Microprocessors and Microsystems 64 (2019) 43–52

Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

Lightweight instruction-level encryption for embedded processors using stream ciphers☆

Thomas Hiscock[a,*], Olivier Savry[a], Louis Goubin[b]

[a] Univ. Grenoble Alpes, CEA, LETI, DSYS, LSOSP/CESTI, F-38000 Grenoble
[b] Laboratoire de Mathématiques de Versailles, UVSQ, CNRS, Université Paris-Saclay, F-78035 Versailles

ARTICLE INFO

Article history:
Received 5 February 2018
Revised 2 August 2018
Accepted 2 October 2018
Available online 11 October 2018

Keywords:
Software encryption
Processor design
Security
FPGA
LLVM

ABSTRACT

Over the last 30 years, a number of secure processor architectures have been proposed to protect software integrity and confidentiality during its distribution and execution. In such architectures, encryption (together with integrity checking) is used extensively, on any data leaving a defined secure boundary.

In this paper, we show how encryption can be achieved at the instruction level using a stream cipher. Thus encryption is more lightweight and efficient, and is maintained deeper in the memory hierarchy than the natural off-chip boundaries considered in most research works. It requires the control flow graph to be used and modified as part of the off-line encryption process, but thanks to the LLVM framework, it can be integrated easily in a compiler pipeline, and be completely transparent to the programmer.

We also describe hardware modifications needed to support this encryption method, the latter were added to a 32-bit MIPS soft core. The synthesis performed on a Altera Cyclone V FPGA shows that encryption requires 26% of extra logic, while slowing-down execution time by an average of 48% in the best setting.

1. Introduction

An increasing number of applications require processor architectures that are both lightweight and able to preserve software confidentiality. The historical motivation was a purely economic one: protect intellectual property and prevent illegitimate duplication. Nowadays, software confidentiality is rather seen from a security perspective, to prevent reverse engineering. The program being more resilient to analysis, the effort needed to discover weaknesses is increased, and critical patches can be deployed without fear of zero-day exploits.

If one can modify the hardware, software encryption is a well-established way to achieve this confidentiality. The program is encrypted using a regular cryptographic primitive, and decryption is done using a hardware implementation of the decryption algorithm in an assumed secure area. The latter is close to the processor executing the software, at least on the same chip. Aside from the confidentiality property, encryption alone brings other interesting properties from a security perspective:

- As each target have different encryption keys, shellcode design is harder, thus exploits cannot be deployed quickly on a large scale.
- It can be used as a building block to provide control flow integrity, as shown recently with the SOFIA [2] architecture.

Current secure processor architectures are mostly concerned with protecting programs and data stored on off-chip memories like Flash, Dynamic RAMs (DRAMs), the chip area being assumed safe. For this reason, and also for performance concerns, decryption is usually done at a cache (level 1, or level 2) memory boundary. As a consequence, data are decrypted by chunks made of one or more cache lines (32B, 64B, 128B or even more). This way, the latency of the decryption algorithm can be almost completely hidden, by overlapping the decryption with memory fetches on a cache miss.

However, to the best of our knowledge, very few works (but [3]) provide methods to achieve a finer encryption granularity, namely at the instruction level. Yet, it is required for applications in which the target processor either do not have cache, or needs to maintain encryption deeper in the memory hierarchy.

In this work we show how encryption can be done at the instruction level using stream ciphers, which are known to be very lightweight and efficient. It requires the control flow graph of the program to be used and restructured as part of the encryption process. The proposed method, developed as a LLVM [4] backend pass, can encrypt almost any given machine code and do not require any

☆ This work is an extended version of the paper [1] published in the Euromicro DSD 2017 proceedings.

* Corresponding author.
E-mail address: thomas.hiscock@wanadoo.fr (T. Hiscock).

https://doi.org/10.1016/j.micpro.2018.10.001
0141-9331/

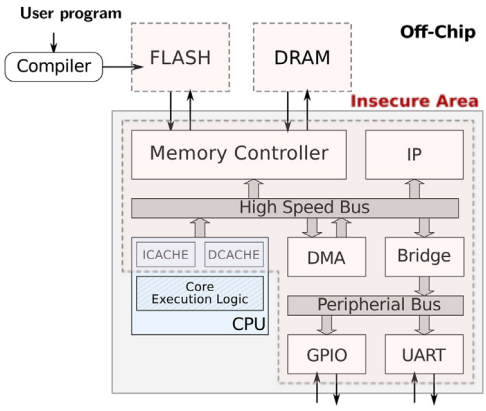


Fig. 1. A typical system considered in this work, with the on-chip insecure area drawn.

modification from the programmer but adding a compiler flag. We stress that we only describe a method for encrypting the software, not verifying its integrity. Of course, integrity is as much as important as encryption in secure processors [5], but its treatment can be somewhat orthogonal, so we decided to focus on the encryption mechanism.

The hardware support, including decryption hardware (based on the Trivium [6] stream cipher) is added into MIPS soft core and deployed on a low-cost Altera Cyclone V Field Programmable Gate Array (FPGA). On this small core, the encryption mechanism requires only 26% of extra logic. The execution slowdown is highly dependent on the compilation profile. In the best setting (performance optimized programs), we measured an average slowdown of 48%, across all benchmarks slowdown is between 29% up to 193%. These results illustrate that a very lightweight and efficient encryption can be achieved to target real-world applications on constrained processors.

The rest of this paper is structured as follows: Section 2 presents in more details the security model used in this work, followed by a survey of related work in Section 3. Our software encryption process as well as its integration in the LLVM framework is described in Section 4. Then, the processor modifications needed to support encryption, and implementation of common software abstractions like exceptions, context switches are presented in Section 5. We conclude this paper by an evaluation of our method in Section 6. The security of this solution is discussed, and result from our practical implementation on FPGA are analyzed.

This work is an extended version of [1]. It provides more details on the implementation of the solution and the handling of exception. It also provides additional results about trade-offs between performance and area that can be achieved with different configurations of the underlying cryptographic primitive.

2. Security model

In this work we consider a standard System on Chip (SoC) system, with a single processor, as the one shown Fig. 1. The processor itself may or may not have instruction and data caches. Secure processors usually draw an insecure boundary at the off-chip memory interfaces. Beyond this boundary, it is assumed that any data can be observed, or tampered with.

Indeed, many popular attacks showed that memory can be easily extracted on a wide range of devices, even with cheap hardware. As an example, "Cold boot" family attacks exploit data persistence in DRAMs [7]: after system reboot some critical part of memory can then be recovered. Direct Memory Access (DMA) components were successfully used to obtain read or write access to CPU's memory through some user accessible peripherals (e.g., Firewire [8]). Even direct probing using FPGA or low cost modchips [9] is feasible on external buses like PCI express.

On the other hand, attacking the internal logic of a processor, say, reading a register value at a given time, is far more challenging and requires advanced physical attack techniques as well as expensive equipment [10,11].

In this work, our goal is to protect the confidentiality of a given machine code. Our insecure boundary is moved deeper inside the chip, between the processor memory interfaces and its execution logic (caches are also considered as insecure). Formally, an adversary is allowed to:

- read any data stored into off-chip memories (the latter will be ciphered),
- read instructions located into the instruction cache or any on-chip memory.

For this purpose, we assume that the CPU's execution logic (Fig. 1) is shielded, physical attacks cannot be performed, such that cryptographic operations can be done safely inside the core. This shielded region includes CPU's internal state, like the program counter (PC), registers, etc.

This work is primarily concerned with software protection. The programs executed on the device are assumed, in the sense that they do not store critical data in memory, or if so, manipulate them using a dedicated secure coprocessor.

3. Related work

Obfuscation techniques are a well-known class of software-only countermeasures [12], but cannot achieve provable security even for restrained models [13] without some secure hardware. Heuristic techniques have proven to increase the time and effort needed to reverse a program, but can be defeated by an experimented adversary. Furthermore, the overhead on both program size and execution time is quite high: depending on obfuscation level, factors between $\times 10$ and $\times 100$ are common [14].

The use of software encryption in processors dates back to Best [15,16], who proposed a series of patents that made up the basis of the Dallas DS5002 [17] secure processor. Early versions of the Dallas DS5002 were defeated by a famous attack performed by Kuhn [5]. He managed to inject instructions and monitor I/O to build a malicious code capable of dumping the whole memory.

Since then, number of researchers proposed hardware-assisted memory encryption [18–23]. A block cipher is used as encryption primitive to perform data authentication and decryption when accessing data from insecure external memory. The plain content is then placed in a processor-close memory, assumed free from tampering (local RAM or a cache). It is well known that decryption latency can quickly become a performance bottleneck in such architectures. Several techniques were proposed to reduce this latency, like predicting decryptions and keep one-time pad in small CPU-internal caches [24,25]. It was shown in [26] that most predictions can be avoided if the compiler adds hints in the code about upcoming decryptions.

The closest approach to ours appears to be Instruction Set Randomization (ISR) [3,27,28], though mainly designed to prevent code injection. It was shown in [3] that ISR can also be used to prevent reverse engineering. They added an additional processor instruction called `rev` to randomize the instruction set on demand. They implemented this software encryption using the Trivium stream cipher on top of a Leon2 (SPARC V8) core. Compared to this work and more generally ISR techniques, our solution do not need any instruction set extension, so it is more transparent to the programmer.

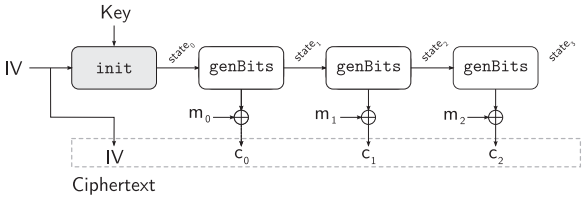


Fig. 2. Encryption using a stream cipher.

4. Static code encryption with stream ciphers

4.1. Background on stream ciphers

Stream ciphers are an efficient class of pseudorandom generators. Unlike block ciphers which provide a fixed-length permutation, they can produce an arbitrary long pseudorandom sequence. Formally speaking, a stream cipher is specified by two functions:

- `init` : $\mathcal{K} \times \mathcal{IV} \rightarrow \mathcal{S}$, which generates an initial state from a secret key and an Initialization Vector (IV). The IV can be made public, while the secret key must be kept private.
- `genBits` : $\mathcal{S} \rightarrow \mathcal{S} \times \mathcal{C}$, which produces the next state and a pseudorandom output.

Once initialized with `init` and an arbitrary IV, pseudorandom bits can be generated on demand and used as a one-time pad to provide an encryption scheme, as shown in Fig. 2. The IV has to be transmitted with the ciphertext to allow the receiver to decrypt. We stress that IVs have to be uniformly distributed to guarantee the full security of the scheme under chosen plaintext attacks (IND-CPA). Furthermore non uniform IV (using counter mode-like construction) might lead to reduced attack complexity through time space trade-off attacks [29].

For most stream ciphers, the `init` function is a costly operation while the `genBits` function is quite fast. As a consequence, stream ciphers are good for generating very long pseudorandom sequences.

Let us note that a stream cipher can be constructed from a block cipher using the output feedback mode of encryption [30]. However, dedicated stream ciphers are far more efficient. In 2008, the eStream [6] competition selected a set of recommended stream ciphers. For the hardware profile, three were selected: Trivium, Grain and Mickey.

4.2. Why stream ciphers?

Block ciphers are a good established way achieve encryption in secure processor architectures [31]. The counter mode of encryption is a good fit for encrypting a processor address space. Indeed, the base address of the data (or a block of data) can be used as counter value, and, unless virtual memory is used [22,32], counter value uniqueness is guaranteed.

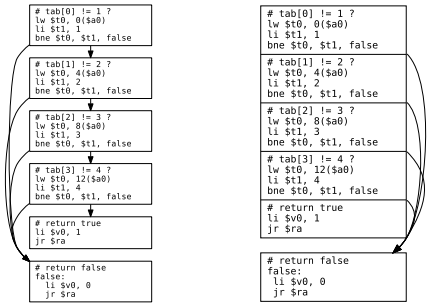
However, block ciphers suffer from two limitations for being used to encrypt at the instruction level. The first one is that block ciphers are intrinsically fixed-length permutations, and for security reason, the minimum recommended block size is 128 bits. On the other hand, in many instruction set architectures, instructions almost never take more than 32 bits [33]. To exploit the full throughput of the block cipher, some sort of complex instruction padding has to be designed.

The second limitation is that the decryption primitive has to be able to work at CPU's execution speed. Of course, a fully pipelined implementation of a block cipher would meet this requirement, but may also significantly increase the hardware footprint. As an illustration, the fully pipelined AES implementation from OpenCores[1], is 3 times the size of our base MIPS processor.

On the other hand, instruction execution is most of the time a very sequential process, at least on in-order processor architectures. As already mentioned, stream ciphers, once initialized, are quite efficient to generate random bits. For this reason they seem to fit quite well with the job of encrypting software at instruction granularity. Once initialized, a stream cipher is usually able to decrypt one instruction per cycle, whereas a block cipher would require several rounds between each instruction.

```
bool f(int tab[]) {
    if (tab[0] != 1) return false;
    if (tab[1] != 2) return false;
    if (tab[2] != 3) return false;
    if (tab[3] != 4) return false;
    return true;
}
```

(a) Original C function



(b) Before basic block merging

(c) After basic block merging

Fig. 3. Illustration of basic block merging opportunities.

4.3. Encrypting machine code

Because of the stateful nature of stream ciphers, it is not straightforward to use them for encrypting instructions. Let first remark that a whole program cannot be encrypted using a unique stream. Indeed, in case of a jump, the stream used at the destination address must be known to keep decrypting instructions. The only way to recover it is to re-compute the entire stream from program start to the target instruction, which would be quite inefficient for long programs. Thus, it seems clear that a finer encryption granularity should be adopted.

For this purpose let us introduce basic blocks (BB), a fundamental structure in compiler construction. A basic block is defined as a sequence of instructions without intermediate incoming or outgoing branches. The only entry point of a BB is its first instruction and its output is its last instruction. As an illustration, the program depicted Fig. 3(b)is made of six BB. Considering in-order processor architectures, a basic block is then always executed sequentially, and can be encrypted using a unique stream cipher sequence. By construction, it is impossible for a branch to fall in the middle of a BB, so the stream cipher state never has to be reconstructed to an arbitrary state.

To encrypt a whole program, independent stream cipher sequences are generated for each basic block, using different initialization vectors (further details on this topic will be given in Section 4.4). For a processor executing such an encrypted software, branch instructions, which notify a BB change, have to trigger a reset of the stream cipher with a new IV.

[1] http://opencores.org/project,tiny_aes

Table 1
MIPS instructions used in this paper and their semantic.

Instruction	Semantic
`li $t0, value`	$\$t0 \leftarrow value$
`lw $t0, N($a0)`	$\$t0 \leftarrow MEM[\$a0 + N]$
`add $t0, $t1, $t2`	$\$t0 \leftarrow \$t1 + \$t2$
`bne $t0, $t1, dst`	if $\$t0 \neq \$t1$, jump to *dst*
`jump dst`	jump to *dst*
`jr $t0`	jump to address in $\$t0$

The rest of this section describes more precisely the encryption process, divided into three stages. To illustrate them, code examples are given in MIPS assembly. Registers are prefixed with a dollar symbol (e.g., `$t0`, `$a1`), all instructions used with their semantic are listed in Table 1. Further details can be found in the instruction set reference [34].

4.3.1. Merging basic blocks for encryption

A limitation of the basic block approach to encryption is that, programs usually have a large number of small basic blocks: for MIPS programs [33], most basic blocks have between 3 and 8 instructions. Of course this metric is highly dependent on the input program and the instruction set architecture. A high number of basic blocks means that the stream cipher `init` function will have to be called very often at the execution, possibly leading to an important slowdown.

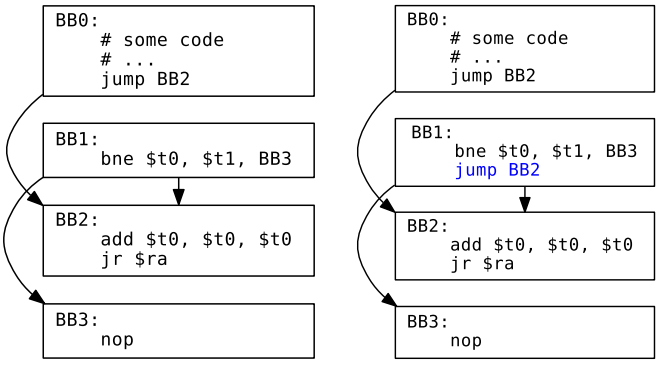
Hopefully longer sequences can be encrypted with the same stream. Indeed, the only requirement for a sequence of instructions to be encrypted with the same stream, is that there is no incoming jump somewhere other than the first instruction. There is absolutely no restriction on the number of outgoing jumps in an encrypted sequence of instructions. It slightly differs from a basic block, which cannot have more than one outgoing jump. This structure will be called an encryptable basic block for the rest of this paper and is defined as follows.

Definition 1. Encryptable Basic Block (EBB): A sequence of instructions which has no incoming jumps other than at its first instruction. It may contain any number of outgoing jumps.

This structure is known in compiler construction as a *superblock* [35]. It is widely used to optimize programs for Very Long Instruction Word (VLIW) architectures. Some powerful techniques are available to create large superblocks: branch target expansion, loop unrolling, common subexpression detection. But an in-depth study of the optimal merging approach would bring us out of the scope of this paper.

Meanwhile, a very simple strategy is applied to merge the basic blocks. Two successive basic blocks (i.e, consecutive in the address space) can be merged for encryption if the second one has no incoming jump (can only be reached from the first one). To perform a full merging, this two-block merging is applied on the control flow graph until a fixed point is reached. Even this simple approach brings performance improvements (7% on average).

The merging occurs very frequently in practice, for instance while translating if-else structures. As an illustration, Fig. 3 is given a very simple function, which checks a set of preconditions on an array and returns false whenever one of them is not satisfied. The control flow graph obtained using a normal compilation process generates lots of basic blocks (Fig. 3b). Without merging basic blocks, it would require six different encryption sequences, one for each basic block. However this control flow graph can be fully merged into just two encryptable basic blocks as shown Fig. 3(c).



(a) Branchless fall through (b) Fixed with a jump instruction

Fig. 4. Illustration of a branchless fall through situation (from BB1 to BB2).

4.3.2. Removing branchless fall through basic blocks

A special case to take care of for encryption correctness is that compilers, as an optimization, usually remove away branches going from two successive basic blocks. Indeed, the compiler assumes that when two BBs are layout successors, then the first one can fall into the second one, without needing an extra jump. This behavior has to be disabled when encrypting programs, otherwise the processor will not detect a sequence change and continue its execution with the wrong stream. As an example, Fig. 4, BB2 has two predecessors (BB0 and BB1), and BB1 falls through BB2 without a jump. To fix this graph for encryption, an explicit direct jump to BB2 is inserted at the end of BB1 (Fig. 4b).

The branch removal appears early in the LLVM compilation pipeline, so instead of modifying it directly, a late pass was implemented, that inserts back these missing branches. Then, this pass can be scheduled after the basic block merging described previously, to insert just the minimum number of branches required to fix the control flow graph.

4.4. Initialization vector selection schemes

The previous section described how the control flow graph can be prepared for the encryption. To fully encrypt a graph of encryptable basic blocks, a unique encryption stream has to be generated for each one of them. The secret key being fixed and hard-wired into the processor, initialization vectors can be used to generate distinct sequences. Unlike the secret key, IVs are public data, there is no need to keep them secret. The only requirement is that they must be unique across the whole program to guarantee security of the one-time pad encryption.

4.4.1. A Counter Mode approach

A first possible approach is to compute IVs from the current program counter value (CTR mode). Formally, a random initialization vector IV_0 is generated for the whole program, then to encrypt an EBB, the IV is computed as $IV = IV_0 + EBB_{addr}$, and instructions are "xored" with the corresponding stream cipher sequence (as shown Table 2). IVs uniqueness is guaranteed if virtual memory isn't used, as for a given program there is only one instruction mapped to a given address.

The benefits of this method are that there is no impact on code size, and it makes decryption dependent on current processor state. This last property can be used to build control flow integrity checking mechanisms [2].

4.4.2. Interleaving IVs in code

Another approach is to interleave IVs within the instructions. The IV for the current EBB can be supplied using a known mem-

Table 2
Basic block encryption using counter mode IVs.

	Instruction	Encryption	Encryption context
			$s_0 \leftarrow \texttt{init}(IV_0 + EBB_{addr})$
i_0	`lw $t0, 0($a0)`	$i_0 \oplus r_0$	$(r_0, s_1) \leftarrow \texttt{genBits}(s_0)$
i_1	`li $t1, 1`	$i_1 \oplus r_1$	$(r_1, s_2) \leftarrow \texttt{genBits}(s_1)$
i_2	`bne $t0, $t1, dst`	$i_2 \oplus r_2$	$(r_2, s_3) \leftarrow \texttt{genBits}(s_2)$
	...	...	
i_n	`j dst`	$i_n \oplus r_n$	$(r_n, s_{n+1}) \leftarrow \texttt{genBits}(s_n)$

Table 3
Basic block encryption with interleaved IVs.

	Instruction	Encryption IV	Encryption context $s_0 \leftarrow \texttt{init}(IV)$
i_0	`lw $t0, 0($a0)`	$i_0 \oplus r_0$	$(r_0, s_1) \leftarrow \texttt{genBits}(s_0)$
i_1	`li $t1, 1`	$i_1 \oplus r_1$	$(r_1, s_2) \leftarrow \texttt{genBits}(s_1)$
i_2	`bne $t0, $t1, dst`	$i_2 \oplus r_2$	$(r_2, s_3) \leftarrow \texttt{genBits}(s_2)$
	...	...	
i_n	`j dst`	$i_n \oplus r_n$	$(r_n, s_{n+1}) \leftarrow \texttt{genBits}(s_n)$

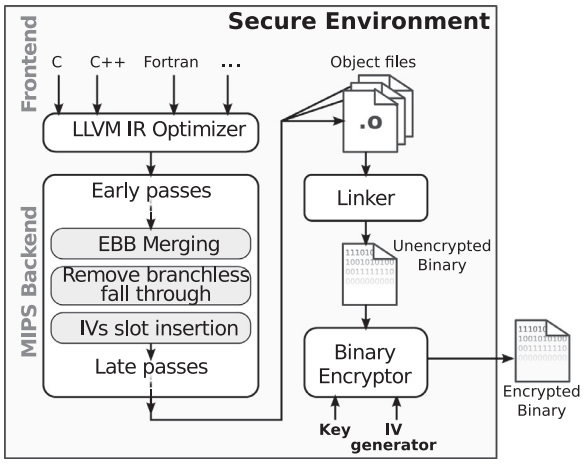


Fig. 5. Compilation flow for an encrypted program.

ory layout, as it would be done for a classic message transmission over an insecure channel. For example, the IV can be inserted at the beginning of each encryptable basic block, as shown in the example Table 3.

Being able to use arbitrary IVs is interesting in terms of security: for instance, in the context of an output feedback mode of encryption (generalized stream cipher), provable IND-CPA [30] encryption can be achieved. This also allows the use other modes of encryption, which require uniformly random IVs, like Cipher Block Chaining mode (CBC). Another interesting application is that it makes code sharing between programs straightforward (solving one of the issues addressed in [32]), so encrypted shared libraries can be generated and dynamically linked to.

However, encryption is not anymore dependent on the address of instructions, hence code can be relocated. Unfortunately, such programs are more prone to code reuse attacks, as any EBB can be moved or called from anywhere.

4.4.3. Combining the two approaches

A third option is to use a combination of the two previous schemes, a random IV is interleaved within the code, and combined with the current program counter value to generate the encryption sequence. This way, the code is more resilient to code reuse attacks, while still allowing other encryption modes to be used.

4.5. LLVM Integration

The full code encryption process is separated in two sequential parts, a control flow restructuring part implemented in the LLVM [4] compiler framework, followed by a second part that does the encryption. The motivation for this two-stage design is to support linkage of encrypted programs, and to statically guarantee IVs uniqueness across the entire program.

Three additional passes are implemented and inserted into LLVM's MIPS code generation backend. It would be much cleaner if they could be done on LLVM intermediate representation (middle-end). Unfortunately these passes make use of basic block placement information, which are generated in early backend passes. That being said, passes are very generic and can be easily ported to other RISC targets.

The first pass does the basic block merging, the second one searches and adds jumps between fall through basic blocks and the optional third one sets up the layout for IV storage. It allocates space in the code where IVs will be stored, at this stage, memory addresses are not computed yet so these slots can be inserted without breaking the address layout. Thanks to LLVM's highly modular design, the code still benefits from late optimization passes, including delay slot filling.

The encryption is done by a standalone program which takes as input a fully linked object file and produces the final encrypted binary that can be distributed securely to the processor. This tool reconstructs the control flow graph and internally runs a software version of the stream cipher (Trivium in our case). It is also responsible for choosing IVs for the whole program and ensures that all of them are unique.

Fig. 5 illustrates the complete compile flow. For the programmer, producing an encrypted binary boils down to: 1. adding a compiler flag while compiling sources to object files, 2. encrypt the final binary. Hence the encryption can be easily integrated in a standard build system like Make.

5. Hardware support

The architectural modifications required for the decryption are shown in Fig. 6. Instead of providing an instruction directly from memory to the decoding logic, this value is unmasked with a stream generated by an internal cipher. Furthermore, the processor branch handling is also modified to handle IV change. Whenever a branch is taken, the processor executes the following steps:

1. Latch the branch destination address in an additional register called PC_{prev}.
2. Compute the IV for the branch destination address. Depending on the IV selection scheme used (discussed 4.4) either read it from instruction memory (and skip several instructions), or compute it from the current program counter value. This step may span over several clock cycles.
3. Reset the stream cipher and wait for initialization to be done.
4. Continue execution as soon as the stream is ready

5.1. Handling exceptions/interruptions

An exception is an unexpected event during program execution (a division by zero, an invalid memory access, an external interruption, ...) which needs a special treatment before continuing. Most processors handle exceptions by jumping to an exception handler and putting the processor into a special mode. Once the exception is handled, the processor resumes its normal execution to the instruction where the exception was triggered.

At first glance, exceptions appear somewhat incompatible with the encryption mechanism described in Section 4.3. Jumping to an exception handler is perfectly fine: one just need to reset the

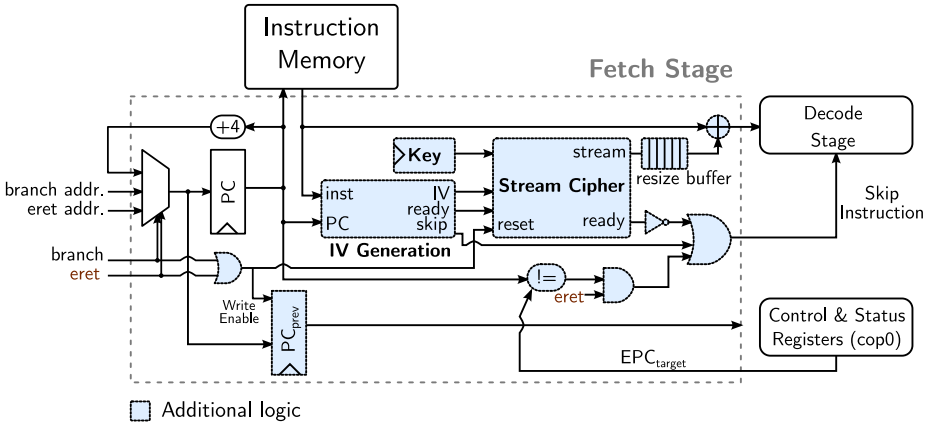


Fig. 6. Modification made to CPU's fetch stage for encryption support.

stream cipher with the IV used to encrypt the exception handler code. However, it might be the case that the execution resumes in the middle of an EBB, which breaks the definition 1.

Then, the processor must be able to restore or re-compute the correct stream cipher state when returning from the exception handler. A straightforward approach would be to just save the whole stream cipher state on exception, but it would require important storage and it is not satisfying in terms of security. Indeed, anyone can compute the full encryption sequence from a given stream cipher state (see Section 4.1), so it must be kept as private as the secret key. Another approach would be to mask exceptions until an EBB boundary is detected. But it seems rather difficult, as one would need additional data for runtime EBB reconstruction and it can be applied only to exception whose treatment can be deferred.

Instead, we investigated a simple strategy to restore the stream cipher state at any location in a program. For this, we remark that only the EBB start address and the current offset in the EBB need to be known to restore the stream cipher state. Indeed, execution can resume at the beginning of the EBB where the exception occurred, and instructions are skipped until the offset is reached. This way, the correct stream cipher state is regenerated and the execution can resume with the correct decryption stream.

The hardware support for this strategy, depicted Fig. 6, is made of three additional registers: PC_{prev} in the fetch stage, together with EPC_{prev} and EPC_{target} both as control and status registers. The register PC_{prev} is latched whenever a branch is taken to the destination address. On exception, PC_{prev} is saved into the control and status register EPC_{prev}. The latter is made readable from the software by overloading MIPS `mfc0` instruction[2] The additional register EPC_{target} is used to detect when the real execution must resume.

5.2. Context switching

In order to support context switching with encryption, the programmer must be able to save and restore the stream cipher state for the current process. As context switching relies on exceptions, the above solution can be reused easily. One just need to save and restore the two additional control and status registers (EPC_{prev} and EPC_{target}) as part of the context switching procedure.

6. Evaluation

6.1. Security analysis

As the CPU machine code is encrypted using a proven IND-CPA[3] encryption scheme, it benefits from security proofs of the underlying scheme. The key must be kept secret inside the processor and assumed free from observation and tampering. Under these hypothesis, an adversary just observing the instruction memory is equivalently viewing encrypted messages, that is, pairs of the form $(IV, \mathtt{Enc}_k(IV, m))$. In practice, it means that an adversary cannot distinguish between the encryption of two different instructions. Even if two instructions are the same or share some common parts like the opcode, the encryption will produce undistinguishable ciphertexts.

However, it is worth saying that encryption alone provide protection only against a very restrained attack model, in particular it does not cover:

- An adversary modifying memory (like Kuhn's attack [5]).
- Dynamic analysis of memory access patterns [36].

6.2. Compatibility with software integrity mechanisms

Software Integrity has been ignored through this work so far. Yet, it is a real concern (see Section 2). In particular when using a one-time pad encryption, ciphertexts can be easily tampered with. For instance, a destination register can be changed just by "xoring" the correct field in a ciphered instruction.

Fortunately, most integrity checking mechanisms from other secure processor architectures [31] can be applied on top of our encryption method (encrypt-then-authenticate paradigm). However instruction level integrity does not seem realistic: a tag would have to be associated with each instruction, resulting in a huge code size increase. The best solution seems to authenticate data per block of fixed length, either a cache line or a buffer of instructions if the system does not have a cache. To be fully effective, integrity checking has to be done before executing a complete block of instruction, to prevent any unchecked instruction from executing.

6.3. Hardware implementation

The hardware support described in Section 5 is implemented on a 32-bit MIPS [34] soft core. The processor itself is an integer only, in-order, five-stage pipeline, with 32 KB of read only in-

[2] The MIPS instructions `mfc0` and `mtc0` move data from and to control and status registers.

[3] The scheme used is IND-CPA under the assumption that the function F defined by $F_k(IV) = \mathtt{genBits}(\mathtt{init}(k, IV))$ is a pseudorandom function [30]

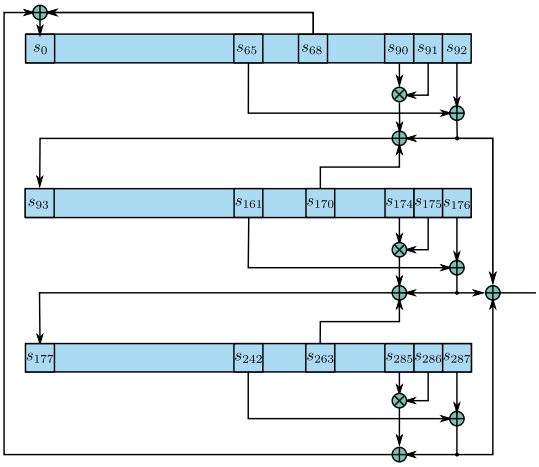


Fig. 7. Trivium stream cipher (one bit version).

Table 4
Synthesis results on Altera Cyclone V (5CEBA4F23C7N).

	Adaptative Logic Module (ALM)	f_{max}
tiny AES 128	3403 (18%)	189 MHz
Trivium_x1	148 (0.8%)	456 MHz
Trivium_x32	237 (1.2%)	360 MHz
Trivium_x64	288 (1.5%)	344 MHz
Trivium_x128	731 (3.9%)	227 MHz
CPU base	1094 (5.9%)	108 MHz
CPU enc	1379 (7.46%)	108 MHz

struction memory, and 32 KB of data RAM, both are single cycle latency memory (implemented using FPGA's BRAMs). The syntheses are done on a low-cost Altera FGPA from the Cyclone V family (5CEBA4F23C7N).

6.3.1. Trivium

Trivium, an eStream [6] hardware profile finalist, is used as the underlying stream cipher for code encryption because of its simplicity and efficiency. The one bit version depicted Fig. 7, is made up of a 288-bit register as its internal state with a few combinatorial gates. Trivium supports 80 bit-length key and IV. The initialization is done by loading the key and the IV in the internal state followed by several rounds (1152 for the one bit version) without outputting bits.

Obviously, the one bit version of this stream cipher is not well-suited for code encryption. Fortunately, Trivium can be unrolled so that several bits are generated per clock cycle. As a side effect, it also reduces the initialization latency. Thanks to the data dependencies of Trivium's combinatorial path, it can be unrolled up to 65 bits without increasing its circuit depth. Further unrolling can still be done, but it would increase the critical path and decrease the maximum frequency.

We synthesized Trivium for unrolling levels ranging from 1 up to 192 and measured the FPGA occupancy (in ALM) together with the initialization latency. The latter is computed using the formula $init_{latency}(p) = f_{\max} * 1152/p$, where p is the unrolling level. The results are shown in Fig. 8. The area usage seems linear in the unrolling level while the initialization latency stabilizes around 25 ns.

6.3.2. Processor with encryption support

Table 4 provides a set of synthesis results as well as maximum frequency obtained through static timing analysis. One can observe that Trivium, even unrolled, has a very small footprint and a very high maximum frequency.

As a comparison, a fully pipelined 128 bit AES implementation found on OpenCores is also synthesized, it has an initial latency of 21 cycles (very close to the 18 cycles needed by Trivium x64). Although it achieves a higher throughput than the Trivium implementations, it uses far more FPGA resources. It is more than three times bigger than our base processor, and more than ten times bigger than the Trivium x64. This illustrates why stream ciphers are such good candidates for the encryption.

The complete encryption mechanism increases overall FPGA occupancy by 1.5%, and the size of the core by 26%, mainly because of Trivium circuit and little additional control hardware. Interestingly, the encryption hardware does not affect CPU's critical path, hence, the maximum frequency of the circuit is unchanged.

6.4. Performance analysis

The following evaluation methodology is used: an input program is compiled using some constant compiler flags, with and without encryption. The two resulting programs are then compared under two criteria, code size and execution time (measured in CPU cycles). The flags used are `-O3`, which optimizes the input program for execution speed, and `-Oz`, which optimizes for size.

These measurements are done for the two different IV selection schemes described in Section 4.4: IVs computed only from program counter, and IVs interleaved in code. A set of relevant programs was selected to run the above measurements. Most of them are based on open-source libraries and can be easily ported on any embedded processors. Raw results are given in Table 5.

6.4.1. Code size overhead

Results for size overhead are represented in Fig. 9. Interestingly, even when storing IVs at the beginning of each basic block, the binary size does not increase by more than 40%. When IVs are not interleaved there is still a binary size increase due the basic block restructuring, but the observed increase does not exceed 11%.

Compiler options have a clear impact on the results. Our interpretation is that speed optimizations (enabled with option `-O3`) perform aggressive inlining and unrolling which increase basic blocks sizes, hence reduce the impact of encryption (in particular if IVs are interleaved).

6.4.2. Execution time overhead

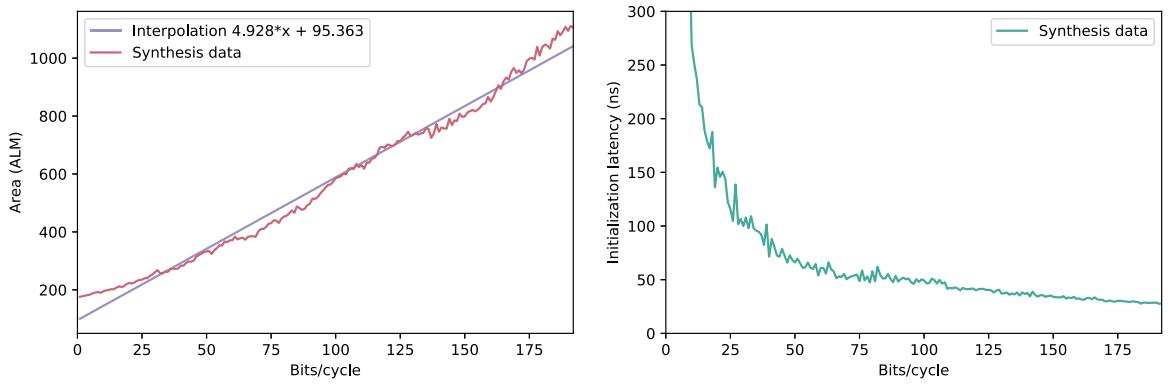
The encryption mechanism also introduces a runtime overhead, more precisely latency is added for each branch taken, because the stream cipher has to be initialized with a new IV. The CPU cannot be fed with new instructions while the stream is not ready. With the stream cipher used in these experiments, trivium_x64 running at CPU's clock, this latency is 18 cycles.

The results given in Fig. 10 show that the slowdown ranges from 29% up to 193%. As a comparison, the authors in [26] observed an average overhead of 60% for decryption in a processor without cache. As expected, the overhead is minimized with performance optimizations (`-O3`) and when IVs are computed from PC. In that case, the average slowdown is of 48%. When IVs are interleaved in code, the processor has to skip some (3 in our case) instructions at the beginning of each basic block, so performances are further reduced. This overhead is likely to stay reasonable unless critical loops contain an important number of jumps.

We stress that our experimental processor does not include any cache memory, hence only the effect of the encryption is taken into account. Performances are expected to be better with an instruction cache, as the fetch on a cache miss can overlap with stream cipher initialization.

6.5. Pitfalls and further work

Our solution still suffers some pitfalls that we tried to identify as best as possible. First, a small hardware support must be



(a) Area (b) Initialization delay

Fig. 8. Evolution of Trivium area and initialization latency for different unrolling levels.

Table 5
Performance and Size Overhead Results.

	IVs from PC				IVs interleaved in code			
	LLVM -Oz		LLVM -O3		LLVM -Oz		LLVM -O3	
Benchmark	size	time	size	time	size	time	size	time
AES	+5 %	x2.39	+2.3%	x1.29	27.4%	x2.93	7.8 %	x1.41
SHA1	+6.6%	x2.10	+6.8%	x1.56	35.3%	x2.52	28.6%	x1.76
Quicksort	+9.5%	x2.26	+11 %	x1.59	35.6%	x2.75	30.1%	x1.82
uECC	+7.4%	x2.16	+7.2%	x1.5	38.1%	x2.61	35.7%	x1.70
Average	+7.12%	x2.23	+6.8%	x1.48	34.1%	x2.70	25.5%	x1.67

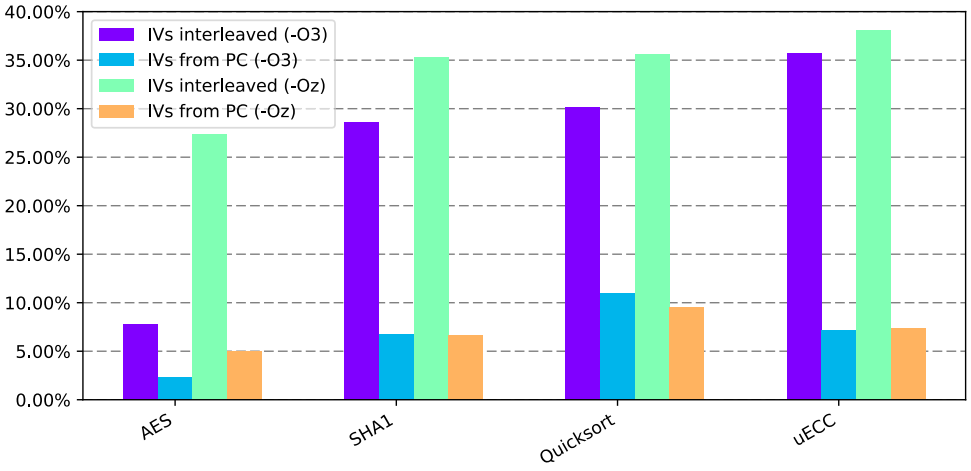


Fig. 9. Overhead factor on program size.

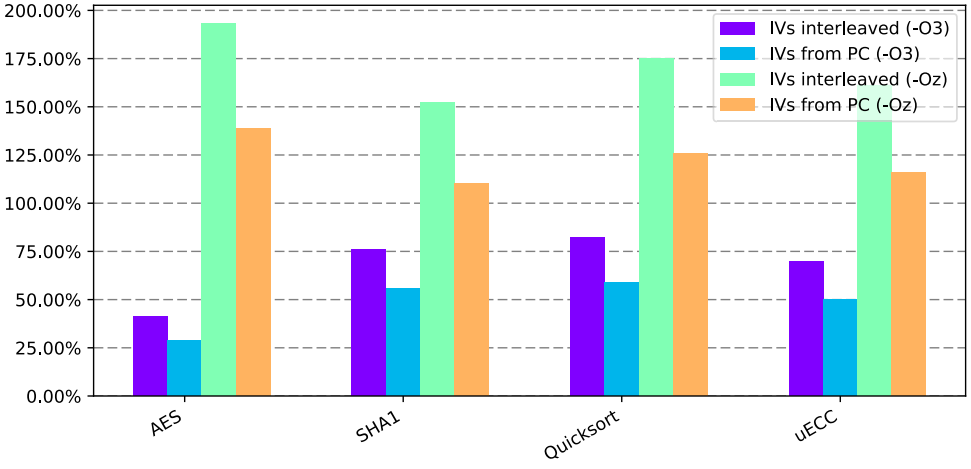


Fig. 10. Overhead factor on execution time.

added on-chip, which restrains the range of target systems. From a security perspective, the protection is static, and does not cover dynamic aspects, e.g., an adversary that would track and analyze memory access patterns (control flow analysis).

Further work could be done to improve performances. We saw that this kind of encryption is highly compiler dependent, this suggests that the basic block placement algorithm could be modified to maximize the basic block merging. The hardware implementation described in this paper is simple, significant speed-up is likely to be achieved with a more evolved architecture. This could be a stream cipher working at higher frequency than the CPU's clock, or to couple the stream cipher with branch prediction to begin initialization ahead.

7. Conclusion

This paper describes an efficient method to encrypt a binary program with a stream cipher. The decryption is so fast and lightweight that it can be performed very deeply in the processor, so that plain instructions remain only in processor execution logic. The method requires slight hardware and software modification, which are implemented and evaluated on FPGA. The results are promising and open interesting perspectives in order to improve performances and increase the range of applications.

Acknowledgements

The authors would like to thanks the anonymous reviewers of the DSD conference and MICPRO journal for their useful comments and suggestions that helped us to improve this work.

References

[1] T. Hiscock, O. Savry, L. Goubin, Lightweight software encryption for embedded processors, in: 2017 Euromicro Conference on Digital System Design (DSD), IEEE, 2017, pp. 213–220.
[2] R. De Clercq, R. De Keulenaer, B. Coppens, B. Yang, P. Maene, K. De Bosschere, B. Preneel, B. De Sutter, I. Verbauwhede, Sofia: software and control flow integrity architecture, in: Design, Automation & Test in Europe Conference & Exhibition (DATE), 2016, IEEE, 2016, pp. 1172–1177.
[3] J.-L. Danger, S. Guilley, F. Praden, Hardware-enforced protection against software reverse-engineering based on an instruction set encoding, in: Proceedings of ACM SIGPLAN on Program Protection and Reverse Engineering Workshop 2014, in: PPREW'14, ACM, 2014, pp. 5:1–5:11, doi:10.1145/2556464.2556469.
[4] C. Lattner, V. Adve, LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation, in: Proceedings of the 2004 International Symposium on Code Generation and Optimization (CGO'04), 2004. Palo Alto, California
[5] M.G. Kuhn, Cipher instruction search attack on the bus-encryption security microcontroller DS5002fp, IEEE Trans. Comput. (10) (1998) 1153–1157.
[6] M. Rogawski, Hardware evaluation of eSTREAM candidates (2007).
[7] M. Gruhn, T. Muller, On the practicability of cold boot attacks, in: Availability, Reliability and Security (ARES), 2013 Eighth International Conference on, 2013, pp. 390–397, doi:10.1109/ARES.2013.52.
[8] A. Boileau, Hit by a bus : physical access attacks with firewire, 2006. Available at https://www.security-assessment.com/files/presentations/ab_firewire_rux2k6-final.pdf.
[9] A. Huang, et al., Keeping secrets in hardware: the microsoft XBox™case study, Cryptographic Hardware and Embedded Systems (CHES) 2523 (2002) 213–227.
[10] O. Kömmerling, M.G. Kuhn, Design principles for tamper-resistant smartcard processors, USENIX workshop on Smartcard Technology, 1999.
[11] R. Anderson, M. Bond, J. Clulow, S. Skorobogatov, Cryptographic processors a survey, Proc. IEEE (2006).
[12] C. Linn, S. Debray, Obfuscation of executable code to improve resistance to static disassembly, in: Proceedings of the 10th ACM Conference on Computer and Communications Security, 2003.
[13] B. Barak, O. Goldreich, R. Impagliazzo, S. Rudich, A. Sahai, S. Vadhan, K. Yang, On the (im) possibility of obfuscating programs, in: Annual International Cryptology Conference, Springer, 2001, pp. 1–18.
[14] P. Junod, J. Rinaldini, J. Wehrli, J. Michielin, Obfuscator-LLVM – software protection for the masses, in: Proceedings of the IEEE/ACM 1st International Workshop on Software Protection, SPRO'15, 2015.
[15] R. Best, Microprocessor for executing enciphered programs, 1979, US Patent 4,168,396. Patent https://patents.google.com/patent/US4278837.
[16] R. Best, Crypto microprocessor for executing enciphered programs, 1981, US Patent 4,278,837. Patent https://patents.google.com/patent/US4278837.
[17] D. Semiconductor, DS5002FP secure microprocessor chip.
[18] D. Lie, C. Thekkath, M. Mitchell, P. Lincoln, D. Boneh, J. Mitchell, M. Horowitz, Architectural support for copy and tamper resistant software, ACM SIGPLAN Not. 35 (11) (2000) 168–177.
[19] X. Zhuang, T. Zhang, S. Pande, Hide: an infrastructure for efficiently protecting information leakage on the address bus, in: ACM SIGPLAN Notices, 39, ACM, 2004, pp. 72–84.
[20] G.E. Suh, C.W. O'Donnell, S. Devadas, AEGIS: a single-chip secure processor, Inf. Secur. Tech. Rep. 10 (2) (2005) 63–73.
[21] G. Duc, R. Keryell, Cryptopage: an efficient secure architecture with memory encryption, integrity and information leakage protection, in: Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual, IEEE, 2006, pp. 483–492.
[22] B. Rogers, S. Chhabra, M. Prvulovic, Y. Solihin, Using address independent seed encryption and bonsai merkle trees to make secure processors os-and performance-friendly, in: Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture, IEEE Computer Society, 2007, pp. 183–196.
[23] W.-j. HUO, Z.-l. LIU, X.-c. ZOU, Pem: a lightweight program memory encryption mechanism for embedded processor, J. China Univ. Post. Telecommun. 17 (1) (2010) 77–84.
[24] W. Shi, H.-H.S. Lee, M. Ghosh, C. Lu, A. Boldyreva, High efficiency counter mode security architecture via prediction and precomputation, in: ACM SIGARCH Computer Architecture News, 33, IEEE Computer Society, 2005, pp. 14–24.
[25] J. Yang, Y. Zhang, L. Gao, Fast secure processor for inhibiting software piracy and tampering, in: Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture, IEEE Computer Society, 2003, p. 351.
[26] V. Nagarajan, R. Gupta, A. Krishnaswamy, Compiler-assisted memory encryption for embedded processors, in: International Conference on High-Performance Embedded Architectures and Compilers, Springer, 2007, pp. 7–22.
[27] A. Papadogiannakis, L. Loutsis, V. Papaefstathiou, S. Ioannidis, ASIST: architectural support for instruction set randomization, in: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, ACM, 2013, pp. 981–992.
[28] G.S. Kc, A.D. Keromytis, V. Prevelakis, Countering code-injection attacks with instruction-set randomization, in: Proceedings of the 10th ACM conference on Computer and communications security, ACM, 2003, pp. 272–280.
[29] D. McGrew, Counter mode security: analysis and recommendations, Cisco Syst. Novemb. 2 (2002) 4.
[30] J. Katz, Y. Lindell, Introduction to modern cryptography, 2nd, Chapman & Hall/CRC, 2014.
[31] M. Henson, S. Taylor, Memory encryption: a survey of existing techniques, ACM Comput. Surv. (CSUR) (2014).
[32] S. Chhabra, B. Rogers, Y. Solihin, M. Prvulovic, Making secure processors os-and performance-friendly, ACM Trans. Arch. Code Optim. (TACO) 5 (4) (2009) 16.
[33] J.L. Hennessy, D.A. Patterson, Computer architecture, fifth edition: A Quantitative approach, 5th, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.
[34] I. Technologies, MIPS32 instruction set reference v5.04, available at https://www.mips.com/?do-download=the-mips32-instruction-set-v5-04.
[35] W.-M.W. Hwu, S.A. Mahlke, W.Y. Chen, P.P. Chang, N.J. Warter, R.A. Bringmann, R.G. Ouellette, R.E. Hank, T. Kiyohara, G.E. Haab, et al., The superblock: an effective technique for VLIW and superscalar compilation, in: Instruction-Level Parallelism, Springer, 1993, pp. 229–248.
[36] M.S. Islam, M. Kuzu, M. Kantarcioglu, Access pattern disclosure on searchable encryption: ramification, attack and mitigation., in: NDSS, 20, 2012, p. 12.

Thomas Hiscock is graduated from Grenoble-INP Phelma in 2014. He obtained his PhD in 2017 in computer science. He is currently working at CEA-LETI, on research topics including hardware security, embedded systems and secure processor design.

Dr. Olivier Savry benefits from a PhD in microelectronics that followed a diploma from an electrical engineering schools ENSE3/ INP Grenoble France. He also received three MSc degrees in Signal Processing, Microelectronics from Universities of Grenoble and Astrophysics from ENS Lyon. Olivier has been working at CEA LETI for several years where he contributed to projects in the field of security of communicating objects especially around RFID and its privacy protection issue. He was involved in European projects like Discreet, IoT-A, Sociotal and many others national projects trying to mitigate hardware security flaws and developing new countermeasures based on cryptography and physical solutions in IoT. Its field of expertise relies also in the security analysis of cyber physical systems and the design of their best required architecture for industrial partners in domains like Automotive, Energy, Manufacturing.

Louis Goubin is a Professor at Versailles St-Quentin-en-Yvelines University. A former student of the \'Ecole normale supérieure (Paris), he holds a PhD in Pure Mathematics from Paris XI University (1995) and an Habilitation to supervise research from Paris VII University (2003). He published more than 80 papers, about the design of new asymmetric cryptosystems, cryptanalysis of existing algorithms and protocols, and the protection of software implementations against physical attacks. He has also filled more than 20 patents on practical cryptology and smart cards.