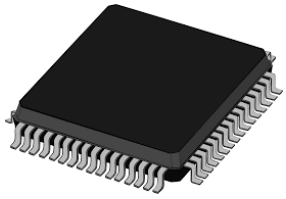
Computer Organization

Lab and Exercise 3



EITF70
EITA15

Low Level Programming

Goals

- Get an understanding of how a CPU works
- Understanding the basics of assembly programming
- Be able to mix C and assembly code
- Understanding when and why assembly language should be used

Christoffer Cederberg
Jonathan Sönnerup
2019

Contents

Introduction

Back in the days, almost all embedded systems were programmed in an assembly language due to heavy timing and memory constraints. Nowadays, memory is relatively cheap and the clock frequency of the CPUs is high. This means that the programmer does not have to program as carefully as before. However, this does not mean that assembly programming is useless. There are still plenty of applications (IoT devices, OS kernels, real-time systems etc.) today having constraints not fulfilled by a compiler, hence assembly must be used.

In this lab, you will learn the basics of the AVR assembly language, and how to use assembly in general. After completing the lab, you should have an understanding when to use assembly and when not to.

Lab Equipment

During this lab, most of the assignment is about looking inside the processor. However, to be a bit more fun, some I/O devices will be used. The external device that will be interfaced is shown in Figure 1 and 2. In Figure 1 the used I/O devices are encapsulated with a red rectangle and labeled with a number and in Figure 2 the corresponding blocks are coloured black.

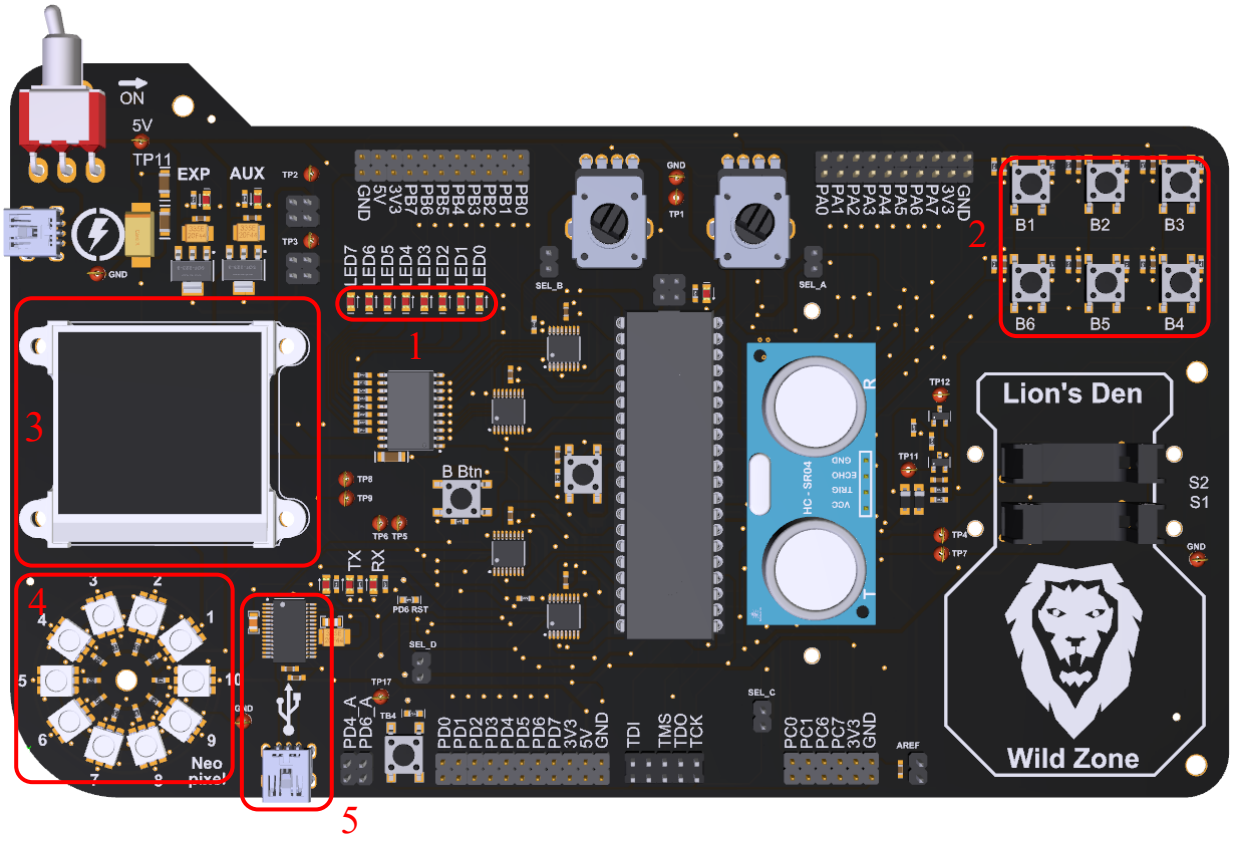


Figure 1: A picture of the circuit board with the used IO devices marked.

1 - LED's Eight LEDs that are connected to port B. By setting a pin to high (3.3 V) the corresponding LED emits photons with the energy 2.8^{-19} J ;).

2 - Buttons There are six button on the circuit broad. They are connected to port A on the on the microcontroller.

3 - OLED The single most expensive part on the circuit board, but also the coolest. The uOLED-128G2 is an OLED screen programmable via UART. It has a resolution of 128x128 pixels with 65K colors. `https://www.4dsystems.com.au/product/uOLED_128_G2/`

4 - Neopixels Neopixels are LEDs with adjustable RGB colors controlled via a serial protocol.

5 - USB-UART interface This integrated circuit translates UART to USB (and the reverse) and enables the microcontroller to communicate with a PC. In order to send data back and forth to a PC, a USB cable needs to be connected between the USB mini connector on the circuit board and the PC.

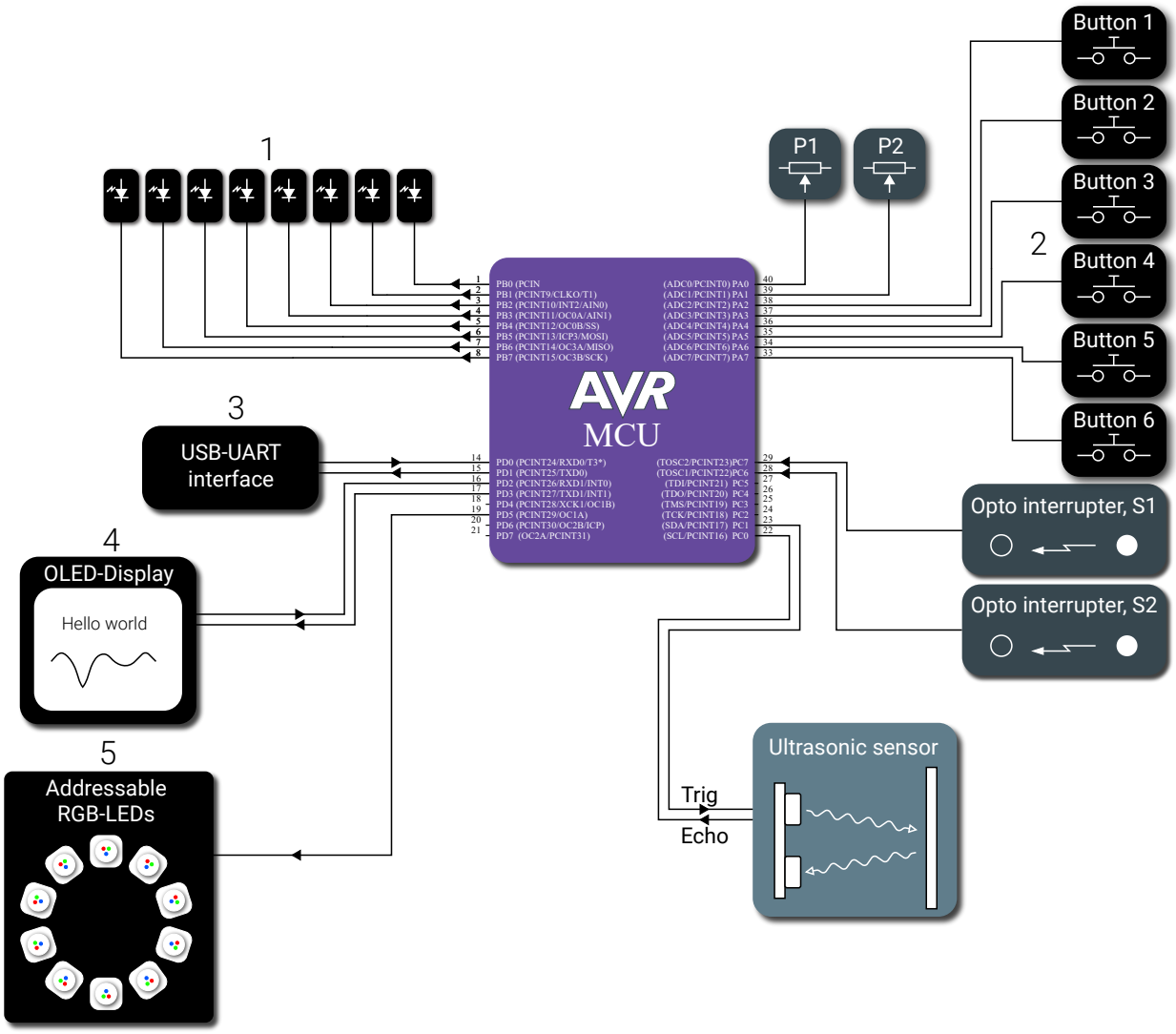


Figure 2: A picture of the circuit board with the used I/O devices marked.

Chapter 1

Assembly Programming

Just like any programming language, we have a certain amount of instructions available in assembly. A simple but working example in AVR is shown in Listing 1.1. For a listing of useful instructions available in the AVR processor, see Appendix **??**.

```
start:
    ldi r16, 12      ; load 12 into register 16
    ldi r17, 13      ; load 13 into register 17
    add r16, r17     ; add r16 and r17, save result in r16
    cpi r16, 26      ; compare r16 with the values 26
    breq end         ; jump to end if true
    add r16, 1       ; add 1 to r16, r16 now holds the value 26
                     ; after this instruction, we will execute the rjmp end
end:
    rjmp end         ; infinite loop
```

Listing 1.1: Simple AVR assembly program.

As seen in the code, there is no `main` function. In AVR assembly, the name of the main function is `start`. The names `start` and `end` are known as *labels*. This is just a name of the actual address where the code resides, to make it easy for a programmer. A label can be used in two ways: to create a place to jump to, like a for-loop or similar, but also to create a function (subroutine) where the label must be called with the `call` instruction. The `end` label in Listing 1.1, together with the "`rjmp end`" instruction, is used to create an infinite loop, just like a `while(1)` loop in C.

1.1 Conditional Expressions

Every mathematical operation affect a so called status register, `SREG`. The status register contains flags such as overflow, carry, and zero. For example, the zero flag, `Z`, is set to 1 if the result after an operation is 0. This can be used to create conditional expressions, such as if statements et cetera. In Listing 1.1, we have such a case. The `cpi` instruction compares the value of a register, `r16`, with the value 26. The compare operation is really a subtraction in the ALU. The result is either 0 or not, and the `Z` flag in the status register is set accordingly. Next, the `breq` instruction (branch if equal) automatically checks the `Z` flag to decide whether to jump (branch) or not. In this case, the result from the `cpi` instruction is $25 - 26 = -1$, and the `Z` flag is set to 0 (the result was not zero). Hence, the branch is not taken, and the program continues with the next instruction, in this case, an `add` instruction.

1.2 I/O in Assembly

To communicate with the surrounding circuits, we need to access the I/O registers. For that, we use the `in` and `out` instructions. An example is shown in Listing 1.2. Note that just as in C, we can define names to numbers and addresses to make it easier to remember. In this case, we defined `DDRB` and `PORTB` to avoid writing the addresses in the assembly code. To make it more interesting (and a little confusing), there are basically two ways of accessing I/O ports in assembly. Either by using normal load and store instructions, or by using special I/O instruction such as `in` and `out`. There are 2 differences, one being that the I/O instructions are more efficient (1 clock cycle instead of 2), the other being that the value $0x20$ should be subtracted from the addresses to the I/O ports. The reason is that they are mapped in different ways in hardware. Table 1.1 shows the addresses when using the special I/O instructions. **Note** that using `in` and `out` is preferable to use.

Table 1.1: I/O registers and their corresponding addresses.

Register	Address
PINA	0x00
DDRA	0x01
PORTA	0x02
PINB	0x03
DDRB	0x04
PORTB	0x05
PINC	0x06
DDRC	0x07
PORTC	0x08
PIND	0x09
DDRD	0x0A
PORTD	0x0B

```
#define DDRB    0x04
#define PORTB   0x05

start:
    ldi r18, 0x01
    out DDRB, r18    ; set LED0 as output
    out PORTB, r18   ; turn on LED0. r18 already contains 0x01 ;)
end:
    rjmp end         ; infinite loop
```

Listing 1.2: Simple AVR assembly program using I/O.

1.3 Measuring Execution Time

Every instruction takes a certain amount of clock cycles. We can use that information in order to evaluate the execution time of a program. Some instructions takes different amount of clock cycles based on the outcome. For example, a branch instruction takes 1 clock cycle if the branch is **not** taken, and 2 clock cycles if it is. Analyzing the program in Listing 1.3 yields the result shown in Table 1.2. Note that in order to

analyze the program, we must unroll the loop, counting clock cycles based on if the branch is taken or not. The total execution time (before the infinite loop) is 9 clock cycles. Running at 16 MHz, this yields

$$9 \cdot \frac{1}{16\text{ M}} = 562.5\text{ ns}.$$

```
start:
    ldi r16, 1          ; load 1 into register 16
    ldi r17, 13         ; load 13 into register 17
    loop:
        add r17, r16    ; add r17 and r16, save result in r17 (r17++)
        cpi r17, 15     ; compare r17 with the values 15
        brne loop       ; jump to loop if r17 != 15
end:
    rjmp end            ; infinite loop
```

Listing 1.3: Simple AVR assembly program.

Table 1.2: Caption

Instruction	Execution time
ldi r16, 1	1
ldi r17, 13	1
loop:	
add r17, r16	1
cpi r17, 15	1
brne loop	2 (taken)
add r17, r16	1
cpi r17, 15	1
brne loop	1 (not taken)
end:	
rjmp end	∞

The general formula for calculating the **average** execution time of a **single round** in a loop (say, a for-loop) is

$$\frac{n-1}{n} \cdot p + \frac{1}{n} \cdot q, \tag{1.1}$$

where n is the number of rounds, p is the number of clock cycles in a round when the branch is taken, and q is the number of clock cycles during the last round (branch not taken). Using the formula for the loop in Listing 1.3, we get

$$\frac{2-1}{2} \cdot (1+1+2) + \frac{1}{2} \cdot (1+1+1) = \frac{1}{2} \cdot 4 + \frac{1}{2} \cdot 3 = 3.5.$$

That is, the average execution time is 3.5 clock cycles. Running for 2 rounds yields 7 clock cycles, which matches the number in Table 1.2. For very long loops, the last round does not affect the average time very much. For example, running the same loop 100 times instead would yield

$$\frac{99}{100} \cdot 4 + \frac{1}{100} \cdot 3 = 3.99,$$

which is basically 4 clock cycles per round, which we get if do not take into account that the last branch instruction only takes 1 clock cycle. In general, for large numbers, n, we get

$$\lim_{n\to\infty} \frac{n-1}{n} \cdot p + \frac{1}{n} \cdot q = p.$$

What constitutes as a *large* number depends on the application. Make sure not to approximate too early, or you may end up with a completely wrong estimation.

1.4 Exercises

Answers to the questions can be found in Appendix B.1.

Assembly Basics

1.1 Write an assembly program that loads the values 3 and 4 into register `r4` and `r5`, respectively. Add the two numbers and store the result in register `r5`.

1.2 Assume a user have entered their age in a program. The value is stored in register `r16`. Write a program such that if the age is above or equal to 18, you store the value 1 in register `r24`, otherwise you store the value 2 in `r24`. After the check, you shall just enter an infinite loop.

1.3 Write a program that checks if a button is pressed. If it is, an LED shall be turned on, otherwise the LED shall be turned off. The button is connected on pin 2 (3rd pin) at address 0x12. The LED is connected on pin 1 (2nd pin) at address 0x15. Remember to not change the other pins as they may be used for other devices. The code should run in an infinite loop.

1.4 Translate the following C program into AVR assembly. Assume that the global variable `input` is located in register `r20`, and that `a` is located in register `r24`.

```c
#include <avr/io.h>

char input;

int main()
{
    char a;
    if (input < 12) {
        a = 3;
    } else if (input == 12) {
        a = 2;
    } else {
        a = 5;  // input > 12
    }

    while (1);  // infinite loop
}
```

Listing 1.4: Small C program.

1.5 Your friend Dan recently attended a course in cryptography at LTH. He now believes that he knows enough crypto and decides to implement his own encryption algorithm. He wants to run his code on his smart watch, using an AVR processor, but unfortunately his code (Listing 1.5) is just too slow. He knows that you are taking a course writing super fast code in assembly, hence he asks you for your assistance.

```c
int main()
{
    char msg;
    char msg_enc;

    msg_encrypted = msg;
    for (int i = 0; i < 17; i++) {
        msg_enc |= (msg >> 2);  // shift msg 2 steps, then OR it in with msg_enc
        msg_enc <<= 3;          // shift msg_enc 3 steps to the left
        msg_enc ^= msg;         // xor with msg
    }
}
```

Listing 1.5: Super unsafe encryption.

Help your friend Dan speed the code up by implementing it in AVR assembly. **Note**: You should never implement your own encryption algorithms as they will always fail, i.e., Dan did not learn anything in the crypto course.

Assume that `msg` is located in register `r16` and `msg_enc` in register `r24`.

Performance Analysis

1.6 Write an assembly program that takes exactly 302 clock cycles to execute.

1.7 Given the following program,

```
start:
    ldi r16, 10

    loop1:
        ldi r17, 255
        loop2:
            dec r17
            brne loop2

        dec r16
        brne loop1

end:
    rjmp end
```

Listing 1.6: Assembly loops.

(a) What does the code do?

(b) Write the corresponding C code using 2 loops

(c) Write the corresponding C code using 1 loop

(d) How many clock cycles does it take before `end` is reached? Approximate the clock cycles in the loops. That is, do not use Equation 1.1.

1.8 Given the following program,

```
start:
    ldi r16, 255

    loop1:
        ldi r17, 10
        loop2:
            dec r17
            brne loop2

        dec r16
        brne loop1

end:
    rjmp end
```

Listing 1.7: Assembly loops revisited.

(a) How many clock cycles does the program take now, using approximations in the loops?

(b) How much does it differ from the previous program?

(c) Why is this?

1.5 Lab Exercises

1.5.1 Hello, Assembly

Just as always when we are to try new things, we implement a simple program as a sanity check. Here, we will use the AVR assembly language to blink an LED.

Tasks:

- Create a new project in Atmel Studio, choose "AVR Assembler", under the "Assembler" tab to the left, instead of "C/C++".

  Home Assignment 1.1
  How many clock cycles are needed if a delay of 0.1 s is desired with a clock frequency of 16 MHz?

  Home Assignment 1.2
  How many bits do we need to count to the value above?

  Home Assignment 1.3
  How many registers do we need for the delay of 0.1 s?

  Home Assignment 1.4
  Write a snippet of assembly code that delays the program for roughly 0.1 s. Remember that the jump instructions also take time to execute.

  Home Assignment 1.5
  How do you turn on LED 2 on port B in assembly? Remember to set the direction to an output. Refer to Appendix **??** for instructions.

- Use your delay code to create a program that blinks an LED at 10 Hz.

  Lab Question 1.1
  How can you use your delay code in order to create arbitrary delays? What is the limitation?

- Make the LED blink at 1 Hz instead.

You are now done with this part, show your work to a lab assistant! _____ ☑

Chapter 2

Memory Management and Subroutines

In this chapter a couple of important topics will be addressed. The first section will cover how to read and write data to and from the RAM. After this the stack will be introduced and how one can use it. The last section will treat subroutines.

2.1 Accessing the RAM

The fact that the AVR processor's work registers are merely eight bits long should by this stage be well known. That means that the largest integer that fit in them are 255.

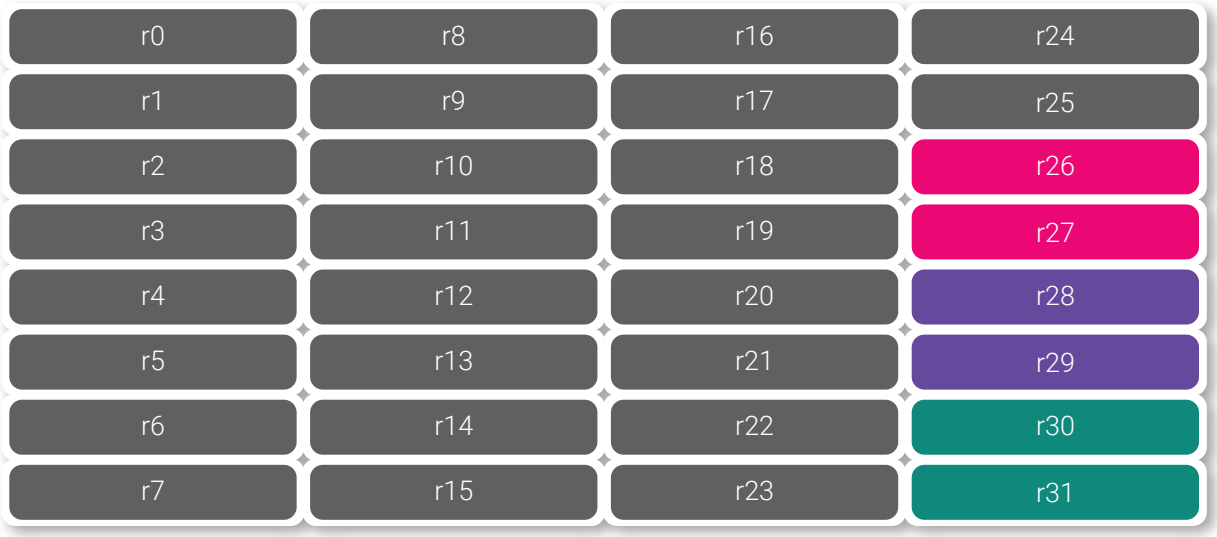


Figure 2.1: The register file in the processor.

2.2 The Stack

The stack is the part of the RAM used for local, temporary, variables. In C, the allocation of local variables is handled automatically by the compiler. In assembly, this must be done by you. The stack grows from higher addresses to lower. This means that when allocating memory on the stack, we must **subtract** the stack pointer with number of bytes we want allocated. When returning the memory, we **add** the same number to the stack pointer. Failing to properly return stack memory will result in program crashes. Recall that when calling subroutines, the return address is being stored on the stack, hence failing to manage the stack makes a subroutine return back to the wrong address.

A stack pointer is used to keep track of the "top". For an AVR processor, the stack pointer is a 16-bit, memory mapped, value located in the memory at address 0x3D (no offset). On start-up, the stack pointer is set to the end of RAM, i.e., the highest address, which is 0x40FF. For each byte added to the stack the stack pointer is decremented. This means that the stack is growing towards lower addresses.

There are two special instructions directly associated with the stack. The first is `push`. With this instruction the content of the specified register will be stored at the address given by the stack pointer. After a `push` instruction is performed the stack pointer is decremented (the stack grows towards lower addresses). The second instruction is the `pop`, which is the opposite. It move the last added byte from the stack to the specified register and increment the stack pointer.

Another way of allocating and deallocating space on the stack is to add or subtrack a given number from to/from the stack pointer. An example of this is shown in Listing 2.1. The code allocates two bytes on the stack and then stores two bytes of data. This is achieved by:

- loading the stack pointer into `r28` and `r29` (later used as the Y register),
- subtract two to `r28` and `r29` (the result is stored in the same registers),
- writing the content of `r28` and `r29` back to the stack pointer register,
- loading `r20` and `r21` with data from address `0x63` and `0x59`, respectively,
- writing `r20` and `r21` to the stack, using the Y register, i.e., `r28` and `r29`.

```
#define STACK_H 0x3E    ; Address to the high byte of the stack pointer
#define STACK_L 0x3D    ; Address to the low byte of the stack pointer
#define N_ALLOC 2

warp:
    .
    .                   ; Code to initialize anti-annihilation chamber
    .
    in r28, STACK_L     ; Load low byte of stack pointer to r28
    in r29, STACK_H     ; Load high byte of stack pointer to r29
    sbiw Y, N_ALLOC     ; Subtract N from the loaded stack pointer
    out STACK_L, r28    ;
    out STACK_H, r29    ; Update stack pointer
    in r20, 0xCC        ; Load r20 with warp core start sequence
    in r21, 0xA1        ; Load r21 with wormhole coordinate system variable
    std Y+1, r20        ; Store content of r20 on the stack
    std Y+2, r21        ; Store content of r21 on the stack
    .
    .                   ; Warp core calibration code
    .
    ret
```

Listing 2.1: A snippet of assembly code that allocates an array with the size of 2 bytes on the stack.

The stack before and after running the snippet of code can be viewed in Figure 2.2. As seen, the stack pointer has been decremented after the execution and the two bytes of data has been added. It should be noted that the stack pointer always points on the next free byte, which is below (lower address) the "top of stack"[1].

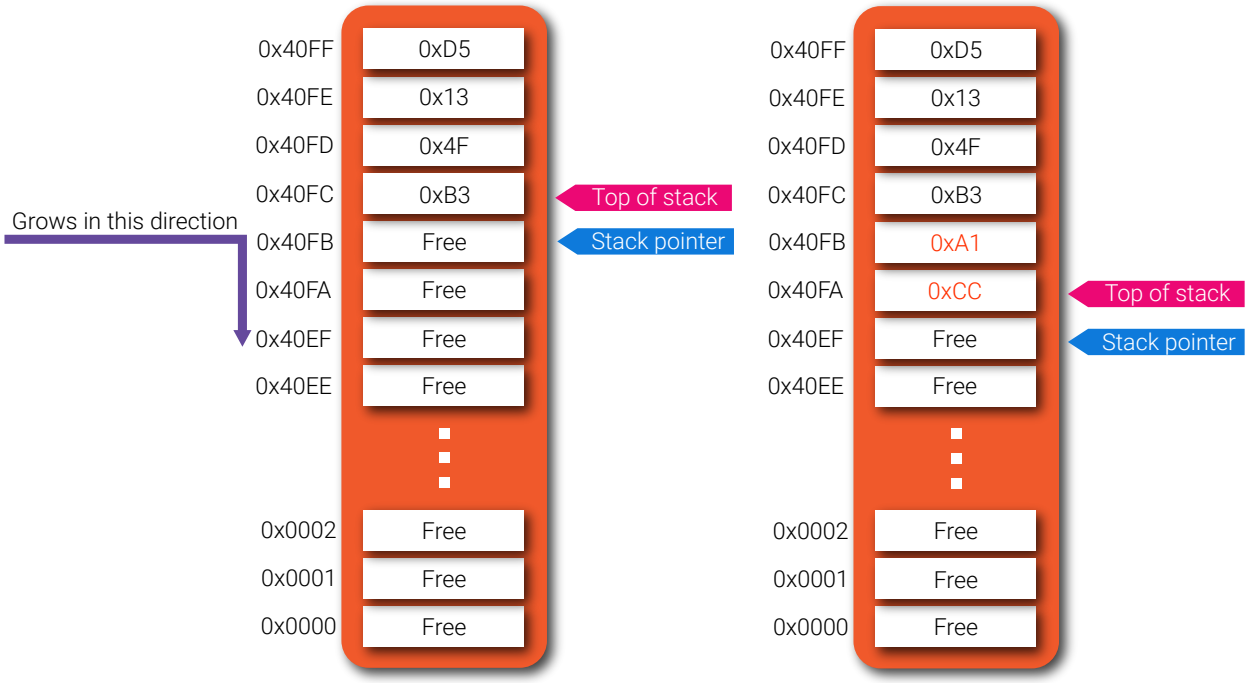


Figure 2.2: The stack before (left) and after running the code in Listing 2.1.

Whenever the stack has been used, that is, the need for storing data is no longer present, the allocated bytes need to be deallocated. If this is not done, the stack will keep growing and soon or later it will collide with the other content stored in the RAM. To return the allocated space, the number of bytes that was allocated should be added to the stack pointer. By doing so the memory will be freed up, see Figure 2.3. Do note that the content of the previously used memory space is still there (but it should be considered to be lost in cyberspace).

[1]The top of stack is purely imaginary, there is no register in the CPU that contains this address.

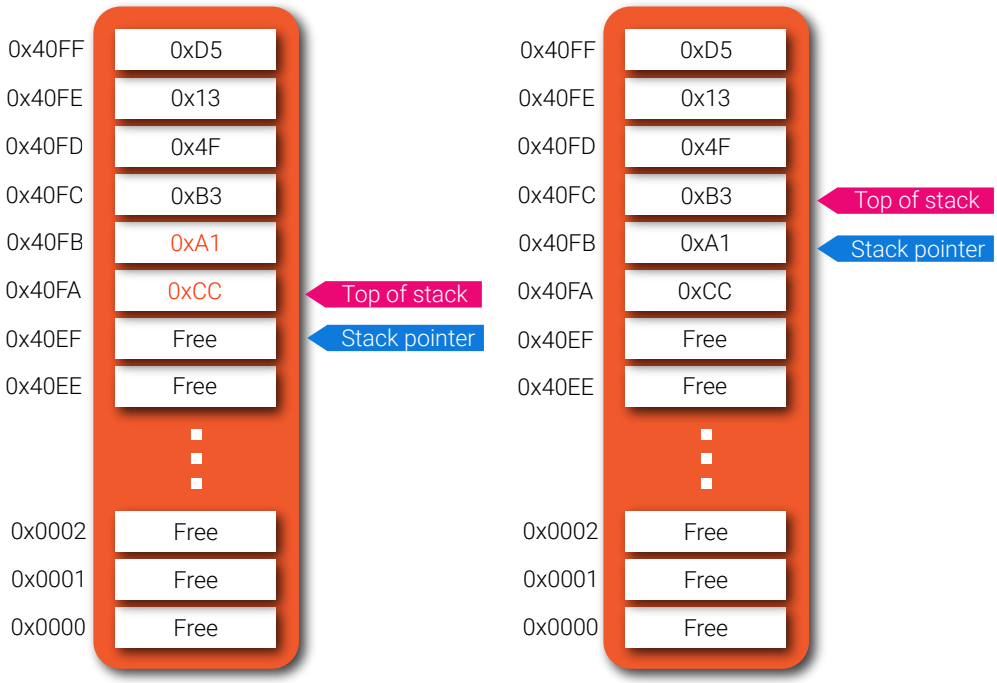


Figure 2.3: The stack before and after (right) the allocated bytes have been "returned".

2.3 Subroutines

Just as we can separate code in different functions in C and Java, assembly lets us separate code in different subroutines, using labels. As before, a label can be used both in loops to jump to, but also to declare a subroutine. The difference is how they are used. To jump to a label (say, in a loop), any branch instruction may be used. When declaring a subroutine, it must be called using the `call` instruction.

When calling a subroutine, using `call`, an important thing happens in the processor. The address of the instruction after the call, which is PC[2] + 1, is pushed on the stack. This is important since we must be able to continue where we left off. An example is shown in Listing 2.2.

```
start:
    call delay          ; delay forever
    rjmp start

delay:
    ldi r16, 100        ; r16 = 100
    loop:
        dec r16         ; r16--
        brne loop       ; loop if r16 != 0
    ret                 ; return from subroutine
```

Listing 2.2: Calling subroutines in assembly.

As shown in the code, we can use labels in both ways at the same time, here used both for declaring a subroutine, `delay`, and also for the loop. Every subroutine must end with a `ret` instruction. This will pop the return address from the stack and jump back to the `start` routine by setting the program counter to the popped value. After this, `delay` will be called again, and again.

For the keen reader a disassembled[3] version of the code can be found in Listing 2.3. And to make life a little bit harder than it already is, it needs to be said that the program memory (the FLASH) is word-address, where one word is 16 bits. This is in contrast to the RAM, which is byte-addressed. Furthermore, some instructions, for instance the `call` instruction, has the length of two words (32-bit), whereas many others only are one word (16-bits) long. Thereof the gap in addresses between the `call` and `rjmp` instruction in Listing 2.3.

```
;Address    Opcode (disassembled code)  Assembly Instruction    Comment
;-------------------------------------------------------------------------------------
                                       start:
00000000    CALL 0x00000003            call delay              ; Calling the delay subroutine
00000002    RJMP PC-0x0002             rjmp start              ; Jump back to start
00000003    LDI R16,0x64               ldi r16, 100            ; Load r16 with 100
00000004    DEC R16                    dec r16                 ; Decrement r16 with 1
                                       loop;
00000005    BRNE PC-0x01               brne loop               ; Branch if not equal to zero
00000006    RET                        ret                     ; Return to address 00000002
```

Listing 2.3: The dissasembled code from Listing 2.3.

[2]This is the program counter which contains the the address to the instruction that is executed at the moment.

[3]A disassembled code is the true code that the processor will execute, i.e., code without labels, `#define`s and so on. The assembly code can not in all cases be direct mapped to raw machine code.

2.3.1 The Anatomy of a Subroutine

A subroutine is a bit more primitive than its high level counterpart in that sense that the programmer needs to do some parts of the compilers job. What the programmer needs to do will be addressed in this section.

A subroutine is divided into three parts, the prologue, body and the epilogue. See Figure 2.4. Each part has its specific purpose.

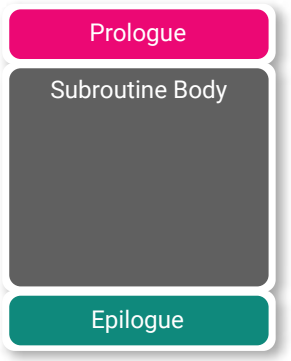


Figure 2.4: The anatomy of a subroutine.

When a subroutine has been called it is important the every register that will be used during the call and which content is of importance to the caller after the subroutine has executed, is being saved on the stack. This should, of course, be done in the beginning of the subroutine. The registers is preferably saved with the `push` instruction. Furthermore, if more space on the stack is needed, it should also be allocated here (for ease of use, after the pushing the registers). This part of the code is referred to as the prologue. In Listing 2.4 an example of this is shown.

```
#define STACK_H 0x3E     ; Address to the high byte of the stack pointer
#define STACK_L 0x3D     ; Address to the low byte of the stack pointer
#define N_ALLOC 5

defense_routine:
    ;==================== Start of the prologue
    push r19
    push r28
    in r28, STACK_L      ; Load low byte of stack pointer to r28
    in r29, STACK_H      ; Load high byte of stack pointer to r29
    sbiw Y, N_ALLOC      ; Subtract N_ALLOC from the loaded stack pointer
    out STACK_L, r28     ;
    out STACK_H, r29     ; Update stack pointer
    ;==================== End of the prologue


    ;==================== Beginning of subroutine body
    .
    .
    .
    .
```

Listing 2.4: The prologue of an assembly subroutine.

In the part of the routine called the subroutine body the routine performs its designated task. When this is

done the epilogue starts. The first thing to do here is to free up all memory that has been allocated on the stack. The second thing is to restore the registers that were push on the stack. This is achieved by adding an appropriate value from the stack pointer. When this is done the stack pointer will be pointing at the last value that was push to the stack in the prologue. When all of this is done the subroutine can return back from where the subroutine was called the (with the `ret` instruction). In Listing 2.5 an example of an epilogue is shown.

```
#define STACK_H 0x3E    ; Address to the high byte of the stack pointer
#define STACK_L 0x3D    ; Address to the low byte of the stack pointer
#define N_ALLOC 5

defense_routine:
    .
    .
    .
    ;==================== End of the subroutine body

    ;==================== Beginning of the epilogue
    in r28, STACK_L     ; Load low byte of stack pointer to r28
    in r29, STACK_H     ; Load high byte of stack pointer to r29
    addiw Y, N_ALLOC    ; Add N_ALLOC to the loaded stack pointer
    out STACK_L, r28    ;
    out STACK_H, r29    ; Update stack pointer
    pop r28
    pop r19
    ret
    ;==================== End of the epilogue and the subroutine
```

Listing 2.5: The epilogue of an assembly subroutine.

2.4 Exercises

2.5 Lab Exercises

2.5.1 Dynamic Memory Allocation

Sometimes, the general purpose working registers are not enough store the data you use. Assume that you are sampling your beautiful singing voice from a microphone over a time period. With only 32 registers in the AVR, we will run out of space pretty quick. In such a case the RAM is utilized. For global variables the `.data` section is used. For local variables (inside a function) the data is usually stored on the stack.

The authors recognize that this assignment is a bit artificial, but it highlights some interesting points students need to know, in a simple way.

Tasks:

- Create a new "Assembler" project in Atmel Studio.

  Home Assignment 2.1
  How do you allocate memory on the stack?

  Home Assignment 2.2
  How do you allocate 10 integers on the stack?

- Write an assembly subroutine that allocates an array of 5 bytes on the stack.

  Lab Question 2.1
  What is the value of the stack pointer directly after allocation? Use the debugger.

  Home Assignment 2.3
  If you allocate an array of bytes on the stack, and the stack pointer after allocation is 0x40E0, what is the address of the value at index 3?

- In the subroutine, sample the state of the buttons B1 to B6 in a loop and save the five latest in the array. The first sample shall be placed at index zero (arr[0]), the fifth sample at index four. After saving the values, simply return from the subroutine.

  Remember that when you borrow something, you must return it.

- In the main function, call your subroutine in a forever-loop.
- Start the debugger and step through each sample to verify that you can read the values of the buttons. Also check the stack to verify that the samples are placed at the correct addresses.

Lab Question 2.2
What happens if we do not return the allocated memory in the subroutine?

Lab Question 2.3
Do we need to store the values in the allocated array? What happens if we just store the values randomly on the stack?

You are now done with this part, show your work to a lab assistant! _____ ☑

Chapter 3

Calling Conventions – Mixing C and Assembly

When calling a subroutine, arguments must be passed to the function and a return value is (sometimes) returned. There are several ways one can pass arguments to a function. These "ways" are called *calling conventions*. It is an agreement of how to pass and return values. There are in practice two general ways this can be done: by putting all arguments on the stack, or by using registers together with the stack.

When writing purely in assembly, the programmer may choose any convention he or she desires, as long as they are being consistent. However, when mixing C and assembly code, certain rules must be followed. The C compiler will follow a specific calling convention and the code written in assembly **must** comply, otherwise the code is likely to crash.

In AVR processors, the calling convention used is the following. Arguments are placed in the registers `r25` to `r8`, using two registers each. If there are additional arguments, they are pushed on the stack. If we want to pass a `char` to a subroutine, we place it in register `r24` and `r25`. Since the size of a `char` is only one byte, we put it `r24` and we let `r25` be zero. Note that we place the lowest significant byte in the register with the lowest value. An example of different function calls and how the values are passed is shown in Table 3.1.

See Section 6 in `http://ww1.microchip.com/downloads/en/appnotes/doc42055.pdf`. Note that there is a typo for register `r22`. It should be `b2`, not `b2b`.

When returning a value from a subroutine, the value is placed in register `r24` and `r25`, the lowest byte placed in `r24`. If the return value is an 8-bit value, you do not have to clear register `r25`.

An example of C code calling a subroutine in assembly is shown in Listing 3.1 and in Listing 3.2 respectively.

```
extern char myadd(char, char);

int main()
{
    char a = 12;
    char b = 24;
    char c = myadd(a, c);
}
```

Listing 3.1: C program that calls a subroutine.

Table 3.1: Function calls with arguments places in registers.

Function call	Register values
`char a = 0x12;` `func(a)`	`r25` = 0x00 `r24` = 0x12
`char a = 0x1234;` `func(a);`	`r25` = 0x12 `r24` = 0x34

`char a = 0x23;`
`char b = 0x42;`
`func(a, b);`
	`r25` = 0x00
	`r24` = 0x23
	`r23` = 0x00
	`r22` = 0x42

```
.global myadd

myadd:
    ; input arguments are in r24 and r22
    add r24, r22     ; add arguments and store the result in r24 (return value)
    ret              ; return from subroutine
```

Listing 3.2: A subroutine adding numbers, being called from C.

Note the two keywords, `extern` in the C code, and `.global` in the assembly code. The `extern` keyword tells the compiler that the function is defined somewhere else. In this case, in the assembly file. The `.global` keyword lets the compiler know that the function shall be accessible outside the assembly file. Without this, the C program will not "see" the function.

3.1 Register Usage

All registers, `r0` to `r31` may be used in a subroutine, but some registers need to be restored because the caller function (C code) expects the values in those registers not to change, see Table 3.2. The registers to be saved shall be pushed on the stack in the subroutine prologue, and popped off the stack, in the epilogue.

For more details, see Section 5 in `http://ww1.microchip.com/downloads/en/appnotes/doc42055.pdf`. Note that there is a typo in the table. The next to last register should be `r30`, not `r0`.

Table 3.2: Register usage in assembly, when called from C code.

Register	**Usage**
`r0`	Save and restore if using
`r1`	Always 0, clear before return if using
`r2-r17`	Save and restore if using
`r28`	Save and restore if using
`r29`	Save and restore if using
`r18-r27`	Can freely use
`30`	Can freely use
`31`	Can freely use

3.2 Exercises

Answers to the questions can be found in Appendix B.3.

Calling Conventions

3.1 Your cannibal friend recently installed a security system, preventing the law enforcement from accessing his *personal belongings*. Help him finish the last piece of code, the compare function checking the password. Implement the subroutine in AVR assembly language, following all conventions. The C code calling the compare routine is shown below. Note that the arguments to the compare function are pointers to the strings. All strings are assumed to be NULL-terminated.

```c
#include <avr/io.h>
#include <stdio.h>

extern uint8_t compare(char *, char *);

char password = "<hidden>";

int main()
{
    char input[32];

    // prompt user for password, over UART
    printf("Enter password: ");

    // read user password and store it in "input".
    fgets(input, 32, stdin);

    // check password
    uint8_t res = compare(input, password);

    if (res == 0) {
        printf("Access granted! Welcome Dr. Lecter.");
    } else {
        printf("Access denied!");
    }
}
```

3.3 Lab Exercises

3.3.1 When Harry Met Sally

Being limited to only write in assembly is daunting for most people (authors excluded), and most would like to write only the critical parts in assembly but the rest in a more user friendly language. Just as we can call Java functions from C, C from Python, and C from Matlab, we can of course call assembly from C (and vice versa).

In this assignment, we will explore the combination of C and assembly.

Tasks:

- Create a new "GCC C Executable" project in Atmel Studio.
- Write a program that blinks an LED if a button is pressed. That is, if a button is pressed, you shall call `led_on()` and `led_off()` with a delay after both functions, use 500 ms. Note that the two functions are not yet implemented, thus you will need to implement them in assembly. On the top of your C-file, add
  - `extern void led_on(char)`,
  - `extern void led_off(char)`,
  - `extern char check_button(char)`.

  This tells the compiler that the functions are defined elsewhere.

Remember to specify the data direction, to make the pins connected to the LEDs outputs.

Remember to define `F_CPU` to 16 MHz before including the <util/delay.h> file.

- Right-click the project name in the solution explorer and select `add -> New Item...` then select `Preprocessing Assembler File (.S)` to add a new assembly file.

Home Assignment 3.1
In AVR assembly, how do you declare a subroutine? How do you call it?

Home Assignment 3.2
In AVR assembly, how are arguments passed to and returned from a subroutine?

- Declare the three subroutines and implement them following all conventions.

Lab Question 3.1
When calling the subroutine, `led_on`, what is being pushed to the stack and why?

Lab Question 3.2
When the `ret` instruction is executed, what happens with the stack pointer?

Lab Question 3.3
In the `check_button` subroutine, comment out the instruction where you place the return value in `r24` and run the code. Does the LED blink? If so, why?

- Run the code in a debugger to verify that the code executes as expected.

When debugging, it is good to comment the `_delay_ms` calls to avoid waiting.

3.3.2 Hello, Mr. Anderson

A NeoPixel, the number 4 in Figure 1, is a single pixel with adjustable RGB colors controlled via a serial protocol. The pixels are chainable, meaning that the output of one pixel can be connected to the input of another pixel in order to have several controllable pixels, not needing more outputs on the microcontroller. This allows for the engineer to create all kinds of shapes and circuits without wasting resources.

Controlling a NeoPixel can be done using only a single pin on the microcontroller. This means that we only can send binary values, ones and zeros. The NeoPixel needs to be able to distinguish between sending data and not sending data, hence sending ones and zeros can not be done by only setting the output pin high or low. Instead, we need to send the data in a pattern, see Figure 3.1.

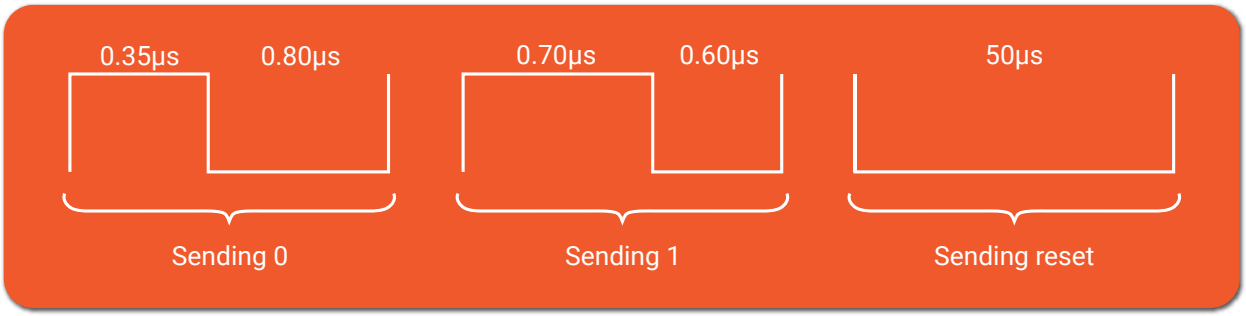


Figure 3.1: A picture of the timings for sending data to a NeoPixel.

As previously stated, sometimes when facing tough timing constraints, a high-level language might not meet the requirements and the programmer must utilize the power of assembly programming. The NeoPixels have such timing constraints. In this part assembly will be used to achieve very accurate timings.

The authors did the heavy lifting and implemented some nice animations. However, they lacked the leet[1] skills to write some of the functions in assembly, which is where you come in.

Tasks:

- Create a new "GCC C Executable" project in Atmel Studio.
- Replace the code in the new C file with the code in `mr_anderson.c` found in folder "Lab 3" on the S: drive.
- Right-click the project name in the solution explorer and select `add -> New Item...` then select `Preprocessing Assembler File (.S)` to add a new assembly file.
- In the assembly file, add two functions, `send_one`, `send_zero` and `send_reset`.

  Home Assignment 3.3
  How many clock cycles are needed for the delays in Figure 3.1.

- In `send_one`, implement the transmission of a logical 1 according to the timing diagram. In `send_zero`, implement the transmission of a logical 0. In `send_reset`, implement the delay according to the specification. Try to be as accurate with the timing as possible, taking into account that all instructions yield some delay.

  The NeoPixels are connected to pin 5 on port D, see Figure 2.

  Remember to only write to the specified pins. There are other circuits connected to the other pins.

- Run the program and check that it works. The NeoPixels should rotate in red.
- Uncomment the line, `read_buttons()`, in the main function, and implement the function in assembly. There is a global variable in the C code, `button_state`, that you shall set in the assembly code. You need to read the state of the buttons and set the value of `button_state` such that is button 1 is pressed, `button_state` = 0x01 (0b00000001). If button 2 is pressed, `button_state` = 0x02 (0b00000010), and so on.

  Remember that the buttons are connected to pin 2 - 7, see Figure 2. That is, you need to shift the values to get the correct values.

[1] Leet, or 1337, is hacker slang for someone being an "elite" or having "elite" skills. Writing 1337 on an old calculator and turning it upside down, the text Leet is seen.

Table 3.3: Actions for the animation.

Button	Action
B1	Set color to red
B2	Set color to green
B3	Set color to blue
B4	Set direction to right
B5	Pause animation
B6	Set direciton to left

To access the `button_state` variable in assembly, add "`.extern button_state`". You can use the name when storing a value to it, e.g., "`sts button_state, r16`", where `r16` contains the value to be written.

- Run the program. See Table 3.3 for the different control options.

You are now done with this part, show your work to a lab assistant! _____ ☑

Chapter 4

Black Magic – The Dark Side of Programming

Using C and/or assembly language gives the programmer (almost) full access to the system. This is sometimes necessary in order to solve specific problems, but may sometimes cause more harm than good. For example, dealing with arrays in C requires more attention of the programmer than in Java. In Java, if an integer array of 10 elements is declared and the programmer tries to add 11 integers, a "`java.lang.ArrayIndexOutOfBoundsException`" exception is thrown. Performing the same test in C results in (usually) nothing. The program seems to work, but sometimes it will crash. There are no security mechanisms preventing the programmer from doing things they should not. What happens is that the array will be *overflowed* and data located next to the array in the memory gets overwritten, without warning.

When a programmer needs some storage to save data, e.g., user names and passwords, they create an array, or a buffer. If the buffer is overflowed, it is called a *buffer overflow*, or a *stack overflow* when it happens to local data stored on the stack.

An example of a stack overflow is shown in Listing 4.1. Here, we see the dangers of a common mistake, an off-by-one error, where the programmer loops one time too many. This results in overwriting the `access` variable, since it is located right after the buffer `buff`. The `access` variable will be given the value 10 after the loop.

```
int main()
{
    char access = 0;
    char buff[10] = { 0 };  // initialize array to 0

    // add values to array, but loops 1 time too many
    for (int i = 0; i <= 10; i++) {
        buff[i] = i;
    }

    if (access == 0) {
        exit(1);    // quit
    } else {
        // release the kraken
        .
        .
        .
    }
```

Listing 4.1: Showing a simple buffer overflow.

4.1 Lab Exercises

4.1.1 Never Trust User Input

Having full access to the memory is both a blessing and a sin. Here, we will explore the dark side of programming.

Tasks:

- Create a new "GCC C Executable" project in Atmel Studio.
- Copy the contents of the file `black_magic.c` from the folder "Lab 3" on the S: drive and paste it in to your newly created C file.
- Make sure to remove both compiler optimization, and the "Garbage collect unused sections" under the Linker optimization tab, in the project settings.

> **Lab Question 4.1**
> What does the code do?

> **Lab Question 4.2**
> **Without** running the code, does it seem like the function `hacked` is ever called?

> **Lab Question 4.3**
> Run the code once, what happens?

- Place a breakpoint at `secure_function` and run the debugger.

> **Lab Question 4.4**
> When calling the `secure_function` function, what will be pushed to the stack and at which addresses?

- Step into the `secure_function` function by clicking "Step Into" or by pressing F11. Then, view the diassembly and step into the `strcpy` function. Continue to step inside `strcpy` and closely follow what happens in the memory.

> **Lab Question 4.5**
> At which addresses do `strcpy` copy the last values? Is this correct? If not, is something overwritten?

Lab Question 4.6
As you exit `strcpy` and executing the `ret` instruction in `secure_function`, what is the return address? Where do we end up?

When debugging, it is good to comment the `_delay_ms` calls to avoid waiting.

You are now done with this part, show your work to a lab assistant! _____ ☑

Appendix A

AVR Instruction Set

An excerpt of AVR assembly instructions is shown in Table A.1. Rd and Rr are registers, usually `r0` to `r31`, and Imm is an immediate value which may be given in decimal, hexadecimal or in binary. One may also use mathematical operations for the immediate value, such as "+", "-", "«" and so on.

Note that not all registers can be used in all instructions. For information regarding valid registers and how many clock cycles the instructions take, see the documentation at `https://www.microchip.com/webdoc/avrassembler/avrassembler.wb_instruction_list.html`.

Table A.1: A short summary of useful instructions.

Category	Instruction	Operation
Arithmetic	`add Rd, Rr`	Rd = Rd + Rr
	`adiw Rd, Imm`	Rd = Rd + Imm
	`sub Rd, Rr`	Rd = Rd - Rr
	`subi Rd, Imm`	Rd = Rd - Imm
	`sbiw Rd, Imm`	Rd = Rd - Imm
	`inc Rd`	Rd = Rd + 1
	`dec Rd`	Rd = Rd - 1
	`and Rd, Rr`	Rd = Rd & Rr
	`andi Rd, Imm`	Rd = Rd & Imm
	`or Rd, Rr`	Rd = Rd | Rr
	`ori Rd, Imm`	Rd = Rd | Imm
	`eor Rd, Rr`	Rd = Rd $\oplus$ Rr
	`lsl Rd`	Rd = Rd << 1
	`lsr Rd`	Rd = Rd >> 1
Branch	`breq label`	Jump to `label` if Rd = Rr
	`brlo label`	Jump to `label` if Rd < Rr
	`brne label`	Jump to `label` if Rd != Rr
	`brsh label`	Jump to `label` if Rd >= Rr
	`rjmp label`	Jump to `label`
Subroutine	`call label`	Call subroutine `label`
	`ret`	Return form subroutine
Memory	`push Rr`	Push value of register Rr on stack
	`pop Rd`	Pop top of stack and store value in `Rd`
	`ld Rd, X/Y/Z`	Load value into `Rd` from address in X, Y or Z
	`ldi Rd, Imm`	Rd = Imm
	`lds Rd, K`	Load value into Rd from address K
	`st Y, Rr`	Store value of Rr to address in Y
	`std Y+k, Rr`	Store value of Rr to addres in Y + k
	`sts Imm, Rr`	Store value of Rr to address Imm
I/O	`in Rd, Imm`	Read from I/O device at address Imm, store in Rd
	`out Imm, Rr`	Write value of Rr to I/O device at address Imm
Other	`mov Rd, Rr`	Copy value from Rr to Rd
	`movw Rd, Rr`	Copy word from Rr to Rd
	`nop`	No operation, do nothing

Appendix B

Answers to Exercise Questions

B.1 Exercise 1

Click here to fast travel back to Section 1.4, price: 1 bit.

1.1 One possible solution is shown below.

```
start:
    ldi r4, 3
    ldi r5, 4
    add r5, r4
end:
    rjmp end
```

1.2 One possible solution is shown below.

```
start:
    cpi r16, 18      ; compare r16 with 18
    brlo lower       ; if r16 < 18, jump to lower
    ldi r24, 1       ; else: store 1 in r24 and jump to end
    rjmp end
lower:
    ldi r24, 2       ; store 2 in r24, then go to end
end:
    rjmp end
```

1.3 One possible solution is shown below.

```
#define BTN 0x12
#define LED 0x15

start:
    in r16, BTN          ; read buttons
    ldi r17, 1 << 2      ; mask for button 2, 0b00000100
    and r16, r17         ; mask out the state of button 2
    breq off             ; if button is not pressed, turn off LED
                         ; else, turn it on
    ldi r16, LED         ; read the LED state
    ldi r17, 1 << 1      ; mask for LED 1
    or r16, r17          ; set bit for LED 1 to high
    out LED, r16         ; write back the value to the output, turning on LED 1
    rjmp start           ; go back to beginning

off:                     ; turn off led
    ldi r16, LED         ; read the LED state
    ldi r17, ~(1 << 1)   ; mask for clearing bit 1
    and r16, r17         ; clear bit 1
    out LED, r16         ; write back to memory
    rjmp start           ; go back to beginning
```

1.4 One possible solution is shown below.

```
start:
    cpi r20, 12
    brlt less
    breq equal
    ldi r24, 5   ; input > 12
    rjmp end
equal:
    ldi r24, 2   ; input == 12
    rjmp end
less:
    ldi r24, 3   ; input < 12

end:
    rjmp end
```

1.5 One possible solution is shown below.

```
start:
    ldi r17, 17      ; loop variable, i = 17
    mov r24, r16     ; copy msg to msg_enc

loop:
    mov r16, r18     ; copy msg to r18
    lsr r18          ;
    lsr r18          ; shift 2 steps to the right (msg >> 2)
    or r24, r18      ; msg_enc |= (msg >> 2)
    lsl r24
    lsl r24          ; shift to the left
    lsl r24          ; msg_enc = msg_enc << 3
    eor r24, r16     ; msg_enc ^= msg

    dec r17          ; i--
    brne loop        ; loop while i > 0

end:
    rjmp end
```

1.6 One possible solution is shown below.

```
start:
    ldi r2, 100
loop:
    dec r2
    brne loop
    nop

end:
    rjmp end
```

1.7 (a) The code loops 2550 times.

(b) The solution is shown below.

```
for (int i = 0; i < 10; i++) {
    for (int j = 0; j < 255; j++) {
        // chill
    }
}
```

(c) The solution is shown below.

```
for (int i = 0; i < 2550; i++) {
    // chill
}
```

B.2 Exercise 2

Click here to fast travel back to Section 2.4, price: 1 bit.

B.3 Exercise 3

Click here to fast travel back to Section 3.2, price: 1 bit.

3.1