"
.LC1:
        .string "world"
.text
.globl main
main:
        PUSHQ   %rbp
        MOVQ    %rsp, %rbp
        SUBQ    $16, %rsp
        MOVQ    %rdi, -8(%rbp)
        MOVQ    %rsi, -16(%rbp)
        MOVQ    $.LC0, %rax
        MOVQ    $.LC1, %rsi
        MOVQ    %rax, %rdi
        MOVQ    $0, %rax
        CALL    printf
        MOVQ    $0, %rax
        LEAVE
        RET
```

(There are many valid ways to compile `hello.c` and so the output of your compiler may be somewhat different.)

Regardless of the CPU architecture, the assembly code has three different kinds of elements:

Directives begin with a dot and indicate structural information useful to the assembler, linker, or debugger, but are not in and of themselves assembly instructions. For example, `.file` simply records the name of the original source file to assist the debugger. `.data` indicates the start of the data segment of the program, while `.text` indicates the start of the program segment. `.string` indicates a string constant within the data section, and `.globl main` indicates that the label `main` is a global symbol that can be accessed by other code modules.

Labels end with a colon and indicate by their position the association between names and locations. For example, the label `.LC0:` indicates that the immediately following string should be called `.LC0`. The label `main:` indicates that the instruction `PUSHQ %rbp` is the first instruction of the main function. By convention, labels beginning with a dot are temporary local labels generated by the compiler, while other symbols are user-visible functions and global variables. The labels do not become part of the resulting machine code *per se*, but they are present in the resulting object code for the purposes of linking, and in the eventual executable, for purposes of debugging.

Instructions are the actual assembly code like (`PUSHQ %rbp`), typically indented to visually distinguish them from directives and labels. Instructions in GNU assembly are not case sensitive, but we will generally uppercase them, for consistency.

To take this `hello.s` and turn it into a runnable program, just run `gcc`, which will figure out that it is an assembly program, assemble it, and link it with the standard library:

```
% gcc hello.s -o hello
% ./hello
hello world
```

It is also interesting to compile the assembly code into object code, and then use the `nm` utility to display the symbols ("names") present in the code:

```
% gcc hello.s -c -o hello.o
% nm hello.o
0000000000000000 T main
                 U printf
```

This display the information available to the linker. `main` is present in the text (`T`) section of the object, at location zero, and `printf` is undefined

(`U`), since it must be obtained from the standard library. But none of the labels like `.LC0` appear because they were not declared as `.globl`.

As you are learning assembly language, take advantage of an existing compiler: write some simple functions to see what `gcc` generates. This can give you some starting points to identify new instructions and techniques to use.

10.3 X86 Assembly Language

X86 is a generic term that refers to the series of microprocessors descended from (or compatible with) the Intel 8088 processor used in the original IBM PC, including the 8086, 80286, '386, '486, and many others. Each generation of CPUs added new instructions and addressing modes from 8-bit to 16-bit to 32-bit, all while retaining backwards compatibility with old code. A variety of competitors (such as AMD) produced compatible chips that implemented the same instruction set.

However, Intel broke with tradition in the 64-bit generation by introducing a new brand (Itanium) and architecture (IA64) that was *not* backwards compatible with old code. Instead, it implemented a new concept known as Very Long Instruction Word (VLIW) in which multiple concurrent operations were encoded into a single word. This had the potential for significant speedups due to instruction-level parallelism but represented a break with the past.

AMD stuck with the old ways and produced a 64-bit architecture (AMD64) that *was* backwards compatible with both Intel and AMD chips. While the technical merits of both approaches were debatable, the AMD approach won in the marketplace, and Intel followed by producing its own 64-bit architecture (Intel64) that was compatible with AMD64 and its own previous generation of chips. X86-64 is the generic name that covers both AMD64 and Intel64 architectures.

X86-64 is a fine example of CISC (complex instruction set computing). There are a very large number of instructions with many different submodes, some of them designed for very narrow tasks. However, a small subset of instructions will let us accomplish a lot.

10.3.1 Registers and Data Types

X86-64 has sixteen (almost) general purpose 64-bit integer registers:

%rax	%rbx	%rcx	%rdx	%rsi	%rdi	%rbp	%rsp
%r8	%r9	%r10	%r11	%r12	%r13	%r14	%r15

These registers are *almost* general purpose because earlier versions of the processors intended for each register to be used for a specific purpose, and not all instructions could be applied to every register. The names of

A Note on AT&T Syntax versus Intel Syntax

Note that the GNU tools use the traditional AT&T syntax, which is used across many processors on Unix-like operating systems, as opposed to the Intel syntax typically used on DOS and Windows systems. The following instruction is given in AT&T syntax:

```
MOVQ %RSP, %RBP
```

`MOVQ` is the name of the instruction, and the percent signs indicate that `RSP` and `RBP` are registers. In the AT&T syntax, the source is always given first, and the destination is always given second.

In other places (such as the Intel manual), you will see the Intel syntax, which (among other things) dispenses with the percent signs and *reverses* the order of the arguments. For example, this is the same instruction in the Intel syntax:

```
MOVQ RBP, RSP
```

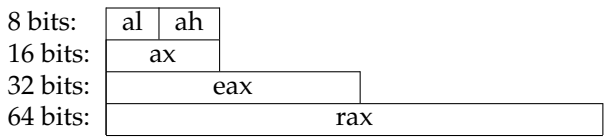
When reading manuals and web pages, be careful to determine whether you are looking at AT&T or Intel syntax: look for the percent signs!

the lower eight registers indicate the purpose for which each was originally intended: for example, %rax is the accumulator.

As the design developed, new instructions and addressing modes were added to make the various registers almost equal. A few remaining instructions, particularly related to string processing, require the use of %rsi and %rdi. In addition, two registers are reserved for use as the stack pointer (%rsp) and the base pointer (%rbp). The final eight registers are numbered and have no specific restrictions.

The architecture has expanded from 8 to 64 bits over the years, and so each register has some internal structure. The lowest 8 bits of the %rax register are an 8-bit register %al, and the next 8 bits are known as %ah. The low 16 bits are collectively known as %ax, the low 32-bits as %eax, and the whole 64 bits as %rax.

Figure 10.1: X86 Register Structure



The numbered registers %r8-%r15 have the same structure, but a slightly different naming scheme:

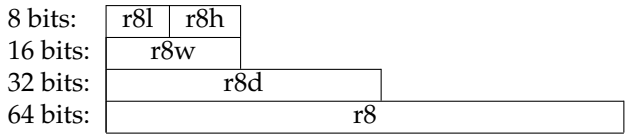


Figure 10.2: X86 Register Structure

To keep things simple, we will focus our attention on the 64-bit registers. However, most production compilers use a mix of modes: a byte can represent a boolean; a longword is usually sufficient for integer arithmetic, since most programs don't need integer values above 2^{32}; and a quadword is needed to represent a memory address, enabling up to 16EB (exa-bytes) of virtual memory.

10.3.2 Addressing Modes

The `MOV` instruction moves data between registers and to and from memory in a variety of different modes. A single letter suffix determines the size of data to be moved:

Suffix	Name	Size	
B	BYTE	1 byte	(8 bits)
W	WORD	2 bytes	(16 bits)
L	LONG	4 bytes	(32 bits)
Q	QUADWORD	8 bytes	(64 bits)

`MOVB` moves a byte, `MOVW` moves a word, `MOVL` moves a long, `MOVQ` moves a quad-word.[1] Generally, the size of the locations you are moving to and from must match the suffix. In some cases, you can leave off the suffix, and the assembler will infer the right size. However, this can have unexpected consequences, so we will make a habit of using the suffix.

The arguments to MOV can have one of several addressing modes.

- A **global value** is simply referred to by an unadorned name such as `x` or `printf`, which the assembler translates into an absolute address or an address computation.
- An **immediate value** is a constant value indicated by a dollar sign such as `$56`, and has a limited range, depending on the instruction in use.
- A **register value** is the name of a register such as `%rbx`.

[1] Careful: These terms are not portable. A *word* has a different size on different machines.

- An **indirect value** refers to a value by the address contained in a register. For example, `(%rsp)` refers to the value pointed to by %rsp.
- A **base-relative** value is given by adding a constant to the name of a register. For example, `-16(%rcx)` refers to the value at the memory location sixteen bytes below the address indicated by `%rcx`. This mode is important for manipulating stacks, local values, and function parameters, where the start of an object is given by a register.
- A **complex** address is of the form $D(R_A, R_B, C)$ which refers to the value at address $R_A + R_B * C + D$. Both R_A and R_B are general purpose registers, while C can have the value 1, 2, 4, or 8, and D can be any integer displacement. This mode is used to select an item within an array, where R_A gives the base of the array, R_B gives the index into the array, C gives the size of the items in the array, and D is an offset relative to that item.

Here is an example of using each kind of addressing mode to load a 64-bit value into `%rax`:

Mode	Example
Global Symbol	`MOVQ x, %rax`
Immediate	`MOVQ $56, %rax`
Register	`MOVQ %rbx, %rax`
Indirect	`MOVQ (%rsp), %rax`
Base-Relative	`MOVQ -8(%rbp), %rax`
Complex	`MOVQ -16(%rbx,%rcx,8), %rax`

For the most part, the same addressing modes may be used to store data into registers and memory locations. There are some exceptions. For example, it is not possible to use base-relative for both arguments of MOV: `MOVQ -8(%rbx), -8(%rbx)`. To see exactly what combinations of addressing modes are supported, you must read the manual pages for the instruction in question.

In some cases, you may want to load the *address* of a variable instead of its value. This is handy when working with strings or arrays. For this purpose, use the `LEA` (load effective address) instruction, which can perform the same address computations as `MOV`:

Mode	Example
Global Symbol	`LEAQ x, %rax`
Base-Relative	`LEAQ -8(%rbp), %rax`
Complex	`LEAQ -16(%rbx,%rcx,8), %rax`

10.3.3 Basic Arithmetic

You will need four basic arithmetic instructions for your compiler: integer addition, subtraction, multiplication, and division.

ADD and SUB have two operands: a source and a destructive target. For example, this instruction:

```
ADDQ %rbx, %rax
```

adds %rbx to %rax, and places the result in %rax, overwriting what might have been there before. This requires a little care, so that you don't accidentally clobber a value that you might want to use later. For example, you could translate `c = a+b+b;` like this:

```
MOVQ a, %rax
MOVQ b, %rbx
ADDQ %rbx, %rax
ADDQ %rbx, %rax
MOVQ %rax, c
```

The `IMUL` instruction is a little unusual, because multiplying two 64-bit integers results in a 128-bit integer, in the general case. `IMUL` takes its argument, multiplies it by the contents of %rax, and then places the low 64 bits of the result in %rax and the high 64 bits in %rdx. (This is implicit: %rdx is not mentioned in the instruction.)

For example, suppose that you wish to translate `c = b*(b+a);`, where `a`, and `b`, and `c` are global integers. Here is one possible translation:

```
MOVQ  a, %rax
MOVQ  b, %rbx
ADDQ  %rbx, %rax
IMULQ %rbx
MOVQ  %rax, c
```

The IDIV instruction does the same thing, except backwards: it starts with a 128 bit integer value whose low 64 bits are in %rax and high 64 bits in %rdx, and divides it by the value given in the instruction. The quotient is placed in %rax and the remainder in %rdx. (If you want to implement the modulus instruction instead of division, just use the value of %rdx.)

To set up a division, you must make sure that both registers have the necessary sign-extended value. If the dividend fits in the lower 64 bits, but is negative, then the upper 64 bits must all be ones to complete the twos-complement representation. The CQO instruction serves the very specific purpose of sign-extending %rax into %rdx for divsion.

For example, to divide a by five:

```
MOVQ a,  %rax     # set the low 64 bits of the dividend
CQO               # sign-extend %rax into %rdx
IDIVQ $5          # divide %rdx:%rax by 5,
                  #          leaving result in %rax
```

The instructions INC and DEC increment and decrement a register destructively. For example, the statement `a = ++b` could be translated as:

```
MOVQ b, %rax
INCQ %rax
MOVQ %rax,b
MOVQ %rax, a
```

The instructions AND, OR, and XOR perform destructive *bitwise* boolean operations on two values. Bitwise means that the operation is applied to each individual bit in the operands, and stored in the result.
So, `AND $0101B $0110B` would yield the result `$0100B`. In a similar way, the NOT instruction inverts each bit in the operand. For example, the bitwise C expression `c = (a & ~b);` could be translated like this:

```
MOVQ a, %rax
MOVQ b, %rbx
NOTQ %rbx
ANDQ %rax, %rbx
MOVQ %rbx, c
```

Be careful here: these instructions *do not* implement logical boolean operations according to the C representation that you are probably familiar with. For example, if you define "false" to be the integer zero, and true to be any non-zero value. In that case, `$0001` is true, but `NOT $0001B` is `$1110B`, which is also true! To implement that correctly, you need to use CMP with conditionals described below.[2]

Like the MOV instruction, the various arithmetic instructions can work on a variety of addressing modes. However, for your compiler project, you will likely find it most convenient to use MOV to load values in and out of registers, and then use only registers to perform arithmetic.

[2] Alternatively, you could could use the bitwise operators as logical operators if you give `true` the integer value -1 (all ones) and `false` the integer value zero.

10.3.4 Comparisons and Jumps

Using the JMP instruction, we may create a simple infinite loop that counts up from zero using the %rax register:

```
      MOVQ $0, %rax
loop: INCQ %rax
      JMP loop
```

To define more useful structures such as terminating loops and if-then statements, we must have a mechanism for evaluating values and changing program flow. In most assembly languages, these are handled by two different kinds of instructions: compares and jumps.

All comparisons are done with the CMP instruction. CMP compares two different registers and then sets a few bits in the internal EFLAGS register, recording whether the values are the same, greater, or lesser. You don't need to look at the EFLAGS register directly. Instead a selection of conditional jumps examine the EFLAGS register and jump appropriately:

Instruction	Meaning
JE	Jump if Equal
JNE	Jump if Not Equal
JL	Jump if Less
JLE	Jump if Less or Equal
JG	Jump if Greater
JGE	Jump if Greater or Equal

For example, here is a loop to count %rax from zero to five:

```
        MOVQ $0, %rax
loop:   INCQ %rax
        CMPQ $5, %rax
        JLE  loop
```

And here is a conditional assignment: if global variable x is greater than zero, then global variable y gets ten, else twenty:

```
        MOVQ x, %rax
        CMPQ $0, %rax
        JLE  .L1
.L0:
        MOVQ $10, $rbx
        JMP  .L2
.L1:
        MOVQ $20, $rbx
.L2:
        MOVQ %rbx, y
```

Note that jumps require the compiler to define target labels. These labels must be unique and private within one assembly file, but cannot be seen outside the file unless a `.globl` directive is given. Labels like `.L0`, `.L1`, etc, can be generated by the compiler on demand.

10.3.5 The Stack

The stack is an auxiliary data structure used primarily to record the function call history of the program along with local variables that do not fit in registers. By convention, the stack grows *downward* from high values to low values. The %rsp register is known as the **stack pointer** and keeps track of the bottom-most item on the stack.

To push %rax onto the stack, we must subtract 8 (the size of %rax in bytes) from %rsp and then write to the location pointed to by %rsp:

```
SUBQ $8, %rsp
MOVQ %rax, (%rsp)
```

Popping a value from the stack involves the opposite:

```
MOVQ (%rsp), %rax
ADDQ $8, %rsp
```

To discard the most recent value from the stack, just move the stack pointer the appropriate number of bytes :

```
ADDQ $8, %rsp
```

Of course, pushing to and popping from the stack referred to by %rsp is so common, that the two operations have their own instructions that behave exactly as above:

```
PUSHQ %rax
POPQ  %rax
```

Note that, in 64-bit code, `PUSH` and `POP` are limited to working with 64-bit values, so a manual `MOV` and `ADD` must be used if it is necessary to move smaller items to/from the stack.

10.3.6 Calling a Function

Prior to the 64-bit architecture described here, a simple stack calling convention was used: arguments were pushed on the stack in reverse order, then the function was invoked with `CALL`. The called function looked for the arguments on the stack, did its work, and returned the result in `%eax`. The caller then removed the arguments from the stack.

However, 64-bit code uses a register calling convention, in order to exploit the larger number of available registers in the X86-64 architecture. [3] This convention is known as the **System V ABI** [2] and is written out in a lengthy technical document. The complete convention is quite complicated, but this summary handles the basic cases:

Figure 10.3: Summary of System V ABI Calling Convention

- The first six integer arguments (including pointers and other types that can be stored as integers) are placed in the registers %rdi, %rsi, %rdx, %rcx, %r8, and %r9, in that order.
- The first eight floating point arguments are placed in the registers %xmm0-%xmm7, in that order.
- Arguments in excess of those registers are pushed onto the stack.
- If the function takes a variable number of arguments (like `printf`) then the %rax register must be set to the number of floating point arguments.
- The return value of the function is placed in %rax.

In addition, we also need to know how the remaining registers are handled. A few are **caller saved**, meaning that the calling function must save those values before invoking another function. Others are **callee saved**, meaning that a function, when called, must save the values of those registers, and restore them on return. The argument and result registers need not be saved at all. Figure 10.4 shows these requirements.

To invoke a function, we must first compute the arguments and place them in the desired registers. Then, we must push the two caller-saved registers (`%r10` and `%r11`) on the stack, to save their values. We then issue the `CALL` instruction, which pushes the current instruction pointer on to the stack then jumps to the code location of the function. Upon return from the function, we pop the two caller-saved registers off of the stack, and look for the return value of the function in the `%rax` register.

[3] Note that there is nothing *stopping* you from writing a compiler that uses a stack calling convention. But if you want to invoke functions compiled by others (like the standard library) then you need to stick to the convention already in use.

Here is an example. The following C program:

```
int x=0;
int y=10;

int main()
{
    x = printf("value: %d",y);
}
```

could be translated to this:

```
.data
x:
      .quad 0
y:
      .quad 10
str:
      .string "value: %d\n"

.text
.globl main
main:
      MOVQ  $str, %rdi  # first argument in %rdi: string
      MOVQ  y,    %rsi  # second argument in %rsi: y
      MOVQ  $0,   %rax  # there are zero float args

      PUSHQ %r10        # save the caller-saved regs
      PUSHQ %r11

      CALL  printf      # invoke printf

      POPQ  %r11        # restore the caller-saved regs
      POPQ  %r10

      MOVQ  %rax, x     # save the result in x

      RET               # return from main function
```

Figure 10.4: System V ABI Register Assignments

Register	Purpose	Who Saves?
%rax	result	not saved
%rbx	scratch	callee saves
%rcx	argument 4	not saved
%rdx	argument 3	not saved
%rsi	argument 2	not saved
%rdi	argument 1	not saved
%rbp	base pointer	callee saves
%rsp	stack pointer	callee saves
%r8	argument 5	not saved
%r9	argument 6	not saved
%r10	scratch	CALLER saves
%r11	scratch	CALLER saves
%r12	scratch	callee saves
%r13	scratch	callee saves
%r14	scratch	callee saves
%r15	scratch	callee saves

10.3.7 Defining a Leaf Function

Because function arguments are passed in registers, it is easy to write a **leaf function** that computes a value without calling any other functions. For example, code for the following function:

```
square: function integer ( x: integer ) =
{
    return x*x;
}
```

Could be as simple as this:

```
.global square
square:
    MOVQ  %rdi, %rax  # copy first argument to %rax
    IMULQ %rax        # multiply it by itself
                      # result is already in %rax
    RET               # return to caller
```

Unfortunately, this won't work for a function that wants to invoke other functions, because we haven't set up the stack properly. A more complex approach is needed for the general case.

10.3.8 Defining a Complex Function

A complex function must be able to invoke other functions and compute expressions of arbitrary complexity, and then return to the caller with the original state intact. Consider the following recipe for a function that accepts three arguments and uses two local variables:

```
.globl func
func:
    pushq %rbp          # save the base pointer
    movq  %rsp, %rbp    # set new base pointer

    pushq %rdi          # save first argument on the stack
    pushq %rsi          # save second argument on the stack
    pushq %rdx          # save third argument on the stack

    subq  $16, %rsp     # allocate two more local variables

    pushq %rbx          # save callee-saved registers
    pushq %r12
    pushq %r13
    pushq %r14
    pushq %r15

    ### body of function goes here ###

    popq %r15            # restore callee-saved registers
    popq %r14
    popq %r13
    popq %r12
    popq %rbx

    movq   %rbp, %rsp    # reset stack pointer
    popq   %rbp          # recover previous base pointer
    ret                  # return to the caller
```

There is a lot to keep track of here: the arguments given to the function, the information necessary to return, and space for local computations. For this purpose, we use the base register pointer %rbp. Whereas the stack pointer %rsp points to the end of the stack where new data will be pushed, the base pointer %rbp points to the start of the values used by this function. The space between %rbp and %rsp is known as the **stack frame** for this function call.

There is one more complication: each function needs to use a selection of registers to perform computations. However, what happens when one function is called in the middle of another? We do not want any registers currently in use by the caller to be clobbered by the called function. To prevent this, each function must save and restore all of the registers that it uses by pushing them onto the stack at the beginning, and popping them off of the stack before returning. According to Figure 10.4, each function must preserve the values of %rsp, %rbp, %rbx, and %r12-%r15 when it completes.

Here is the stack layout for `func`, defined above:

Contents	Address	
old %rip register	8(%rbp)	
old %rbp register	(%rbp)	← %rbp points here
argument 0	-8(%rbp)	
argument 1	-16(%rbp)	
argument 2	-24(%rbp)	
local variable 0	-32(%rbp)	
local variable 1	-40(%rbp)	
saved register %rbx	-48(%rbp)	
saved register %r12	-56(%rbp)	
saved register %r13	-64(%rbp)	
saved register %r14	-72(%rbp)	
saved register %r15	-80(%rbp)	← %rsp points here

Figure 10.5: Example X86-64 Stack Layout

Note that the base pointer (`%rbp`) locates the start of the stack frame. So, within the body of the function, we may use base-relative addressing against the base pointer to refer to both arguments and locals. The arguments to the function follow the base pointer, so argument zero is at -8(%rbp), argument one at -16(%rbp), and so forth. Past those are local variables to the function at -32(%rbp) and then saved registers at -48(%rbp). The stack pointer points to the last item on the stack. If we use the stack for additional purposes, data will be pushed to further negative values. (Note that we have assumed all arguments and variables are 8 bytes large: different types would result in different offsets.)

Here is a complete example that puts it all together. Suppose that you have a C-minor function defined as follows:

```
compute: function integer
         ( a: integer, b: integer, c: integer )
{
    int x, y;
    x = a+b+c;
    y = x*5;
    return y;
}
```

A complete translation of the function is on the next page. The code given is correct, but rather conservative. As it turned out, this particular function didn't need to use registers `%rbx`–`%r15`, so it wasn't necessary to save and restore them. In a similar way, we could have kept the arguments in registers without saving them to the stack. The result could have been computed directly into `%rax` rather than saving it to a local variable. These optimizations are easy to make when writing code by hand, but not so easy when writing a compiler.

For your first attempt at building a compiler, your code created will (probably) not be very efficient if each statement is translated independently. The preamble to a function must save all the registers, because it does not know *a priori* which registers will be used later. Likewise, a statement that computes a value must save it back to a local variable, because it does not know beforehand whether the local will be used as a return value. We will explore these issues later in Chapter 12 on optimization.

Figure 10.6: Complete X86 Example

```
.globl compute
compute:
#################### preamble of function sets up stack
pushq %rbp          # save the base pointer
movq  %rsp, %rbp    # set new base pointer to rsp

pushq %rdi          # save first argument (a) on the stack
pushq %rsi          # save second argument (b) on the stack
pushq %rdx          # save third argument (c) on the stack

subq  $16, %rsp     # allocate two more local variables

pushq %rbx          # save callee-saved registers
pushq %r12
pushq %r13
pushq %r14
pushq %r15

######################## body of function starts here
movq  -8(%rbp),  %rbx   # load each arg into a register
movq  -16(%rbp), %rcx
movq  -24(%rbp), %rdx

addq  %rdx, %rcx        # add the args together
addq  %rcx, %rbx
movq  %rbx, -32(%rbp)   # store the result into local 0 (x)

movq  -32(%rbp), %rbx   # load local 0 (x) into a register.
movq  $5, %rcx          # load 5 into a register
movq  %rbx, %rax        # move argument in rax
imulq %rcx              # multiply them together
movq  %rax, -40(%rbp)   # store the result in local 1 (y)

movq  -40(%rbp), %rax   # move local 1 (y) into the result

################### epilogue of function restores the stack
popq %r15          # restore callee-saved registers
popq %r14
popq %r13
popq %r12
popq %rbx

movq %rbp, %rsp    # reset stack to base pointer.
popq %rbp          # restore the old base pointer

ret                # return to caller
```

10.4 ARM Assembly

The ARM processor architecture has a history almost as long as the X86 architecture. It originated as the 32-bit **Acorn RISC Machine** used in the **Acorn Archimedes** personal computer in 1987, and has since evolved into a wide-used low-power CPU employed in many embedded and mobile systems, now known as the **Advanced RISC Machine (ARM)**. The evolving architecture has been implemented by a number of chip vendors working from a common architecture definition. The most recent versions of the architecture are ARMv7-A (32-bit) and ARMv8-A (64-bit.) This chapter will focus on the 32-bit architecture, with some remarks on differences in the 64-bit architecture at the end.

ARM is an example of a **Reduced Instruction Set Computer (RISC)** rather than a **Complex Instruction Set Computer**CISC. Compared to X86, ARM relies on a smaller set of instructions which are simpler to pipeline or run in parallel, reducing the complexity and energy consumption of the chip. ARM is sometimes considered "partially" RISC due to a few exceptions. For example, the difference in the time to execute some ARM instruction makes the pipeline imperfect, the inclusion of a barrel shifter for preprocessing brings forward more complex instructions, and conditional execution decreases some of the potential instructions executed and lead to less branching instructions so less energy used by the processor. We will mainly be focusing on the elements of the instruction set which allow us to do the most in a compiler, leaving the more complex aspects and optimizations of the programming language for later.

10.4.1 Registers and Data Types

ARM-32 has a set of 16 general purpose registers, `r0-r15`, with the following conventions for use:

Name	Alias	Purpose
r0		General Purpose Register
r1		General Purpose Register
...		. . .
r10		General Purpose Register
r11	fp	Frame Pointer
r12	ip	Intra-Procedure-Call Scratch Register
r13	sp	Stack Pointer
r14	lr	Link Register (Return Address)
r15	pc	Program Counter

In addition to general purpose registers, there are two registers that cannot be directly accessed: the Current Program Status Register (CPSR) and the Saved Program Status Register (SPSR). These hold the results of comparison operations, as well as privileged data regarding the process

state. A user-level program cannot access these directly, but they can be set as a side effect of some operations.

ARM uses the following suffixes to indicate data sizes. Note that these have *different* meanings than in X86 assembly! If no suffix is given, the assembler assumes an unsigned word operand. Signed types are used to provide appropriate sign-extension when loading a small data type into a larger register. There is no register naming structure for anything below a word.

Data Type	**Suffix**	**Size**
Byte	B	8 bits
Halfword	H	16 bits
Word	W	32 bits
Double Word	-	64 bits
Signed Byte	SB	8 bits
Signed Halfword	SH	16 bits
Signed Word	SW	32 bits
Double Word	-	64 bits

10.4.2 Addressing Modes

ARM makes the use of two different classes of instructions to move data between registers and between registers and memeory. `MOV` copies data and constants between registers, while `LDR` (load) and `STR` (store) instructions are used to move data between registers and memory.

The `MOV` instruction is used to move a known immediate value into a given register or move another register into the first register. In ARM, immediate values are denoted by a `#`. However, these immediate values must be 16-bits or less. If they are greater, the `LDR` instruction must be used instead. Most ARM instructions indicate the destination register on the left and the source register on the left. (`STR` is an exception.) So for moving data between immediate values and registers we would have the following:

Mode	**Example**
Immediate	`MOV r0, #3`
Register	`MOV r1, r0`

The mnemonic letter for each data type can be appended to the `MOV` instruction allowing us to be sure of which is being transferred and how that data is being transferred. If not, the assembler assumes an entire word.

To move values in and out of memory, use the Load (`LDR`) and Store (`STR`) instructions, both of which indicate the source or destination register as the first argument, and the source or destination memory address as the second argument. In the simplest case, the address is given by a register and indicated in brackets:

Figure 10.7: ARM Addressing Modes

Address Mode	Example
Literal	`LDR Rd, =0xABCD1234`
Absolute Address	`LDR Rd, =label`
Register Indirect	`LDR Rd, [Ra]`
Preindexing - Immediate	`LDR Rd, [Ra, #4]`
Preindexing - Register	`LDR Rd, [Ra, Ro]`
Preindexing - Immediate & Writeback	`LDR Rd, [Ra, #4]!`
Preindexing - Register & Writeback	`LDR Rd, [Ra, Ro]!`
Postindexing - Immediate	`LDR Rd, [Ra], #4`
Postindexing - Register	`LDR Rd, [Ra], Ro`

```
LDR Rd, [Ra]
STR Rs, [Ra]
```

In this case, Rd denotes the destination register, Rs denotes the source register and Ra denotes the register containing the address. (Note that the memory address must be aligned to the data type: a byte can load a value from any address, a half-word from every even address, and so on.)

Both `LDR` and `STR` support a variety of addressing modes, shown in Figure 10.7. First, `LDR` can be used to load a literal value (or label address) of a full 32-bits into a register. (See the next section for a full explanation of this.) Unlike X86, there is no single instruction that loads a value from a given memory address. To accomplish this, you must first load the address into a register, and then perform a register-indirect load:

```
LDR r1, =x
LDR r2, [r1]
```

A number of more complex addressing modes are available which facilitate the implementation of pointers, arrays, and structures in high level programs. Preindexing modes add a constant (or register) to a base register, and then load from the computed address:

```
LDR r1, [r2, #4] ; Load from address r2 + 4
LDR r1, [r2, r3] ; Load from address r2 + r3
```

It is also possible to write-back to the base register by appending a bang (!) character. This indicates that the computed address should be saved in the base register after the address is loaded:

```
LDR r1, [r2, #4]! ; Load from r2 + 4 then r2 += 4
LDR r1, [r2, r3]! ; Load from r2 + r3 then r2 += r3
```

Post-indexing modes do the same thing, but in the reverse order. First, the load is performed from the base register, then the base register is incremented:

```
LDR r1, [r2], #4 ; Load from r2 then r2 += 4
LDR r1, [r2], r3 ; Load from r2 then r2 += r3
```

These complex pre-indexing and post-indexing modes make it possible to have single-instruction implementations of idiomatic C operations like `b = a++`. The corresponding modes are also available to the `STR` instruction.

Absolute addresses (and other large literals) are somewhat more complicated in ARM. Because every instruction must fit into a 32-bit word, it is not possible to fit a 32-bit address into an instruction, alongside the opcode. Instead, large literals must be stored in a **literal pool**, which is a small region of data inside the code section of the program. A literal can be loaded from a pool with a PC-relative load instruction, which can reference ± 4096 bytes from the loading instruction. This results in several small literal pools being scattered throughout the program, so that each one is close to the load instruction that uses it.

The ARM assembler generally hides this complexity from the user. When a label or large literal is prefixed with an equals sign, this indicates to the assembler that the marked value should be placed into a literal pool, and a corresponding PC-relative instruction emitted instead.

For example, the following instructions, which load the address of `x` into `r1`, then the value of `x` into `r2`:

```
LDR r1, =x
LDR r2, [r1]
```

Will be expanded into this load of the address of `x` from an adjacent literal pool, followed by loading the value of `x`:

```
LDR r1, .L1
LDR r2, [r1]
B   .end
.L1:
    .word x
.end:
```

10.4.3 Basic Arithmetic

ARM provides three-address arithmetic instructions on registers. The `ADD` and `SUB` instructions specify the result register as the first argument, and compute on the second and third arguments. The third operand may be an 8-bit constant, or a register with an optional shift applied. The variants with carry enabled will set or clear the C bit of the CPSR on overflow, as appropriate.

Instruction	Example
Add without carry	`ADD Rd, Rm, Rn`
Add with carry	`ADC Rd, Rm, Rn`
Subtract without carry	`SUB Rd, Rm, Rn`
Subtract with carry	`SBC Rd, Rm, Rn`

Multiplication works much the same way, except that multiplying two 32-bit numbers could result in a 64-bit value. The ordinary `MUL` discards the high bits of the results, while `UMULL` puts the 64-bit result in two 32-bit registers. The signed variant `SMULL` will sign extend the high register as needed.

Instruction	Example
Multiplication	`MUL Rd, Rm, Rn`
Unsigned Long Multiplication	`UMULL RdHi, RdLo, Rm, Rn`
Signed Long Multiplication	`SMULL RdHi, RdLo, Rm, Rn`

There is no division instruction in ARM, because it cannot be carried out in a single pipelined cycle. Instead, when division is needed, it is accomplisehd by invoking an external function in a standard library. This is left as an exercise for the reader.

The logical instructions are very similar in structure to the arithmetic instructions. We have the bitwise-and, bitwise-or, bitwise-exclusive-or and bitwise-bit-clear, which is the equivalent of a bitwise-and of the first value and the inverted second value. The move-not `MVN` instruction performs a bitwise-not while moving from one register to another.

Instruction	Example
Bitwise-And	`AND Rd, Rm, Rn`
Bitwise-Or	`ORR Rd, Rm, Rn`
Bitwise-Xor	`EOR Rd, Rm, Rn`
Bitwise-Bit-Clear	`BIC Rd, RM, Rn`
Move-Not	`MVN Rd, Rn`

10.4.4 Comparisons and Branches

The `CMP` instruction compares two values and sets the N (negative) and Z (zero) flags in the CPSR, to be read by later instructions. In the case of comparing a register and an immediate value, the register must be on the right had side of the comma:

```
CMP Rd, Rn
CMP Rd, #imm
```

In addition, an "S" can be appended to the arithmetic instructions to update the CPSR in a similar way. For example, `SUBS` will subtract two values, store the result, and update the CPSR.

Figure 10.8: ARM Branch Instructions

Opcode	Meaning		
B	Branch Always	BL	Branch and Link
BX	Branch and Exchange	BLX	Branch-Link-Exchange
BEQ	Equal	BVS	Overflow Set
BNE	Not Equal	BVC	Overflow Clear
BGT	Greater Than	BHI	Higher (unsigned >)
BGE	Greater Than or Equal	BHS	Higher or Same (uns. >=)
BLT	Less Than	BLO	Lower (unsigned <)
BLE	Less Than or Equal	BLS	Lower or Same (uns. <=)
BMI	Negative	BPL	Positive or Zero

A variety of branch instructions consult the previously-set values of the CPSR, and then jump to a given label, if the appropriate flags are set. An unconditional branch is specified with simply `B`.

For example, to count from zero to five:

```
        MOV r0, #0
loop:   ADD r0, r0, 1
        CMP r0, #5
        BLT loop
```

And to conditionally assign a global variable y ten if x is greater than 0 and 20 if it is not

```
        LDR r0, =x
        LDR r0, [r0]
        CMP r0, #0
        BGT .L1
.L0:
        MOV r0, #20
        B   .L2
.L1:
        MOV r0, #10
.L2:
        LDR r1, =y
        STR r0, [r1]
```

The branch-and-link (`BL`) instruction, is used to implement function calls. `BL` sets the link register to be the address of the next instruction, and then jumps to the given label. The link register is then used as the return address at the conclusion of the function. The `BX` instruction branches to the address given in a register, and is most frequently used to return from a

function call by branching to the link register. `BLX` performs a branch-and-link to the address given by a register, and can be used to invoke function pointers, virtual methods, or any other indirect jump.

A special feature of the ARM instruction set is **conditional execution**. A 4-bit field in each instruction word indicates one of 16 possible conditions that must hold, otherwise the instruction is ignored. The various types of conditional branch shown above are simply a plain branch (`B`) instruction with the various conditions applied. The same two letter suffixes can be applied to almost any instruction.

For example, suppose we have the following code fragment, which increments either `a` or `b`, depending on which one is smaller:

```
if(a<b) { a++; } else { b++; }
```

Instead of implementing this as control flow using branches and labels, we can simply make each of the two additions conditional upon a previous comparison. Whichever condition holds true will be executed, and the other skipped. Assuming that `a` and `b` are held in `r0` and `r1` respectively:

```
CMP   r0, r1
ADDLT r0, r0, #1
ADDGE r1, r1, #1
```

10.4.5 The Stack

The stack is an auxiliary data structure used primarily to record the function call history of the program along with local variables that do not fit in registers. By convention, the stack grows *downward* from high values to low values. The `sp` register is known as the **stack pointer** and keeps track of the bottom-most item on the stack.

To push the `r0` register onto the stack, we must subtract the size of the register from `sp`, and then store `r0` into the location pointed to by `sp`:

```
SUB sp, sp, #4
STR r0, [sp]
```

Alternatively, this can be done with a single instruction making use of pre-indexing and writeback:

```
STR r0, [sp, #-4]!
```

The `PUSH` pseudo-instruction accomplishes the same thing, but can also move any number of registers (encoded as a bitmask) to the stack. Curly braces are used to indicate the list of registers to push:

```
PUSH {r0,r1,r2}
```

Popping a value off the stack involves the opposite:

Figure 10.9: Summary of ARM Calling Convention

- The first four arguments are placed in registers r0, r1, r2 and r3.
- Additional arguments are pushed on the stack in reverse order.
- The caller must save r0-r3 and r12, if needed.
- the caller must always save r14, the link register.
- The **callee** must save r4-r11, if needed.
- The result is placed in r0.

Figure 10.10: ARM Register Assignments

Register	Purpose	Who Saves?
r0	argument 0 / result	not saved
r1	argument 1	CALLER saves
r2	argument 2	CALLER saves
r3	argument 3	CALLER saves
r4	scratch	callee saves
...	...	...
r10	scratch	callee saves
r11	frame pointer	callee saves
r12	intraprocedure	CALLER saves
r13	stack pointer	callee saves
r14	link register	CALLER saves
r15	program counter	saved in link register

```
LDR r0, [sp]
ADD sp, sp, #4
```

Once again this can be done with a single instruction:

```
LDR r0, [sp], #4
```

And, to pop a set of registers all at once:

```
POP {r0,r1,r2}
```

Unlike X86, any data items ranging from a byte to a double-word can be pushed on to the stack, as long as data alignment is respected.

10.4.6 *Calling a Function*

ARM uses a register calling convention described by the ARM-Thumb Procedure Call Standard (ATPCS) [4], which is summarized in Figure 10.9.

To call a function, place the desired arguments in the registers `r0-r3`, save the (current) value of the link register, and then use the `BL` instruction to jump to the function. Upon return, restore the previous value of the link register, and examine the result in register `r0`.

For example, the following C function:

```
int x=0;
int y=10;
int main() {
    x = printf("value: \%d\n",y);
}
```

Would become the following in ARM:

```
.data
  x:  .word 0
  y:  .word 10
  S0: .ascii "value: \%d\012\000"
.text
  main:
    LDR r0, =S0  @ Load address of S0
    LDR r1, =y   @ Load address of y
    LDR r1, [r1] @ Load value of y

    PUSH {ip,lr} @ Save registers
    BL  printf   @ Call printf
    POP  {ip,lr} @ Restore registers

    LDR r1, =x   @ Load address of x
    STR r0, [r1] @ Store return value in x
.end
```

10.4.7 *Defining a Leaf Function*

Because function arguments are passed in registers, it is easy to write a **leaf function** that computes a value without calling any other functions. For example, code for the following function:

```
square: function integer ( x: integer ) =
{
    return x*x;
}
```

Could be as simple as this:

```
.global square
square:
    MUL  r0, r0, r0   @ multiply argument by itself
    BX   lr           @ return to caller
```

Unfortunately, this won't work for a function that wants to invoke other functions, because we haven't set up the stack properly. A more complex approach is needed for the general case.

10.4.8 Defining a Complex Function

A complex function must be able to invoke other functions and compute expressions of arbitrary complexity, and then return to the caller with the original state intact. Consider the following recipe for a function that accepts three arguments and uses two local variables:

```
func:
    PUSH {fp}        @ save the frame pointer
    MOV  fp, sp      @ set the new frame pointer
    PUSH {r0,r1,r2}  @ save the arguments on the stack
    SUB  sp, sp, #8  @ allocate two more local variables
    PUSH {r4-r10}    @ save callee-saved registers

    @@@ body of function goes here @@@

    POP  {r4-r10}     @ restore callee saved registers
    MOV  sp, fp       @ reset stack pointer
    POP  {fp}         @ recover previous frame pointer
    BX   lr           @ return to the caller
```

Through this method, we ensure that we are able to save all the values in the registers into the stack and ensure that no data will be lost. The stack, once this has been done, looks very similar to that of the X86 stack, just with some extra callee-saved registers stored on the stack.

Here is a complete example that puts it all together. Suppose that you have a C-minor function defined as follows:

```
compute: function integer
         ( a: integer, b: integer, c: integer )
{
    int x, y;
    x = a+b+c;
    y = x*5;
    return y;
}
```

Figure 10.11: Example ARM Stack Frame

Contents	Address	
Saved r12	[fp, #8]	
Old LR	[fp, #4]	
Old Frame Pointer	[fp]	← fp points here
Argument 2	[fp, #-4]	
Argument 1	[fp, #-8]	
Argument 0	[fp, #-12]	
Local Variable 1	[fp, #-16]	
Local Variable 0	[fp, #-20]	
Saved r10	[fp, #-24]	
Saved r9	[fp, #-28]	
Saved r8	[fp, #-32]	
Saved r7	[fp, #-36]	
Saved r6	[fp, #-40]	
Saved r5	[fp, #-44]	
Saved r4	[fp, #-48]	← sp points here

A complete translation of the function is on the next page. Note that this is one of many ways to construct a valid stack frame for a function definition. Other approaches are valid, as long as the function uses the stack frame consistently. For example, the callee could first push all of the argument and scratch registers on the stack, and then allocate space for local variables below that. (The reverse process must be used on function exit, of course.)

Another common approach is for the callee `PUSH {fp,ip,lr,pc}` on to the stack, before pushing arguments and local variables. While not strictly required for implementing the function, it provides additional debugging information in the form of a **stack backtrace** so that a debugger can look backwards through the call stack and easily reconstruct the current execution state of the program.

The code given is correct, but rather conservative. As it turned out, this particular function didn't need to use registers `r4-r5`, so it wasn't necessary to save and restore them. In a similar way, we could have kept the arguments in registers without saving them to the stack. The result could have been computed directly into `r0` rather than saving it to a local variable. These optimizations are easy to make when writing code by hand, but not so easy when writing a compiler.

For your first attempt at building a compiler, your code created will (probably) not be very efficient if each statement is translated independently. The preamble to a function must save all the registers, because it does not know *a priori* which registers will be used later. Likewise, a statement that computes a value must save it back to a local variable, because

Figure 10.12: Complete ARM Example

```
.globl compute
compute:
@@@@@@@@@@@@@@@@@@@ preamble of function sets up stack
PUSH {fp}         @ save the frame pointer
MOV  fp, sp       @ set the new frame pointer
PUSH {r0,r1,r2}   @ save the arguments on the stack
SUB  sp, sp, #8   @ allocate two more local variables
PUSH {r4-r10}     @ save callee-saved registers

@@@@@@@@@@@@@@@@@@@@@@@@@ body of function starts here
LDR  r0, [fp,#-12]      @ load argument 0 (a) into r0
LDR  r1, [fp,#-8]       @ load argument 1 (b) into r1
LDR  r2, [fp,#-4]       @ load argument 2 (c) into r2
ADD  r1, r1, r2         @ add the args together
ADD  r0, r0, r1
STR  r0, [fp,#-20]      @ store the result into local 0 (x)
LDR  r0, [fp,#-20]      @ load local 0 (x) into a register.
MOV  r1, #5             @ move 5 into a register
MUL  r2, r0, r1         @ multiply both into r2
STR  r2, [fp,#-16]      @ store the result in local 1 (y)
LDR  r0, [fp,#-16]      @ move local 1 (y) into the result

@@@@@@@@@@@@@@@@@@@ epilogue of function restores the stack
POP  {r4-r10}      @ restore callee saved registers
MOV  sp, fp        @ reset stack pointer
POP  {fp}          @ recover previous frame pointer
BX   lr            @ return to the caller
```

it does not know beforehand whether the local will be used as a return value. We will explore these issues later in Chapter 12 on optimization.

10.4.9 64-bit Differences

The 64-bit ARMv8-A architecture provides two execution modes: The A32 mode supports the 32-bit instruction set described above, and the A64 mode supports a new 64-bit execution model. This permits a 64-bit CPU with a supporting operating system to execute a mix of 32-bit and 64-bit programs simultaneously. Though not binary compatible with A32, the A64 model follows much of the same architectural principles, with a few key changes:

Word Size. A64 instructions are still a fixed size of 32 bits, however, registers and address computations are 64 bits.

Registers. A64 has 32 64-bit registers, named `x0`–`x31`. `x0` is a dedicated zero register: when read, it always provides the value zero, when written, there is no effect. `x1`–`x15` are general purpose registers, `x16` and `x17` are for interprocess communication, `x29` is the frame pointer, `x30` is the link register and `x31` is the stack pointer. (The program counter is not directly accessible from user code.) Instead of using a data type suffix, a 32-bit value may be indicated by naming a register as `w#`.

Instructions. A64 instructions are largely the same as A32, using the same mnemonics, with a few differences. Conditional predicates are no longer part of every instruction. Instead, all conditional codes must perform an explicit `CMP` and then a conditional branch. The `LDM`/`STM` instructions and pseudo-instructions `PUSH`/`POP` are not available and must be replaced with a sequence of explicit loads and stores. (This can be made more efficient by using `LDP`/`STP` which load and store pairs of registers.

Calling Convention. To invoke a function, the first eight arguments are placed in registers `x0`–`x7`, and the remainder are pushed on to the stack. The caller must preserve registers `x9`–`x15` and `x30` while the callee must preserve `x19`–`x29`. The (scalar) return value is placed in `x0`, while extended return values are pointed to by `x8`.

10.5 Further Reading

This chapter has given you a brief orientation to the core features of the X86 and ARM architectures, enough to write simple programs in the most direct way. However, you will certainly need to look up specific details of individual instructions in order to better understand their options and limitations. Now you are ready to read the detailed reference manuals and keep them handy while you construct your compiler:

1. Intel64 and IA-32 Architectures Software Developer Manuals. Intel Corp., 2017.http://www.intel.com/content/www/us/en/processors/architectures-software-developer-manuals.html

2. System V Application Binary Interface, Jan Hubicka, Andreas Jaeger, Michael Matz, and Mark Mitchell (editors), 2013. https://software.intel.com/sites/default/files/article/402129/mpx-linux64-abi.pdf

3. ARM Architecture Reference Manual ARMv8. ARM Limited, 2017. https://static.docs.arm.com/ddi0487/bb/DDI0487B_b_armv8_arm.pdf.

4. The ARM-THUMB Procedure Call Standard. ARM Limited, 2000. http://infocenter.arm.com/help/topic/com.arm.doc.espc0002/ATPCS.pdf.