

SoK: Where’s the “up”?! A Comprehensive (bottom-up) Study on the Security of Arm Cortex-M Systems

Xi Tan
CactiLab, University at Buffalo

Zheyuan Ma
CactiLab, University at Buffalo

Sandro Pinto
Universidade do Minho

Le Guan
University of Georgia

Ning Zhang
Washington University in St. Louis

Jun Xu
University of Utah

Zhiqiang Lin
Ohio State University

Hongxin Hu
University at Buffalo

Ziming Zhao
CactiLab, University at Buffalo

Abstract

Arm Cortex-M processors are the most widely used 32-bit microcontrollers among embedded and Internet-of-Things devices. Despite the widespread usage, there has been little effort in summarizing their hardware security features, characterizing the limitations and vulnerabilities of their hardware and software stack, and systematizing the research on securing these systems. The goals and contributions of this paper are multi-fold. First, we analyze the hardware security limitations and issues of Cortex-M systems. Second, we conducted a deep study of the software stack designed for Cortex-M and revealed its limitations, which is accompanied by an empirical analysis of 1,797 real-world firmware. Third, we categorize the reported bugs in Cortex-M software systems. Finally, we systematize the efforts that aim at securing Cortex-M systems and evaluate them in terms of the protections they offer, run-time performance, required hardware features, etc. Based on the insights, we develop a set of recommendations for the research community and MCU software developers.

1 Introduction

Microcontroller units (MCUs) are small computers designed for embedded and Internet of Things (IoT) applications in contrast to microprocessors used in smartphones, personal computers, and servers. They operate at frequencies ranging from several kHz to several hundred MHz. The sizes of their ROMs and RAMs are small and usually fall into the range of several hundred bytes to several megabytes. Even though MCUs are general-purpose computers, they are commonly employed for running specialized software and firmware tailored to specific applications.

The Arm Cortex-M family, which has three major architectures and 12 processors as of 2023, is the most popular 32-bit MCU architecture without a memory management unit (MMU) on the market. More than 80 hardware vendors have licensed Cortex-M cores [1]. 4.4 billion Cortex-M MCUs were shipped in the 4th quarter of 2020 alone [2], and it is es-

timated that Cortex-M MCUs account for almost 100 billion deployed embedded and IoT devices in 2021 [3].

Given the sheer volume of deployed Cortex-M systems, one would anticipate that the security of their hardware and software stack has been thoroughly studied and systematized. Unfortunately, this is not the case. To bridge the knowledge gap that hinders the users and researchers, we seek to answer the following questions regarding their security states:

- *Q1 - What are the security features, limitations, and issues at the Cortex-M microarchitecture, instruction set architecture (ISA), and beyond?* The answer helps understand the constraints in securing software on Cortex-M.

To address this question, we analyze the hardware security limitations of Cortex-M by comparing its offerings with microprocessors. *Our main observation (§3) is that Cortex-M processors lack support for memory virtualization and provide only basic memory protection mechanisms. Additionally, their other security features, e.g., TrustZone, are streamlined compared to their Cortex-A counterparts and introduce new vulnerabilities.*

- *Q2 - What are the security mechanisms and flaws of Cortex-M based software systems?* The answer helps understand the status of Cortex-M software security in real-world systems.

To answer this question, we compile a dataset of 1,797 real-world Cortex-M firmware samples, including 1,003 newly collected ones, and perform by far the largest empirical analysis on the adoption of security mechanisms on real-world Cortex-M systems. In particular, we summarize the software architectures found in these samples and other research projects. We develop binary analysis tools to verify if the collected samples leverage the security mechanisms that have been widely deployed on microprocessor-based systems, e.g., privilege separation and stack canaries. *We uncovered that (§4) despite extensive research on more secure architectures for microcontroller-based systems, these advancements are rarely implemented in real-world firmware. Moreover, the hardware security features offered by Cortex-M processors are seldom utilized in the majority of the assessed firmware; hence, where is the “up”?! Furthermore, existing compiler-based mitiga-*

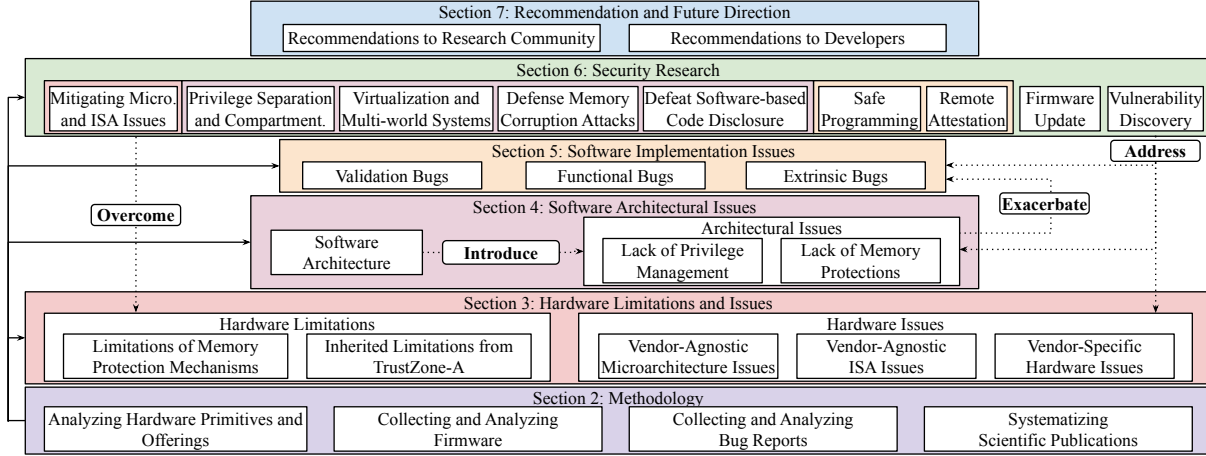


Figure 1: Overview of the organization and contributions of this paper

tions designed for process-based operating systems (e.g., stack canaries) prove ineffective when operating within a single physical address space.

- *Q3 - What are the nature and severity of the publicly disclosed vulnerabilities in the Cortex-M based software systems?* The answer helps find out software bugs that are more likely to be exploited in such systems.

To tackle this question, we analyze 310 Cortex-M related software bug reports spanning nearly six years, from 2017 until 2023. Our analysis includes systems developed by nine hardware vendors, e.g., Nordic and NXP, and seven real-time operating systems (RTOS), e.g., FreeRTOS. We further categorize the software implementation issues into validation, functional, and extrinsic bugs, a taxonomy adopted in a recent work studying the vulnerabilities in Cortex-A systems [4]. *Our insights (§5) include that these systems not only exhibit memory corruption vulnerabilities but also display weaknesses in their protocol and cryptographic implementations.*

- *Q4 - What defenses for Cortex-M systems have been explored in the literature, and what are their limitations?* Together with the previous answers, this helps shed light on new research directions to secure Cortex-M systems.

To address this question, we create a taxonomy and comparative evaluation of over 50 papers spanning nearly nine years. Our evaluation framework considers the defenses each solution offers, the hardware units it relies on, and their run-time overhead in terms of memory size, performance, etc. *Our major observations (§6) include the research community not only shifts the exact same defenses from microprocessor-based systems on Cortex-M systems, e.g., enforcing isolation and confinement, stack integrity, and control flow integrity, but also develops solutions intrinsically linked to the MCU characteristics, e.g., peripheral-oriented fuzzing.*

Based on the insights, we develop a set of recommendations for the research community and MCU software developers (§7). Figure 1 provides an overview of the organization and contributions of this paper. We have open-sourced our source

code, dataset, and supplementary materials ¹.

2 Methodology

2.1 Adversarial Model

In general, we consider the security limitations and issues of the microarchitecture, ISA, and above. In particular, we assume an adversary can perform (i) microarchitecture side-channel attacks, e.g., bus interconnect; (ii) glitching, e.g., voltage fault injection; (iii) remote attacks via a network; (iv) nearby wireless attacks via BLE, ZigBee, etc.; (v) local attacks through peripherals and debug ports; and (vi) software side-channel attacks. On systems without TrustZone-M, we consider an adversary with one or more of the following objectives: (i) to obtain secrets from the flash, e.g., intellectual property (IP) theft and RAM; (ii) to tamper sensitive data; (iii) code execution and privilege escalation, e.g., control-flow hijacking. On systems with TrustZone-M, we assume all components in the non-secure state are untrusted and consider an adversary with all aforementioned goals as well as compromising the secure state.

2.2 Analyzing Hardware Offerings

We provide a detailed analysis of the hardware security limitations and issues. Due to the page limit, a detailed walk-through of the Cortex-M architecture is not included in this paper. Interested readers please refer to our supplementary materials, which consolidate information from various official sources [5–14]. To aid in research for the community, we have developed an open-source code suite, demonstrating the use of Cortex-M security features.

2.3 Collecting and Analyzing Firmware

Collecting Firmware. The process of collecting and decoding Cortex-M firmware was far from straightforward and re-

¹<https://github.com/CactiLab/SoK-Cortex-M>

Table 1: Manufacturer distribution of the compiled real-world firmware dataset. *Italic* represents newly collected sample that were not publicly released before.

HW Vendor	Nordic [15]	<i>Other Nordic</i>	TI [15]	<i>Telink</i>	<i>Dialog</i>	<i>NXP</i>	<i>Cypress</i>	ST [16]	Total
# Firmware	768	690	22	192	53	1	67	4	1,797
# Devices	513	-	20	120	36	1	-	-	689

sulted in the accumulation of significant amounts of unusable data. We used three approaches to collect firmware: (i) we filtered Cortex-M firmware from publicly available embedded system datasets [15, 17–21]; (ii) adopting an analogous methodology as described in [15], we developed scripts to analyze/unpack mobile apps and extract potential Cortex-M firmware. Using this approach, we collected 4,693 potential samples from six silicon vendors. These samples are in various formats, e.g., S-record for NXP, cyacd format for Cypress, and proprietary format of Qualcomm; (iii) we crawled websites for 25 silicon and device vendors known for embedded and IoT devices. This effort resulted in 1,687 potential samples, but none of them turned out to be Cortex-M firmware. This aligns with the findings in FirmXRay [15], which noted that vendors seldom make their firmware available online.

As shown in Table 1, our firmware collection endeavor ended up with 1,797 unique Cortex-M firmware from seven hardware vendors. Among these, the FirmXRay dataset includes 790 firmware samples, representing 533 distinct devices from two vendors (768 from Nordic [22] and 22 from Texas Instruments [23]). Additionally, the HEAPSTER dataset [16] encompasses four Cortex-M binaries from STMicroelectronics (ST) [24]. Furthermore, we have gathered 1,003 firmware from other vendors, including Nordic (690), Telink [25] (192 firmware for 120 unique devices), Dialog [26] (53 firmware for 36 devices), NXP [27] (1), and Cypress [28] (67). These samples have not been publicly shared before. The firmware in our collection is in raw binary format, lacking symbolic information.

Analyzing Firmware. We used FirmXRay [15] to recognize the base address of each firmware. Scripts were then developed to identify the Cortex-M vector table and perform recursive disassembly with Ghidra [29]. We also applied scripts to filter a portion of firmware samples that contain device information, ensuring that they are from distinct devices. We conducted an analysis of the disassembled samples using the following heuristics: (i) to identify if firmware uses any RTOS, we performed binary function recognition [30] and string searches for ten popular RTOSs; (ii) for firmware that uses an RTOS, we analyzed if task stack overflow checks are performed. To this end, we checked if the task stack overflow handling functions, e.g., `osRtxKernelErrorNotify()` with the parameter `osRtxErrorStackOverflow` in CMSIS RTOS2 [31], are called by other functions in the firmware; (iii) we analyzed if and how the `CONTROL` register is changed and how the `SVC` instruction is used to determine privilege separation and stack usages; (iv) to check if there are stack

canaries, we analyzed function prologues and epilogues for specific instruction patterns derived from canary-protected functions generated by three compilers. In addition, we searched if the firmware has the hard-coded libc error message “*** stack smashing detected ***” and whether the function printing out this message is called by other functions, which is a practice used before [32].

2.4 Collecting and Analyzing Bug Reports

We retrieved over 500 hardware and software bug reports related to Cortex-M systems from 2017 to 2023 [33], which shows a growing trend. Besides “Arm”, we included in our list of keywords the names of top hardware vendors [34], popular RTOSs [35], and embedded SSL libraries, e.g., Mbed TLS [36] and wolfSSL [37]). We manually confirmed the bug reports indeed affect Cortex-M systems, including verifying the affected chips and inspecting the source code. Two researchers worked together to categorize each bug into a relevant subclass, which was verified by a third researcher.

2.5 Systematizing Scientific Publications

We collected over 30 papers on Cortex-M security from top conferences². In addition, we supplement our list of surveyed papers with another over 20 articles that are highly relevant to the topic but published in other venues. Note that our systematization focuses on the works explicitly designed for and implemented on Cortex-M. Nevertheless, we discuss related works that were designed for or implemented on other architectures but may be applied to Cortex-M in §6.10.

2.6 Threats to Validity

Our analysis of firmware may be subject to biases and imprecision due to the limited number of firmware. There is a risk of over-representing systems from specific vendors. Most firmware in our dataset (57.3%) are raw binaries and lack detailed device and architecture information, making it difficult to confirm their intended use cases and resulting in a potential bias in analyzing similar firmware samples. Additionally, the lack of proof-of-concept exploits and vague CVE descriptions introduces imprecisions in the classification of vulnerabilities. Furthermore, our analysis focuses on publicly disclosed vulnerabilities. Undiscovered vulnerabilities could unveil additional fundamental issues in Cortex-M systems.

3 Hardware Limitations and Issues

3.1 Hardware Limitations

Hardware limitations are missing or constrained hardware security features, which are typically non-patchable. Compared

²<https://csrankings.org/>

with Cortex-A, Cortex-M features distinct design elements, particularly in its memory protection mechanisms and the TrustZone extension (TrustZone-M versus TrustZone-A).

Limitations of Memory Protection Mechanisms

L01. No memory virtualization: No hardware-supported memory virtualization is available on Cortex-M due to the absence of a memory management unit (MMU). Instead, software modules share the same physical address space. Such lack of memory virtualization also implies a small address space (4GB), which presents challenges to effective address space layout randomization (ASLR) due to low entropy.

L02. No input-output memory management unit: Besides MMU, input-output memory management unit (IOMMU) or its equivalents, i.e., IOMPU, that provide memory protection from malicious direct memory access (DMA)-capable peripherals are also missing on Cortex-M. Some hardware vendors implement their own IOMPU, i.e., the resource domain controller on NXP i.MX RT [38, 39], but they are only found in some of the latest devices.

L03. A small number of MPU regions and limited sizes: Cortex-M only supports a small number of memory protection unit (MPU) regions, and the size of regions must be a multiple of 32 bytes. Compared to the page-based memory access control on microprocessors, the granularity of MPU-based is coarse-grained, and it is insufficient to implement fine-grained isolation that requires a large number of regions.

L04. A small number of secure/non-secure memory regions: The number of regions supported by secure attribute unit (SAU) is small, e.g., up to 8 regions on Cortex-M33, resulting in limited design choices in splitting the secure and non-secure address space. To alleviate this issue, silicon vendors use the implementation defined attribution unit (IDAU), which supports up to 256 regions, to create more partitions. However, if more than 256 partitions are needed or the device has many peripherals, this may not be enough [40].

Inherited Limitations from TrustZone-A

L05. No intrinsic encryption to protect the secure state memory: TrustZone-M does not encrypt the secure state memory. Consequently, cold boot attacks [41] can dump the secure state memory. There could also be information leakage when a memory protection controller (MPC) assigns a memory region from the secure state to the non-secure state at run-time, which we will discuss in [105](#).

L06. Lack of intrinsic support for multiple trusted execution environments: TrustZone-M only provides *one* isolated execution environment in which the trusted firmware executes, resulting in a large software trusted computing base (TCB). For instance, TF-M [42] has over 117K lines of code.

L07. Lack of hardware-based remote attestation in TrustZone-M: Same as Cortex-A [4], Cortex-M TrustZone lacks a hardware-based integrity reporting mechanism, so it

cannot provide a hardware-based remote attestation as Intel software guard extensions (SGX) does. For example, the Arm platform security architecture (PSA) introduces a weakened software-based attestation method [43, 44].

Insights

- The Cortex-M architecture offers weaker memory management interfaces than popular microprocessors, creating challenges to enforce memory isolation and security.
- TrustZone-M inherits hardware limitations of TrustZone-A and introduces more constraints.

3.2 Hardware Issues

Hardware issues discuss vulnerable hardware components and hardware-supported operations.

Vendor-Agnostic Microarchitecture Issues

I01. Vulnerable to microarchitectural side-channel attacks: Although most Cortex-M processors lack a cache or branch predictor at the microarchitectural level, there are other side channels that can leak information.

Information leakage through power analysis: ELMO [45] demonstrates the feasibility of reversing AES S-Box output code sequences through power analysis on the Cortex-M0 processor. Furthermore, Vafa et al. [46] successfully applied a power analysis attack to recover running instructions on the Cortex-M3 processor.

Information leakage through timing side-channels: MCU bus interconnect arbitration logic involves delays when multiple bus masters, such as the CPU and DMA, simultaneously access a shared secondary port, like a memory controller. As demonstrated in BUSTed [47], the attacker can successfully bypass protections provided by the MPU and TrustZone by exploiting these timing differences.

Information leakage through long-term data remanence: UnTrustZone [48] reveals that static random-access memory (SRAM) can be manipulated to imprint and expose on-chip secrets by accelerating analog-domain changes in SRAM. Using this method, UnTrustZone successfully extracts AES keys and proprietary firmware from various Cortex-M devices protected by TrustZone.

I02. Vulnerable to fault injections: A fault injection attack involves deliberately causing errors in a system's hardware (e.g., voltage, clock, electromagnetic) to disrupt its normal operations of a digital circuit and exploit these induced faults for malicious purposes. Johannes Obermaier and Marc Schink et al. discussed how to escalate the debug interface permissions or execute arbitrary code by injecting faults into voltage [49], Quad-SPI bus [50], and electromagnetic [51] at boot time on Cortex-M0/3/4 devices. μ -Glitch [52] entails injecting multiple, coordinated voltage faults into Cortex-M devices to bypass the TrustZone protection, allowing leaking secrets stored in secure memory into non-secure code.

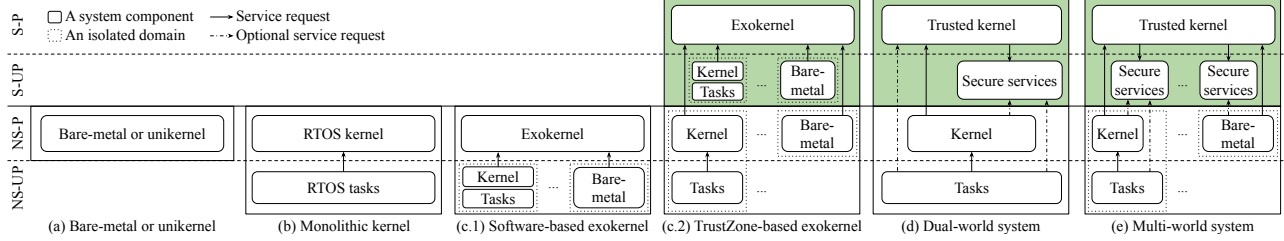


Figure 2: Identified Cortex-M software architectures in the collected dataset and in the literature. NS-UP: non-secure unprivileged, NS-P: non-secure privileged, S-UP: secure unprivileged, S-P: secure privileged.

Vendor-Agnostic ISA Issues

I03. Fast state switch mechanism exploitable for privilege escalation: Cortex-M TrustZone uses the fast state switch technique to allow direct cross-state transitions from any privilege level without the need for a higher privileged secure monitor mode like Cortex-A TrustZone. Although this feature makes cross-state transitions more efficient, it exposes vulnerabilities to a recently discovered attack known as ret2ns [53]. This attack leverages critical system registers and instructions used by the fast state switch to escalate privilege in the non-secure state, potentially leading to arbitrary code execution.

I04. Improper privilege management for inter-processor debugging: CVE-2018-18068 reports that the debugging host's privilege level is ignored in the inter-processor debugging mode, allowing the non-secure state on both TrustZone-M and TrustZone-A to gain access to the secure state resources via the ETM [54, 55].

I05. Information leakage to the non-secure state due to state switches: This could happen through memory and general-purpose and special registers: (i) if a region used by the secure state is re-mapped by MPC into the non-secure state without proper sanitization, sensitive information will be leaked; (ii) information leakage could happen if the general-purpose registers are not cleared when switching to the non-secure state. To address this issue, Arm recommends general-purpose registers that are not used to pass arguments should be cleared before state switches [7]; (iii) CVE-2021-35465 reports an issue of the floating-point lazy load multiple (VLLDM) instruction, which allows the non-secure code to access secure state floating-point registers.

Vendor-Specific Hardware Issues

I06. Improper privilege management in vendor-specific hardware features: Some hardware vendors introduce over-powerful hardware features that can be exploited to gain full control of the system. For example, NXP LPC55S6x MCUs include a ROM patch controller to fix bugs in the ROM after fabrication. CVE-2021-31532 reports that even attackers in the non-secure state and unprivileged level can utilize the ROM patch controller to reconfigure the SAU regions to gain privilege escalation. CVE-2022-22819 shows that the ROM patch controller firmware also has a buffer overflow bug that can lead to arbitrary code execution at the privileged level.

I07. Bypassable vendor-specific readback protection:

Only M55 and M85 have the execute-only memory (XOM) feature, which prevents software or a hardware debugger from reading execute-only memory [56]. For MCUs before M55, some hardware vendors implement their own hardware units to prevent reading from the debug interface, a feature known as readback protection. For instance, the Nordic nRF51 series implements a mechanism to prevent debuggers from directly accessing flash and RAM address ranges. Notwithstanding, we found that only 32 out of the 1,458 Nordic samples in our dataset enable this feature. This protection, however, can be easily bypassed through arbitrary register read and write and single stepping in debugging [57]. Though the mechanism was improved in the nRF52 series [58], CVE-2020-27211 reports that a voltage glitch attack can still bypass it [51]. Similar mechanisms implemented by ST [59], NXP [60], and TI [61] are also bypassable by inferring instructions from the observed state transitions [62].

Insights

- Streamlined hardware mechanisms in Cortex-M, e.g., fast state switch, lead to new privilege management vulnerabilities and information leakage.
- The fragmentation of the Cortex-M ecosystem has brought in new security challenges: vendors aggressively introduce over-powerful hardware, which can undermine Cortex-M systems security if not properly designed.

4 Software Architectural Issues

4.1 Software Architectures

As shown in Figure 2, we identified two (i.e., a and b) software architectures in the collected firmware dataset and another three (i.e., c, d, and e) in the literature. **Bare-metal systems and unikernels** (a) run directly on the hardware at the highest (non-secure) privilege level. The RTOSs in such systems are only linked as a library OS, e.g., Mbed OS bare-metal profile [63]. We will discuss in I08 that over 99.44% of the 1,797 firmware belong to this category, including 66 firmware samples that use FreeRTOS and another 13 firmware use Mbed OS. **Monolithic kernels** (b) are the most common organization in microprocessor-based systems, e.g., Linux and Windows. Such systems run the kernel entirely at the

Table 2: Empirical Analysis of Security Features Adopted in Real-world Firmware

Hardware Vendor	Nordic (FirmXRay)				Other Nordic		TI		Telink		Dialog		NXP	Cypress	ST	Total
Security Feature	#F		#D		#F		#F	#D	#F	#D	#F	#D	#F	#F	#F	#F
Readback Protection (I07)	17	2.21%	9	1.75%	15	2.17%	0	-	0	-	0	-	0	0%	0	32
Privilege Separation (I08)	8	1.04%	5	0.97%	2	0.29%	0	0%	0	0%	0	0%	0	0%	0	10
SVC for Library Call (I09)	753	98.04%	500	97.47%	690	100%	2	9.09%	1	5%	17	8.85%	17	14.17%	2	1,466
Stack Separation (I10)	49	6.38%	34	6.63%	82	11.88%	0	0%	0	0%	0	0%	3	5.66%	1	134
Stack Limit Register Usage (I10)	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%	0	0
Task Stack Ovf. Guard* (I10)	59	96.72%	4	80%	9	32.14%	-	-	-	-	-	-	-	-	-	68
Memory Access Control (MPU) (I12)	0	0%	0	0%	4	0.58%	0	0%	0	0%	0	0%	1	100%	0	5
Memory Access Control (sMPU) (I12)	19	2.47%	17	3.31%	0	0%	-	-	-	-	-	-	-	-	-	19
Stack Canaries (I13)	0	0%	0	0%	1	0.14%	0	0%	0	0%	0	0%	0	0%	0	1
Proper Instruction Sync. Barriers† (I14)	30	36.59%	16	27.12%	68	40%	-	-	-	-	0	0%	0	0%	-	98

#F: Number of firmware, #D: Number of devices, -: Not applicable, *: The percentage is only based on firmware that use RTOS, †: The percentage is only based on firmware that update CONTROL with the MSR instruction.

privileged level, and applications run in (unprivileged) user space. However, only 0.56% of the firmware samples in our dataset fall into this category. **Exokernels** (c) run at the highest privilege level, virtualizing and allocating resources to RTOSs or bare-metal applications running at a lower privilege level. We will discuss two software-based exokernel projects, Hermes [64] and MultiZone [65], and two Cortex-M TrustZone-based exokernel projects, ILTZVisor [66, 67] and SBIs [68], in **D05. Dual-world systems** (d), which are enabled by TrustZone-M, run RTOSs and applications in the non-secure state, whereas secure OS/services run in the secure state. The Trusted Firmware for Cortex-M (TF-M) [69] is a reference implementation of this architecture. **Multi-world systems** (e) enable multiple equally-secure TEEs. We will discuss uTango [70], one prominent example of a multi-world TEE implementation leveraging TrustZone-M in **D06**.

Insights

- Despite the research progress towards more secure architectures for Cortex-M systems, a large number of the real-world firmware in our dataset are simply bare-metal systems and unikernels.

4.2 Architectural Issues

Software architectural issues refer to common limitations and flaws we found in real-world firmware.

Lack of Privilege Management

I08. No or weak privilege separation: As shown in Table 2, only 10 out of 1,797 samples in our dataset execute some code at the unprivileged level, and the others execute entirely at the privileged level. Due to the lack of spatial isolation and privilege separation, a bug anywhere may compromise the whole system, even reverting MPU settings.

I09. SVC repurposing: The SVC instruction is designed to escalate the execution level; however, executing this instruction at the privileged level also transfers the control to the SVC handler. Surprisingly, we find that 1,466 (81.58%) samples run everything at the privileged level and repurpose this feature to call library APIs, e.g., Nordic SoftDevice [71], in-

stead of privilege escalation. The behavior is consistent across vendors, e.g., Nordic, TI, Telink, Cypress, and ST.

Lack of Memory Protections

I10. No or weak stack separation: RTOSs, such as FreeRTOS [72] and Zephyr [73], support multi-tasking, so each task has its own stack. However, stack separation between the kernel and application is rarely used in bare-metal firmware. Armv8-M also introduces stack limit registers (PSPLIM and MSPLIM) to delimit the boundaries of stacks. However, no firmware in our dataset has been used them.

RTOS Implementations: We found that only a few RTOSs protect tasks' stacks, and only Zephyr optionally supports using stack limit registers. When stack guard is enabled, FreeRTOS [74] and Mbed OS [75] insert a predefined delimiter to mark the boundary of each task's stack. Zephyr can use either PSPLIM or an MPU-configured memory guard to prevent overwriting beyond a task's stack [76].

Empirical Analysis on Real-world Firmware: 10 samples that adopt privilege separation (discussed in **I08**) leverage both the MSP- and PSP-based stacks. In addition, another 124 samples use both the MSP- and PSP-based stacks without privilege separation. All other samples (1,663; 92.54%) only adopt a single MSP-based stack. 59 of the 66 FreeRTOS-based firmware samples and 7 of the 13 Mbed OS-based firmware samples use task stack overflow guards.

I11. Secure state exception stack frame manipulation: CVE-2020-16273 shows that the non-secure state software may manipulate the secure stacks and hijack the secure control flow if the secure software does not properly initialize the secure stacks. To this end, an attacker creates a fake exception return stack frame to deprivilege an interrupt.

I12. No or weak memory access control; executable stack: Despite the presence of MPU, previous research suggests that it is rarely utilized in most real-world systems [77–79]. We confirm that 1,773 of the 1,797 firmware in our dataset do not use MPU, which means the *code*, *SRAM*, and *RAM* regions are executable and malicious code can read and write arbitrary memory. Out of the 24 firmware that use MPU in our dataset, five use the MPU defined by Arm. The remaining 19 use a vendor-specific implementation, i.e., Nordic's simplified

MPU (sMPU) [80], which only supports a subset of MPU features. Specifically, sMPU only supports read and write permissions with two protection domains.

I13. No or weak stack canary: Stack canary implementation involves initializing the canary value, runtime verification, and handling mismatches. The compiler and libraries manage the latter two, with the system initializing the canary value. In the standard C libraries (libc), the value of the stack canary is taken from a global variable `__stack_chk_guard`. In modern OSs, the value of the canary is randomly initialized when a process is created. However, embedded systems often use a fixed canary value post-compilation or boot [81]. Notably, there is only one `__stack_chk_guard` for the entire physical address space. We found that only *one* of the 1,797 firmware samples in our dataset adopts it.

I14. Missing barrier instructions: Barrier instructions, including data memory barrier (DMB), data synchronization barrier (DSB), and instruction synchronization barrier (ISB), guarantee that system configurations take effect before any memory operations [82]. The omission of them is unlikely to cause any issues on most Cortex-M MCUs because they do not have out-of-order execution and branch prediction capabilities. For MCUs that do have such capabilities, e.g., M7, this may lead to similar vulnerabilities that were discovered on microprocessors [83–85]. To check if barriers are set in firmware, for any CONTROL register update, we verify if there is an ISB instruction in its ten subsequent instructions. Our analysis shows that only 98 of the 281 firmware samples (34.88%) that update the CONTROL register use the ISB instruction thereafter. However, as we cannot confirm which architecture those firmware are using, it is unclear whether the missing barrier instructions will cause issues or not.

Insights

- The real-world firmware samples in our dataset barely use the security features of Cortex-M and largely lack the security mitigations that are widely deployed on modern microprocessor-based systems.
- Some software- and compiler-based mitigations, e.g., stack canaries, are less effective on MCU-based systems and should be redesigned.

5 Software Implementation Issues

Table 3 presents the numbers of Cortex-M related CVEs affecting nine hardware vendors, seven RTOSs, and two TLS libraries. We break down the number based on CVSS scores [86]. As shown in Table 3, the majority of CVEs (53.85%) affecting hardware vendors are classified as “medium” severity, while the majority of CVEs affecting RTOSs (78.07%) are categorized as either “critical” or “high”. We use a bug classification system proposed in [4] to characterize them into three major classes, i.e., validation, functional, and extrinsic. We summarize the results in Table 4, where we

Table 3: Distribution of disclosed Cortex-M related CVEs (2017 - 2023)

HW Vendor/RTOS/Lib	Critical	High	Medium	Low	Total
Arm	0 0%	4 57.14%	2 28.57%	1 14.29%	7 1.99%
Microchip Technology	1 14.29%	2 28.57%	4 57.14%	0 0%	7 1.99%
Silicon Labs	6 40.00%	2 13.33%	6 40.00%	1 6.67%	15 4.27%
NXP Semiconductors	1 7.69%	6 46.15%	6 46.15%	0 0%	13 3.70%
ST Microelectronics	2 12.50%	2 12.50%	12 75.00%	0 0%	16 4.56%
Cypress Semiconductor	0 0%	6 50.00%	6 50.00%	0 0%	12 3.42%
Gigadevice	0 0%	0 0%	6 100.00%	0 0%	6 1.71%
Texas Instruments	0 0%	6 54.55%	5 45.45%	0 0%	11 3.13%
Nordic	0 0%	2 50.00%	2 50.00%	0 0%	4 1.14%
Subtotal (HW vendors)	10 10.99%	30 32.97%	49 53.85%	2 2.20%	91 25.93%
FreeRTOS	3 15.79%	9 47.39%	7 36.84%	0 0%	19 5.41%
CMSIS RTOS2	1 100.00%	0 0%	0 0%	0 0%	1 0.28%
Mbed OS	6 60.00%	4 40.00%	0 0%	0 0%	10 2.85%
Zephyr	17 23.61%	36 50.00%	18 25.00%	1 1.39%	72 20.51%
RIOT-OS	10 33.33%	18 60.00%	2 6.67%	0 0%	30 8.55%
Contiki-ng	16 39.02%	18 43.90%	7 17.07%	0 0%	41 11.68%
Azure	5 35.71%	3 21.43%	5 35.71%	1 7.14%	14 3.99%
Subtotal (RTOSs)	58 31.01%	88 47.06%	39 20.86%	2 1.07%	187 53.28%
Mbed TLS	6 20.69%	12 41.38%	11 37.93%	0 0%	29 8.26%
WolfSSL	10 22.73%	14 31.82%	20 45.45%	0 0%	44 12.54%
Subtotal (Libs)	16 21.92%	26 35.62%	31 42.47%	0 0%	73 20.80%
Total	84 23.93%	144 41.03%	119 33.90%	4 1.14%	351

further provide a breakdown of bugs based on the functionality and the software components.

5.1 Validation bugs

Validation bugs refer to bugs that mishandle or improperly validate input and output data. Examples are out-of-bounds read and write and improper parameter validation. They are frequently exploited for arbitrary write and read, allowing attackers to steal/overwrite sensitive information, execute remote code, or cause a denial of service.

I15. Validation bugs in communication components: Table 4 shows that 57.78% of validation bugs affect communication stacks, e.g., Bluetooth and TCP/IP implementations. For instance, FreeRTOS has a DNS poisoning bug that does not check if a DNS answer matches an outgoing query (CVE-2018-16598). Open-source libraries that are heavily used by Cortex-M systems, such as Mbed TLS or WolfSSL, also have 42 validation bugs.

I16. Validation bugs in device drivers: Device drivers are exposed to attackers through physically-accessible peripherals, e.g., the USB interface. We found 25 bugs that affect two hardware vendors and two RTOSs in this category. For instance, the buffer overread bug of the NXP Kinetis K82 USB driver can be leveraged to access the flash (CVE-2021-44479). The USB driver in Zephyr also has a buffer overflow bug that allows a USB-connected host to cause possible remote code execution (CVE-2020-10019).

I17. Validation bugs in dynamic memory allocations: Embedded systems commonly implement custom allocators rather than using the standard heap implementations in the Libc [16]. Bugs in heap management can result in a system crash or arbitrary code execution. For example, NXP’s SDK, RIOT-OS, Mbed OS, and CMSIS RTOS are vulnerable to integer overflows in their allocator functions [87].

Table 4: Distribution of Cortex-M software CVEs in different classes

Bug Class	Functions	Affected HW Vendors' SDKs	Affected RTOSs / TLS libs	#Bugs	
Validation	Communication	NXP (2), Microchip (5), ST (1), TI (9), Cypress (10), Silicon Labs (8), Nordic (3)	FreeRTOS (11), RIOT-OS (24), Mbed OS (7), Zephyr (32), Contiki-ng (39), Mbed TLS (14), wolfSSL (28)	193	57.78%
	Device Driver	TF-M (1), NXP (4), ST (7)	Zephyr (8), Azure (5)	25	7.48%
	Memory Allocation	NXP (1)	FreeRTOS (2), RIOT-OS (2), Mbed OS (2), CMSIS RTOS2 (1), Zephyr (2)	10	2.99%
	Context Switch	TF-M (2)	FreeRTOS(1), Zephyr (3)	6	1.79%
	Others	Silicon Labs(5), NXP (2), Microchip (1)	Contiki-ng (1), Zephyr (10), Azure (9)	28	6.59%
Functional	Protocol Implementation	TI (1), Cypress (2), Silicon Labs (2)	FreeRTOS (3), RIOT-OS (4), Zephyr (13), Mbed OS (1), Mbed TLS (3), wolfSSL (9)	38	11.38%
	Memory Access Control	TF-M (1), NXP (1), ST (1)	FreeRTOS (2), Zephyr (4), Contiki-ng (1)	10	2.99%
	Cryptography Primitive	TF-M (2), Microchip (1), ST (1)	Mbed TLS (4), wolfSSL (4)	12	3.59%
Extrinsic	Software Side-Channel	ST (1)	Mbed TLS (8), wolfSSL (5)	14	4.19%

I18. Validation bugs in context switch components: Bugs in these components have been exploited for privilege escalation. Zephyr uses signed integer comparison to validate the syscall number, so a negative number leads to privilege escalation (CVE-2020-10027). TF-M has a bug allowing for out-of-bounds write in an NSC function, which can lead to data leakage from the secure state (CVE-2021-27562).

I19. Validation bugs in other components: As discussed in I08, many systems execute entirely at the privileged level, and bugs in any component could lead to severe consequences. For example, a buffer overflow in FreeRTOS' shell can cause privileged code execution (CVE-2020-10023). Microchip's SDK has integer overflows that can be leveraged to access flash memory (CVE-2019-16127).

5.2 Functional bugs

Functional bugs refer to programming errors that do not correctly implement the intended design.

I20. Functional bugs in protocol implementations: 11.38% of the functional bugs are related to protocol implementations. For instance, the Bluetooth controller in the Cypress SDK uses a much shorter random number (than 128 bits) as the paring number, allowing the brute force of the random number to perform a man-in-the-middle attack during BLE pairing (CVE-2020-11957).

I21. Functional bugs in memory access control: Incorrect memory access control configurations, including for MPU and TrustZone, compromise isolation. We found eight bugs affecting one hardware vendor and two RTOSs in this category. For example, FreeRTOS has a bug that allows any code to set the system privilege level (CVE-2021-43997).

I22. Functional bugs in cryptography primitives: We found four bug reports in this category. For instance, RIOT-OS has a nonce reuse bug in its encryption function (CVE-2021-41061) and TF-M has a functional bug when cleaning up the memory allocated for a multi-part cryptographic operation, resulting in a memory leak (CVE-2021-32032). The implementations of PKCS #1 v1.5 padding for RSA in the ST (CVE-2020-20949) and Microchip (CVE-2020-20950) SDKs are vulnerable to the Bleichenbacher attack [88]. This vulnerability relies on the use of error messages or responses

from the server to gain information about the validity of the padding after decryption attempts.

5.3 Extrinsic bugs

Extrinsic bugs refer to defects that do not belong to the validation bugs or functional errors.

I23. Software side-channels: The Lucky 13 attack in Mbed TLS (CVE-2020-16150 and CVE-2020-36423) enables an attacker to deduce secret key information by exploiting time variations in the decryption process. This vulnerability, specifically found in Cipher Block Chaining (CBC) mode, is based on the time differences associated with padding length.

Insights

- Most Cortex-M based production systems are written in memory-unsafe languages, e.g., C [89], and they suffer from memory corruption vulnerabilities.
- Microcontrollers lack security mechanisms present in microprocessors for decades, such as privilege separation. Microcontroller developers may not realize the absence of features like an MMU can pose greater risks than microprocessors. Without privilege separation, any bug can be critical and compromise the entire system.

6 Security Research

We present a taxonomy of the security research projects on Cortex-M systems. Figure 3 depicts and summarizes the relationships among limitations, issues, and mitigations at different layers. Table 5 presents a comparative evaluation.

Addressing Hardware Issues

6.1 Addressing Microarchitectural and ISA Issues

D01. Mitigating microarchitectural attacks: To mitigate information leakage through timing side-channels (I01), BUSTed [47] recommends disabling DMA during sensitive execution, and introducing random delays. To counter information leakage through long-term data remanence, UnTrustZone [48] suggests initializing SRAM at startup. To mitigate

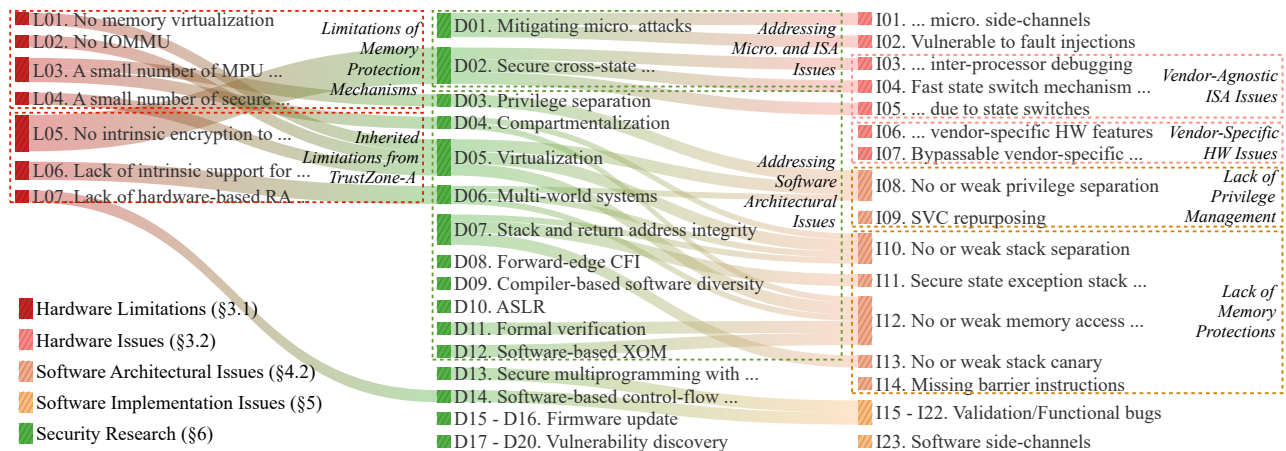


Figure 3: The relationships among the systematized Cortex-M related limitations, issues, and mitigations. The connections indicate the issues a research direction attempts to address and the limitations it needs to overcome. For instance, to address the issue of *no or weak privilege separation* (I08), mitigations (D03, D05, and D06) have been proposed, and they overcome some limitations (L01, L02, and L03). An interactive version of this figure can be accessed at our anonymized repo.

fault injection attacks (I02), one strategy is the use of duplicate security-critical registers [131]. μ Glitch suggests introducing random delays in the execution code to complicate the parameter determination process for fault injections.

D02. Secure cross-state control and data interactions: One effective way to counteract privilege escalation through fast state switching (I03) is to add additional privilege checks. Ret2ns [53] suggests using address masking and MPU configuration checks to limit return targets from secure to non-secure state at the non-secure unprivileged level. In improving privilege management for inter-processor debugging (I04), Nailgun [55] employs MPU to restrict low-privilege access to debug registers. To mitigate information leakage during cross-state switches (I05), one approach is to implement authentication and authorization between the two states, as SeCRet [132] does for TrustZone-A. Secure Informer [95] and ShieldD [96] authenticate secure service calls from the non-secure state by verifying non-secure MPU configurations.

Addressing Software Architectural Issues

6.2 Separation of Privilege

Projects in this category provide different levels of granularity in isolating and confining software modules of *one* bare-metal system or *one* RTOS to address I08.

D03. Privilege separation: Solutions were proposed to automatically relegate RTOS tasks and bare-metal systems to the unprivileged level and use MPU to govern memory access. SAFER SLOTH [97] dispatches tasks as interrupt handlers and lowers the privilege level in the interrupt service routine. EPOXY [77] automatically identifies operations requiring privileged execution (e.g., MSR, move to system registers from general-purpose registers) in bare-metal systems. It then relegates the whole bare-metal system to the unprivileged level and instruments privilege escalation and relegation instructions around the operations requiring privileged execution.

These privilege separation approaches only introduce a small number of context switches, introducing low overhead.

D04. Compartmentalization: The projects on privilege separation (D03) only split a program into privileged and unprivileged parts. However, software modules at the same privilege level still reside in the same security and fault domain, resulting in coarse-grained memory access control (I12). Several compartmentalization solutions attempt to address this issue.

Compartmentalization with heavy context switches: uSFI compiler [98] instruments an entry function for each module and changes cross-module procedure calls to SVC instructions. ACES [79] instruments binaries to enforce inter-component isolation. MINION [99] automatically identifies the reachable memory regions of tasks through static analysis and enforces run-time memory access control. Because there are limited available MPU regions (L03), ACES and MINION propose schemes to merge the compartments. Compared to D03, compartmentalization introduces more context switches between modules; hence, the overhead is higher.

Compartmentalization with reduced context switches: To reduce the overhead introduced by compartmentalization, OPEC [100] leverages global variable shadowing to minimize the need for MPU regions and compartmentalizes programs to include only essential functions. EC [101] uses a formally verified microkernel and intra-kernel isolation to achieve compartmentalization. CRT-C [102] compartments an RTOS into kernel, threads, and device drivers and utilizes CheckedC [133] to restrict their programming capabilities.

DMA-enabled compartmentalization: The aforementioned compartmentalization solutions do not support DMA, leaving the system vulnerable to malicious DMA-capable devices due to the absence of an IOMMU (L02). D-Box [103] addresses this issue by introducing more secure MPU configurations and kernel extensions with explicit support for DMA operations.

Table 5: Comparative evaluation of system isolation and attack mitigation projects for Cortex-M (§6.2 - §6.8). The first column of the table lists the major defense mechanism proposed or adopted in a project.

	Project	Year	Venue	Input (S: source code; B: binary)	Target (B: bare-metal; R: RTOS)	Prototype Implementation (ISA)	Limit. of Memory Prot. Mechanisms	Inherited Limit. from TrustZone-A	Vendor-Agnostic Micro. Issues	Vendor-Agnostic ISA Issues	Vendor-specific Hardware Issues	Lack of Privilege Management	Lack Memory Protections	Validation Bugs	Functional Bugs	Extrinsic Bugs	MPU	Unprivileged Store/Load Instructions	TrustZone	DWT	FPB	Code, Binary Size Increase	Memory Overhead	Energy Consumption Overhead	Bare-metal Applications	RTOSs	BEEBS [90]	CoreMark [91]	CoreMark-Pro [92]	Dhrystone [93]	Embench [94]
D01	BUSted [47]	2023	S&P	-	-	v8		✓									+														
	UnTrustZone [48]	2023	S&P	-	-	-		✓									+														
	μGlitch [52]	2023	USENIX	-	-	-		✓																							
D02	Nailgun [55]	2021	TDSC	S	R	v7			✓			✓					+														
	ret2ns [53]	2023	DAC	S	R	v8			✓			✓					+	+													
	Secure Informer [95]	2022	CPSS	S	R	v8	✓		✓			✓					+	+				<.01				3.5					
	ShieLD [96]	2022	TDSC	S	R	v8	✓		✓			✓					+	+				.04				2600					
D03	SAFER SLOTH [97]	2014	RTAS	S	R	v7						✓	✓				+										>100				
	EPOXY [77]	2017	S&P	S	B	v7						✓	✓				+					29	29	2.6	2.4			1.6			
D04	uSFI [98]	2018	DATE	S	R	v7						✓	✓				+	+				10					1.1				
	ACES [79]	2018	USENIX	S	B	v7	✓					✓	✓				+					70				13					
	MINION [99]	2018	NDSS	S	R	v7	✓					✓	✓				+					-71.3	-98.86			6.13					
	OPEC [100]	2022	EuroSys	S	B	v7	✓					✓	✓				+					1.79	5.53			.23					
	EC [101]	2023	S&P	S	B/R	v7	✓					✓	✓				+		+				2.57								
	CRT-C [102]	2023	S&P	S	R	v7						✓	✓										1.75				2.63				
	D-Box [103]	2022	NDSS	S	R	v7	✓					✓	✓				+					-12	-.07	-18.2			2				
	Hermes [64]	2018	MCSA	S	B/R	v7	✓					✓	✓				+										.01				
	MultiZone [65]	2020	EW	B	B	v7	✓					✓	✓				+										.6				
	ILTZVisor [66,67]	2018	RTAS	S	B/R	v8	✓					✓	✓					+	+												
D05	SBIs [68]	2022	RTAS	S	B/R	v8	✓					✓	✓				+	+													
	RT-TEE [104]	2022	S&P	S	R	v8	✓	✓				✓	✓				+	+													
D06	SafeTEE [105]	2022	DATE	S	R	v8	✓	✓				✓	✓				+	+													2.5
	uTango [70]	2022	Access	B	B/R	v8	✓	✓				✓	✓				+	+				4.6									.05
D07	CaRE [106]	2017	RAID	B	B	v8						✓	✓				+	+				14.5				369					513
	Silhouette [107]	2020	USENIX	S	B	v7						✓	✓				+	+				8.9						3.4		14.14	
	TZmCFI [108]	2020	IJPP	S	R	v8						✓	✓				+	+									84				
	Kage [109]	2022	USENIX	S	R	v7						✓	✓				+	+				49.8					5.2				
	SUM [110]	2023	C&S	S	B	v7						✓	✓				+						8.33				2.77	2.63			
D08	SHERLOC [111]	2023	CCS	S	B/R	v8						✓	✓				+	+	+								123	1106			
	μRAI [112]	2020	NDSS	S	B	v7						✓	✓				+					54.1	15.2				.1			8.1	
	RIO [113]	2023	Access	S	B	v8						✓	✓				+	+				29.9				16.83					
	Randezvous [114]	2022	EuroS&P	S	B	v7/8			✓			✓	✓				+	+				13.6	24.5			0.6		6.9		7.0	
D09	HARM [115]	2022	EuroS&P	B	B/R	v8						✓	✓				+	+				15.49				5.8	21	28			
	fASLR [116,117]	2022	ESORICS	S	B	v8						✓	✓				+	+					4.73			9.65					
D10	Pip-MPU [118]	2023	IJESA	S	B/R	v7/8						✓	✓				+														
D11	uXOM [119]	2019	USENIX	S	R	v7	✓					✓	✓				+	+				15.7			7.5			7.3			
	PicoXOM [120]	2020	SecDev	S	B/R	v7	✓					✓	✓				+	+				5.89				.02		0.46			-11
D12	Tock [121]	2017	SOSP	S	R	v7	✓					✓	✓				+														
	DIAT [122]	2019	NDSS	S	R	v7	✓					✓	✓				+										400				
	LAPE [123]	2020	HPCC	S	B	v7						✓	✓				+					38	8.2			2.2					
	ISC-FLAT [124]	2023	RTAS	S	B	v8	✓					✓	✓				+	+							17.5		35.1				
	ARI [125]	2023	USENIX	S	R	v7	✓					✓	✓				+	+								10.7					
D13	ASSURED [126]	2018	TCAD	B	R	v8						✓	✓	✓			+					80									
	DisPatch [127]	2022	MobiSys	B	R	v7						✓	✓	✓								.53					1.48				
	Shimware [128]	2023	RAID	B	B/R	v7						✓	✓	✓																	
D14	HERA [129]	2021	NDSS	S	R	v7						✓	✓	✓																	
	RapidPatch [130]	2022	USENIX	S	B/R	v7						✓	✓	✓				+									1.5				

v7: Armv7-M, v8: Armv8-M. ✓: Implemented defense techniques to address at least one issue or overcome one or more limitations in the corresponding category. +: Need specific hardware support. -: Not applicable. ↓ and ↓ represent small and big steps towards a similar goal, respectively.

6.3 Virtualization and Multi-world Systems

Solutions in this category enable or secure *multiple* bare-metal systems and RTOSs to run in an isolated fashion on *one* MCU.

D05. Virtualization: This technique can be used to support privilege separation (see [108]).

Software-based virtualization: In those solutions, bare-metal systems and RTOSs execute at the unprivileged level

and the exokernel or an exception handler runs at the privileged level, as shown in Figure 2(c.1). A challenge is that the MSR and MRS (move to general-purpose registers from system registers) instructions fail silently without triggering any exceptions when executing at the unprivileged level, which can be addressed by replacing them with undefined instructions. Examples are Hermes [64] and MultiZone [65].

TrustZone-based virtualization: As shown in Figure 2(c.2), the exokernel or hypervisor runs at the highest privilege level (privileged secure state), and bare-metal systems and RTOSs can execute at the other three privilege levels. Prominent examples include ILTZVisor [66, 67] and SBIs [68].

D06. Multi-world systems: Multiple isolation environments enhance the isolation between system components.

Real-time and secure TrustZone-assisted dual-world system: De facto Cortex-M TEE solutions, e.g., TF-M [69], have availability and security issues, e.g., CVE-2021-32032. To address these issues, RT-TEE [104] ensures the real-time availability of both computation and I/O by adopting a policy-based event-driven hierarchical scheduler. SafeTEE [105] targets multi-core Cortex-M devices and isolates applications by assigning cores exclusively to them.

TrustZone-assisted multi-world system: As shown in Figure 2(e), TrustZone-assisted multi-world systems create multiple secure execution environments within the non-secure state to overcome L06. The uTango [70] kernel runs in the secure state at the privileged level, while other applications, services, and OSs are isolated in their non-secure state domains. Each domain has its own SAU configuration, which is only accessible by the uTango kernel.

6.4 Defeating Memory Corruption Attacks

The quest to defeat memory corruption attacks on Cortex-M systems (I15 - I19) largely includes adapting the security solutions for microprocessor-based systems to the resource and power constraint platforms.

D07. Stack and return address integrity: Stack and return addresses are a major attack vector (I10 and I11). Besides stack canaries (I13), there have been many attempts to maintain stack integrity on Cortex-M.

SafeStack: SafeStack [134] keeps unsafe local variables in a separate unsafe stack while keeping the return address in the regular stack. EPOXY implements an adapted SafeStack by (i) putting the unsafe stack on top of the RAM, (ii) making the stack grow up, and (iii) placing a region guard between the unsafe stack and other memory regions.

Shadow stack: Shadow stack [135] stores protected copies of return addresses. CaRE [106] and TZmCFI [108] use TrustZone-M and place the shadow stack in the secure state. To achieve low overhead, Silhouette [107] and Kage [109] restrict the writes to the shadow stack by transforming regular store instructions to unprivileged ones (STR*T). SUM [110] restricts unauthorized access to the shadow stack via the MPU.

Return address integrity: μ RAI [112] enforces the property of return address integrity by removing the need to spill return addresses to the stack. Rio [113] encrypts all return instructions in the firmware and instruments a runtime module to decrypt and execute these instructions. SHERLOC [111] introduces a reconstructed call stack (RCS) approach to ensure the matching of function calls and returns.

ROP gadget removal: Thumb-2 instruction set [136] allows the creation of ROP gadgets by jumping into the middle of a 32-bit instruction. To replace exploitable instructions, uSFI [98] and uXOM [119] convert all 32-bit instructions to equivalent 16-bit instruction sequences.

Stack sealing: To secure the secure world stack exception frame (I11), Arm recommends adding an integrity signature to the bottom of the secure exception stack frame [137].

D08. Forward-edge control-flow integrity (CFI): TZmCFI adopts LLVM's forward-edge CFI [138]. CaRE calculates the absolute target addresses, stores them in a branch table, and replaces all indirect branches with SVC instructions for run-time checking. Silhouette and Kage insert fixed CFI labels at the beginning of every address-taken function and check the label before the jump or the function call executes. SHERLOC maintains an indirect branch table to constrain the forward target within a predetermined CFG. InsectACIDE [170] retrieves a set of offline-computed legitimate transfer targets to validate the forward-edge transfers.

D09. Compiler-based software diversity: This technique randomizes the code and data of programs [139] to offer weakened probabilistic protection from code reuse attacks and data corruption attacks. However, the system memory layout remains the same after compilation. For instance, EPOXY [77] and Rendezvous [114] randomize the function order and add dummy variables to the .data and .bss regions.

D10. Address space layout randomization (ASLR): Without an MMU (L01) and the dynamic loading of programs, an ASLR solution on Cortex-M needs to increase entropy and decide when to perform the randomization. Both HARM [115] and fASLR [116] copy code from flash to RAM for execution and conduct randomization at the function level to increase entropy. HARM triggers randomization periodically by SysTick exceptions, while fASLR copies the function to a random location of RAM when it is called for the first time.

D11. Formal verification: Pip-MPU [118] introduces a formally verified kernel for Cortex-M. It features user-defined, MPU-guarded multiple isolation levels and is a refactored version of the MMU-based Pip protokernel [140]. It disables exceptions and puts the kernel inside the privileged level.

6.5 Defeating Software-based Code Disclosure

Projects in this category explore software-based XOM. Note that these efforts cannot address I07, in which a hardware debugger can disclose the contents in memory.

D12. Software-based XOM: uXOM [119] converts memory access instructions, excluding those that need privilege, into unprivileged ones (STR*T/LDR*T) and sets the code region as privileged access only. For the instructions that are not converted, uXOM instruments verification before them. PicoXOM [120] implements XOM by utilizing the address range matching feature of DWT with a much lower overhead.

The DWT, however, only has up to four comparators, which limits the number of configurable XOM regions.

Addressing Software Implementation Issues

6.6 Memory-safe Programming

Developing software in a manner that inherently reduces the likelihood of bugs and errors, thereby enhancing the overall safety and reliability of the system (I15 - I23).

D13. Secure multiprogramming with memory-safe languages: Tock [121] takes advantage of MPU and the type-safety features of Rust to build a multiprogramming system on Cortex-M. Rust encapsulates a large fraction of the Tock kernel with granular and type-safe interfaces.

6.7 Remote Attestation

Compared to the attack mitigation discussed in §6.4, remote attestation only detects adversarial presence.

D14. Software-based control-flow and data integrity attestation: Control-flow attestation (CFA) extends static attestation of code to run-time control-flow paths. DIAT [122] provides data integrity attestation and CFA of the code that generates and processes the data. LAPE [123] provides a coarse-grained CFA by grouping functions into compartments and attests the inter-compartment control-flow transfers. ISC-FLAT [124] extends the aforementioned approaches to support interrupts, and ARI [125] formulates the property of real-time mission execution integrity.

Addressing Other Issues

6.8 Firmware Update

D15. Secure software update: ASSURED [126] allows a device to authenticate the source of firmware updates. DisPatch [127] allows end users to write patches in a domain-specific language, which DisPatch then automatically injects into the binary firmware. Shimware [128] investigates the challenges of updating monolithic firmware images with new security features. It automates finding safe injection locations and implementing self-checks to prevent modifications.

D16. Firmware hotpatching: While updating the whole firmware requires interrupting its normal execution (D15), hotpatching can fix minor issues at run-time. HERA [129] uses flash patch and breakpoint (FPB) to insert hardware breakpoints and redirects the instructions at breakpoints to the patch codes on RAM. However, FPB is only supported on M3 and M4 MCUs. To address this issue, RapidPatch [130] utilizes other hardware mechanisms, e.g., DWT.

6.9 Vulnerability Discovery

D17. Full firmware rehosting: One main challenge in emulating firmware on a desktop is how to model peripherals.

P²IM [141] observes the MMIO access pattern of each peripheral during firmware emulation. DICE [142] improves P²IM by additionally modeling DMA. Symbolic execution that models the return value of an MMIO read as a symbolic value has also been used in firmware emulation. Examples include Laelaps [143], μ Emu [144], Jetset [145], and Fuz-zware [21]. SEmu [146] extracts the condition-action rules to dynamically synthesize peripheral models. To sidestep the challenges in peripheral modeling, HALucinator [147] detects and replaces hardware abstraction layer functions of major chip vendors with host implementations. SAFIREFUZZ [148] executes embedded firmware as a Linux userspace process on systems sharing the same instruction set family as the targeted device. HOEDUR [149] employs multi-stream inputs, restructuring the traditional approach of firmware fuzzing into multiple, strictly typed, and cohesive streams, thereby enhancing mutation effectiveness and coverage.

D18. Hardware-in-the-loop rehosting: Full firmware rehosting techniques cannot accurately model more complex peripherals, such as the USB. Hardware-in-the-loop approaches address this challenge by redirecting I/O interactions to the physical hardware. The pioneer in this direction is Avatar [150], which is followed by its variants [151–154]. Instead of redirecting I/O interactions, Frankenstein [155] directly uses dumped firmware images from real devices to re-establish emulator states.

D19. On-device fuzzing: Existing rehosting solutions fall short in testing low-level drivers, either because they cannot provide the needed emulation fidelity or completely sidestep driver emulation. μ AFL [156] supports on-device fuzzing with the help of a debug dongle and ETM. Moreover, over-the-air fuzzing has been explored to find bugs in Bluetooth controllers [157, 158]. Lastly, to make bugs observable during fuzzing, μ SBS uses binary rewriting to instrument the firmware for sanitization checks [159]. SyzTrust [160] combines ETM for direct fuzzing on IoT devices with non-invasive state and code coverage tracking.

D20. Static methods: Static methods are typically geared toward detecting a particular type of bug. For instance, PASAN [161] considers concurrency issues with peripheral access. FirmXRay [15] aims to detect Bluetooth link layer vulnerabilities from bare-metal firmware. HEAPSTER [16] inspects common classes of heap vulnerabilities in Cortex-M monolithic firmware images.

6.10 Other research

Solutions and ideas for other architectures may be ported to or optimized for Cortex-M with proper modifications. For instance, the ideas of control-flow attestation (C-FLAT [162]) and operation execution integrity (OAT [163]) apply to Cortex-M naturally but were only implemented on Cortex-A. In addition, on Arm Cortex-A, pointer authentication code (PAC) has been utilized to enforce spatial (e.g., return ad-

dresses [164] and all pointers [165]) and temporal [166, 167] memory safety on userspace programs and the kernel [168].

7 Recommendations and Future Directions

7.1 Recommendations to research community

R01. Explore the pros and cons of new hardware features for security: The hardware features of Cortex-M exhibit streamlining and differences from its Cortex-A counterparts. This distinction spans from the microarchitectural layer to the ISA. For instance, TrustZone-M is a streamlined version of TrustZone-A, and the key management for PAC [14] on Cortex-M significantly differs from that on Cortex-A. All of these differences pose new challenges and opportunities in discovering their limitations and utilizing them for protections that were not possible before.

R02. Explore diverse IoT attack models and scenarios to identify new research problems and challenges: The application scenarios of Cortex-M systems, e.g., (i) deployed in the field and (ii) functionality implemented in privileged mode, present unique trust models and security research opportunities, which must be addressed with extra consideration for performance, memory, and energy cost [169, 170]. Future research should not only port the same defenses from microprocessor systems to Cortex-M systems but also address the challenges specific to MCUs.

R03. Investigate how to facilitate the practical adoption of academic research results: Compared to security research on Cortex-M, its deployment significantly lags behind. Operational research may focus on bridging the gap between security research outcomes and practical implementation. Such research may involve how to foster collaborations between researchers and industry practitioners, how to advocate for best practices, and how to promote educational programs to raise awareness about the importance of timely security deployment in Cortex-M systems.

7.2 Recommendations to developers

R04. Securing the network communications: As discussed in section §5, network protocol implementations often expose many vulnerabilities including validation and functional bugs. This is because these protocols are designed to work with microcontroller- and microprocessor-based systems, where developers may prioritize functionalities rather than security. Microprocessor-based systems have advanced security mechanisms like ASLR and DEP, which can handle most security issues. However, employing vulnerable protocols on microcontroller-based systems can lead to severe problems. Thus, microcontroller system developers should pay extra attention to security improvements, such as validating the input and output, utilizing security mechanisms discussed in section §6, and assessing the security of protocols before using them.

R05. Implement privilege separation or employ RTOSs with distinct privilege levels: We have observed that numerous real-world firmware was built upon vendor-supplied project templates, lacking privilege separation. We strongly recommend developers opt for templates incorporating essential security features or, alternatively, adopt RTOSs with different privilege levels as the foundational framework for their development.

R06. (Partially) Transition into memory-safe languages: A full transition into memory-safe languages, e.g., Rust, may not be immediately feasible for all Cortex-M projects due to factors like existing codebase, expertise, and project timelines [171]. Partial adoption of memory-safe languages, which provides a pragmatic and manageable approach toward embracing memory-safe languages' advantages within existing projects, can be highly valuable for enhancing the overall system robustness by mitigating memory-related issues like buffer overflows and null pointer dereferences.

R07. Enhance the synergy between developers and the security research community: During our efforts to systematize security research, we noticed that some issues lack corresponding defense mechanisms (Figure 3). This could be due to incomplete publication collections, as we primarily focused on security conferences. Nonetheless, similar to the varying levels of collaboration observed between the hacker community and academia [172], if developers and the security research community unite to share findings and insights, the security of microcontroller-based systems may be significantly improved.

8 Conclusion

We present a comprehensive systematization study of the hardware and software security of Cortex-M systems. It covers the Cortex-M hardware architectures, security-related features, limitations, and issues. The study includes by far the largest empirical analysis of real-world Cortex-M firmware, characterization of reported software bugs, and an overview of state-of-the-art security research in this area. Based on the insights, we develop a set of recommendations for the research community and MCU software developers.

Acknowledgment

This material is based upon work supported in part by National Science Foundation (NSF) grants (2237238, 2329704, 2207202, 2238264), a National Centers of Academic Excellence in Cybersecurity grant (H98230-22-1-0307), FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope UIDB/00319/2020, and a Cisco University Research Program Fund (71858473). Any opinions and findings expressed in this material are those of the author(s) and do not necessarily reflect the views of United States Government or any agency thereof.

References

- [1] Arm, “Arm Partner Ecosystem Catalog,” [https://www.arm.com/partners/catalog/results#sort=date%20descending&f:armip=\[Cortex-M\]](https://www.arm.com/partners/catalog/results#sort=date%20descending&f:armip=[Cortex-M]).
- [2] Arm, “The Arm ecosystem ships a record 6.7 billion Arm-based chips in a single quarter,” <https://www.arm.com/company/news/2021/02/arm-ecosystem-ships-record-6-billion-arm-based-chips-in-a-single-quarter>.
- [3] —, “Arm Partners Have Shipped 200 Billion Chips,” <https://www.arm.com/blogs/blueprint/200bn-arm-chips>.
- [4] D. Cerdeira, N. Santos, P. Fonseca, and S. Pinto, “Sok: Understanding the prevailing security vulnerabilities in trustzone-assisted tee systems,” in *IEEE Symposium on Security and Privacy (S&P)*, 2020.
- [5] Arm, “Armv6-M Architecture Reference Manual,” <https://developer.arm.com/documentation/ddi0419/c/Application-Level-Architecture/The-Armv6-M-Instruction-Set/About-the-instruction-set>.
- [6] —, “Armv7-M Architecture Reference Manual,” <https://developer.arm.com/documentation/ddi0403/ed>.
- [7] —, “Armv8-M Architecture Reference Manual,” https://developer.arm.com/documentation/ddi0553/latest?_ga=2.1957362.2138159006.1623856318-792272022.1611588763.
- [8] J. Yiu, “ARMv8-M architecture technical overview,” *ARM white paper*, 2015.
- [9] Arm, “Armv8-M Memory Protection Unit,” <https://developer.arm.com/documentation/100699/0100>.
- [10] —, “Cortex-M23 Technical Reference Manual,” <https://developer.arm.com/documentation/ddi0550/>.
- [11] —, “Cortex-M33 Technical Reference Manual,” <https://developer.arm.com/documentation/100230/>.
- [12] —, “Cortex-M55 Technical Reference Manual,” <https://developer.arm.com/documentation/101051/>.
- [13] —, “TrustZone technology for the Armv8-M architecture Version 2.1,” <https://developer.arm.com/documentation/100690/latest/>.
- [14] —, “Armv8.1-M Pointer Authentication and Branch Target Identification Extension,” <https://community.arm.com/developer/ip-products/processors/b/processor-s-ip-blog/posts/armv8-1-m-pointer-authentication-and-branch-target-identification-extension>.
- [15] H. Wen, Z. Lin, and Y. Zhang, “FirmXRay: Detecting Bluetooth Link Layer Vulnerabilities From Bare-Metal Firmware,” in *ACM Conference on Computer and Communications Security (CCS)*, 2020.
- [16] F. Gritti, F. Pagani, I. Grishchenko, L. Dresel, N. Redini, C. Kruegel, and G. Vigna, “HEAPSTER: Analyzing the Security of Dynamic Allocators for Monolithic Firmware Images,” in *IEEE Symposium on Security and Privacy (S&P)*, 2022.
- [17] “ucsb-seclab/monolithic-firmware-collection,” <https://github.com/ucsb-seclab/monolithic-firmware-collection>.
- [18] “ThePBone/GalaxyBudsFirmwareDownloader,” https://github.com/ThePBone/GalaxyBudsFirmwareDownloader/tree/master/firmware_archive.
- [19] “grant-h/ShannonBaseband,” <https://github.com/grant-h/ShannonBaseband/tree/master/firmware>.
- [20] J. Friebertshäuser, F. Kosterhon, J. Classen, and M. Hollick, “Polypyus—the firmware historian,” in *Workshop on Binary Analysis Research (BAR)*, vol. 2021, 2021, p. 21.
- [21] T. Scharnowski, N. Bars, M. Schloegel, E. Gustafson, M. Muench, G. Vigna, C. Kruegel, T. Holz, and A. Abbasi, “Fuzzware: Using Precise MMIO Modeling for Effective Firmware Fuzzing,” in *USENIX Security Symposium*, 2022.
- [22] “Nordic semiconductor,” <https://www.nordicsemi.com/>.
- [23] “Texas instruments,” <https://www.ti.com/>.
- [24] “STMicroelectronics,” https://www.st.com/content/zt_com/en.html.
- [25] “Telink Semiconductor,” <https://www.telink-semi.com/>.
- [26] “Dialog Semiconductor,” <https://www.dialog-semiconductor.com/>.
- [27] “NXP Semiconductors,” <https://www.nxp.com/>.
- [28] “Cypress Semiconductor,” <https://www.infineon.com/>.
- [29] “Ghidra,” <https://ghidra-sre.org/>.
- [30] T. Bao, J. Burket, M. Woo, R. Turner, and D. Brumley, “ByteWeight: Learning to recognize functions in binary code,” in *USENIX Security Symposium*, 2014.
- [31] Arm, “ARM CMSIS RTOS2,” https://github.com/ARM-software/CMSIS_5/blob/2ccc9e92637fe80f50d5e8b9d503bb715112fe69/CMSIS/RTOS2/RTX/RTX5.scvd.

- [32] R. Yu, F. Del Nin, Y. Zhang, S. Huang, P. Kaliyar, S. Zarko, M. Conti, G. Portokalidis, and J. Xu, "Building Embedded Systems Like It's 1996," in *Network and Distributed System Security Symposium (NDSS)*, 2022.
- [33] MITRE, "CVE database," <https://cve.mitre.org/>.
- [34] "MCU Market Size In 2022 By Fastest Growing Companies," <https://www.marketwatch.com/press-release/iot-microcontroller-mcu-market-size-2022-industry-analysis-by-growth-share-trends-demand-segment-s-opportunities-and-forecast-2028-2022-09-19>.
- [35] Market Growth Reports, "United States IoT Operating Systems Market Report & Forecast 2021-2027," <https://www.marketgrowthreports.com/united-states-iot-operating-systems-market-19250528>.
- [36] Arm, "Mbed OS TLS," <https://tls.mbed.org/>.
- [37] wolfSSL, "wolfSSL," <https://www.wolfssl.com/>.
- [38] NXP Semiconductors, "i.MX RT Crossover MCUs," <https://www.nxp.com/products/processors-and-microcontrollers/arm-microcontrollers/i-mx-rt-crossover-mcus:IMX-RT-SERIES>.
- [39] —, "MCUXpresso SDK API Reference Manual," https://mcuxpresso.nxp.com/api_doc/dev/1411/a00057.html.
- [40] J. Y. Afonso Santos, "SAU, IDAU, MPC and PPC. What's the difference?" <https://community.arm.com/support-forums/f/architectures-and-processors-forum/12065/sau-idauc-mpc-and-ppc-what-s-the-difference/34873>.
- [41] J. A. Halderman, S. D. Schoen, N. Heninger, W. Clarkson, W. Paul, J. A. Calandrino, A. J. Feldman, J. Appelbaum, and E. W. Felten, "Lest we remember: cold-boot attacks on encryption keys," *Communications of the ACM*, 2009.
- [42] Linaro, "Trusted Firmware M (TFM) v1.3.0 source code," <https://git.trustedfirmware.org/TF-M/trusted-firmware-m.git/tag/?h=TF-Mv1.3.0>.
- [43] "Arm Platform Security Architecture Security Model," https://armkeil.blob.core.windows.net/developer/Files/pdf/PlatformSecurityArchitecture/Architect/DEN0079-PSA_SM_ALPHA-02.pdf.
- [44] "PSA Attestation API," https://armkeil.blob.core.windows.net/developer/Files/pdf/PlatformSecurityArchitecture/Implement/IHI0085-PSA_Attestation_API-1.0.1-2.pdf.
- [45] D. McCann, C. Whittall, and E. Oswald, "ELMO: Emulating Leaks for the Arm Cortex-M0 without Access to a Side Channel Lab," *IACR Cryptol. ePrint Arch.*, 2016.
- [46] S. Vafa, M. Masoumi, and A. Amini, "An efficient profiling attack to real codes of PIC16F690 and Arm Cortex-M3," *IEEE Access*, 2020.
- [47] C. Rodrigues, D. Oliveira, and S. Pinto, "BUSTed!!! Microarchitectural Side-Channel Attacks on the MCU Bus Interconnect," in *IEEE Symposium on Security and Privacy (S&P)*, 2023.
- [48] J. Mahmood and M. Hicks, "UnTrustZone: Systematic Accelerated Aging to Expose On-chip Secrets," in *IEEE Symposium on Security and Privacy (S&P)*, 2023.
- [49] J. Obermaier and S. Tatschner, "Shedding too much Light on a Microcontroller's Firmware Protection," in *USENIX Workshop on Offensive Technologies (WOOT)*, 2017.
- [50] J. Obermaier, M. Schink, and K. Moczek, "One exploit to rule them all? on the security of drop-in replacement and counterfeit microcontrollers," in *USENIX Workshop on Offensive Technologies (WOOT)*, 2020.
- [51] M. Schink, A. Wagner, F. Unterstein, and J. Heyszl, "Security and Trust in Open Source Security Tokens," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2021.
- [52] X. M. Saß, R. Mitev, and A.-R. Sadeghi, "Oops..! I Glitched It Again! How to Multi-Glitch the Glitching-Protections on ARM TrustZone-M," *USENIX Security*, 2023.
- [53] Z. Ma, X. Tan, L. Ziarek, N. Zhang, H. Hu, and Z. Zhao, "Return-to-Non-Secure Vulnerabilities on ARM Cortex-M TrustZone: Attack and Defense," in *ACM/IEEE Design Automation Conference*, 2023.
- [54] Z. Ning and F. Zhang, "Understanding the Security of Arm Debugging Features," in *IEEE Symposium on Security and Privacy (S&P)*, 2019.
- [55] Z. Ning, C. Wang, Y. Chen, F. Zhang, and J. Cao, "Revisiting arm debugging features: Nailgun and its defense," *Transactions on Dependable and Secure Computing (TDSC)*, 2021.
- [56] Sultan Qasim Khan, "Whitepaper: Microcontroller Readback Protection: Bypasses and Defenses," *Technical Report*, 2020.

- [57] Kris Brosch, "Firmware dumping technique for an Arm Cortex-M0 SoC," <https://blog.includesecurity.com/2015/11/firmware-dumping-technique-for-an-arm-cortex-m0-soc/>.
- [58] Nordic Semiconductor, "nRF52832 Objective Product Specification," https://infocenter.nordicsemi.com/pdf/nRF52832_OPS_v0.6.3.pdf.
- [59] STMicroelectronics, "Proprietary code read-out protection on microcontrollers of the STM32F4 Series," https://www.st.com/resource/en/application_note/an4701-proprietary-code-readout-protection-on-microcontrollers-of-the-stm32f4-series-stmicroelectronics.pdf.
- [60] NXP Semiconductors, "Using the Kinetis Flash Execute-Only Access Control Feature," <https://www.nxp.com/docs/en/application-note/AN5112.pdf>.
- [61] Texas Instruments, "Tiv TM4C123GH6PM Microcontroller," <https://www.ti.com/lit/ds/symlink/tm4c123gh6pm.pdf>.
- [62] M. Schink and J. Obermaier, "Taking a Look into Execute-Only Memory," in *Workshop on Offensive Technologies (WOOT)*, 2019.
- [63] "Mbed OS," <https://os.mbed.com/mbed-os/>.
- [64] N. Klingensmith and S. Banerjee, "Hermes: A real time hypervisor for mobile and iot systems," in *International Workshop on Mobile Computing Systems & Applications*, 2018.
- [65] S. Pinto and C. Garlati, "Multi zone security for arm cortex-m devices," in *Embedded World Conference*, 2020.
- [66] H. M. E. Araújo, "ILTZVisor: a lightweight TrustZone-assisted hypervisor for low-end Arm devices," Ph.D. dissertation, University of Minho, 2018.
- [67] S. Pinto, H. Araujo, D. Oliveira, J. Martins, and A. Tavares, "Virtualization on trustzone-enabled microcontrollers? voilà!" in *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2019.
- [68] R. Pan and G. Parmer, "SBIs: Application Access to Safe, Baremetal Interrupt Latencies," in *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2022.
- [69] "Trusted Firmware-M," <https://www.trustedfirmware.org/projects/tf-m>.
- [70] D. Oliveira, T. Gomes, and S. Pinto, "uTango: an open-source TEE for IoT devices," *IEEE Access*, 2022.
- [71] Nordic Semiconductor, "SoftDevices," https://infocenter.nordicsemi.com/topic/ug_gsg_ses/UG/gsg/softdevices.html.
- [72] FreeRTOS, "RTOS Fundamentals - Context Switching," <https://www.freertos.org/implementation/a00006.html>.
- [73] Zephyr Project Documentation, "Arm Cortex-M Developer Guide - Thread context switching," https://docs.zephyrproject.org/3.0.0/guides/arch/arm_cortex_m.html#thread-context-switching.
- [74] "The FreeRTOS Kernel," <https://www.freertos.org/RTOS.html>.
- [75] Arm, "API and RTX Reference Implementation - Configure RTX v5," https://www.keil.com/pack/doc/CMSIS/RTOS2/html/config_rtx5.html.
- [76] Zephyr Project Documentation, "Arm Cortex-M Developer Guide - Memory protection features," https://docs.zephyrproject.org/3.0.0/guides/arch/arm_cortex_m.html#memory-protection-features.
- [77] A. A. Clements, N. S. Almakhdhub, K. S. Saab, P. Srivastava, J. Koo, S. Bagchi, and M. Payer, "Protecting bare-metal embedded systems with privilege overlays," in *IEEE Symposium on Security and Privacy (S&P)*, 2017.
- [78] W. Zhou, L. Guan, P. Liu, and Y. Zhang, "Good Motive but Bad Design: Why Arm MPU Has Become an Outcast in Embedded Systems," *arXiv preprint arXiv:1908.03638*, 2019.
- [79] A. A. Clements, N. S. Almakhdhub, S. Bagchi, and M. Payer, "ACES: Automatic Compartments for Embedded Systems," in *USENIX Security Symposium*, 2018.
- [80] Nordic Semiconductor, "nRF51 Series Reference Manual," https://infocenter.nordicsemi.com/pdf/nRF51_RM_v3.0.pdf.
- [81] X. Tan, S. Mohan, M. Armanuzzaman, Z. Ma, G. Liu, A. Eastman, H. Hu, and Z. Zhao, "Is the Canary Dead? On the Effectiveness of Stack Canaries on Microcontroller Systems," in *ACM/SIGAPP Symposium On Applied Computing (SAC)*, 2024.
- [82] "Arm Cortex-M Programming Guide to Memory Barrier Instructions," <https://developer.arm.com/documentation/dai0321/latest/>.
- [83] M. Lipp, M. Schwarz, D. Gruss, T. Prescher, W. Haas, A. Fogh, J. Horn, S. Mangard, P. Kocher, D. Genkin *et al.*, "Meltdown: Reading kernel memory from user space," in *USENIX Security Symposium*, 2018.

- [84] P. Kocher, J. Horn, A. Fogh, , D. Genkin, D. Gruss, W. Haas, M. Hamburg, M. Lipp, S. Mangard, T. Prescher, M. Schwarz, and Y. Yarom, "Spectre Attacks: Exploiting Speculative Execution," in *IEEE Symposium on Security and Privacy (S&P)*, 2019.
- [85] J. Ravichandran, W. T. Na, J. Lang, and M. Yan, "PACMAN: attacking Arm pointer authentication with speculative execution," in *International Symposium on Computer Architecture (ISCA)*, 2022.
- [86] METRE, "Common Vulnerability Scoring System v3.1: User Guide," <https://www.first.org/cvss/v3.1/user-guide>.
- [87] "Multiple RTOS (Update E) | CISA," <https://www.cisa.gov/uscert/ics/advisories/icsa-21-119-04>.
- [88] D. Bleichenbacher, "Chosen ciphertext attacks against protocols based on the RSA encryption standard PKCS# 1," in *Annual International Cryptology Conference*. Springer, 1998.
- [89] Embedded by AspenCore, "2019 embedded markets study," https://www.embedded.com/wp-content/uploads/2019/11/EETimes_Embedded_2019_Embedded_Markets_Study.pdf.
- [90] J. Pallister, S. Hollis, and J. Bennett, "BEEBS: open benchmarks for energy measurements on embedded platforms," *arXiv preprint arXiv:1308.5174*, 2013.
- [91] "CoreMark," <https://www.eembc.org/coremark>.
- [92] "CoreMark-Pro," <https://www.eembc.org/coremark-pro/>.
- [93] R. P. Weicker, "Dhrystone: a synthetic systems programming benchmark," *Communications of the ACM*, 1984.
- [94] "Embench: A Modern Embedded Benchmark Suite," <https://www.embench.org/>.
- [95] A. K. Iannillo, S. Rivera, D. Suciu, R. Sion, and R. State, "An REE-independent Approach to Identify Callers of TEEs in TrustZone-enabled Cortex-M Devices," in *ACM Cyber-Physical System Security Workshop (CPSS)*, 2022.
- [96] A. Khurshid, S. D. Yalaw, M. Aslam, and S. Raza, "ShieLD: Shielding Cross-zone Communication within Limited-resourced IoT Devices running Vulnerable Software Stack," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2022.
- [97] D. Danner, R. Müller, W. Schröder-Preikschat, W. Hofer, and D. Lohmann, "Safer Sloth: Efficient, hardware-tailored memory protection," in *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2014.
- [98] Z. B. Aweke and T. Austin, "uSFI: Ultra-lightweight software fault isolation for IoT-class devices," in *IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2018.
- [99] C. H. Kim, T. Kim, H. Choi, Z. Gu, B. Lee, X. Zhang, and D. Xu, "Securing Real-Time Microcontroller Systems through Customized Memory View Switching," in *Network and Distributed System Security Symposium (NDSS)*, 2018.
- [100] X. Zhou, J. Li, W. Zhang, Y. Zhou, W. Shen, and K. Ren, "OPEC: operation-based security isolation for bare-metal embedded systems," in *European Conference on Computer Systems*, 2022.
- [101] A. Khan, D. Xu, and D. Tian, "Ec: Embedded systems compartmentalization via intra-kernel isolation," in *Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2023.
- [102] —, "Low-cost privilege separation with compile time compartmentalization for embedded systems," in *Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2023.
- [103] A. Mera, Y. H. Chen, R. Sun, E. Kirda, and L. Lu, "D-Box: DMA-enabled Compartmentalization for Embedded Applications," in *Network and Distributed System Security Symposium (NDSS)*, 2022.
- [104] J. Wang, A. Li, H. Li, C. Lu, and N. Zhang, "RT-TEE: Real-time System Availability for Cyber-physical Systems using Arm TrustZone," in *IEEE Symposium on Security and Privacy (SP)*, 2022.
- [105] M. Schönstedt, F. Brasser, P. Jauernig, E. Stapf, and A.-R. Sadeghi, "SafeTEE: combining safety and security on ARM-based microcontrollers," in *IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2022.
- [106] T. Nyman, J.-E. Ekberg, L. Davi, and N. Asokan, "CFI CaRE: Hardware-supported call and return enforcement for commercial microcontrollers," in *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 2017.
- [107] J. Zhou, Y. Du, Z. Shen, L. Ma, J. Criswell, and R. J. Walls, "Silhouette: Efficient protected shadow stacks for embedded systems," in *USENIX Security Symposium*, 2020.

- [108] T. Kawada, S. Honda, Y. Matsubara, and H. Takada, "TZmCFI: RTOS-Aware Control-Flow Integrity Using TrustZone for Armv8-M," *International Journal of Parallel Programming*, 2020.
- [109] Y. Du, Z. Shen, K. Dharsee, J. Zhou, R. J. Walls, and J. Criswell, "Holistic Control-Flow Protection on Real-Time Embedded Systems with Kage," in *USENIX Security Symposium*, 2022.
- [110] W. Choi, M. Seo, S. Lee, and B. B. Kang, "SuM: Efficient Shadow Stack Protection on ARM Cortex-M," *Computers & Security*, 2023.
- [111] X. Tan and Z. Zhao, "SHERLOC: Secure and Holistic Control-Flow Violation Detection on Embedded Systems," in *ACM Conference on Computer and Communications Security (CCS)*, 2023.
- [112] N. S. Almahdhub, A. A. Clements, S. Bagchi, and M. Payer, "μRAI: Securing embedded systems with return address integrity," in *Network and Distributed System Security Symposium (NDSS)*, 2020.
- [113] B. Kim, K. Lee, W. Park, J. Cho, and B. Lee, "RIO: Return Instruction Obfuscation for Bare-metal IoT Devices," *IEEE Access*.
- [114] Z. Shen, K. Dharsee, and J. Criswell, "Rendezvous: Making Randomization Effective on MCUs," in *Annual Computer Security Applications Conference (ACSAC)*, 2022.
- [115] J. Shi, L. Guan, W. Li, D. Zhang, P. Chen, and N. Zhang, "HARM: Hardware-Assisted Continuous Re-randomization for Microcontrollers," in *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2022.
- [116] L. Luo, X. Shao, Z. Ling, H. Yan, Y. Wei, and X. Fu, "fASLR: Function-Based ASLR via TrustZone-M and MPU for Resource-Constrained IoT Systems," *IEEE Internet of Things Journal*, 2022.
- [117] X. Shao, L. Luo, Z. Ling, H. Yan, Y. Wei, and X. Fu, "faslr: Function-based aslr for resource-constrained iot systems," in *European Symposium on Research in Computer Security (ESORICS)*, 2022.
- [118] N. Dejon, C. Gaber, and G. Grimaud, "Pip-MPU: Formal verification of an MPU-based separation kernel for constrained devices," *International Journal of Embedded Systems and Applications*, 2023.
- [119] D. Kwon, J. Shin, G. Kim, B. Lee, Y. Cho, and Y. Paek, "uXOM: Efficient eExecute-Only Memory on Arm Cortex-M," in *USENIX Security Symposium*, 2019.
- [120] Z. Shen, K. Dharsee, and J. Criswell, "Fast Execute-Only Memory for Embedded Systems," in *IEEE Secure Development (SecDev)*, 2020.
- [121] A. Levy, B. Campbell, B. Ghena, D. B. Giffin, P. Panunto, P. Dutta, and P. Levis, "Multiprogramming a 64kb computer safely and efficiently," in *ACM SIGOPS symposium on Operating systems principles (SOSP)*, 2017.
- [122] T. Abera, R. Bahmani, F. Brasser, A. Ibrahim, A.-R. Sadeghi, and M. Schunter, "DIAT: Data Integrity Attestation for Resilient Collaboration of Autonomous Systems," in *Network and Distributed System Security Symposium (NDSS)*, 2019.
- [123] D. Huo, Y. Wang, C. Liu, M. Li, Y. Wang, and Z. Xu, "LAPE: A Lightweight Attestation of Program Execution Scheme for Bare-Metal Systems," in *IEEE HPC-C/SmartCity/DSS*, 2020.
- [124] A. J. Neto and I. d. O. Nunes, "ISC-FLAT: On the Conflict Between Control Flow Attestation and Real-Time Operations," 2023.
- [125] J. Wang, Y. Wang, A. Li, Y. Xiao, R. Zhang, W. Lou, Y. T. Hou, and N. Zhang, "Ari: Attestation of real-time mission execution integrity," 2023.
- [126] N. Asokan, T. Nyman, N. Rattanavipanon, A.-R. Sadeghi, and G. Tsudik, "ASSURED: Architecture for secure software update of realistic embedded devices," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, 2018.
- [127] T. Kim, A. Ding, S. Etigowni, P. Sun, J. Chen, L. Garcia, S. Zonouz, D. Xu, and D. Tian, "Reverse engineering and retrofitting robotic aerial vehicle control firmware using dispatch," in *International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2022.
- [128] E. Gustafson, P. Grosen, N. Redini, S. Jha, A. Continella, R. Wang, K. Fu, S. Rampazzi, C. Kruegel, and G. Vigna, "Shimware: Toward Practical Security Retrofitting for Monolithic Firmware Images," in *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, 2023.
- [129] C. Niesler, S. Surminski, and L. Davi, "HERA: Hot-patching of Embedded Real-time Applications," in *Network and Distributed System Security Symposium (NDSS)*, 2021.
- [130] Y. He, Z. Zou, K. Sun, Z. Liu, K. Xu, Q. Wang, C. Shen, Z. Wang, and Q. Li, "RapidPatch: Firmware Hotpatching for Real-Time Embedded Devices," in *USENIX Security Symposium*, 2022.

- [131] A. Barenghi, L. Breveglieri, I. Koren, G. Pelosi, and F. Regazzoni, “Countermeasures against fault attacks on software implemented AES: effectiveness and cost,” in *Workshop on Embedded Systems Security (WESS)*, 2010.
- [132] J. S. Jang, S. Kong, M. Kim, D. Kim, and B. B. Kang, “SeCRiT: Secure Channel between Rich Execution Environment and Trusted Execution Environment,” in *Network and Distributed System Security Symposium (NDSS)*, 2015.
- [133] A. S. Elliott, A. Ruef, M. Hicks, and D. Tarditi, “Checked C: making C safe by extension,” in *Cybersecurity Development (SecDev)*. IEEE, 2018.
- [134] P. Larsen and A.-R. Sadeghi, *The Continuing Arms Race: Code-Reuse Attacks and Defenses*. Association for Computing Machinery and Morgan & Claypool, 2018, ch. Code-pointer integrity.
- [135] N. Burow, X. Zhang, and M. Payer, “SoK: Shining light on shadow stacks,” in *IEEE Symposium on Security and Privacy (S&P)*, 2019.
- [136] Arm, “Arm Architecture Reference Manual Thumb-2 Supplement,” <https://class.ece.iastate.edu/cpre288/resources/docs/Thumb-2SupplementReferenceManual.pdf>.
- [137] —, “Armv8-M Stack Sealing Vulnerability,” <https://developer.arm.com/support/arm-security-updates/armv8-m-stack-sealing>.
- [138] C. Tice, T. Roeder, P. Collingbourne, S. Checkoway, Ú. Erlingsson, L. Lozano, and G. Pike, “Enforcing forward-edge control-flow integrity in GCC & LLVM,” in *USENIX Security Symposium*, 2014.
- [139] P. Larsen, A. Homescu, S. Brunthaler, and M. Franz, “SoK: Automated software diversity,” in *IEEE Symposium on Security and Privacy (S&P)*, 2014.
- [140] N. Jomaa, D. Nowak, and P. Torrini, “Formal Development of the Pip Protokernel,” *ENTROPY*, 2018.
- [141] B. Feng, A. Mera, and L. Lu, “P²IM: Scalable and Hardware-independent Firmware Testing via Automatic Peripheral Interface Modeling,” in *USENIX Security Symposium*, 2020.
- [142] A. Mera, B. Feng, L. Lu, E. Kirda, and W. Robertson, “DICE: Automatic Emulation of DMA Input Channels for Dynamic Firmware Analysis,” in *IEEE Symposium on Security and Privacy (S&P)*, 2021.
- [143] C. Cao, L. Guan, J. Ming, and P. Liu, “Device-agnostic firmware execution is possible: A concolic execution approach for peripheral emulation,” in *Annual Computer Security Applications Conference (ACSAC)*, 2020.
- [144] W. Zhou, L. Guan, P. Liu, and Y. Zhang, “Automatic Firmware Emulation through Invalidity-guided Knowledge Inference,” in *USENIX Security Symposium*, 2021.
- [145] E. Johnson, M. Bland, Y. Zhu, J. Mason, S. Checkoway, S. Savage, and K. Levchenko, “Jetset: Targeted Firmware Rehosting for Embedded Systems,” in *USENIX Security Symposium*, 2021.
- [146] W. Zhou, L. Zhang, L. Guan, P. Liu, and Y. Zhang, “What Your Firmware Tells You Is Not How You Should Emulate It: A Specification-Guided Approach for Firmware Emulation,” in *ACM SIGSAC Conference on Computer and Communications Security*, 2022.
- [147] A. A. Clements, E. Gustafson, T. Scharnowski, P. Grosen, D. Fritz, C. Kruegel, G. Vigna, S. Bagchi, and M. Payer, “HALucinator: Firmware Re-hosting Through Abstraction Layer Emulation,” in *USENIX Security Symposium*, 2020.
- [148] L. Seidel, D. Maier, and M. Muench, “Forming faster firmware fuzzers,” in *USENIX Conference on Security Symposium*, 2023.
- [149] T. Scharnowski, S. Wörner, F. Buchmann, N. Bars, M. Schloegel, and T. Holz, “HOEDUR: embedded firmware fuzzing using multi-stream inputs,” in *USENIX Conference on Security Symposium*, 2023.
- [150] J. Zaddach, L. Bruno, A. Francillon, D. Balzarotti *et al.*, “AVATAR: A Framework to Support Dynamic Security Analysis of Embedded Systems’ Firmwares,” in *Network and Distributed System Security (NDSS)*, 2014.
- [151] M. Muench, A. Francillon, and D. Balzarotti, “Avatar²: A Multi-target Orchestration Platform,” in *Workshop on Binary Analysis Research*, 2018.
- [152] K. Koscher, T. Kohno, and D. Molnar, “SURROGATES: Enabling Near-Real-Time Dynamic Analyses of Embedded Systems,” in *USENIX Workshop on Offensive Technologies (WOOT 15)*, 2015.
- [153] Corteggiani, Nassim and Camurati, Giovanni and Francillon, Aurélien, “Inception: System-wide security testing of real-world embedded systems software,” in *USENIX Security Symposium*, 2018.
- [154] N. Corteggiani and A. Francillon, “HardSnap: Leveraging Hardware Snapshotting for Embedded Systems Security Testing,” in *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2020.

- [155] J. Ruge, J. Classen, F. Gringoli, and M. Hollick, "Frankenstein: Advanced Wireless Fuzzing to Exploit New Bluetooth Escalation Targets," in *USENIX Security Symposium*, 2020.
- [156] W. Li, J. Shi, F. Li, J. Lin, W. Wang, and L. Guan, "μAFL: Non-intrusive Feedback-driven Fuzzing for Microcontroller Firmware," in *IEEE/ACM International Conference on Software Engineering*, 2022.
- [157] M. E. Garbelini, C. Wang, S. Chattopadhyay, S. Sumei, and E. Kurniawan, "SweynTooth: Unleashing Mayhem over Bluetooth Low Energy," in *USENIX Annual Technical Conference*, 2020.
- [158] M. E. Garbelini, V. Bedi, S. Chattopadhyay, S. Sun, and E. Kurniawan, "BRAKTOOTH: Causing Havoc on Bluetooth Link Manager via Directed Fuzzing," in *USENIX Security Symposium*, 2022.
- [159] M. Salehi, D. Hughes, and B. Crispo, "μSBS: Static binary sanitization of bare-metal embedded devices for fault observability," in *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, 2020, pp. 381–395.
- [160] Q. Wang, B. Chang, S. Ji, Y. Tian, X. Zhang, B. Zhao, G. Pan, C. Lyu, M. Payer, W. Wang *et al.*, "SyzTrust: State-aware Fuzzing on Trusted OS Designed for IoT Devices," 2023.
- [161] T. Kim, V. Kumar, J. Rhee, J. Chen, K. Kim, C. H. Kim, D. Xu, and D. J. Tian, "PASAN: Detecting Peripheral Access Concurrency Bugs within Bare-Metal Embedded Applications," in *USENIX Security Symposium*, 2021.
- [162] T. Abera, N. Asokan, L. Davi, J.-E. Ekberg, T. Nyman, A. Paverd, A.-R. Sadeghi, and G. Tsudik, "C-FLAT: control-flow attestation for embedded systems software," in *ACM Conference on Computer and Communications Security*, 2016.
- [163] Z. Sun, B. Feng, L. Lu, and S. Jha, "OAT: Attesting operation integrity of embedded devices," in *IEEE Symposium on Security and Privacy (S&P)*, 2020.
- [164] H. Liljestrand, T. Nyman, L. J. Gunn, J.-E. Ekberg, and N. Asokan, "PACStack: an Authenticated Call Stack," in *USENIX Security Symposium*, 2021.
- [165] H. Liljestrand, T. Nyman, K. Wang, C. C. Perez, J.-E. Ekberg, and N. Asokan, "PAC it up: Towards pointer integrity using ARM pointer authentication," in *USENIX Security Symposium*, 2019.
- [166] R. M. Farkhani, M. Ahmadi, and L. Lu, "PTAuth: Temporal Memory Safety via Robust Points-to Authentication," in *USENIX Security Symposium*, 2021.
- [167] Y. Li, W. Tan, Z. Lv, S. Yang, M. Payer, Y. Liu, and C. Zhang, "PACMem: Enforcing Spatial and Temporal Memory Safety via ARM Pointer Authentication," in *ACM SIGSAC Conference on Computer and Communications Security*, 2022.
- [168] S. Yoo, J. Park, S. Kim, Y. Kim, and T. Kim, "In-Kernel Control-Flow Integrity on Commodity OSes using ARM Pointer Authentication," in *USENIX Security Symposium*, 2022.
- [169] Z. Zhao, M. Armanuzzaman, X. Tan, and Z. Ma, "Trusted Execution Environments in Embedded and IoT Systems: A CactiLab Perspective," in *IEEE International Symposium on Secure and Private Execution Environment Design (SEED)*, 2024.
- [170] Y. Wang, C. Lemieux Mack, X. Tan, N. Zhang, Z. Zhao, S. Baruah, and B. C. Ward, "InsectACIDE: Debugger-Based Holistic Asynchronous CFI for Embedded System," in *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2024.
- [171] A. Sharma, S. Sharma, S. Torres-Arias, and A. Machiry, "Rust for Embedded Systems: Current State, Challenges and Open Problems," *arXiv preprint arXiv:2311.05063*, 2023.
- [172] H. Bos, "NDSS 2024 Keynote - Corruption of Memory: Those who don't know history are doomed to repeat it," <https://www.youtube.com/watch?v=vhj2We2vjqs>.

Appendix

Our open-source repository contains extra information for researchers:

- A Cortex-M firmware analysis tool (in the `firmware_analysis` folder).
- A Cortex-M firmware database (in the `firmware_analysis` folder).
- Cortex-M hardware feature test suites (in the `hw_feature_test_suites` folder).
- Supplementary Material 1: Cortex-M Architecture in a Nutshell (`Background.pdf`).
- An interactive figure showcasing the relationships between Cortex-M limitations, issues, and mitigations (download `relations_interactive_fig.html`).
- A collection of Cortex-M-related CVEs in Google Spreadsheet.