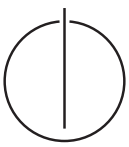
DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Evaluation of WebAssembly IoT Runtimes on a ESP32 Microcontroller

Lukas Heddendorp



DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Evaluation of WebAssembly IoT Runtimes on a ESP32 Microcontroller

Evaluation von WebAssembly IoT Runtimes auf einem ESP32 Microcontroller

Author:	Lukas Heddendorp
Supervisor:	Prof. Dr.-Ing. Jörg Ott
Advisor:	M.Sc. Teemu Kärkkäinen
Submission Date:	16.03.2020

I confirm that this bachelor's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 16.03.2020 Lukas Heddendorp

Abstract

Microcontrollers are all around us and are being used in many different devices. They fulfill particular tasks and are subject to many constraints, such as small memory and limited processing power. In this thesis, we will look at the feasibility of running WebAssembly on the ESP32 microcontroller. WebAssembly is a newly developed bytecode format meant to serve as a compilation target that can be used in any browser to execute optimized code at near-native speeds. Recently the interest around running WebAssmbly on embedded devices has picked up, and we want to evaluate how WebAssembly can be run on the ESP32 microcontroller. For this, we found the WASM3 runtime that can interpret and execute WebAssembly on the ESP32 and performs better than all other currently known WebAssembly interpreters. To test the execution, we designed a collection of workloads modeled after requirements that programs on a microcontroller might have. We ran them while measuring the execution time of native code compared to WebAssembly code. Our tests show that the execution times increase by up to 90 times when interpreting the code as WebAssmbly. While the lower performance and limited support for system interaction pose severe drawbacks to using WebAssembly, there are also significant advantages. Languages that support WebAssembly as a compilation target can be used without being explicitly supported by the platform and modules can be dynamically loaded over the air and executed on the microcontroller without the need for flashing the system.

Contents

1 Introduction

WebAssembly is enabling new experiences on the web and could become a widely used universal bytecode outside of browsers as well. After bringing near-native performance to the web, new use-cases for WebAssembly emerge in the embedded market. In late 2019 the first runtimes for WebAssembly on microcontrollers became publicly available and made it an interesting technology for the IoT market. As defined by Gartner [17]

> The Internet of Things (IoT) is the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment.

Spending on the Internet of Things is prognosed to increase from $646 B in 2018 to $1,100 B in 2023 [20]. As visible in figure 1.1 the number of connected IoT devices is rising steadily as well. This very big market "is moving beyond the hype"[18] and will continue to see a lot of innovation and interest.

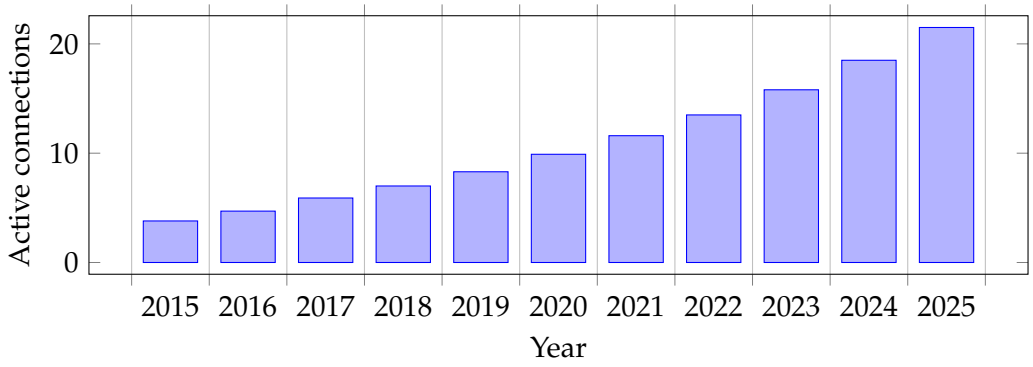


Figure 1.1: IoT active device connections in billions [3]

If WebAssembly can be used on IoT devices, it opens the programming of embedded devices up for new deployment models and new languages that were previously not supported. Being a universal bytecode that is optimized for being transmitted over the network, it can be used to send the same program to many different devices over the air.

There is not much research available that concerns the usage of WebAssembly in embedded devices or the ESP32 specifically. Most of the research is currently focused on

the applications inside the browser. In this thesis, we want to layout a method to assess the feasibility of running WebAssembly and explore its drawbacks and advantages. Before we explain our approach, we will give some background on the essential parts of the work.

Microcontrollers Meant for executing specific tasks, microcontrollers are small computers with minimal resources. They are designed with the aim to have just enough resources while keeping costs low. A popular system on a chip in this class is the ESP32 family. They are very affordable and can be used from experimentation and prototyping to production products. Their connectivity options and CPU performance makes them a great fit for IoT devices. We will test running WebAssembly on the ESP32 in this thesis, which uses the realtime operating system FreeRTOS.

WebAssembly Since its beginning with static pages, the web has evolved to become a universal platform for applications, available on many different devices. However, even though browser engines have made significant progress at optimizing JavaScript, the only natively supported language on the web, the performance is still not reliable and depends on the optimizations deployed by the browser engine. To solve this problem, WebAssembly was created. It is a new, low-level bytecode format that allows running optimized code on browsers at near-native speeds. Being adopted by all major browser vendors, it is now almost universally available [11].

However, since WebAssembly has no explicit dependencies on the web platform, its attributes such as portability, safety, and speed, making it very useful outside of the browser too. Runtimes meant to be used on embedded devices have become available recently and might open exciting new options for programming a microcontroller.

Assessing WebAssembly

In order to assess the current state of WebAssembly on the ESP32, we found a runtime, WASM3, which has support for the ESP32 running FreeRTOS. While other runtimes are available already, most of them only target desktop PCs. The only other runtime for embedded use, the WebAssembly micro runtime, does not support the ESP32 operating system. WASM3 also achieves the best execution speeds amongst WebAssembly runtimes in benchmarks [31].

The comparison we are interested in is between the execution of code compiled to WebAssembly and compiled to native code. For this, we designed a collection of Workloads[1] that are inspired by real-world applications. We ran those tests as

[1] All tests can be accessed at `https://github.com/Isigiel/bsc-thesis/tree/master/code/platforms`

WebAssembly and native code and measured the different behavior to gain more insight into the drawbacks and advantages of running WebAssembly.

After running the test and interpreting results, we will draw some learnings from our measurements and research that can help developers when considering it for a new IoT project. Apart from that, all tests are meant to be reproducible on other hardware platforms or with other engines once they are released to compare the performance and asses WebAssembly in other setups. If a new runtime for embedded devices is published, they can be rerun with minor changes to have an immediate comparison.

2 Background

Since WebAssembly on embedded devices is a new topic that just recently surfaced, we will introduce some concepts that are important to follow the thesis. While both microcontrollers and WebAssembly are widely available, there are specific details about them that are essential to be aware of in order to assess the use of WebAssembly on the ESP32 microcontroller.

2.1 Microcontrollers

A microcontroller (**MCU** for microcontroller unit) is a small computer, meant to fulfill a particular requirement without a complex operating system. They are designed for embedded applications from implantable medical devices to toys and very prominently in IoT devices. Bigger devices will often have multiple microcontrollers, each responsible for a particular function. A car, for example, could include, amongst others, an MCU to control the mirror adjustments, one to handle fuel injection and another one for traction control.

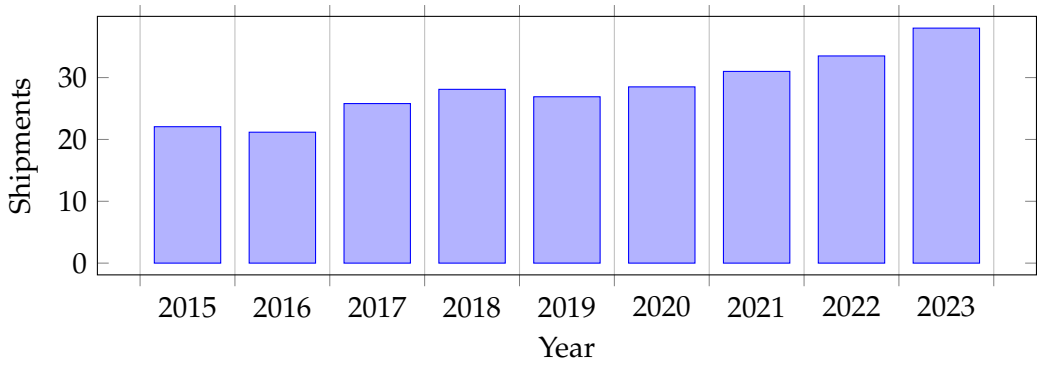


Figure 2.1: MCU shipments worldwide from 2015 to 2023 (in billions) [21]

Core elements of an MCU are the processor, memory, and I/O peripherals. The Processor (CPU) can be thought of as the brain of the MCU. It performs basic arithmetic, logic, and I/O operations. Memory is where any data is stored the processor needs to fulfill its tasks. Lastly, the I/O peripherals are the controller's connection to the outside

world; they allow the receiving and sending of information, such as receiving a signal from a switch and turning on a light in response.

2.1.1 ESP32

For this thesis, we are not looking at big devices using multiple MCUs but instead at the ESP32 system on a chip (SoC) specifically. The ESP32 is a very popular low-cost, low power series of microcontrollers with integrated WiFi and Bluetooth. Developed by the Shanghai-based company Espressif Systems this successor to the ESP8266 offers a great platform for IoT and embedded projects [14]. Compared to many other MCUs and its predecessor, the ESP32 has the additional processing power and I/O options that make it a great platform for developing secure IoT devices. The additional resources allow the use of more complex encryption without impacting the performance too much. It gained popularity fast after being released in September of 2016 [33].

The ESP32 systems family provides an excellent base for many IoT applications. There are multiple versions available from ones very well suited for hobbyists to ones used for industrial manufacturers. With a low price point and small area footprint, they still provide significant performance and many operational features [27]. Examples for the ESP32 being used in IoT products include the light bracelets given at Alibaba's annual meeting in 2017 [2] and DingTalks's biometric attendance monitor M1 [12].

To reach price and power consumption targets, the ESP32 has significant hardware limitations. This introduces some constraints when working with the platform, such as 4MB flash memory and 520KB RAM installed. Compared to popular desktop operating systems like Windows and Linux, the used FreeRTOS is much more specialized. This will be explained further in the following passage.

2.1.2 FreeRTOS

Many MCUs are used in applications where throughput is less important than a guaranteed performance. This is why the ESP32 uses FreeRTOS, a real-time operating system (RTOS), specifically intended to be used in time-critical situations. A key characteristic of such an operating system is the predictable behavior of the scheduler, the part of the operating system that decides which task should be run by the CPU at any given time. Most real-time schedulers allow the user to set priorities for tasks in order to decide which task should be run next.

FreeRTOS is the leading RTOS amongst MCUs and is designed to be small enough to run on a microcontroller [15]. Since most applications in which MCUs are used do not warrant the use of a full RTOS, FreeRTOS only provides the core scheduling functionality, timing, and synchronization primitives. This allows basic process management

and synchronization. It can, however, be extended by using addon components, for example, to make use of a specific networking stack. FreeRTOS also built a significant developer community and support for many platforms in its 15-year development.

In more recent history, Amazon has taken over stewardship of FreeRTOs and also offers its own extension a:FreeRTOS [26]. This version additionally comes with some direct integration into Amazon's AWS service [16]. It is supposed to make the development of new IoT devices easier, especially when using Amazon's platform for server-side processing. The core of FreeRTOS remains open-source.

2.2 WebAssembly

Beginning with static HTML pages, the web has since developed into a universal application platform, accessible from many different devices running various operating systems. JavaScript is the only natively supported language on the web. However, even though it is universally used and made impressive progress in modern implementations, it still has some problems as a compilation target. The main issue being that JavaScript performance heavily depends on optimizations made by the browser, making it somewhat unpredictable and unreliable. WebAssembly addresses these issues and provides a compilation target for the web [9].

WebAssembly (WASM) was first announced in June 2015 [34] and reached a cross-browser consensus in March 2017 [35]. Its goal was to provide near-native performance for browser-based applications, which could only be written in JavaScript for a long time. Since reaching consensus on 2017, browser adoption makes it currently usable for about 90% of global internet users [11]. More recently, the interest picked up around usage outside of the browser, which is also the primary concern of this thesis.

Being designed for the web, WebAssembly was developed with specific goals in mind that give the format some unique attributes. It has to be safe since, on the web, code is loaded mainly from untrusted sources. It has to be fast as the primary motivation to introduce WebAssembly was to provide a compile target on the web with reliable performance. Other than the usual low-level code such as regular assembly, WebAssembly has to be portable and work in all the different environments the web is currently used in. Lastly, because the code is transmitted over the network, it has to be as small as possible to reduce bandwidth and improve latency [29].

WASM is a bytecode format, designed to be a portable target for high-level languages like C/C++ or Rust. It is executed on a stack-based virtual machine on which it executes in near-native speed due to its low-level design. Still, it runs in a memory-safe environment inside the browser and is subject to the same security policies as JavaScript code would be. WebAssembly modules are loaded with the application and provide

bindings to JavaScript that make them usable in the browser. They use exports to provide functions that can be called from the JavaScript context, integrating the module into web applications.

Together with the binary format of WebAssembly, there is a text format that defines a programming language with syntax and structure. Every WASM binary is a self-contained module with functions, globals, tables, imports, and exports. This concept provides both encapsulation and sandboxing since modules can only interact with their environment using imports, and the client can only access the specified exports. Inside the module, the code is organized in functions that can call each other even recursively.

WebAssembly is designed to be run on a stack machine, meaning that most of its instructions operate on a pushdown stack instead of registers. This leads to compact code as the location of parameters does not have to be specified; they are always expected to be on the stack. This model was chosen for WebAssembly due to its efficient encoding, compilation, and interpretation [10].

Other than most stack machines, WebAssembly provides structured control flow instead of arbitrary jumps [22]. This limitation simplifies one-pass verification and manipulation of WebAssembly code by other tools. Figure 2.2 shows the control flow syntax in WebAssembly, which shows that all block, loop, and if constructs must be terminated with the end instruction and have to be correctly nested. They bracket nested sequences of instructions referred to as blocks in WASM; each block has an implicit label, which is the target for branch instructions.

$$
\begin{array}{lcl}
\mathit{instr} & ::= & \ldots \\
& | & \mathsf{nop} \\
& | & \mathsf{unreachable} \\
& | & \mathsf{block}\ \mathit{resulttype}\ \mathit{instr}^*\ \mathsf{end} \\
& | & \mathsf{loop}\ \mathit{resulttype}\ \mathit{instr}^*\ \mathsf{end} \\
& | & \mathsf{if}\ \mathit{resulttype}\ \mathit{instr}^*\ \mathsf{else}\ \mathit{instr}^*\ \mathsf{end} \\
& | & \mathsf{br}\ \mathit{labelidx} \\
& | & \mathsf{br\_if}\ \mathit{labelidx} \\
& | & \mathsf{br\_table}\ \mathit{vec}(\mathit{labelidx})\ \mathit{labelidx} \\
& | & \mathsf{return} \\
& | & \mathsf{call}\ \mathit{funcidx} \\
& | & \mathsf{call\_indirect}\ \mathit{typeidx}
\end{array}
$$

Figure 2.2: WebAssembly control flow syntax

Since they are used later on, we want to explain the `br_if`, `return` and `call` instruc-

tions. The first two are branch instructions that will reference a specific block label. `br_if` performs a conditional branch and is used, for example, to execute if statements. `return` is a special unconditional branch that always branches to the outermost block. The `call` instruction invokes another function by taking the necessary arguments from the stack and returning the result of the execution of the function.

The WebAssembly runtime keeps all the global state that can be manipulated by the module in the store. As shown in Figure 2.3, the store consists of all instances of functions, tables, memories, and globals that have been allocated while the abstract machine was running. Besides the store, almost all instructions interact with an implicit stack. This stack contains values, labels, and call frames of active function calls. Thus, if a function returns a result, it will be the first entry on the stack after the execution has finished [30].

$$
\begin{array}{lcll}
\textit{store} & ::= & \{ & \mathsf{funcs}\ \ \textit{funcinst}^*, \\
 & & & \mathsf{tables}\ \ \textit{tableinst}^*, \\
 & & & \mathsf{mems}\ \ \textit{meminst}^*, \\
 & & & \mathsf{globals}\ \textit{globalinst}^*\ \}
\end{array}
$$

Figure 2.3: WebAssembly store contents

2.2.1 WebAssembly for IoT

While WASM was developed for the web, it carefully avoids any dependencies on the web. It is meant to be an open standard that can be embedded in a variety of ways [28]. The goals mentioned above, which WebAssembly achieves, make it an exciting format to explore on embedded devices. Due to its aim to be universal, it would allow the use of languages on MCUs that were not previously supported, and since it is already meant to be transmitted over the network, over the air updates of code running on the controller are also possible. To achieve portability, the source level interface libraries would have to map the host environments' capabilities either at build time or runtime.

While using WebAssembly in an IoT context is an up-and-coming concept, the development has only just begun. The best support for WASM right now is in the browser, of course, but out of browser runtimes and compilers are being developed, implemented in various languages such as Python, JavaScript, C, Rust, and even OCaml [1]. Runtimes meant to be used on MCUs are much rarer and not as mature yet. Given the significant interest in the idea, though, and the working groups' avoidance of web dependencies, it can be assumed that this situation will change in the future.

2.3 Interpreters

With our specific use-case in mind, the WASM3 [38] engine was chosen because one of its main goals is to run WebAssembly on MCUs. Contrary to many other engines, it does not follow a just-in-time (JIT) compilation pattern though but instead acts as an interpreter.

Interpreters are computer programs that execute a program. They pose a different concept to compiled execution, where a program would be translated to machine code before being run directly on the CPU. While offering multiple advantages, the main drawback is the execution speed compared to native code execution, which is often slower by order of magnitude and sometimes more. The overhead is generated by the interpreter having to analyze the program code before it can be executed.

Interpreters thus offer benefits in development speed since the code does not have to be recompiled in order to run and in portability because the same code could be run on multiple platform-specific interpreters without the need to compile it into the native machine code of multiple platforms. For our use-case, the interpreter executes WASM instructions, allowing the dynamic loading of modules and running them in the chosen environment.

2.4 Microbenchmarking

Benchmarking is any form of measurement used to qualify the behavior of a system. The most obvious examples would be measuring performance, energy, or memory consumption, but also reliability and temperature stability could contribute to a benchmark. Building useful benchmarks is hard because a program has to be created that yields repeatable and consistent results.

Having a reliable method for measuring microcontroller performance is a critical industry need [24] as developers need to know how their system will behave in certain situations [25]. For this thesis, we are looking at synthetic benchmarks, which are developed to measure specific parameters. A notable effort in the world of microcontroller benchmarking is the embedded microprocessor benchmark consortium (EEMBC) [13].

The EEMBC is an industry association that has been designing benchmarks for over 20 years. The consortium offers multiple benchmarks, all meant to cover specific use-cases of embedded controllers from ultra-low-power IoT applications, over processor performance measurements to a recent benchmark designed to assess machine learning performance [5]. They strive to provide standardized benchmarks that allow vendors to compare platforms across the industry reliably.

2.5 Summary

Microcontrollers are tiny, and restricted computers meant to perform a specific task. They are part of our everyday life and are used in a great variety of applications. The ESP32 system on a chip is a prevalent family of Microcontrollers that are very well suited for use in the Internet of Things devices. They are affordable but still offer WiFi and Bluetooth connectivity.

WebAssembly is a new bytecode format that was designed for browsers, with the aim of allowing developers to run optimized low-level code at near-native speeds. It is supported by all major browsers and in active development. However, people are not only interested in WebAssembly on the web itself but also recognize its potential as a new universal and portable bytecode outside of the browser.

3 Methodology

The goal of this thesis is to evaluate the use of WebAssembly for programming the ESP 32 microcontroller. After we have introduced the concepts around MCUs and WebAssembly, we will explain our approach of measuring the impact of running WebAssembly on MCUs.

3.1 Running WebAssembly

First, as explained earlier, WebAssembly always needs a runtime, which is usually provided by the browser. In this case, we do not need all the features a browser would provide, just a way to execute WASM. Since the momentum around running WASM outside of a browser environment has been picking up recently, more and more runtimes are becoming available. A big push for WebAssembly on new platforms came in November of 2019 in the form of the Bytecode Alliance [6]. An open-source community dedicated to creating the foundations needed to run WASM on multiple platforms in a secure way.

The Bytecode Alliance maintains a couple of different compilers and runtimes for WebAssembly. This project includes the WebAssembly Micro Runtime (WAMR), which is an interpreter based runtime, specifically meant to run on embedded devices such as the ESP32 [7]. Unfortunately, the OS used by the ESP32 (FreeRTOS) was not yet supported, with no current plans to change that.

While a couple of other WebAssembly runtimes are available [1], WAMR used to be the only one capable of running on an embedded device. In late 2019 the second runtime for embedded devices was released in WASM3 [38]. This runtime is the first one we know of to support the ESP32 and FreeRTOS. It also performs significantly better than WAMR in benchmarks [31]. Thus we decided that using WASM3 was the way to go about running WASM on the ESP32.

According to the developers' measurements [31], WASM3 is currently the fastest available WebAssembly interpreter. It is about 4x slower than current just in time compiling runtimes and about 12x slower than native execution. Because of the strict constraints that embedded devices have, WASM3 uses an interpreter model, which is more memory efficient and provides better startup times than JIT compilation. This

approach also makes portability and security much more comfortable to achieve and maintain while developing the runtime.

The speed of WASM3 is impressive, considering that even in browsers, a performance loss of up to 3x can be experienced when comparing WebAssembly to native code execution [23]. For our tests, it is crucial to keep in mind that the ESP32 is not a reference platform due to its limited capacities and features.

While the most basic interpreter can be thought of as a loop around a big switch statement that matches all possible instructions in the interpreted code, WASM3 follows a model dubbed M3 [32]. In WASM3, the bytecode first gets compiled into operations for a meta machine, which is traversed by one operation calling the next, which relies on tail-call optimization by the compiler. This leads to an efficient and elegant execution model for their virtual machine.

3.2 Comparing the Platform

After finding a way to run WebAssembly on the ESP32 and verifying it with basic tests, we had to find a way to test how well WASM execution on the MCU worked. A popular tool to compare different platforms is benchmarking, in which the same workload is run on multiple platforms to generate metrics usable for comparing those platforms.

While most benchmarks are meant to provide a comparison of two hardware platforms, in this specific case, we are not interested in the performance of the platform. Instead, we are interested in the performance of different execution models on the same platform, being the default native execution of code explicitly compiled for the ESP32 and the interpreted execution of the WASM code that could run anywhere.

Our desired comparison makes the test setup quite simple. The basic idea is to run the same code on the ESP32 twice, only once compiled to WASM. This approach has worked for the most part, with small detours being made when testing the import of outside functions into WebAssembly. All tests consist of a `run` method, which is once called from the main file and once loaded into the engine and run as a WASM function.

In line with how benchmarks work, we set out to design a couple of workloads we could run in both the native and the WebAssembly environments and compare the way they execute. In order to generate meaningful results, we tried to find simple tests that are not too far from what an MCU would execute. Keeping the tests limited and simple also allows us to look at the WASM output and understand the exact instructions in some instances.

3.2.1 Specialized Workloads

The first and most basic test is recursively calculating a Fibonacci number, while extensive recursive calls are not a part of most applications, function calls in general are. This also shows some WebAssembly specific features since it does not only have jumps available but allows functions and function calls in the assembler code.

Secondly, switch and if statements are an integral part of any application. So comparing the performance of a switch statement is another indicator of how good applications would perform.

Of course, every application needs memory access, so we decided to implement two memory tests—one using direct access and one using random access to see if there is any impact on performance. Combining memory access and calculation, we also ran a matrix multiplication. This algorithm is the basis for many more complex algorithms, and the tight inner loop offers itself for optimization on the hardware and during compilation.

Of course, it is essential for embedded devices also to have hardware access. Currently, WebAssembly does not have a model of specific hardware features, network stack, or even CPU cores. All this functionality is assumed to be in the browser environment. The runtime we selected offers a mechanism to link external functions that can then be called from the WASM code, so we designed tests to see if outside calls came with a significant overhead that would impact applications using them.

Lastly, we implemented the Fibonacci test again but in AssemblyScript instead of C++. AssemblyScript [4] compiles a subset of TypeScript to WebAssembly. This is exciting for developers with a background in web development, as TypeScript is probably already familiar to them. We implemented this test to show the new options using WebAssembly opens on the ESP32, which does not natively support a way to run TypeScript. TypeScript itself is an extension of JavaScript that supports strict types.

Optimization Trying to design benchmarks that cannot be optimized is essential for the tests we are running. Since the test workloads are constructed and run code that is not strictly necessary for the testing program, the compiler tries to optimize them. Returning values, even though they are not needed, is a crucial step to make sure the test code is not skipped. Also, the optimizations are less aggressive in the WebAssembly code since the compiler cannot infer the input arguments. Occasionally we were not able to construct a test that was not optimized within the time constraints of this thesis. Deactivating the compiler would also falsify the result since any production program would be compiled with optimizations enabled. So some tests have special instructions or structure that is meant to prevent too much optimization.

3.3 Running Tests

All tests were run on the ESP-WROOM-32 using the ESP-devkit provided by espressif. The same C++ code was compiled to WebAssembly and also imported into the test program to allow for native execution. Then the test code was run multiple times to generate statistically significant results. The results from these tests will be explained in more detail in a specific section for each test.

3.3.1 Testing Setup

All tests share a very similar main program to execute and time the tests, which we would like to explain now.

Listing 3.1: Main testing method

```
1  extern "C" void app_main(void) {
2      // Variable initialization
3
4      setup_wasm(); // Explained in Listing 3.2
5
6      for (long long &wasm_time : wasm_times) {
7          int64_t start_time = esp_timer_get_time();
8          for (int j = 0; j < 10; ++j) {
9              run_wasm("20"); // Explained in Listing 3.3
10         }
11         int64_t end_time = esp_timer_get_time();
12         wasm_time = (end_time - start_time) / 10;
13     }
14
15     for (long long &native_time : native_times) {
16         int64_t start_time = esp_timer_get_time();
17         for (int j = 0; j < 10; ++j) {
18             long value = run(20);
19         }
20         int64_t end_time = esp_timer_get_time();
21         native_time = (end_time - start_time) / 10;
22     }
23
24     printf("|Run|WASM|NATIVE|\n|---|---|---|\n");
25     for (int i = 0; i < sizeof(wasm_times) / sizeof(wasm_times[0]); ++i) {
```

```
26          printf("|%d|%lld|%lld|\n", i + 1, wasm_times[i], native_times[i]);
27      }
28      sleep(100);
29      printf("Restarting...\n\n\n");
30      esp_restart();
31  }
```

The primary testing method in Listing 3.1 starts with setting up the WebAssembly runtime in line 4. This process is timed to see how much overhead the runtime initialization introduces. Also, there are two arrays set up to hold the test results. In line 6 to line 13 the times of WASM test functions are gathered by running the function call in line 9 ten times and taking the average execution time. We are running this 100 times to generate statistically relevant results.

The process for gathering the times of the native call in line 18 is very similar and can be seen in lines 18 through 22. After taking the times the results are printed in the loop in line 25, this is deferred as not to measure the console output as well. Eventually the controller is restarted in line 30.

It is important to note that with this setup, the runtime will be loaded in memory during both tests. Also, the native code is shipped together with the WASM test code. We do not expect interference due to the minimal footprint of both test cases.

Listing 3.2: Runtime setup

```
1  IM3Environment env;
2  IM3Runtime runtime;
3  IM3Module module;
4  IM3Function f;
5
6  static void setup_wasm() {
7          M3Result result = m3Err_none;
8
9          auto *wasm = (uint8_t *) wasm_test_cpp_wasm;
10         uint32_t fsize = wasm_test_cpp_wasm_len - 1;
11
12         env = m3_NewEnvironment(); // Error output omitted
13         runtime = m3_NewRuntime(env, 2048, NULL);
14         result = m3_ParseModule(env, &module, wasm, fsize);
15         result = m3_LoadModule(runtime, module);
16         result = LinkThesis(runtime);
17         result = m3_FindFunction(&f, runtime, "run");
18     }
```

```
19 }
```

Setting up the runtime is pretty straight forward, initially, the WASM module is imported from a header file, together with its length in lines 9 and 10. Then the environment and runtime are created, followed by parsing the module. If the runtime has to provide functions that the WASM module relies upon, they are linked after loading the module in line 16. An example where outside functions are linked can be found in Section 3.3.6. Once the runtime is fully set up, the function itself is searched for in line 17. In our case, the function's name is always "run".

Listing 3.3: Code to execute the WASM function

```
1  static void run_wasm(const char *input1) {
2      M3Result result = m3Err_none;
3
4      const char *i_argv[2] = {input1, NULL};
5      result = m3_CallWithArgs(f, 1, i_argv);
6
7      if (result) FATAL("m3_CallWithArgs: %s", result);
8
9      long value = *(uint64_t *) (runtime->stack);
10 }
```

The execution of the WASM function is a matter of calling the previously found function with the runtimes `m3_CallWithArgs()` method and supplying it with the input arguments. The return value of the operation can be found on the virtual machines stack afterward and is retrieved in line 9.

3.3.2 Testing Recursive Calls

The first test we designed is the recursive calculation of a Fibonacci number. Its simple design allows us to take a look at the generated WebAssembly as well and understand the specific instructions.

Listing 3.4: Recursive calling test

```
1  uint32_t run(uint32_t n) {
2      if (n < 2) {
3          return n;
4      }
5      return run(n - 1) + run(n - 2);
6  }
```

Listing 3.5: Recursive calls WASM code excerpt

```
1  (module
2    (type $t1 (func (param i32) (result i32)))
3    (func $run (export "run") (type $t1) (param $p0 i32) (result i32)
4      (block $B0
5        (br_if $B0
6          (i32.lt_u
7            (local.get $p0)
8            (i32.const 2)))
9        (return
10         (i32.add
11           (call $run
12             (i32.add
13               (local.get $p0)
14               (i32.const -1)))
15           (call $run
16             (i32.add
17               (local.get $p0)
18               (i32.const -2))))))
19     (local.get $p0)))
```

Listing 3.5 shows the WebAssembly text format generated for the Fibonacci function in Listing 3.4 and we will take a closer look at what the WASM module looks like for this specific example. After the module opening, the type of our `uint32_t run(uint32_t n)` function is defined and reused in the function definition in line 3, in this line the functions input and return types are also defined. The input is assigned to the $p0 variable for later use.

In line 4 the block `$B0` is started, it contains the main function body. In line 5, we can see a `br_if` instruction; this is a conditional branch that breaks the execution of the passed block if the condition is true. The condition, in this case, is the rest of the instructions included in the parentheses. Namely the comparison of the accepted parameter with 2 to see if it is smaller if that is the case the remaining block is skipped and code execution would continue in line 19, where the parameter is pushed on the stack, as the topmost value of the stack after the execution is the return value of a WASM function.

Alternatively, the execution could continue in line 9 with the `return` instruction, which executes the instructions inside the parentheses and prevents any further code execution after that, mirroring the common return instruction in C. The Value return is the result of the two recursive calls of line 5 in the listing 3.4. The `run` function is called

again by using the `call` instruction in lines 11 and 15.

It is important to note that this text format is not strictly WebAssembly, but one version to make it readable for humans. To make it more similar to the look of common programming languages for this Listing, code folding was applied. Code folding reorganizes the instructions and adds parentheses and indention to show related instructions better. To make the difference visible, Listing 3.6 shows the WASM code without folding.

Listing 3.6: Recursive calls WASM code without folding

```
1  (func $run (export "run") (type $t1) (param $p0 i32) (result i32)
2    block $B0
3      local.get $p0
4      i32.const 2
5      i32.lt_u
6      br_if $B0
7      local.get $p0
8      i32.const -1
9      i32.add
10     call $run
11     local.get $p0
12     i32.const -2
13     i32.add
14     call $run
15     i32.add
16     return
17   end
18   local.get $p0)
```

3.3.3 Testing Switch Statements

As previously mentioned, switch statements are a widespread occurrence in software running on microcontrollers. Use-cases range from interpreters to the correct handling of communication protocols. The test method for this, as shown in Listing 3.7, is a big switch statement that is looped over. To prevent optimization, the return value is updated from an array and returned by the function.

Listing 3.7: Switch statement test code

```
1 uint32_t run(uint32_t n) {
2     uint32_t array1[20] = {1, /*some entries omitted*/ 20};
```

```
3     uint32_t result = 200;
4     for (int i = 0; i < n; ++i) {
5         uint32_t compare = i % 20;
6         switch (compare) {
7             case 0:
8                 result = array1[0];
9                 break;
10            case 1:
11                result = array1[1];
12                break;
13            // Some cases omitted
14            case 18:
15                result = array1[18];
16                break;
17            case 19:
18                result = array1[19];
19                break;
20            default:
21                result = 100;
22                break;
23        }
24    }
25    return result;
26 }
```

3.3.4 Testing Memory Performance

Of course, memory performance is a critical aspect of any computing platform, so we designed a test to compare it between native, and WASM execution, the C++ code of the test is shown in Listing 3.8.

Listing 3.8: Linear memory test

```
1 uint32_t run(uint32_t n) {
2     uint32_t array1[n];
3
4     for (int i = 0; i < n; ++i) {
5         array1[i] = i+1;
6     }
7     array1[n-1] = 0;
```

```
8     uint32_t nextStep = 1;
9     while(nextStep){
10        nextStep = array1[nextStep];
11    }
12    return nextStep;
13 }
```

The setup is relatively simple; an array is created and filled with the indices of the respective following elements, imitating a linked list. This version was chosen due to `malloc` not being available when compiling for WebAssembly. Then it is read from, starting at index one and saving whatever index was found there into a variable that defines the next index to be read until finally, the next index ends up being 0. This leads to the entire array being read but prevents optimization that would have happened with a loop. Lastly, the next index is returned to prevent compiler optimization. We also designed a test for random memory access, which is available online in the sources for all tests created for the thesis[1]. Since it did not lead to any additional insight while running it we decided to not include the details. A functioning system interface would have allowed us to follow standard linked list implementation and compare random vs. linear memory access.

3.3.5 Testing Matrix Multiplication

To test performance during matrix multiplication, we use the code of Listing 3.9. The test function receives the matrix size as an argument in line 1 and allocates two-dimensional arrays to hold the matrix values. With the loop started in line 4, those arrays are filled with values based on the current loop iteration, to generate data that can be multiplied afterward.

The actual multiplication then happens in lines 11 through 18 and is saved a thrid array. Finally, in line 20 one value from the resulting matrix is return to assure, that the compiler does not optimize the multiplication too much.

Listing 3.9: Matrix multiply test

```
1 uint32_t run(uint32_t n) {
2     uint32_t a[n][n], b[n][n], mul[n][n];
3
4     for (uint32_t i = 0; i < sizeof(a) / sizeof(a[0]); ++i) {
5         for (uint32_t j = 0; j < sizeof(a[0]) / sizeof(a[0][0]); ++j) {
6             a[i][j]=i+1;
```

[1] https://github.com/Isigiel/bsc-thesis/tree/master/code/platforms/memory-array-random

```
7             b[i][j]=i+2;
8         }
9     }
10
11    for (uint32_t i = 0; i < sizeof(a) / sizeof(a[0]); ++i) {
12        for (uint32_t j = 0; j < sizeof(a[0]) / sizeof(a[0][0]); ++j) {
13            mul[i][j] = 0;
14            for (uint32_t k = 0; k < sizeof(a[0]) / sizeof(a[0][0]); ++k) {
15                mul[i][j] += a[i][k] * b[k][j];
16            }
17        }
18    }
19
20    return mul[n-1][n-1];
21 }
```

3.3.6 Testing Native Calls

A vital function of the runtime is to expose outside functions to the WASM module and allow the interaction with other libraries from within the WASM code. This is often needed for I/O libraries and interaction with peripherals. For this, we designed two reasonably simple tests that call functions not defined in the WASM code.

Listing 3.10: Outside call test code

```
1 #include "test_api.h"
2
3 WASM_EXPORT
4 void run(uint32_t n) {
5     mark();
6 }
```

As is obvious from listing 3.10, the test code just calls the outside `mark()` function. There is also an import of the test API header in which the external function is defined to make the test code compile.

Listing 3.11: Test api defintiion for native calls

```
1 extern "C" {
2
3     WASM_IMPORT("thesis", "sendValue") uint32_t sendValue (void);
4     WASM_IMPORT("thesis", "mark") void mark (void);
```

```
5
6 }
```

The resulting WASM code does not include the `mark` method, but instead imports it from the `thesis` module that is expected to be available at runtime. To provide this function, we import the setup shown in Listing 3.11, the second function declared in that listing `sendValue` is used in our other native test, which also has a return value.

Listing 3.12: WASM code for the outside call

```
1 (module
2   (type $t0 (func))
3   (type $t1 (func (param i32)))
4   (import "thesis" "mark" (func $thesis.mark (type $t0)))
5   (func $run (export "run") (type $t1) (param $p0 i32)
6     (call $thesis.mark)))
```

As displayed in listing 3.12 line 4 the mark function from the thesis module is imported as defined in listing 3.11. The run function then just calls the imported function in line 6.

In order to provide this imported function at runtime, the setup for our tests has to be changed slightly; namely, it has to be linked during the runtime setup in listing 3.2.

Listing 3.13: Code to link functions into the runtime

```
1  int64_t native_timestamp;
2
3  m3ApiRawFunction(m3_thesis_mark) {
4      native_timestamp = esp_timer_get_time();
5      m3ApiSuccess();
6  }
7
8  M3Result LinkThesis(IM3Runtime runtime) {
9      IM3Module module = runtime->modules;
10     const char *thesis = "thesis";
11
12     m3_LinkRawFunction(module, thesis, "mark", "i()", &m3_thesis_mark);
13     return m3Err_none;
14 }
15
16 void mark() {
17     native_timestamp = esp_timer_get_time();
18 }
```

In listing 3.13 line 1 we introduce a variable to hold a timestamp after the `mark` function was called. In line 3, we define the function, which just saves the current timestamp and ends with success. This is then linked into the runtime in line 12. To compare native execution this time, we can not call the exact same function since it was not compiled to WASM at all, so we implement a similar function in line 16 that is called during the test of native execution. A similar code was written for the native test case with a return value [2].

3.3.7 Running TypeScript

In listing 3.14 the same function as for 3.3.2 is implemented but in TypeScript instead of C++. This is done to explore the new options opened by WebAssembly on the ESP32. Previously there was no way to compile any other language into a format that could run on the ESP32. The test is then run the same way the other tests work. We are once again comparing the execution performance with the native C++ code.

Listing 3.14: TypeScript testing code

```
1 export function run(n: i32): i32 {
2   if(n < 2 ) {
3     return n;
4   }
5   return run(n - 1) + run(n - 2);
6 }
```

The generated WebAssembly code is almost identical to the bytecode generated for the C++ version that can be seen in listing 3.15. In line 4 however `if` block is used instead of a conditional branch. That block is closed with the `then` statement in line 8. Another difference is the use of the `sub` instruction instead of adding a negative number in line 13. The differences are most likely caused by different compilers being used for C++ code and TypeScript code.

Listing 3.15: WASM code excerpt for fib in TypeScript

```
1 (module
2   (type $t0 (func (param i32) (result i32)))
3   (func $run (export "run") (type $t0) (param $p0 i32) (result i32)
4     (if $I0
5       (i32.lt_s
6         (local.get $p0)
7         (i32.const 2))
```

[2] https://github.com/Isigiel/bsc-thesis/tree/master/code/platforms/native-return

```
8       (then
9         (return
10          (local.get $p0))))
11    (i32.add
12      (call $run
13        (i32.sub
14          (local.get $p0)
15          (i32.const 1)))
16      (call $run
17        (i32.sub
18          (local.get $p0)
19          (i32.const 2)))))
```

3.4 Summary

In order to assess WebAssembly on the ESP32, we used a new runtime from late 2019, WASM3. Our survey of existing runtimes showed that it is only the second runtime meant to run on embedded devices, and the first one to date that supports FreeRTOS. It is an interpreter for WebAssembly that follows a meta machine pattern and achieves excellent speeds compared to other WebAssembly runtimes currently available.

Secondly, we designed a collection of workloads that can be used to compare the WebAssembly execution to the native code execution. We modeled the tests after requirements, which applications on an MCU will typically have. These include function calls, memory access, computation, and the use of outside functions to interact with peripherals, for example. Additionally, we tested the ability of WebAssembly to be a target for new languages, not natively available on the controller, and wrote one test in TypeScript.

4 Evaluation

4.1 Running Benchmarks

When implementing the workloads for our benchmarks, we noticed that using WebAssembly as of now imposes many constraints. Since we are not targeting a browser platform, which offers a wide variety of APIs meant to handle system calls, code targeting WASM on the ESP32 can not make use of APIs available in the browser environment such as memory allocation.

In order to run the test, we had to compile the C++ test code into WASM bytecode. An established tool for this is the emscripten compiler. After using this for the initial tests, though, we noticed problems, since it is meant for compiling WASM that will run in the browser. This leads to modules that expect all the available functions of the browser environment since we are running the tests on an embedded device; those functions are not available to us.

Because emscripten is meant for use in browsers after compilation, we switched to the `wasmc++` compiler, which is part of wasienv, a toolchain for compiling C into WebAssembly [37]. This project provides a couple of utilities to compile C code into WASM modules. `wasmc++` wraps clang++ with the correct configuration for WebAssembly already applied and makes compiling very easy. For our tests, the code was compiled by running `wasmc++ -Os -Wl,--strip-all -nostdlib wasm/test.cpp -o wasm/test.cpp.wasm`. The flags are needed to enable optimization and create a minimal WebAssembly file that only includes the test function.

Lastly, the WASM3 runtime expects the bytecode to be loaded into an array. In Linux, the `xxd` utility is exactly what is needed to achieve that. After compilation, the WASM code can be converted into a C++ header file by running `xxd -i wasm/test.cpp.wasm > main/test.wasm.h`.

After designing tests as described in 3.3.1, we ran all of them on the ESP32 to report times. In general, the tests showed a significant slowdown in execution speed when running the workloads in a WebAssembly context compared to running them compiled natively. For all tests, we could observe the variance of the measured time to be very low; this is to be expected since the tests ran without any other load on the MCU and the deterministic nature of FreeRTOS, as mentioned in section 2.1.2.

It is important to mention when looking at the results of these tests that WASM3 heavily relies on tail-call elimination, which currently is not performed by the ESP32 compiler. This leads to excessive use of the native stack and lower performance. Tail-call optimization is a technique deployed by compilers that optimizes recursive calls, in which the last instruction is the next recursion step in a way that already clears the call frame from the stack when calling the next function. Collaborators and authors of WASM3 are currently exploring solutions that would make the runtime faster and more efficient on the platform soon [19].

4.2 Test results

In order to generate statistically significant results, every test was run 1000 times. We took the average execution times of ten runs and generated 100 values for each workload this way. Table 4.1 shows the average execution times and standard deviations for every test.

Test case	WASM Execution	Native Exectuion	Slowdown factor
Recursive Calls	41,766 µs (σ=0.79)	1,000 µs (σ=0.28)	42
Switch statement	1,021 µs (σ=0.71)	0 µs (σ=0.00)	N.A.
Memory Access	1,802 µs (σ=0.74)	55 µs (σ=0.26)	33
Matrix Multiplication	26,277 µs (σ=0.60)	281 µs (σ=0.58)	93
Native calls[1]	33 µs (σ=5.15)	8 µs (σ=0.64)	4
TypeScript execution	41,493 µs (σ=0.85)	1,000 µs (σ=0.24)	41

Table 4.1: Testing results

Additionally, we timed how long the runtime needed to load for every individual test. Those times can be seen in table 4.2, together with the size of the compiled WebAssembly module for every test.

Test case	Module size	Setup Time
Recursive Calls	212 B	1,260 µs
Switch statement	234 B	1,299 µs
Memory Access	288 B	1,662 µs
Matrix Multiplication	552 B	1,649 µs
Native calls	208 B	1,352 µs
TypeScript execution	110 B	698 µs

Table 4.2: Runtime setup times

4.2.1 Recursive Calls

Calling functions is an integral capability of any application, so this is the first test to compare WASM with the native execution of our code. We are using the test described in Section 3.3.2. This test has almost no instructions but produces many function calls that could be troubling to handle for the runtime.

Even though most applications on an MCU might not run recursive calculations, this test does show the cost of calling many functions. It is apparent from the numbers in Table 4.1 that running the code in the interpreter takes about 41x longer than the native execution.

As seen in the WASM code of listing 3.5, it is straightforward but requires the runtime to manage the execution of the same functions many times over. Compared to some of the following tests, 41x is not a very high slowdown.

To see if both the runtime and native execution behave similarly for different inputs, we ran the test multiple times. Figure 4.1 shows the change in execution time for increasingly big input numbers. As expected, the execution time grows exponentially, but the runtime and native execution maintain their 41x difference in speed.

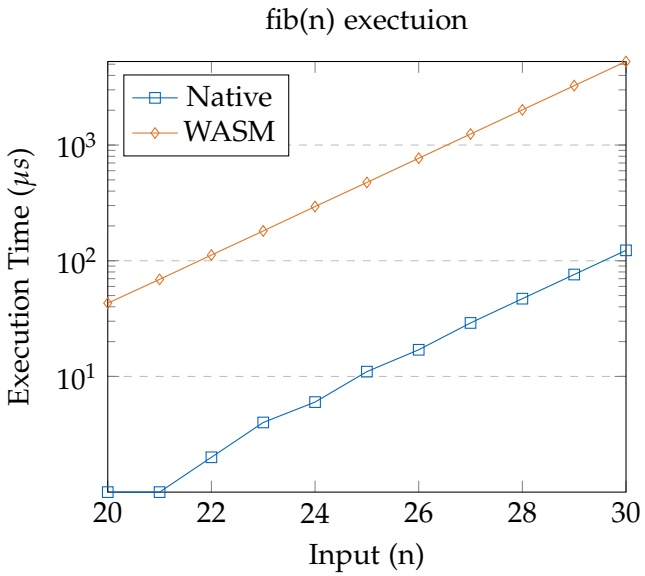


Figure 4.1: Recursive call times for different inputs

4.2.2 Switch Statements

Next, we tested the performance of switch statements in the runtime by using the function from Section 3.3.3. The times listed in Table 4.1 lead to the conclusion, that the native compiler optimized our switch statement very well, effectively skipping over what we tried to test. Otherwise, the native instruction should take a while longer. This problem was previously described in more detail in Section 3.2.1. In the timeframe of the thesis, we were not able to develop a test that avoided optimization.

Since there are no values to compare this statement with, we cannot pass judgment on how much of a performance hit switch statements take from being executed in the runtime. Since the WASM code never executed more than 100x slower than native code, it would be reasonable to expect a result in the same range for switch statements.

4.2.3 Memory Access

Of course, every application requires memory access, so we ran tests that perform linear reads on the memory were described in Section 3.3.4. The runtime caused a longer execution time of around 33x, as can be seen in the measured times in table 4.1. Since WASM is always run in a virtual machine, memory access optimization is not achieved in the source code but instead taken care of by the runtime. In a browser, for example, the WASM memory is a JavaScript ArrayBuffer.

We were not able to directly compare linear memory access with random memory access. However, when running the test meant to assess random access, we saw the performance loss increase to about 73x. This could be caused by several things, such as the generation of pseudo-random numbers in the test code. We can not make a precise determination if random access has any impact on the performance of the WASM code.

4.2.4 Matrix Multiplication

The matrix multiplication test combines both memory access and calculations and experiences the most significant increase of execution time in any of our tests. As the numbers in table 4.1 show, the interpreted code runs more than 90x slower than the natively executed code.

This test does show that more extended calculations and more sophisticated algorithms will take a significant performance hit when being run as WebAssembly. Even in contact with the contributors of the runtime, we were not able to identify specific tasks that are very expensive in the interpreted environment. Nevertheless, as a general rule, it is not advisable to implement large complex workloads in WebAssembly at this point.

4.2.5 Calling of Native Code

Of course, not all functionality can be included in the WASM module; mainly, native platform features will have to make use of functions only available outside of the module. When measuring the performance of calls outside the test-module, we are taking the total time of ten runs instead of taking the average over ten runs. The test we used was introduced in Section 3.3.6. The timings in table 4.3 show that the overhead introduced by the runtime for such a simple test case is relatively small. The interpreted code took less than 10x longer than the native calls.

In a second test, which also returned a value to the WASM module, we did not see an additional increase in execution time, making this the best performing test in our experiments. This test can also be found in the repository[2].

Run	WASM Execution	Native Exectuion
15	33 µs	7 µs
16	33 µs	8 µs
17	33 µs	7 µs
18	32 µs	8 µs
19	33 µs	8 µs
20	32 µs	8 µs
21	33 µs	8 µs
22	32 µs	7 µs
23	33 µs	8 µs
24	32 µs	7 µs
25	33 µs	8 µs

Table 4.3: Exerpt of the measured times for external calls

4.2.6 Typescript Execution

Closely related to 4.2.1 is the test case running TypeScript since it is the same test function and almost the same WASM code being tested. The observed results in table 4.1 are very similar to the times measured when executing WebAssembly code generated from C++. This test shows the exciting potential of running languages on the ESP32 that are not natively supported. Also, it is interesting to note that, at least for this example, there is no performance loss by using TypeScript instead of C++.

Of course, the compilation of TypeScript to WebAssembly requires a different compiler than `wasmc++`. We used the `asc` compiler that is part of

[2] `https://github.com/Isigiel/bsc-thesis/tree/master/code/platforms/native-return`

the AssemblyScript toolchain and compiled the TypeScript code by running `asc index.ts -b test.wasm --validate -O3z --runtime none --noAssert`. After compilation, we followed the same steps as with the other tests to run the comparison.

4.3 Learnings

4.3.1 Drawbacks of WASM Execution

The very obvious drawback of executing WebAssembly on a microcontroller is the performance loss that comes with it. We have shown this with all our tests, and even though the slowdown varies from test to test, code compiled to WASM and executed by an interpreter will run an order of magnitude slower than the same code directly compiled into the main program and executed on the MCUs CPU.

This puts WebAssembly into an unexpected position on embedded devices since it is meant initially to provide better performance. In browsers, WebAssembly made it possible to run adobe lightroom on the web, allows Facebook to compress images before uploading them and Wikipedia to play videos the users' browser does not support [36]. However, even though it reaches near-native speeds on browsers, it performs much worse than the current alternative on embedded devices.

It is fair to assume that this performance decrease causes the MCU to consume much more energy in order to make the same computation. For devices running on a battery, this could pose a challenge. Further experiments could be made to quantify the impact of running WASM and also compare it to other energy-consuming tasks such as network IO.

Apart from the impact on the running of program code on the MCU, we also noticed other limitations. The availability of certain expected functionality in WASM is minimal. For example, the use of `malloc()` is not possible if compiling code to WASM and running it in our runtime. In web browsers, this functionality is available for import from the environment and implemented such that it can be used as expected. WASM3, however, does not offer any functions for import, and implementing dynamic memory management would be a significant effort.

Since the problem of system access outside the browser is prevalent, a subgroup of the WASM community group is working on specifying a system interface for WASM[8]. The WebAssembly system interface (WASI) is meant to provide a foundation for developers to build upon when targeting non-browser platforms. Once specified, code compiled for WASI will run in any WASI-compliant runtime, truly enabling WASMs portability.

4.3.2 Potential of WebAssembly on Embedded Devices

As we demonstrated in our tests, the WebAssembly code is interpreted at runtime; this means that it could also be loaded from the network instead of being included in the code. Dynamically loading code and executing it allows the deployment of new behavior to an MCU without having to perform a flash but rather in the form of an over the air update.

Since WebAssembly support in the browser is excellent already, all the tests we developed can alternatively be run in the browser. This allows embedded developers to test their programs locally in development and be able to make sure everything works as expected even before deploying it to an MCU for the first time. For the ESP32, for example, a browser emulator could be built, which provides all the native functions expected on the platform in JavaScript.

Additionally, we showed in the last test that WebAssembly could open the doors for new languages that are not natively supported on the MCU. In our example, we were able to program the ESP32 by using TypeScript, which usually compiles to JavaScript. However, in our example, we compiled it to WebAssembly and were able to run it on the ESP32. This enables developers without previous to get into embedded programming from their current field of work.

4.3.3 Usecase Examples on the ESP32

For embedded use-cases that are CPU bound, WebAssembly could pose a big problem since the performance is much worse than current native execution. Making up for this with multiple devices could be a way of mitigating that problem, but since we have observed a performance loss of close to 100x, that seems like a costly way of overcoming this problem.

Should the most time-consuming things not be calculations and similar tasks, though, but waiting for a slow sensor read or a signal from the outside, the performance loss might not be as significant. Execution time also matters less if the task is performed periodically, and the device has much idle time.

A very promising use-case of WASM is the customization of device behavior. This example is often given and also supported by the results of our tests and relies on the interoperation of the native code with the loaded WASM module. An IoT device could, for example, have an extensive API and allow the user to deploy logic in the form of a WebAssembly module making use of all the API functions. Here, other features of WebAssembly also come into play, such as it is security guarantees, which make it safer to execute unknown code on the device.

4.4 Summary

To run the Benchmarks, we compiled the testing code to WebAssembly and set up a test function that can be run on the microcontroller. The test function initialized the runtime and loaded the WebAssembly code. Thereafter it runs both the WASM function and the native implementation 1000 times, taking the average runtime of ten runs. Every test aimed to run the same code once compiled to WebAssembly and once being part of the regular native compilation.

Our tests show a significant slowdown of up to 93x when running the code in WebAssembly, which leads to significant disadvantages when running CPU bound applications on the MCU. With applications that are not limited by the CPU performance but by a slow sensor readout or network interaction, for example, it would not pose as much of a problem. Also, we saw that the TypeScript code, compiled to WebAssembly, ran at a very similar performance to the C++ code compiled to WASM, showing that new languages can be used without additional performance decrease.

5 Conclusion

This thesis showed the current status of running WebAssembly on the ESP32 and evaluated its viability. We explored the ecosystem around WebAssembly on embedded devices and established the necessary background to show the significance of our findings. We ran a selection of tests using the current state of the art technology available on the ESP32.

Process

To run WebAssembly on the ESP32, we found the WASM3 runtime. It performs excellenty compared to other WASM runtimes currently available and can be used on MCUs. We were not able to find a second runtime that would be usable on the ESP32 at this point. To run the tests, we set up a test environment, including compiling the test code to WASM.

The test workloads we designed are modeled after tasks, which production application on an MCU would perform and show the behavior of the runtime in different circumstances. We tested function calls, memory access, matric multiplication, and the calling of outside functions. Those might be needed to control the peripherals of the microcontroller. Additionally, we also ran one test that was written in TypeScript to see the experience of languages, not natively supported by the ESP32.

Findings

Our tests showed that the interpreted execution of WebAssembly takes 40 - 90x longer than executing the same function natively compiled. Except for outside calls, where we noticed a slowdown lower than 4x. Also, programming for a WASM target is very limited right now since there is no unified system interface available that would allow the code to interact with the underlying OS.

This significant performance decrease makes it a bad fit for CPU bound applications. Mitigating the lost performance with more devices would be very expensive and require much coordination. We also suspect the longer execution times to make the MCU use

more energy than it does when running the same computation in native code, which could pose a problem to ultra low powered devices and should be researched further.

However, our tests also showed the potential of WebAssembly. The test that used TypeScript to implement the same testing function that was originally written in C++ ran at a similar performance. This shows the advantage of WebAssembly being universal and supported as a target for many languages, that are not typically seen in embedded programming.

Also, due to it is specific attributes, WebAssembly is well fitted for network transmission and dynamic execution. It also guarantees memory safety, which makes executing unknown code a much smaller risk. These properties allow network-connected devices to receive new instruction over the air and run them without a flash being required.

Evaluation WebAssembly on the ESP32 shows excellent potential for new ways of developing for embedded platforms. It opens systems to new languages and deployment strategies. Due to the performance decrease and missing system interface, it is certainly not a good fit for all applications right now. However, with the WebAssembly system interface being actively worked on and more embedded runtimes becoming available, we expect the attractiveness of WebAssembly to increase further. It stands to see if WebAssembly becomes the new universal bytecode, but we see powerful potential, even in the early stages.

Listings

List of Figures

List of Tables

Bibliography

[1] S. Akinyemi. *appcypher/awesome-wasm-runtimes*. original-date: 2018-10-08T10:48:10Z. Mar. 5, 2020. URL: https://github.com/appcypher/awesome-wasm-runtimes (visited on 03/06/2020).

[2] *Alibaba's IoT Wrist Bands Based on ESP32 | Espressif Systems*. Library Catalog: www.espressif.com. Sept. 30, 2017. URL: https://www.espressif.com/en/media_overview/news/alibaba%E2%80%99s-iot-wrist-bands-based-esp32 (visited on 03/15/2020).

[3] I. Analytics. *Internet of Things (IoT) active device connections installed base worldwide from 2015 to 2025*. Statista, Aug. 2018. URL: https://www.statista.com/statistics/1101442/iot-number-of-connected-devices-worldwide/ (visited on 03/14/2020).

[4] *AssemblyScript/assemblyscript*. original-date: 2017-09-28T11:06:50Z. Mar. 13, 2020. URL: https://github.com/AssemblyScript/assemblyscript (visited on 03/13/2020).

[5] *Benchmark Product List - EEMBC - Embedded Microprocessor Benchmark Consortium*. Benchmark Product List - EEMBC - Embedded Microprocessor Benchmark Consortium. URL: https://www.eembc.org/products/ (visited on 03/13/2020).

[6] *Bytecode Alliance*. Bytecode Alliance. Library Catalog: bytecodealliance.org. URL: https://bytecodealliance.org/ (visited on 03/06/2020).

[7] *bytecodealliance/wasm-micro-runtime*. original-date: 2019-05-02T21:32:09Z. Mar. 6, 2020. URL: https://github.com/bytecodealliance/wasm-micro-runtime (visited on 03/06/2020).

[8] L. Clark. *Standardizing WASI: A system interface to run WebAssembly outside the web – Mozilla Hacks - the Web developer blog*. Mozilla Hacks – the Web developer blog. Library Catalog: hacks.mozilla.org. Mar. 27, 2019. URL: https://hacks.mozilla.org/2019/03/standardizing-wasi-a-webassembly-system-interface (visited on 03/04/2020).

[9] M. contributors. *WebAssembly*. MDN Web Docs. Library Catalog: developer.mozilla.org. URL: https://developer.mozilla.org/en-US/docs/WebAssembly (visited on 03/13/2020).

[10] *Design Rationale - WebAssembly*. URL: https://webassembly.org/docs/rationale/ (visited on 03/15/2020).

[11] A. Deveria. *Can I use... Support tables for HTML5, CSS3, etc*. Ca I use. URL: https://caniuse.com/#feat=wasm (visited on 03/14/2020).

[12] *DingTalk's New Biometric Attendance Monitor Based on ESP32 | Espressif Systems*. Library Catalog: www.espressif.com. June 2, 2017. URL: https://www.espressif.com/en/media_overview/news/dingtalk%E2%80%99s-new-biometric-attendance-monitor-based-esp32 (visited on 03/15/2020).

[13] *Embedded Microprocessor Benchmark Consortium*. Embedded Microprocessor Benchmark Consortium. URL: https://www.eembc.org/ (visited on 03/13/2020).

[14] *ESP32 Overview | Espressif Systems*. URL: https://www.espressif.com/en/products/hardware/esp32/overview (visited on 03/02/2020).

[15] *FreeRTOS - Market leading RTOS (Real Time Operating System) for embedded systems with Internet of Things extensions*. FreeRTOS. URL: https://www.freertos.org/ (visited on 03/02/2020).

[16] *FreeRTOS - Real-time operating system for microcontrollers - AWS*. Amazon Web Services, Inc. Library Catalog: aws.amazon.com. URL: https://aws.amazon.com/freertos/ (visited on 03/02/2020).

[17] Gartner. *Internet Of Things (iot)*. Gartner Glossary. Library Catalog: www.gartner.com. URL: https://www.gartner.com/en/information-technology/glossary/internet-of-things (visited on 03/14/2020).

[18] Gartner. *Internet of Things: The Gartner Perspective*. Gartner. Library Catalog: www.gartner.com. URL: https://www.gartner.com/en/information-technology/insights/internet-of-things (visited on 03/14/2020).

[19] I. Grokhotkov. *esp32-idf example fails when -DESP32 is defined · Issue #28 · wasm3/wasm3*. GitHub. Library Catalog: github.com. URL: https://github.com/wasm3/wasm3/issues/28 (visited on 03/12/2020).

[20] IDC. *Prognosis of worldwide spending on the Internet of Things (IoT) from 2018 to 2022 (in billion U.S. dollars)*. June 13, 2019. URL: https://www.statista.com/statistics/668996/worldwide-expenditures-for-the-internet-of-things/.

[21] I. Insights and S. estimates. *Microcontroller unit (MCU) shipments worldwide from 2015 to 2023*. Sept. 2019. URL: https://www-statista-com.eaccess.ub.tum.de/statistics/935382/worldwide-microcontroller-unit-shipments/ (visited on 03/06/2020).

[22] *Instructions — WebAssembly 1.0*. URL: https://webassembly.github.io/spec/core/syntax/instructions.html#syntax-instr-control (visited on 03/13/2020).

[23] A. Jangda, B. Powers, E. Berger, and A. Guha. "Not So Fast: Analyzing the Performance of WebAssembly vs. Native Code". In: *arXiv:1901.09056 [cs]* (May 31, 2019). arXiv: 1901.09056. URL: http://arxiv.org/abs/1901.09056 (visited on 03/13/2020).

[24] J. Knox. "Atmel Joins EEMBC to Participate in Microcontroller Benchmark Development and to Advocate Low Power". In: *Automotive Industries* 193.2 (Feb. 2013). Publisher: Automotive Industries, pp. 133–134. ISSN: 10994130. URL: http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=85966010&site=ehost-live (visited on 03/15/2020).

[25] K.-D. Kramer, T. Stolze, and T. Banse. "Benchmarks to Find the Optimal Microcontroller-Architecture". In: *2009 WRI World Congress on Computer Science and Information Engineering*. 2009 WRI World Congress on Computer Science and Information Engineering. Vol. 2. ISSN: null. Mar. 2009, pp. 102–105. DOI: 10.1109/CSIE.2009.928.

[26] F. Lardinois. *Amazon FreeRTOS is a new operating system for microcontroller-based IoT devices*. TechCrunch. URL: http://social.techcrunch.com/2017/11/29/amazon-freertos-is-a-new-operating-system-for-microcontroller-based-iot-devices/ (visited on 03/02/2020).

[27] A. Maier, A. Sharp, and Y. Vagapov. "Comparative analysis and practical implementation of the ESP32 microcontroller module for the internet of things". In: *2017 Internet Technologies and Applications (ITA)*. 2017 Internet Technologies and Applications (ITA). Wrexham: IEEE, Sept. 2017, pp. 143–148. ISBN: 978-1-5090-4815-1. DOI: 10.1109/ITECHA.2017.8101926. URL: http://ieeexplore.ieee.org/document/8101926/ (visited on 03/02/2020).

[28] *Non-Web Embeddings - WebAssembly*. URL: https://webassembly.org/docs/non-web/ (visited on 03/13/2020).

[29] A. Rossberg, B. Titzer, A. Haas, D. Schuff, D. Gohmann, L. Wagner, A. Zakai, J. Bastien, and M. Holman. "Bringing the web up to speed with WebAssembly". In: *Communications of the ACM* 61 (Nov. 20, 2018), pp. 107–115. ISSN: 00010782. (Visited on 03/13/2020).

[30] *Runtime Structure — WebAssembly 1.0*. URL: https://webassembly.github.io/spec/core/exec/runtime.html#store (visited on 03/13/2020).

[31] V. Shymanskyy. *WASM3 Performance*. Library Catalog: github.com. Feb. 3, 2020. URL: https://github.com/wasm3/wasm3/blob/master/docs/Performance.md (visited on 03/06/2020).

[32] V. Shymanskyy. *wasm3/Interpreter*. GitHub. Library Catalog: github.com. Jan. 23, 2020. URL: https://github.com/wasm3/wasm3/blob/master/docs/Interpreter.md (visited on 03/13/2020).

[33] E. Systems. *Espressif Announces the Launch of ESP32 Cloud on Chip and Funding by Fosun Group | Espressif Systems*. Library Catalog: www.espressif.com. Sept. 7, 2016. URL: https://www.espressif.com/en/media_overview/news/espressif-announces-launch-esp32-cloud-chip-and-funding-fosun-group (visited on 03/15/2020).

[34] L. Wagner. *WebAssembly | Luke Wagner's Blog*. June 17, 2015. URL: https://blog.mozilla.org/luke/2015/06/17/webassembly/ (visited on 03/13/2020).

[35] L. Wagner. *WebAssembly consensus and end of Browser Preview from Luke Wagner on 2017-02-28 (public-webassembly@w3.org from February 2017)*. Feb. 28, 2017. URL: https://lists.w3.org/Archives/Public/public-webassembly/2017Feb/0002.html (visited on 03/13/2020).

[36] L. Wagner. *WebAssembly Will Finally Let You Run High-Performance Applications in Your Browser - IEEE Spectrum*. IEEE Spectrum: Technology, Engineering, and Science News. Library Catalog: spectrum.ieee.org. Sept. 21, 2017. URL: https://spectrum.ieee.org/computing/software/webassembly-will-finally-let-you-run-highperformance-applications-in-your-browser (visited on 03/12/2020).

[37] *wasienv/wasienv*. original-date: 2019-10-16T19:19:48Z. Mar. 10, 2020. URL: https://github.com/wasienv/wasienv (visited on 03/13/2020).

[38] *wasm3/wasm3*. original-date: 2019-10-01T17:06:03Z. Mar. 6, 2020. URL: https://github.com/wasm3/wasm3 (visited on 03/06/2020).