The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/204493

Please be advised that this information was generated on 2019-10-04 and may be subject to change.
INVITED: In Hardware We Trust
Gains and Pains of Hardware-assisted Security

Lejla Batina
lejla@cs.ru.nl
Radboud University, The Netherlands

Patrick Jauernig
patrick.jauernig@trust.tu-darmstadt.de
TU Darmstadt, Germany

Nele Mentens
nele.mentens@kuleuven.be
KU Leuven, Belgium

Ahmad-Reza Sadeghi
ahmad.sadeghi@trust.tu-darmstadt.de
TU Darmstadt, Germany

Emmanuel Stapf
emmanuel.stapf@trust.tu-darmstadt.de
TU Darmstadt, Germany

ABSTRACT

Data processing and communication in almost all electronic systems are based on Central Processing Units (CPUs). In order to guarantee confidentiality and integrity of the software running on a CPU, hardware-assisted security architectures are used. However, both the threat model and the non-functional platform requirements, i.e., performance and energy budget, differ when we go from high-end desktop computers and servers to low-end embedded devices that populate the internet of things (IoT). For high-end platforms, a relatively large energy budget is available to protect software against attacks. However, measures to optimize performance give rise to microarchitectural side-channel attacks. IoT devices, in contrast, are constrained in terms of energy consumption and do not incorporate the performance enhancements found in high-end CPUs. Hence, they are less likely to be susceptible to microarchitectural attacks, but give rise to physical attacks, exploiting, e.g., leakage in power consumption or through fault injection. Whereas previous work mostly concentrates on a specific architecture, this paper covers the whole spectrum of computing systems, comparing the corresponding hardware architectures, and most relevant threats.

ACM Reference Format:

1 INTRODUCTION

Software attacks can lead to unauthorized access or physical damage to infected devices and networks. To limit potential attacks to only the code containing the flaw, operating system kernels make use of process isolation. However, flaws in the kernel itself can be used to undermine process isolation. Trusted Execution Environments (TEEs) and hardware-assisted approaches try to guarantee a level of security that mitigations for vulnerable software have failed to provide. The basic assumption for hardware-assisted security mechanisms is that hardware is less likely to have software-exploitatable vulnerabilities. Building on those primitives, the complexity of the software parts in mitigations can be reduced. As a consequence, industry is advocating hardware-assisted approaches on all fronts: desktop computers (Intel SGX [16], AMD SEV [20]), smartphones (ARM TrustZone [2]), down to low-energy embedded devices (ARM TrustZone-M [43]). In parallel, academia proposed many security solutions based on those industrial hardware trust anchors. A promising approach is to manage hardware by a small software Trusted Computing Base (TCB) to create TEEs, e.g., protecting sensitive user-space code on mobile devices based on ARM TrustZone [7], or improving Intel SGX on RISC-V [11]. However, other types of attacks, besides software attacks, are exploiting the microarchitectural details of the CPU or the physical leakage from cryptographic implementations in hardware or software. Traditional physical attacks either maliciously inject faults [5] or exploit physical leakage from side channels such as timing [23], power consumption [25] or electromagnetic emanations [14] to infer secret information from cryptographic algorithms. Microarchitectural attacks are side-channel attacks that rely on the CPU’s architectural specifics [15, 17, 24, 29, 38].

In this paper, we survey different security architectures based on the gains they inherit from underlying hardware security primitives, and present pains caused by subtle state-of-the-art attacks on these security architectures that exploit microarchitectural effects of modern hardware performance enhancements.

2 ADVERSARY MODEL AND REQUIREMENTS

We rely on the classification presented in [1] and distinguish the following types of adversaries: (1) remote adversary, capable of inserting malicious software; (2) local adversary, capable of remote attacks, and controlling and eavesdropping on the communication; (3) physical adversary, capable of performing (non-intrusive physical attacks—we distinguish between microarchitectural side-channel analysis (SCA) attacks and classical physical attacks. Depending on the computing platform, the importance of each attack model differs, as illustrated in Figure 1. Remote and local attacks are applicable to all types of computing platforms. Classical physical attacks, discussed in Section 5 are not considered a main threat in servers and desktop computers, while they are prominent on IoT devices that allow potential adversaries in close proximity. Today’s microarchitectural attacks, on the other hand, are a consequence of performance enhancements in high-end CPUs and are thus mostly applicable to servers and desktop computers, as explained in Section 4.
3 HARDWARE-ASSISTED SECURITY ARCHITECTURES

We categorize existing hardware-assisted security architectures depending on the performance capabilities of the devices on which they are implemented.

3.1 Stationary High-Performance Devices

We present the most popular industrial and academic hardware-assisted security architectures, namely, Intel SGX for x86, and Sanctum for the open RISC-V architecture.

Intel SGX. SGX [10, 16] provides TEEs, also called enclaves, on recent Intel processors. An enclave is used to execute sensitive program code in user space, isolated from a potentially malicious OS or hypervisor. Each enclave is bundled with a regular non-sensitive host application which invokes the enclave. During the enclave setup, the integrity of the enclave code is attested locally or remotely using authenticated measurements of the code. The enclave memory management, exception handling and I/O is performed by the untrusted OS. However, hardware features protect the enclave memory from an unauthorized access by the untrusted OS or hypervisor. SGX’s TCB only comprises the CPU hardware and its microcode. SGX encrypts all enclave code and data leaving the CPU. This allows SGX to protect enclaves from DMA attacks and to persistently store the state of an enclave.

Sanctum. Sanctum [11] was developed for the open RISC-V architecture and resembles Intel SGX regarding its high-level concept. However, SGX’s microcode TCB is substituted by a privileged software component called the monitor code. The isolation of the enclave memory is enforced by introducing small hardware changes around the page table walker of the CPU’s Memory Management Unit (MMU). In contrast to SGX, Sanctum does not encrypt the enclave code and data in the main memory. However, Sanctum provides a basic DMA attack protection by modifying the memory controller. Unlike SGX, Sanctum implements cache partitioning in the shared last-level cache through memory page coloring. This allows Sanctum to assign cache lines exclusively to an enclave.

3.2 Mobile High-Performance Devices

ARM TrustZone. ARM TrustZone [2] is used to implement security architectures on ARM-based devices. TrustZone separates the system into two worlds, normal world and secure world. The secure world represents the single enclave (or TEE) of the system. All sensitive apps run in the secure world but are only separated by a software-based isolation. TrustZone can, in contrast to SGX and Sanctum, establish secure channels between peripherals and sensitive apps. The world switch is performed by a privileged component called the monitor code which also verifies all secure world code during boot using digital signatures. The monitor code represents the software TCB of the system, together with all code in the secure world. The separation between both worlds is enforced in hardware by a set of security enhancements of the CPU and other SoC components.

Sanctum. Sanctum [7] solves the main problem of currently deployed TrustZone-based architectures by providing an arbitrary number of user-space enclaves without introducing new hardware components. In Sanctum, sensitive apps are not included in the secure world but temporarily isolated on physical cores. The secure world only contains security primitives provided by the device vendor. Therefore, the system TCB only consists of device vendor code. No trust relationships between device vendor and app developers need to be established. The security primitives provide TrustZone functionalities to Sanctum (e.g. DMA access control or secure channels to peripherals). The isolation of the physical cores is enforced by exploiting a feature of ARM’s TrustZone-enabled address space controller. In contrast to Sanctum, Sanctuary cannot provide cache partitioning of the shared last-level cache. However, software cache side-channel attacks on sensitive apps are prevented by excluding enclave memory from the shared caches.

3.3 Embedded Devices

Embedded systems have a tight energy budget. Thus, instead of integrating fully-fledged MMUs, these systems use primitive access controllers that limit secure and trusted execution capabilities.

SMART. SMART [12] establishes a dynamic root of trust using hardware-assisted attestation. SMART supports remote attestation but not code isolation. To do so, SMART leverages read-only memory (ROM) containing the attestation code and a secure key that can only be accessed if the program counter is pointing to the ROM region. In detail, attestation is invoked by an untrusted entity that wants to attest code in regular memory. First, the attestation code disables interrupts. Then, the secret key is used to compute an attestation report containing the HMAC of the memory region, input parameters, a nonce and an after-attestation destination address. Then, the report is copied to regular memory, SMART cleans up attestation traces, and jumps to the destination address in the attested code. The attested code can enable interrupts again, execute, and send the attestation report to a verifier. Because of disabled interrupts, SMART is not suitable for real-time applications, and does not consider side-channel attacks or DMA attacks [31] in its threat model. Sancus [33] reduces SMART’s TCB to pure hardware.
TrustLite. In contrast to SMART and Sancus, TrustLite [26] provides a fully-fledged TEE for embedded devices. TrustLite leverages an (extended) execution-aware Memory Protection Unit (EA-MPU) and generalizes the concept of a read-only attestation code to freely-configurable regions, called Trustlets, protected in terms of confidentiality and integrity. First, the Secure Loader, stored in ROM, loads the Trustlets into memory and configures the EA-MPU to enforce isolation for each Trustlet’s memory. Second, EA-MPU configuration is locked, thus, protection regions are static and a cleanup as in SMART is not needed anymore. Finally, the OS is started. TrustLite provides more flexibility in protecting enclaves than SMART and Sancus. Still, side-channel and DMA attacks are not part of the attacker model. TyTAN [6], an extension of TrustLite for real-time systems, further adds secure boot and secure storage.

4 MICROARCHITECTURAL ATTACKS

In recent years, security research has shown that CPU performance optimizations like caches or transient execution can be exploited by software attackers to infer information from victim processes.

4.1 Software Cache Side-Channel Attacks

In software-based cache side-channel attacks, the attacker measures the duration of memory accesses to infer if a victim process touched certain data. This information can, e.g., be used to attack cryptographic algorithms [34, 35]. Recently, various cache side-channel attacks were presented, e.g. Evict+Time [34], Prime+Probe [34], and Flush+Reload [42]. Attacks are, however, not limited to memory caches: theoretically, any cache structure shared by the attacker and the victim can be exploited, e.g. the TLB [15] or the BTB [28]. Software countermeasures can be implemented in the algorithms that should be protected, e.g. [3, 34], or in privileged software levels that oversee the attacker and victim processes, e.g. [9, 32]. Hardware countermeasures implemented as part of a security architecture either use some sort of cache partitioning [39] or randomize the mapping from memory addresses to cache lines [40].

Software-induced cache side-channel attacks have only been considered in security architectures most recently. SGX and TrustZone do not provide cache side-channel protection on an architectural level for their enclaves [8, 44]. Sanctuary and Sanctum flush core-exclusive caches on the enclave context switches. In contrast to SGX and TrustZone, Sanctum provides partitioning for the shared last-level cache. Sanctuary relies on the TrustZone and can therefore not provide cache partitioning. However, Sanctuary protects from cache side-channel attacks by excluding the Sanctuary memory from the shared caches, while none of the presented security architectures for embedded devices even considers cache side channels.

4.2 Transient Execution Attacks

Transient instructions are used to implement CPU performance optimization techniques like out-of-order execution or branch prediction. However, recent attacks showed that there are several flaws in the design of transient execution which an attacker can exploit to alter the microarchitectural state normally invisible to the architecture. In this section, we give an overview over these attacks. Spectre. Spectre and its variants [22, 24, 27] use mistraining of branch (or return) prediction to transiently execute code even when it is guarded by an access-restricting branch while bypassing all software defenses like bounds checking or CFI. In addition, on modern CPUs, branch prediction buffers are indexed using virtual addresses of the branch instructions [21], allowing mistraining not only from the same address space, but also from different processes. Meltdown. Meltdown and its variants [22, 29, 36] focus on memory accesses instead of code execution. By exploiting the time window between the cause of an exception and its actual raise at retirement of the causing instructions, Meltdown leaks results of transient instructions that get successively executed during this window. At instruction retirement, the microarchitectural state gets cleaned up, yet, (cache) side channels can be used to extract these results. Foreshadow. Foreshadow [38] is a transient-execution attack on Intel SGX based on Meltdown’s insights. SGX is immune to a plain Meltdown attack as enclave memory usually does not raise memory access exceptions. However, as the OS is in control of all page tables, an attacker can set the present or reserved bit to force the enclave to raise a page fault (dubbed L1 Terminal Fault, L1TF [17], by Intel) [41]. Under those conditions, address translation is aborted, only cache values tagged with the corresponding physical address can be extracted this way. However, arbitrary encrypted enclave pages can be externally forced to be decrypted to the L1 cache using SGX’s secure page swapping, enabling the attacker to proceed analogously to a normal Meltdown attack.

Consequences. Most of the known transient execution attacks have been mitigated. Yet, trust has been shattered irreparably, e.g., Foreshadow was used to extract attestation keys of Intel SGX. In parallel, new transient execution attacks are emerging [4, 18], even on embedded devices [13]. For most of the hardware-assisted security mechanisms presented in this paper, an extensive evaluation of transient execution attacks has not been presented yet.

5 CLASSICAL PHYSICAL ATTACKS

Physical attacks can be categorized into non-intrusive (passive) and intrusive (active) attacks. In the former case, an adversary, e.g., monitors power consumption while a device performs secret key operations [25]. Other sources of side-channel information, such as electromagnetic (EM) emanations from a chip [14] and timings for different operations performed [23] have also shown to be effective [30]. In contrast, intrusive attacks induce faults in the system that lead to secret information being leaked in the system’s output [5]. Physical attacks are performed on both software and hardware implementations of cryptographic algorithms.

Typical countermeasures against passive SCA attacks can be divided into two categories: hiding and masking. Hiding countermeasures aim at breaking the link between the processed data and the side-channel leakage. Masking countermeasures break the link between the actual data and the processed data such that side-channel leakage does not infer the secret information. For an overview, see [30].

While SCA is a passive technique that relies on merely observing leakages, fault attacks are actively provoked and require direct access to the device. Fault attacks aim at exposing sensitive information triggered by irregular conditions that lead to faulty computations. One way to accomplish this is by “glitching” the device, i.e., forcing changes in the values of relevant physical parameters outside the specified intervals. Glitches can be induced through the clock signal, the power supply, EM pulses or optical signals. When
the glitch leads to exploitable fault computations, fault analysis can be performed. The state of the art in fault attacks and countermeasures can be found in [19]. A recent attack worth mentioning, is demonstrated on ARM TrustZone: CLKSCREW forces a processor to operate beyond its Dynamic Voltage and Frequency Scaling (DVFS) limits in order to leak cryptographic keys [37].

### 6 CONCLUSION

Learning from the mistakes of speculative execution, computing architectures should be built with security in mind—not only performance. Similar, as the current trend of shrinking the software TCB by adding hardware security extensions is continuing, careful analysis of these components and their interplay is needed more than ever. Building on this underlying TCB, security solutions like cryptographic primitives or TEEs face multi-faceted challenges in the future. Existing TEEs have unsolved issues and are still not widely-adopted, also because of licensing contracts needed to deploy TEE apps in industry solutions. In addition, mobile devices and IoT devices need to face the challenge of protecting against attackers that are in the vicinity of the devices and thus capable of performing physical attacks through side channels or fault injection. In general, it is important to select the optimal security architecture given the energy and performance budget of the application.

### Acknowledgements

This work has been supported by the projects HWSec, P3 and S7 within ERC 1119 CROSSING (German Research Foundation, DFG), the Intel ICRI-CARS institute, WearTH4EHealth (EU/Interreg/EMR) and in part by the Technology Foundation TTW (Project 13499 TYPHOON) from the Dutch government.

### REFERENCES


