Abstract—Remote code disclosure attacks threaten embedded systems as they allow attackers to steal intellectual property or to find reusable code for use in control-flow hijacking attacks. Execute-only memory (XOM) prevents remote code disclosures, but existing XOM solutions either require a memory management unit that is not available on ARM embedded systems or incur significant overhead.

We present PicoXOM: a fast and novel XOM system for ARMv7-M and ARMv8-M devices which leverages ARM's Data Watchpoint and Tracing unit along with the processor's simplified memory protection hardware. On average, PicoXOM incurs 0.33% performance overhead and 5.89% code size overhead on two benchmark suites and five real-world applications.

I. INTRODUCTION

Remote code disclosure attacks threaten computer systems. Remote attackers exploiting buffer overread vulnerabilities [41] can not only steal intellectual property (e.g., proprietary application code, for reverse engineering), but also leak code to locate gadgets for advanced code reuse attacks [37], thwarting code layout diversification defenses like Address Space Layout Randomization (ASLR) [32]. Embedded Internet-of-Things (IoT) devices exacerbate the situation; many of these microcontroller-based systems have the same Internet connectivity as desktops and servers but rarely employ protections against attacks [23], [36]. Given the ubiquity of these embedded devices in industrial production and in our lives, making them immune to code disclosure attacks is crucial.

Recent research [7]–[9], [11], [14], [18]–[20], [26], [33], [44] implements execute-only memory (XOM) to defend against code disclosure attacks. Despite being unable to prevent code pointer leakage from data regions such as heaps and stacks, XOM enforces memory protection on the code region so that instruction fetching is allowed but reading or writing instructions as data is disallowed. This simple and effective defense, however, is not natively available on low-end microcontrollers. For example, the ARMv7-M and ARMv8-M architectures used in mainstream devices support memory protection but not execute-only (XO) permissions [4], [5]. uXOM [26] implements XOM on ARM embedded systems but incurs significant performance and code size overhead (7.3% and 15.7%, respectively) as it transforms most load instructions into special unprivileged load instructions. Given embedded systems' real-time constraints and limited memory resources, a practically ideal XOM solution should have close-to-zero performance penalty and minimal memory overhead.

This paper presents PicoXOM, a fast and novel XOM system for ARMv7-M and ARMv8-M devices using a memory protection unit (MPU) and the Data Watchpoint and Tracing (DWT) unit [4], [5]. PicoXOM uses the MPU to enforce write protection on code and uses the unique address range matching capability of the DWT unit to control read access to the code region. On a matched access, the DWT unit generates a debug monitor exception indicating an illegal code read, while unmatched accesses execute normally without slowdown. As PicoXOM disallows all read accesses to the code segment, it includes a minimal compiler change that removes all data embedded in the code segment.

We built a prototype of PicoXOM and evaluated it on an ARMv7-M board with two benchmark suites and five real-world embedded applications. Our results show that PicoXOM adds negligible performance overhead of 0.33% and only has a small code size increase of 5.89% while providing strong protection against code disclosure attacks.

To summarize, our contributions are:

- PicoXOM: a novel method of utilizing the ARMv7-M and ARMv8-M debugging facilities to implement XOM.
- To the best of our knowledge, this is the first use of ARM debug features for security purposes.
- A prototype implementation of PicoXOM on ARMv7-M.
- An evaluation of PicoXOM's performance and code size impact on the BEEBS benchmark suite, the CoreMark-Pro benchmark suite, and five real-world embedded applications, showing that PicoXOM only incurs 0.33% runtime overhead and 5.89% code size overhead.

The rest of the paper is organized as follows. Section II provides background information on ARMv7-M and ARMv8-M. Section III describes our threat model and assumptions. Sections IV and V present the design and implementation of PicoXOM, respectively. Section VI reports on our evaluation of PicoXOM. Section VII discusses related work, and Section VIII concludes the paper and discusses future work.

II. BACKGROUND

PicoXOM targets ARMv7-M and ARMv8-M architectures, which cover a wide range of embedded devices on the market, and it leverages unique features of these architectures. This section provides important background material on the instruction sets, execution modes, address space layout, memory protection mechanisms, and on-chip debug support found in ARMv7-M and ARMv8-M.
region holds memory-mapped system registers
region typically contains heaps and stacks, and
region maps
1
Code
SRAM
shows, the address space is generally divided into
System
A
M
DWT comparators
42
S
5
HREAT
0xFFFFFFFF
0x20000000
0x40000000
0x60000000
0xE0000000
power-of-two size from 32 bytes to 4 GB, and separate access
regions, each of which is configurable with a base address, a
management unit (MMU) that supports virtual memory; in-
the buggy application into reading its code content becomes a
such as those in EPOXY [35]. Therefore, remotely tricking
the buggy application into reading its code content becomes a
reasonable choice for the attacker.

A. Instruction Sets and Execution Modes

ARMv7-M [4] and ARMv8-M [5] are the mainstream M-
profile ARM architectures for embedded microcontrollers.
Unlike ARM’s A and R profiles, they only support the Thumb
instruction set which is a mixture of 16-bit and 32-bit densely-
encoded Thumb instructions.

ARMv7-M [4] supports two execution modes: unprivileged
mode and privileged mode. An ARMv7-M processor always
executes exception handlers in privileged mode, while
application code is allowed to execute in either mode. Code
running in unprivileged mode can raise the current execution
mode to privileged using a supervisor call instruction
(SVC). This is typically how ARMv7-M realizes system calls.
However, embedded applications usually run in privileged
mode to reduce the cost of system calls.

ARMv8-M inherits all the features of ARMv7-M and adds
a security extension called TrustZone-M [5] that isolates
software into a secure world and a non-secure world; this
effectively doubles the execution modes as software can be
executing in either world, privileged or unprivileged.

B. Address Space Layout

Both ARMv7-M [4] and ARMv8-M [5] architectures oper-
ate on a single 32-bit physical address space and use memory-
mapped I/O to access external devices and peripherals. As
Figure 1 shows, the address space is generally divided into
eight consecutive 512 MB regions; the Code region maps
flash memory/ROM that contains code and read-only data, the
SRAM region typically contains heaps and stacks, and the
System region holds memory-mapped system registers
including a Private Peripheral Bus (PPB) subregion. The PPB
subregion contains all critical system registers such as MPU
configuration registers and the Vector Table Offset Register
VTOR. All other regions are for memory-mapped peripherals
and external devices. Note that ARMv7-M and ARMv8-M
do not have special privileged instructions to access system
registers mapped in the System region; instead, they can be
modified by regular load and store instructions.

C. Memory Protection Unit

ARMv7-M and ARMv8-M devices do not have a memory
management unit (MMU) that supports virtual memory; in-
stead, they support an optional MPU that can be configured
to enforce region-based access control on physical memory [4],
[5]. A typical ARMv7-M device supports up to 8 MPU
regions, each of which is configurable with a base address, a
power-of-two size from 32 bytes to 4 GB, and separate access
permissions (R, W, and X) for privileged and unprivileged
modes. With TrustZone-M, ARMv8-M has separate MPU
configurations for secure and non-secure worlds [5]. MPU
configuration registers are in the PPB region.

There are, however, limitations on how one can configure
access permissions for an MPU region. First, the privileged
access permission cannot be more restrictive than the unpriv-
eligible one; this prohibits an MPU region with, for exam-
ple, unprivileged read-write and privileged read-only per-
missions. Second, the PPB region is always privileged-accessible,
unprivileged-inaccessible, and non-executable regardless of
the MPU configuration. Third, and most importantly, the
MPU does not have the execute-only permission necessary
to support XOM; an MPU region is executable only if it is
configured as both readable and executable.

D. Debug Support

Debug support is another processor feature that ARMv7-
M and ARMv8-M devices can optionally support. Of all
components in the architecture’s debug support, we focus on
the DWT unit [4], [5] which provides groups of debug registers
called DWT comparators that support instruction/data address
matching, PC value tracing, cycle counters, and many other
functionalities. Most importantly, a DWT comparator enables
monitoring of read accesses over a specified address range;
if the processor reads from or writes to an address within a
specified range, the DWT comparator will halt the software
execution or generate a debug monitor exception. If, instead,
the access does not fall into the specified range, execution
proceeds as normal, and performance is unaffected. When
multiple DWT comparators are configured for data address
range matching, an access that hits any of them will trap.

On ARMv7-M, a DWT comparator can be configured to
match an address range by programming its base address with
a mask that specifies a power-of-two range size [4]. ARMv8-
M implements DWT address range matching by using two
consecutively numbered DWT comparators [5], where the first
one specifies the lower bound of the address range and the
second one specifies the upper bound.

III. THREAT MODEL AND SYSTEM ASSUMPTIONS

We assume a buggy but unmalicious application running
on an embedded device with memory safety vulnerabilities
that allow a remote attacker to read or write arbitrary memory
locations. The attacker wants to either steal proprietary ap-
lication code for purposes like reverse engineering or learn
the application code layout in order to launch code reuse
attacks such as Return-into-libc [42] and Return-Oriented
Programming (ROP) [35] attacks. Physical and offline attacks
are out of scope as we believe such attacks can be stopped by
orthogonal defenses [24], [36]. Our threat model also assumes
the application code and data is diversified, using techniques
such as those in EPOXY [13]. Therefore, remotely tricking
the buggy application into reading its code content becomes a
reasonable choice for the attacker.
PicoXOM Compiler dictates. For ARMv8-A configurations, memory to be executable cannot be also configured as readable [31]. The MPU can configure a memory region as read-only, and RAM regions are readable and writable. Note that the MPU cannot configure memory to be executable but unreadable; the MPU can configure a memory region as executable only if it is also configured as readable [4], [5].

PicoXOM runs application code in privileged mode and configures a background MPU region to allow read and write access to the remainder of the address space such as peripherals. This, however, leaves critical memory-mapped system registers in the PPB (such as MPU configuration registers and VTOR) open to modifications, which can be leveraged by an attacker to turn off MPU protections or, even worse, implant a custom exception handler. Section IV-B discusses how PicoXOM prevents such cases.

### IV. DESIGN

Figure 2 shows PicoXOM’s overall design. PicoXOM consists of three components that together implement a strong and efficient XOM on ARM embedded devices. First, PicoXOM uses a specially-configured DWT configuration to detect read accesses to program code. Second, it utilizes a special MPU configuration that prevents write access to the code region and prevents writable memory from being executable. Third, it employs a small change to the LLVM compiler [27] to eliminate constant data embedded within the code region.

To use PicoXOM, embedded application developers merely compile their code with the PicoXOM compiler and install it on their embedded ARM device. On boot, the PicoXOM run-time configures MPU regions and DWT comparators using PicoXOM’s MPU and DWT configurations and then passes control to the compiled embedded software.

#### A. W@X with MPU

PicoXOM requires that memory either be writeable or executable but not both i.e., the W@X policy [31]; otherwise, an attacker could simply inject code or overwrite code to achieve arbitrary code execution. To enforce W@X, PicoXOM configures the MPU regions at device boot time so that the code region is readable and executable, read-only data is read-only, and RAM regions are readable and writable. Note that the MPU cannot configure memory to be executable but unreadable; the MPU can configure a memory region as executable only if it is also configured as readable [4], [5].

PicoXOM runs application code in privileged mode and configures a background MPU region to allow read and write access to the remainder of the address space such as peripherals. This, however, leaves critical memory-mapped system registers in the PPB (such as MPU configuration registers and VTOR) open to modifications, which can be leveraged by an attacker to turn off MPU protections or, even worse, implant a custom exception handler. Section IV-B discusses how PicoXOM prevents such cases.

#### B. R@X with DWT

PicoXOM leverages ARM’s DWT comparators to watch over the whole code region for read accesses. As Section II-D states, each (pair) of DWT comparators available on an ARM microcontroller can be configured to generate a debug monitor exception when a memory access of a specified type to an address within a specified range occurs. PicoXOM therefore uses one (pair) of the available DWT comparators as follows:

1. At device boot time, PicoXOM configures a DWT comparator register (say DWT_COMP<n>) to hold the lower bound of the code region.
2. PicoXOM then sets the address-matching range by either writing the upper bound of the code region to the next DWT comparator register DWT_COMP<n+1> (for ARMv8-M) or writing the correct mask to the corresponding DWT mask register DWT_MASK<n> (for ARMv7-M).
3. PicoXOM enables the DWT comparator (pair) by configuring the DWT function register DWT_FUNC<n> for data address reads. For ARMv8-M devices, DWT_FUNC<n+1> is also configured in order to form address range matching.
4. Finally, PicoXOM enables the debug monitor exception by setting the MON_EN bit (bit 16) of the Debug Exception and Monitor Control Register DEMCR.

With a DWT comparator (pair) set up for monitoring read accesses to the code region, R@X is effectively enforced. However, as Section IV-A stated, the DWT registers and DEMCR are also memory-mapped system registers which could be modified by vulnerable application code. An attacker could leverage a buffer overflow vulnerability to reconfigure the debug registers to neutralize PicoXOM.

We can address the issue in two ways. One approach is to break the assumption that PicoXOM runs everything in privileged mode. As code running in unprivileged mode has no access to the PPB region regardless of the MPU configuration, the system registers that PicoXOM must protect (e.g., MPU configuration registers, DWT registers, DEMCR, and VTOR) are all in the PPB region and therefore inherently safe from unprivileged tampering. However, this approach requires PicoXOM to implement system calls that support privileged operations which application code could previously perform, incurring expensive context switching between privilege modes. The other approach is to use extra (pairs of) DWT comparators to prevent writes to critical system registers. For example, on ARMv7-M, we can configure one DWT comparator to write-protect the System Control Block SCB (0xE000ED00 – 0xE000ED8F) and DEMCR (0xE000EDFC) by setting the lower bound and the size to 0xE000ED00 and 256 bytes, respectively. Since MPU configuration registers are in the SCB, they are protected as well. DWT registers on ARMv7-M reside in a separate range (0xE0001000 – 0xE0001FFF),
so we can use another DWT comparator to write-protect that range.

C. Constant Island Removal

By default, ARM compilers generate code that has constant data embedded in the code region (so-called "constant islands"). Since PicoXOM prevents the code from reading these constant islands, these programs will fail to execute when used with PicoXOM. PicoXOM therefore transforms these programs so that all data within the program is stored outside of the code region.

We have identified two cases of constant islands generated by LLVM’s ARM code generator: load constants and jump-table jumps. Figures 3 and 4 show examples of the two cases, respectively, as well as their corresponding execute-only versions to which PicoXOM transforms them. Specifically, in the left part of Figure 3, a load constant instruction loads a constant from a PC-relative memory location $L$ into register $r0$. Such instructions are usually generated to quickly load an irregular constant in light of the limited immediate encoding scheme of the Thumb instruction set [4], [5]. PicoXOM transforms such load constants into MOVW and MOVT instructions that encode the 32-bit constant in two 16-bit immediates, as the right part of Figure 3 shows. Jump-table jump instructions (TBB and TBH) [4], [5] are used to implement large switch statements; the second register operand ($r2$ in Figure 4) serves as an index into a jump table pointed to by the first register operand ($pc$ in Figure 4), and a byte/half-word offset is loaded from the jump table to add to the program counter ($pc$) to calculate the target of the jump. Optimizing compilers like GCC and LLVM usually select $pc$ as the first register operand in order to reduce register pressure, forcing the jump table to be located next to the jump-table jump itself. PicoXOM transforms such jump-table jumps into instruction sequences like that shown in the right part of Figure 4; it encodes the original jump table’s contents into a sequence of branch instructions and expands the jump-table jump into a few explicit instructions that calculate which branch instruction to jump to and perform an indirect jump.

V. IMPLEMENTATION

We built our PicoXOM prototype for the ARMv7-M architecture. Our prototype provides MPU and DWT configurations as a run-time component written in C and executed at the end of the device boot sequence. We implemented constant island removal as a simple intermediate representation (IR) pass in the LLVM 10.0 compiler [27]. The constant island removal pass simply uses the existing -mexecute-only option in LLVM’s Clang front-end and passes it along to the link-time optimization (LTO) code generator. Our prototype runs the constant island removal pass when linking the IR of the application, libraries (e.g., newlib and compiler-r0), and MPU and DWT configurations; this ensures that all code has no constant islands. Our prototype adds 88 source lines of C++ code to LLVM and has 177 source lines of C code in the PicoXOM run-time. We leave the PicoXOM implementation on ARMv8-M for future work.

Different ARM microcontrollers support different numbers of MPU regions and DWT comparators, and the maximum ranges of their DWT comparators may vary. Our prototype runs on an STM32F469 Discovery board which supports up to 8 MPU regions [39] and 4 DWT comparators [40]. Each DWT comparator can only watch over a maximum address range of 32 KB (a maximal mask value of 15), limiting our prototype to the following two options:

1) Use all 4 DWT comparators to support a maximum code size of 128 KB; the application must run in unprivileged mode in order for the critical system registers to be write-protected.

2) Configure one DWT comparator to write-protect the DWT registers (0xE0001000 – 0xE0001FFF) and another write-protect the SCB (0xE000ED00 – 0xE000ED8F) and DEMCR (0xE000EDFC). This protects a maximum code size of 64 KB using the remaining 2 DWT comparators.

To accommodate a wider range of applications on our board with less performance loss, our prototype automatically chooses one option over the other based on the application code size. It rejects an application if the code size exceeds our board’s 128 KB limit.

While our PicoXOM prototype only supports single bare-metal embedded applications, PicoXOM can also support multiple applications running on an embedded real-time operating system (RTOS) such as Amazon FreeRTOS [3]. On embedded systems, the application and RTOS kernel code is linked into a single shared code segment. PicoXOM can protect this code segment with little adaptation.

VI. EVALUATION

We evaluated PicoXOM on our STM32F469 Discovery board [40] which has an ARM Cortex-M4 processor implementing the ARMv7-M architecture that can run as fast as 180 MHz. The board comes with 2 MB of flash memory, 384 KB of SRAM, and 16 MB of SDRAM, and has an LCD screen and a microSD card slot. We configured the board to run at its fastest speed to understand the maximum impact that PicoXOM can incur on performance.

<table>
<thead>
<tr>
<th>Instructions</th>
<th>Assemble</th>
<th>Machine Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>ldr r0,=L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L: .word 0x12345678</td>
<td></td>
<td></td>
</tr>
<tr>
<td>movw r0,#0x1234</td>
<td></td>
<td></td>
</tr>
<tr>
<td>movt r0,#0x1234</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Constant Island Removal of a Load Constant

<table>
<thead>
<tr>
<th>Instructions</th>
<th>Assemble</th>
<th>Machine Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>tbb [pc,r2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L0: .byte (L1-L0)/2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.byte (L2-L0)/2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.byte (L3-L0)/2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1: ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2: ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L3: ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4. Constant Island Removal of a Jump-Table Jump
To evaluate PicoXOM’s performance and code size overhead, we used the BEEBS [30] and CoreMark-Pro [16] benchmark suites and five embedded applications (FatFs-RAM, FatFs-uSD, LCD-Animation, LCD-uSD, and PinLock). BEEBS targets energy consumption measurement for embedded platforms and is widely used in evaluating embedded systems including uXOM [26], the state-of-the-art XOM implementation on ARM microcontrollers. It consists of a wide range of programs characterizing different workloads seen on embedded systems, including AES encryption, data compression, and matrix multiplication. Of all 80 benchmarks in BEEBS, we picked 42 benchmarks that have an execution time longer than 500 milliseconds when executed for 10,240 iterations.

CoreMark-Pro is a processor benchmark suite that works on both high-performance processors and low-end microcontrollers, featuring five integer benchmarks (e.g., JPEG image compression, XML parser, and SHA-256) and four floating-point benchmarks (e.g., fast Fourier transform and neural network) that stress the CPU and memory. FatFs-RAM and FatFs-uSD operate a FAT file system on SDRAM and an SD card, respectively. LCD-Animation displays a single animated picture loaded from an SD card. LCD-uSD displays multiple static pictures from an SD card with fading transitions. PinLock simulates a smart lock reading user input from a serial port and deciding whether to unlock (send an I/O signal) based on whether the SHA-256 hashed input matches a precomputed hash. The above five applications represent real-world use cases of embedded devices and were also used to evaluate previous work [2], [12], [13].

We used the LLVM compiler infrastructure [27] to compile benchmarks and applications into the default non-XO format, with MPU and DWT disabled; this is our baseline. We then used PicoXOM’s configuration, i.e. enabling MPU, DWT, and constant island removal. Note that with PicoXOM, none of the benchmarks and applications exceed the code size limitation (128 KB) on our board. Only jpeg rose7preset in CoreMark-Pro has a code size larger than 64 KB and thereby has to run in unprivileged mode; nevertheless, it does not require source code modifications as it does not perform privileged operations.

### A. Performance

We measured PicoXOM’s performance on our benchmarks and applications. We configured each BEEBS benchmark to print the time, in milliseconds, for executing its workload 10,240 times. We ran each BEEBS benchmark 10 times and report the average execution time. Each CoreMark-Pro benchmark is pre-programmed to print out the execution time in a similar way; the difference is that we configure each benchmark to run a minimal number of iterations so that the program takes at least 10 seconds to run for each experimental trial. Again, we ran each benchmark 10 times and report the average execution time. For the real-world applications, we ran FatFs-RAM 10 times and report the average execution time. The other applications exhibit higher variance in their execution times as they access peripherals like an SD card, an LCD screen, and a serial port, so we ran them 20 times and report the average with a standard deviation. All other programs exhibit a standard deviation of zero.

### Tables I and II

Tables I and II and Figure 5 present PicoXOM’s performance on BEEBS, CoreMark-Pro, and the five real-world applications, respectively; Figure 5 shows baseline execution time in milliseconds on top of the Baseline bars. Overall, PicoXOM incurs negligible performance overhead of 0.33%: 0.46% on BEEBS with a maximum of 7.48%, −0.11% on CoreMark-Pro with a maximum of 0.17%, and 0.02% on the applications with a maximum of 0.22%. Thirteen programs

### Table I

**Performance Overhead on BEEBS**

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Baseline (ms)</th>
<th>PicoXOM (% ×)</th>
<th>Baseline (ms)</th>
<th>PicoXOM (% ×)</th>
</tr>
</thead>
<tbody>
<tr>
<td>aha-compress</td>
<td>821.0</td>
<td>0.0000</td>
<td>nettle-arcfour</td>
<td>814.0</td>
</tr>
<tr>
<td>aha-mont64</td>
<td>856.0</td>
<td>0.9988</td>
<td>picogeg</td>
<td>43.864</td>
</tr>
<tr>
<td>bubblesort</td>
<td>4.392</td>
<td>1.0000</td>
<td>qdudino</td>
<td>40.877</td>
</tr>
<tr>
<td>crc32</td>
<td>956.0</td>
<td>1.0000</td>
<td>rjndael</td>
<td>70.024</td>
</tr>
<tr>
<td>elf-string</td>
<td>630.0</td>
<td>1.0000</td>
<td>sglib-arraybin</td>
<td>808.0</td>
</tr>
<tr>
<td>elf-vector</td>
<td>786.0</td>
<td>0.9987</td>
<td>sglib-arrayhea</td>
<td>1.039</td>
</tr>
<tr>
<td>cubic</td>
<td>35.140</td>
<td>1.0005</td>
<td>sglib-arrayqu</td>
<td>735.0</td>
</tr>
<tr>
<td>dijkstra</td>
<td>36.582</td>
<td>1.0000</td>
<td>sglib-dlist</td>
<td>1.800</td>
</tr>
<tr>
<td>dtoa</td>
<td>631.0</td>
<td>1.0127</td>
<td>sglib-hashable</td>
<td>1.302</td>
</tr>
<tr>
<td>edn</td>
<td>3.167</td>
<td>1.0003</td>
<td>sglib-listinsert</td>
<td>2.030</td>
</tr>
<tr>
<td>fasta</td>
<td>16.900</td>
<td>0.9999</td>
<td>sglib-lsort</td>
<td>1.265</td>
</tr>
<tr>
<td>fir</td>
<td>16.048</td>
<td>1.0000</td>
<td>sglib-queue</td>
<td>1.177</td>
</tr>
<tr>
<td>frac</td>
<td>5.658</td>
<td>1.0323</td>
<td>sglib-rbtree</td>
<td>4.808</td>
</tr>
<tr>
<td>hufbench</td>
<td>20.682</td>
<td>0.9995</td>
<td>sfire</td>
<td>2.761</td>
</tr>
<tr>
<td>levenshtein</td>
<td>2.685</td>
<td>1.0000</td>
<td>sqrt</td>
<td>38.506</td>
</tr>
<tr>
<td>matmul-float</td>
<td>1.150</td>
<td>0.9991</td>
<td>std</td>
<td>20.906</td>
</tr>
<tr>
<td>matmul-int</td>
<td>4.532</td>
<td>1.0000</td>
<td>std_perlin</td>
<td>5.132</td>
</tr>
<tr>
<td>mergesort</td>
<td>24.353</td>
<td>1.0062</td>
<td>trio-sprintr</td>
<td>69.7</td>
</tr>
<tr>
<td>nbodies</td>
<td>128.126</td>
<td>1.0090</td>
<td>trio-sscanf</td>
<td>1.064</td>
</tr>
<tr>
<td>ndes</td>
<td>2.039</td>
<td>0.9995</td>
<td>whelson</td>
<td>112.754</td>
</tr>
<tr>
<td>nnet</td>
<td>5.687</td>
<td>0.9989</td>
<td>wksort</td>
<td>113.195</td>
</tr>
</tbody>
</table>

Min (×)                      0.9873
Max (×)                      1.0748
Geomean (×)                  1.0046

### Table II

**Performance Overhead on CoreMark-Pro**

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Baseline (ms)</th>
<th>PicoXOM (% ×)</th>
<th>Baseline (ms)</th>
<th>PicoXOM (% ×)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cjpeg-rose7-...</td>
<td>10.200</td>
<td>1.0001</td>
<td>parser-125k</td>
<td>12.363</td>
</tr>
<tr>
<td>core</td>
<td>83.160</td>
<td>0.9918</td>
<td>radix2-big-64k</td>
<td>21.955</td>
</tr>
<tr>
<td>linear_alg-...</td>
<td>22.902</td>
<td>1.0000</td>
<td>sha-test</td>
<td>25.463</td>
</tr>
<tr>
<td>loops-alg-...</td>
<td>33.830</td>
<td>0.9995</td>
<td>zip-test</td>
<td>23.227</td>
</tr>
</tbody>
</table>

Min (×)                      0.9918
Max (×)                      1.0017
Geomean (×)                  0.9989

![Fig. 5. Performance Overhead on Real-World Applications](image-url)
show the baseline code size and the trio-sscanf 34,156 1.0668 1.0614
1.0627
perlin 31,140 1.0627
sglib-queue 29,988 1.0650
stb 31,124 1.0602
nettle-arcfour 29,988 1.0649
linear_alg-mid-100x100-sqrt 30,372 1.0641

Baseline PicoXOM increases the code size by 6.14% on BEEBS, 4.39% on CoreMark-Pro, and 6.52% on the five real-world applications, respectively. On average, the overhead incurred by PicoXOM on BEEBS, CoreMark-Pro, and five real-world applications, respectively. On average, PicoXOM increases the code size by 6.14% on BEEBS, 4.39% on CoreMark-Pro, and 6.52% on the five real-world applications, with a 5.89% overall overhead. We studied PicoXOM’s code size overhead and discovered that constant island removal caused the majority of the code size overhead, especially for programs with relatively large code bases like CoreMark-Pro. In fact, the additional code that sets up the MPU and DWT only contributes a minor part of the overhead (1.22% and 0.53% on average, respectively).

VII. RELATED WORK

Two other XOM implementations exist for ARM microcontrollers. uXOM [26] provides XOM for ARM Cortex-M systems by transforming loads into special unprivileged load instructions and configuring the MPU to make the code region unreadable by unprivileged loads. uXOM similarly transforms stores to protect the memory-mapped MPU configuration registers. Since some loads and stores do not have unprivileged counterparts, transforming them requires the compiler to insert additional instructions, causing the majority of uXOM’s overhead. PicoXOM is more efficient in both performance (0.33% compared to uXOM’s 7.3%) and code size (5.89% compared to uXOM’s 15.7%) as no such transformation is needed. A trade-off for PicoXOM is the code size limit on some ARMv7-M devices; we envision no such limit on ARMv8-M. PCROP [38] is a programmable feature of the flash memory which prevents the flash memory from being read out and modified by application code but still allows code in the flash memory to execute. However, PCROP is only available on some STMicroelectronics devices and cannot be used for other types of memory. In contrast, PicoXOM relies on the MPU and DWT features [4], [5] which can be found on most conforming devices and can protect code stored in any type of memory.

Hardware-assisted XOM has been explored on other architectures. The AArch64 [6] and RISC-V [34] page tables natively support XO permissions. NORAX [11] enables XOM for commercial-off-the-shelf binaries on AArch64 that have constant islands using static binary instrumentation and runtime monitoring. Various approaches [9], [14], [18]–[20], [44] leverage features of the MMU on Intel x86 processors [22] to implement XOM. None of these approaches are applicable on ARM embedded devices lacking an MMU. Lie et al. [28] proposed an architecture with memory encryption to mimic XOM, but it only provides probabilistic guarantees and cannot be directly applied to current embedded systems. Compared to solutions for systems lacking native hardware XOM support, PicoXOM is faster as it has nearly no overhead.

Software can emulate XOM. XnR [7] maintains a sliding window of currently executing code pages and keeps only pages accessible. It still allows read accesses to a subset of code pages and may incur higher overhead for a smaller sliding window size due to frequent page permission changes. LR [8] and kR*X [33] instrument all load instructions to prevent them from reading the code segment. While these software XOM approaches can generally be ported to embedded devices, they can be bypassed by attacker-manipulated control flow and are less efficient than hardware-assisted XOM [26].

There are also methods of hardening embedded systems. Early versions of SAFECode [15] enforced spatial and temporal memory safety on embedded applications, and
nesCheck [29] uses static analysis to build spatial memory safety for simple nesC [17] applications running on TinyOS [21]. PicoXOM enforces weaker protection than memory safety but supports arbitrary C programs (unlike SAFE-Code and nesCheck) and does not rely on heavy static analysis like nesCheck. RECFISH [43], μRAI [2], and Silhouette [45] mitigate control-flow hijacking attacks on embedded systems. They protect forward-edge control flow using coarse-grained CFI [1] and backward-edge control flow by using either a protected shadow stack [10] or a return address encoding mechanism. EPOXY [13] randomizes the order of functions and the location of a modified safe stack from CPI [25] to resist control-flow hijacking attacks on bare-metal microcontrollers. These systems do not enforce XOM and are still vulnerable to forward-edge corruptions; they can incorporate PicoXOM’s techniques to mitigate forward-edge attacks with negligible additional overhead.

VIII. CONCLUSIONS AND FUTURE WORK

This paper presented PicoXOM: a fast and novel XOM system for ARMv7-M and ARMv8-M devices which leverages ARM’s MPU and DWT unit. PicoXOM incurs an average performance overhead of 0.33% and an average code size overhead of 5.89% on the BEEBS and CoreMark-Pro benchmark suites and five real-world applications.

In future work, we will investigate techniques to ensure that randomization techniques utilizing PicoXOM are effective against brute-force attacks. Embedded systems have limited code placement options for code layout randomization, motivating us to investigate whether the entropy is sufficient and develop techniques to strengthen code randomization if necessary. We will also explore how to leverage debug support like DWT to enforce other security policies with low overhead.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their insightful comments. This work was funded by ONR Award N00014-17-1-2996.

REFERENCES


[38] AN4701 Application Note: Proprietary Code Read-Out Protection on Microcontrollers of the STM32F446FCT6, STMicroelectronics, November 2016, DocID027893 Rev 3.


