Time Is Money, Friend!
Timing Side-channel Attack against Garbled Circuit Frameworks

Mohammad Hashemi
mhashemi@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

Domenic Forte
dforte@ece.ufl.edu
University of Florida
Gainesville, FL, USA

Fatemeh Ganji
fganji@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

ABSTRACT

With the advent of secure function evaluation (SFE), distrustful parties can jointly compute on their private inputs without disclosing anything besides the results. Yao's garbled circuit protocol has become an integral part of secure computation thanks to considerable efforts made to make it feasible, practical, and more efficient. These efforts have resulted in multiple optimizations on this primitive to enhance its performance by orders of magnitude over the last years. The advancement in protocols has also led to the development of general-purpose compilers and tools made available to academia and industry. For decades, the security of protocols offered in those tools has been assured with regard to sound proofs and the promise that during the computation, no information on parties' input would be leaking.

In a parallel effort, however, side-channel analysis (SCA) has gained momentum in connection with the real-world implementation of cryptographic primitives. Timing side-channel attacks have proven themselves effective in retrieving secrets from implementations, even through remote access to them. Nevertheless, the vulnerability of garbled circuit frameworks to timing attacks has, surprisingly, never been discussed in the literature. This paper introduces Goblin, the first timing attack against commonly employed garbled circuit frameworks. Goblin is a machine learning-assisted, non-profiling, single-trace timing SCA, which successfully recovers the garbler's input during the computation under different scenarios, including various GC frameworks, benchmark functions, and the number of garbler's input bits. Furthermore, we discuss Goblin's success factors and countermeasures against that. In doing so, Goblin hopefully paves the way for further research in this matter.

CCS CONCEPTS

• **Security and privacy → Side-channel analysis and countermeasures**;

KEYWORDS

Multiparty Computation; Garbled Circuits; Timing Side-channel Analysis; Clustering; Non-profiling Attack; Single-trace Attack.

1 INTRODUCTION

Secure function evaluation (SFE) has had an immense impact on the field of cryptography. Practical implementations of general SFE have been proposed and flourished after the introduction of garbled circuits (GCs) by Yao [108]. It has found several applications including secure multi-party computation [13, 33, 34, 68, 108], functional encryption [38, 39, 92], key-dependent message security [7, 9], homomorphic encryption [37, 87], and recently, quantum circuits [16]. The key premise of GCs is that it allows two parties to evaluate any (known) function on their respective inputs x and y without violating their privacy. Besides real-world applications foreseen for GCs traditionally (e.g., credit evaluation function, background- and medical history checking, privacy-preserving database querying, etc. [64, 94]), nowadays GCs have found applications in privacy-preserving genome analysis [53], email spam filtering [47], image processing [19] and machine learning and statistical analysis [22, 35, 80, 85], just to name a few. To become practical, GCs have undergone dramatic improvements and optimization to reduce the computation and communication costs (proportional to the size of the circuit). Thanks to the increase in hardware performance and the improvement in GC algorithms themselves, one of the main bottlenecks for these protocols has become the network bandwidth needed to transmit the garbled gates [12, 112].

To face obstacles preventing further adoption of GCs in real-world systems, optimization techniques have been developed, aiming to reduce communication and computation costs. Here we focus on two of the most acknowledged methods, namely free-XOR [64] and half-gates [112]. Similar to other optimization mechanisms, the main argument put forward by these techniques is that security is not compromised for the sake of being efficient. However, the question is whether this holds true when implementing these protocols. This becomes even more critical since today's applications of GCs (or potential ones) encompass services run on distributed computing systems, cloud services, connected devices, etc.

Timing side-channel analysis. Irrespective of what cryptographic functions are embedded in programmable instruction set processors, such systems can exhibit observable features and data-dependent behavior that leak information about users' data/keys from the implementation. Side-channel attacks leverage this information through analyzing execution time [17, 28], power consumption [62], instruction or data cache behavior [1, 14, 88, 109], branch predictor behavior [5], pipeline instruction and execution behavior as well as pipeline speculation behavior [105].

Timing side-channel analysis (SCA) has been launched to deduce information on the secret, e.g., keys, security tokens, and passwords [63, 74, 81, 104]. In general, timing side channels can be observed when the time taken to execute a piece of code depends on the secret variables. The temporal behavior of a code may depend on the control flow of a program, on its data flow, and on its contention over resources that the program has to share with other running programs [24]; therefore, any of these depends on the secret, timing SCA can be launched. In this regard, two broad categories of timing side channels can be identified: instruction-related and cache-related cf. [107]. The former refers to the number or type of instructions executed along a path that can differ depending on the values of secret variables, leading to differences in

the number of computer processor unit (CPU) cycles. On the other hand, cache-related timing side channels correspond to the case, where the memory subsystem may behave differently based on the values of secret variables. In both categories, CPU instruction execution, specifically the branch prediction, memory access, and data caches, have been exhibited to leak adequate timing side-channel information and launch successful SCA on the cryptographic systems cf. [3, 14, 31, 40, 89]. As prime examples, the branch predictions [5] and memory accesses [1] are dependent on the inputs of the job/process, i.e., the execution time of the same job/process is input-dependent [78]. Recently, the security of open-source cryptographic libraries and implementations of protocols (excluding GC) has just been evaluated in an extensive study [55], where the vulnerability of some of those libraries to timing SCA has been demonstrated. In this regard, more interesting and inspiring from the perspective of this work is the gap between academic research and cryptographic engineering when it comes to timing SCA.

SCA against GC constructions. Despite the achievements made to define the adversary models, prove the security, and construct numerous GC schemes, there is a gap between what theoretical findings have suggested and what observations can be made by parties involved in executing an GC protocol. The only example of studies addressing this gap is a recent attack proposed by Levi et al., which leverages the side-channel leakage as a result of three main shortcomings as enumerated in [66]: (1) a secret, global value is used to perform free-XOR, and (2) power consumption of the garbler's device can be linked to this secret value. Although multiple assumptions have been made to launch the attack, their attack has successfully disclosed the global value used to perform free-XOR optimization. For this, they have taken advantage of the leakage from the garbler's transmitted labels along with the power leakage from non-linear gates. Now that the possibility of SCA against free-XOR-optimized GC implementation has been indicated, the question is whether one can go even beyond that attack and perform timing SCA and whether some of the assumptions made in [66] can be relaxed in that case.

Generally speaking, timing attacks feature outstanding properties that make them more interesting compared to other types of SCA, e.g., power and electromagnetic (EM) attack cf. [55]. First and foremost, timing attacks can be launched remotely, including cases of running code in parallel to the victim code without the need for local access to the target computer; hence, restricting physical access to the target machine cannot prevent timing attacks. Second, timing attacks can be carried out covertly. In light of this state of affairs and of the fact that timing attacks against GC construction have never been discussed in the literature, this work attempts to answer the following question: *Is it possible to reveal parties' input by observing the timing information leaking when executing an GC protocol?* More specifically, we answer this question positively for free-XOR- and half-gates-optimized constructions. The contribution of our work is as follows.

Contributions. Our contributions are summarized as follows.

(1) We introduce *Goblin*, the first non-profiling, single-trace timing SCA that successfully extract the user's input, which by definition, should have been kept secret. To better demonstrate the power of our attack, we compare it with the recent attack in [66]. The power SCA in [66] has successfully extracted the global secret used in free-XOR optimization, whereas Goblin focuses entirely on the recovery of the garbler's input. Needless to say that even with the help of the disclosed secret, the garbler's input could not be fully recovered. Moreover, in contrast to [66], Goblin's effectiveness is limited to neither circuits with a minimum number of input gates nor gate types (XOR or AND).
(2) Goblin is machine-learning assisted in disclosing the garbler's input, regardless of its size. For this purpose, k-means clustering is applied, where no manual tuning or heuristic leakage models are needed. It is, of course, advantageous to the attacker and allows for scalable and efficient attacks.
(3) Last but not least, our paper highlights the vulnerabilities of multiple available garbling tools to timing SCA. We believe that this constitutes a basis for studying the SCA with respect to GC.

2 BACKGROUND

Notations. We follow a standard notation typically used in SFE-related literature. $\in_R$ denotes uniform sampling, $\|$ is used to show concatenation of bit strings. $\langle a, b\rangle$ represents a vector with two components a and b, whereas $a \parallel b$ is its bit string representation. A *gate* is denoted by $W_c = g(W_a, W_b)$ with input wires W_a and W_b, output wire W_c and $g: \{0,1\}^2 \rightarrow \{0,1\}$.

2.1 Yao's Garbled Circuit (GC)

One of the most widely studied SFE approaches, designed to meet the needs of Boolean circuits, is garbling [67, 69]. This section gives a brief overview of GC building blocks.

Oblivious transfer (OT). The first protocol within the context of GC is OT. We consider 1-out-of-2 OT, which is a two-party protocol with the following definition. The sender P_1 posses two secret messages m_0, and m_1, and the receiver P_2 has a selection bit $i \in \{0,1\}$. By executing the protocol, P_2 learns m_i, but not m_{1-i}, while the sender P_1 does not learn anything about i.

Garbling. One of the main components of GC is the primitive associated with the cryptographic operation, often referred to as "encryption," namely hashing or symmetric key operations, e.g., fixed-key block cipher. The protocol execution begins with garbling the circuit C, where the garbler (P_1) randomly chooses secrets w_i^j with the garbled value of $j \in \{0,1\}$ on each wire W_i. Needless to say that it is expected that w_i^j does *not* reveal any information about j. Practical implementations of Yao's GC, e.g., [99] considered in this paper, represent each of the logical "0" and "1" values with n-bit values, where n is often referred to as the security parameter. In this sense, w_i^j (so-called token) is the encryption of the concatenation of j and $(n-1)$-bit values drawn uniformly. After generating the tokens, the garbler creates a garbled table T_i for each gate G_i, where each row of the gate truth table is encrypted output with regard to the tokens, and the output of the gate is called a "ciphertext," illustrated in Figure 1.(a) as the output of the operand $E(\cdot)$, i.e., the encryption operation. Since the table rows can reveal information about the internal wire values, they are permuted so that the recovery of the output labels does not result in uncovering the garbler input. The main property of T_i is that its output can be recovered given a set of garbled inputs, while this process does not leak any

information about the garbler's and evaluator's (P_2) inputs. For this, along with T_i's, the token corresponding to the garbler's input value is obliviously transferred to P_2 through OT. P_2 is then able to obtain the garbled output by evaluating the garbled circuit gate by gate using the tables T_i and receiving j for the output wire from P_1 cf. [100]. Garbling of the output wires of the circuit can be skipped so that two parties learn (only) the output of the circuit [64].

2.2 Optimizations of Yao's GC

Reducing the computation and communication costs of SFE protocols has been an objective of numerous studies. The legacy construction of GCs requires four garbled values per gate, corresponding to the combinations of values on the input wires. To reduce this cost on the evaluator and/or garbler side, various optimization methods have been introduced in the literature.

Among optimization techniques introduced in the literature, **free-XOR** has attracted considerable attention since it reduces the cost on the garbler side effectively, namely by 25%. To reduce garbler's cost, the wire values are garbled as presented in Figure 1.(b). For any gate G_i, $w_i^1 = w_i^0 \oplus R$ for some secret, global $R \in_R \{0,1\}^n$[1]. Here, for the sake of simplicity, let $(A, A \oplus R)$ and $(B, B \oplus R)$ denote the wire labels.

half-gates protocol complements the free-XOR protocol in the sense that not only are XOR gates evaluated for free, but also AND gates are garbled using only two ciphertexts (see Figure 1.(c)). Since Goblin is interested in recovering the garbler's input, in Figure 1.(c), we show how the half-gates are generated on the garbler's side, where garbler knows which inputs she wants to garble (for more information about the whole process, see [112]).

2.3 *k*-means Algorithm

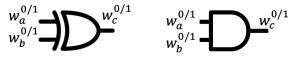
In general, clustering algorithms are mostly in the category of unsupervised machine learning. The main goal of clustering algorithms is to group samples of a set with some common features into subsets, i.e., *clusters*. The similarity of all members of each cluster is measured based on the pairwise distances of values (so-called features) [48]. With regard to the pairwise distances, clusters are then made around the mean vectors, which are called *centroids* [106].

k-means is a clustering algorithm that aims to partition N members of a set into k clusters in a way that each member of a cluster has a close value to the centroid of the cluster [106]. The process of clustering data seems to be simple, but in fact, it is an NP-hard problem [106]. Hence, there exist heuristic algorithms that find a local optimum centroid over a series of finding iterations [106]. To be more specific, k-means finds partitions (clusters) $p = \{p_1, p_2, \cdots, p_k\}$ for the dataset $c = \{c_i\}_{i=1}^n$ that minimizes the total cluster variance [50]:

$$\min_{p,\{\mu_j\}_1^k} \sum_{j=1}^{k} \sum_{c_i \in p_j} ||c_i - \mu_j||^2, \tag{1}$$

where μ_j is the mean of all examples assigned to j^{th} centroid. Here the squared Euclidean distance is one of the commonly applied distance measures applied to minimize the total cluster variance [98].

[1]For specifics of the encryption function in the free-XOR protocol, see [21, 45].



Input	Garbled Input	XOR Output	Garbled XOR Output	AND Output	Garbled AND Output
0,0	w_a^0, w_b^0	0	$E_{w_a^0,w_b^0}(w_c^0)$	0	$E_{w_a^0,w_b^0}(w_c^0)$
0,1	w_a^0, w_b^1	1	$E_{w_a^0,w_b^1}(w_c^1)$	0	$E_{w_a^0,w_b^1}(w_c^0)$
1,0	w_a^1, w_b^0	1	$E_{w_a^1,w_b^0}(w_c^1)$	0	$E_{w_a^1,w_b^0}(w_c^0)$
1,1	w_a^1, w_b^1	0	$E_{w_a^1,w_b^1}(w_c^0)$	1	$E_{w_a^1,w_b^1}(w_c^1)$

(a)

Input	Garbled Input	XOR Output	Garbled XOR Output	AND Output	Garbled AND Output
0,0	A_i^0, B_i^0	0	$w_c^0 = A_i^0 \oplus B_i^0$	0	$E_{A_i^0,B_i^0}(w_c^0)$
0,1	$A_i^0, B_i^0 \oplus R$	1	$w_c^0 \oplus R$	0	$E_{A_i^0,B_i^0 \oplus R}(w_c^0)$
1,0	$A_i^0 \oplus R, B_i^0$	1	$w_c^0 \oplus R$	0	$E_{A_i^0 \oplus R,B_i^0}(w_c^0)$
1,1	$A_i^0 \oplus R, B_i^0 \oplus R$	0	w_c^0	1	$E_{A_i^0 \oplus R,B_i^0 \oplus R}(w_c^0 \oplus R)$

(b)

Known input	Other input	Garbled Input	XOR output	Garbled XOR output	AND Output	Garbled AND output
0	0	A_i^0, B_i^0	0	$w_c^0 = A_i^0 \oplus B_i^0$	0	$E_{B_i^0}(w_c^0)$
	1	$A_i^0, B_i^0 \oplus R$	1	$w_c^0 \oplus R$		
1	0	$A_i^0 \oplus R, B_i^0$	1	$w_c^0 \oplus R$	0	
	1	$A_i^0 \oplus R, B_i^0 \oplus R$	0	w_c^0	1	$E_{B_i^0 \oplus R}(w_c^0 \oplus R)$

(c)

Figure 1: Garbled gates look-up table with (a) no optimization, (b) free-XOR optimization, and (c) half-gate optimization.

2.4 Cache Architecture

Modern x86 processors include three layers of cache [83]: L1, L2, and L3. All cache levels are mutually inclusive, which means every available data on L1 is also available in L2 and L3 [36]. Figure 2 illustrates Intel core-i7 cache architecture. Each central processing unit's (CPU) core has dedicated L1 and L2 caches, where the L1 cache is divided into a data and an instruction cache with 32 kilobytes (KB) capacity each in Intel core-i7 CPU [83]. Moreover, the L2 cache is shared between all the CPU threads; therefore, it has a bigger capacity (256 KB [83]) in Intel core-i7 CPU. All CPU cores can access shared cache L3, the largest CPU cache with 8 megabytes (MB) capacity [83].

Memory access time variations. Each procedure of x86 processors contains two main instructions categories: (1) read/write data from/to memory and (2) process the data (so-called control flow) [23, 36]. In the latter case, the execution time depends on the instruction type and equals the number of an arithmetic-logic unit (ALU) calls during the instruction execution [23]. However, memory access time depends on either the instruction tries to access random-access memory (RAM) or any level of cache [43, 75, 82, 93, 109]. For instance, an additional instruction between data a and b takes fewer clock cycles when both data are available in the L1 cache compared to the case when either of the data is in a higher level cache or the worst case, is stored in the RAM.

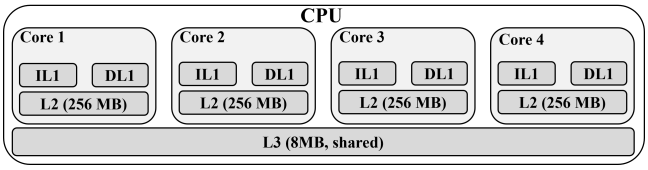


Figure 2: Intel core-i7 cache architecture [83].

Memory storage hierarchy. To tackle the issue of the CPU's insufficient storage, one of the CPU's primary duties is efficiently managing instructions' access to memory, especially the cache. When the CPU executes an instruction over a data, the resulting data is first stored in L1, and inclusively in L2 and L3, and cache [65, 114]. Moreover, the CPU stores an instantiation of the data on the RAM. From then on, if any task on the same core as the data is generated requests the data, the CPU first checks the availability of data in the core's L1 cache level. If the allocated data address memory cells are overwritten in the L1 cache by other tasks assigned to the core, the CPU checks the availability of the data in the L2 and L3 cache levels. Finally, if the data is unavailable in the cache, the CPU fetches the data from RAM down to the L1 level of the cache assigned to the core on which the task requests the data [65].

3 ADVERSARY MODEL

Before giving details about our adversary model, it is necessary to review the models discussed for GC protocols.

Adversary models in the context of SFE. After reviewing how SCA, in particular, timing SCA, has been studied in other domains, we shift our focus to how this topic is relevant in the context of SFE. The security of GCs has been considered in two main paradigms, namely honest-but-curious and malicious adversary models. The latter reflects the situation, where a party potentially cheats by corrupting the function to be jointly computed or, generally, adopting an arbitrary attack strategy. On the other hand, honest-but-curious parties follow the protocol honestly, although they may attempt to learn additional information from the execution. From this definition and what has been discussed about SCA, it is clear that an adversary capable of performing SCA can be classified as an honest-but-curious one. This has also been well-formulated in [11], where it is suggested that Yao's GC reveals no side-information beyond the function being computed, i.e., no information about parties' inputs leaks. Nevertheless, it should be noted that these adversary models were developed to reflect the conventional setting assumed for SFE, namely, parties' symmetric computation power.

To encompass all aspects of real-world applications of SFE, the notion of server-aided or cloud-assisted SFE has been introduced. In this setting, the standard SFE protocol is run with the help of a server (or a small set of them), which does not contribute to running the protocol by giving inputs, but by making their computational resources available to the parties cf. [15, 25, 26, 59]. In spite of all difficulties in deploying even *single*-server-aided protocols relying on Yao's GC, it has been demonstrated that it is indeed possible to construct practical ones [59]. In the proposed setting, the server is instantiated by a public cloud service provider, where parties who need more computational power (e.g., the garbler) can outsource their computations. An intrinsic part of the proposed methodology is the application of optimization techniques, namely free-XOR. [59] has indeed well explained how any two-party SFE can be converted to a server-aided protocol if the server and the garbler are not simultaneously *malicious*. The proposed protocols have been proven secure under various circumstances, where parties and the server can be of different adversary types, including honest-but-curious server [60]. Nonetheless, as further explained in Section 3.1, although the adversary model assumed in studies devoted to server-aided protocols might seem the most relevant to ours, any -even unprivileged- access to the CPU running the SFE protocol can be enjoyed by the adversary.

Privileged vs. unprivileged access for timing SCA. To access fine-grained timing side-channel observations, both privileged and unprivileged adversaries have been considered in the literature; see, e.g., [20]. In the former case, one possible scenario constitutes privileged attackers in control of the operating system. In fact, such an attacker can be capable of launching more direct attacks than side-channel attacks [78]; however, platforms supporting shielded execution, e.g., Intel Software Guard eXtension (SGX), may not come under attack by even most privileged adversaries. Hence, SCA in the presence of privileged attacker, for instance, through intercepting the control flows, inferring page table and last-level cache, has become an important research direction [20, 30, 86]. Nevertheless, unprivileged attackers still attract attention as they can mount attacks in various settings, e.g., an unprivileged process in a native environment, an unprivileged process in a virtual machine, and a sandboxed process cf. [20, 42]. Examples of such attacks include [1, 2, 84, 90, 103, 113]. In this respect, `rdtsc` exemplifies the instructions used to obtain fine-grained timing information by providing unprivileged access to a model-specific register, which stores the current cycle count, commonly used for cache attacks on, e.g., Intel CPUs cf. [72].

3.1 Our Adversary Model

The attack model taken into consideration in this paper is the semi-honest model, where parties follow the protocol, but there is an attempt to learn information from the execution of the protocol. In doing so, even though the (single) server-aided SFE seems to be the most relevant adversary model from the SFE protocols' perspective, we consider the challenging setting where the server itself is honest-but-curious. Moreover, similar to models taken into account in [32, 58, 59], we assume that the parties and the server are independent, meaning that none of them collude. In practice, given the consequences in terms of losing the reputation and legal actions, it is reasonable to assume that the server will not collude with the parties. Note that although throughout the paper, we refer to the server as the entity collecting the timing information, this does not rule out the fact that any entity having access to the timing information can launch the attack. The main requirement for Goblin to be launched successfully is the ability to acquire fine-grained timing information when the garbling process is run. The attacker aims to take advantage of this information to extract the garbler's input. Next we explain how this can be possible in real-world implementations.

4 TIMING SIDE-CHANNEL LEAKAGE IN GARBLING TOOLS: AN OBSERVATION

Broadly speaking, timing side-channels leak due to the dependency of the time taken to execute a piece of software code on the values of secret variables. Here, two types of timing side-channels are of interest, namely instruction-related and cache-related ones. The former indicates that the number or type of instructions executed along a path depends on the values of secret variables. In contrast, cache-related timing side channels refer to the difference due to the memory subsystem behavior depending on the values of secret variables, e.g., a cache hit takes a few CPU cycles. Still, a miss takes hundreds of cycles cf. [107]. By analyzing the code line-by-line, the adversary can find and further exploit such vulnerabilities. Nevertheless, manual analysis of the timing characteristics of a code is challenging as it requires thorough knowledge of the code and the platform on which it is executed. The broad range of existing tools for automatically checking timing side-channel leakage can help pinpoint such vulnerabilities. In doing so, we select a recent tool recommended in the literature [54], namely SC-Eliminator [107]. Among the most important features of SC-Eliminator is the fact that, in view of available garbling protocols, it can analyze codes written in C/C++. To this end, using an LLVM compiler performs static analyses to identify the sensitive variables and timing leakage associated with them, given a program and a list of secret inputs.

GC tools. To explore whether GC frameworks would be vulnerable to timing SCA, we selected 5 open-source tools written in C/C++, which mostly support AES-NI (Advanced Encryption Standard New Instruction) instruction set (for more features of these tools cf. [49]). As a result, they have made computing AES encryptions on modern processors efficient, and consequently, the computation cost of GC is reduced drastically. **JustGarble** [10] is a library for garbling and evaluating circuits licensed under GNU GPL v3 license; however, JustGarble does not support communication or circuit generation and is, therefore, not a general-purpose framework. Nevertheless, it has become a cornerstone of various frameworks, e.g., [41, 44, 46, 57, 79, 100]. The reason behind JustGarble's efficiency is its ability to make only one AES call per garbled-gate evaluation which makes it far faster than any prior reported results [10]. JustGarble exploits the cryptographic permutations realizable by fixed-key AES acting like a public random permutation [10]. Although this might be a strong assumption cf. [44, 46], thanks to its efficiency and the theoretical foundation laid for JustGarble, it has been used in a wide variety of MPC and GC frameworks cf. [41, 79].

Songhori et al. [99, 100] proceeded JustGarble and proposed **TinyGarble**, a highly compressed and scalable sequential GC, which is a self-contained framework that can directly be used in MPC applications [49]. Three steps are taken in TinyGarble, namely converting a function defined in Verilog to a netlist format, converting that netlist to a custom circuit description (SCD), and finally, securely evaluating the resulting Boolean circuit using a garbled circuit protocol. This flow has been considered a strict improvement over JustGarble as TinyGarble further includes recent protocol and circuit optimizations. Nevertheless, and irrespective of the flexibility of TinyGarble for producing hardware circuits, changes made to JustGarble have introduced timing side-channel leakage, as will be discussed in Sections 6-7.

Table 1: The number of leaky IF conditions (IF) in various frameworks (for a detailed report, refer to Appendix A).

Framework	IF
TinyGarble [99] (half-gate)	4
TinyGarble [99] (free-XOR)	7
JustGarble [52]	11
EMP-toolkit [77]	0
Obliv-C [111]	4
ABY [27]	0

In contrast to TinyGarble, which is an extension of Verilog, **Obliv-C** is an extension of C that executes a GC protocol in a two-party setting [110]. The C language is extended by adding an obliv qualifier that is applied to C types and constructs. By enforcing typing rules, obliv types remain secret unless explicitly revealed. In doing so, it is suggested that oblivious functions and conditionals could modify public data, if they are executed within a qualified obliv block, where the code is always executed cf. [110, 111]. In addition to the data security achieved by means of these rules, modular libraries can be easily developed when using Obliv-C. Thanks to this property, Obliv-C has found application in, e.g., linear regression [35], decentralized certificate authorities [56], aggregated private machine-learning models [102], classification of encrypted emails [47] and stable matching [29].

Besides the frameworks mentioned above, we also took EMP-toolkit [77] and ABY [27], libraries developed in C++, into account. EMP-toolkit is composed of multiple MPC frameworks and allows for executing circuit-based protocols due to the available circuit generation and cryptographic libraries. ABY library offers a mechanism for mixing protocols, including optimized versions of Yao's garbled circuit protocol.

Our observations. As mentioned earlier, as a first, we examined the possibility of mounting timing SCA against GC frameworks enumerated above. In such an attack scenario, the adversary attempt to take advantage of possible unbalance if-else statements (branches). The adversary can assume that different operations performed to generate garbled inputs in free-XOR and half-gate optimized Yao's GC protocols (see Figure 1) can result in leakage if neither a constant-time implementation nor branch-less assignments are used for sensitive branches. To examine this, SC-Eliminator [107] is applied against TinyGarble [99], JustGarble [52], EMP-toolkit [77], Obliv-C [111], and ABY [27]. Table 1 contains the number of leaky IFs for this experiment. When taking a close look at the list of leaky IFs (Table 3 in Appendix A), among the set of leaky IFs, we observed unbalanced IF statements in the garbled-input generation, i.e., garbled inputs were generated in a secret-dependent manner. The existence of these unbalanced IFs demonstrates the likelihood of timing attacks to be successfully mounted against them. According to the results in Table 1, EMP-toolkit [77] and ABY [27] do not have any leaky IFs. Nevertheless, we should stress that although SC-Eliminator does not find any vulnerability in terms of leaky IFs in these frameworks, this does not rule out the possibility of other attacks. Next, we introduce our attack, Goblin, to leverage the timing side-channel leaking from existing unbalanced IF statements.

5 BUILDING BLOCKS OF GOBLIN

The main steps in Goblin's flow are: (1) filling the cache with junks. This step aims to maximize the CPU core's access time to the global secret (R) from the cache and capture the CPU cycles corresponding to each gate connected to input wires (i.e., gates in the input layer); (2) measuring the time on the CPU, including the time taken to generate garbler token, linked to the input size; (3) recovering the garbler's secret (i.e., garbler's input) after pre-processing the acquired CPU cycles and running a clustering algorithm.

5.1 Junk Generator

Goblin is only interested in the execution time of secret-dependent instructions to guess the input. Based on our observation, the input bit-dependant execution time varies for "0" and "1", although it can be subtle in some cases (see Section 6 for more information). Therefore, the idea behind the junk generator (JG) is to fill the cache repeatedly with some junk data to force the CPU to either fetch the requested data into the L1 cache from the RAM to maximize the time consumption of the read data instructions from the cache (see Section 7 for a discussion about memory management). We emphasize that if the input bit-dependant execution time differs significantly, this step can be skipped. In fact, the time differences of read instructions caused by the availability of data on cache or fetching the data from the RAM has contributed to the success of multiple attacks [43, 65, 109, 114]. As a prime example, in Flush+Reload attack [109], the attacker first flushes the cache, reloads the flushed data, and then reveals secret information of schemes according to the time consumption of reload instruction. These attack scenarios, however, are not applicable against ARM architecture [8]. This inability is because the cache evicts instruction on ARM architecture requires privileged access [109]. However, our approach is not limited to a specific architecture as JG utilizes two structures that help it to fill most lines of the L3 cache level and to avoid overwriting the generated junk.

The first structure is to use only registers in the source file and avoid using pointers. Using registers guarantees that the CPU allocates a part of the cache to the data, whereas when using pointers, the CPU allocates an address in the RAM to the data and the cache. Moreover, using pointers increase the probability of overwriting unused data in the cache as its instantiation is available in the RAM. The second structure utilizes a recursive algorithm to avoid overwriting the previously generated data to maximize the cache utilization by JG. To do so, JG first generates a 4 random long integer, then in the first iteration ($n = 1$), it calculates the summation of each two of them, resulting in 6 long integers (including the 4 previously generated randoms). From this point on, in each iteration, JG generates a sequence of summations from $n + 1$ previously generated long integers, where n is the number of the iteration. Hence, JG forces the CPU to overwrite R with junk since its task requests more memory than the task on which the garbling is taken care of; therefore, using JG leads to the fact that R will be overwritten with a high probability and must be fetched from RAM when requested by any core (refer to Section 7 for more details). The process mentioned above requires a considerable amount of memory; therefore, JG uses registers to avoid storing this massive amount of data on the RAM; otherwise, the system crashes due to insufficient available memory space.

5.2 Measuring Time on CPUs

After the JG boosts the difference between the input bit-dependent execution times, the time can be measured. According to Martin et al. [78], to measure the time without breaking the software, there are three main sources to take advantage of cf. [76]: (1) internal, hardware time sources, e.g., timestamp counters; (2) external time sources, e.g., external interrupts; and (3) creating a virtual clock, for instance, the virtual clock implementation on multi-processor systems with shared memory [90]. Without loss of generality, we focus on how timing information can be retrieved using the first option, namely `rdtsc`. The Read Timestamp Counter `rdtsc` is an x86 instruction that returns the value of the CPU timestamp counter (TSC) register. In general, the `TSC` register is shared with every user with any level of privileged access [76]; therefore, it can be accessed by: (1) a privileged/non-privileged user who has complete control over the CPU; (2) a service provider who shares the processor with the victim, such as cloud servers [78]; (3) a virtual-machine user with a privileged/non-privileged access level, who runs a process on a shared processor with the victim (e.g., cross-virtual machine attacks) [76]. Hence, the adversary can have either privileged/non-privileged access to (1) the CPU on which the garbling scheme is running, (2) the CPU of the service provider's system, or (3) a cross-virtual machine to share the processor with the victim running the garbling scheme. What could make a difference is that an unprivileged attacker cannot precisely control the garbler's execution and interrupt it, unlike a privileged attacker. Nevertheless, if the attacker can figure out when the garbling process begins, or use a trigger signal such as a cache-based side channel [95], then the collected traces can be aligned based on that timing information [73]. Therefore, without loss of generality, we assume that the attacker can align the timing measurements to mount the attack.

Resolution of timing measurements. The timestamps provided by `rdtsc` often have a resolution between 1 and 3 cycles on modern CPUs cf. [72]. For example, on AMD CPUs until the Zen microarchitecture, a cycle-accurate resolution can be obtained; however, more recent generations come with a significantly lower resolution as the register is only updated every 20 to 35 cycles. Another example is Intel Core $i7-7700$ Processors, i.e., what has been used in this study, where the `rdtsc` register is updated every cycle [51]. Nevertheless, although it might be thought that lower resolutions might make performing attacks more challenging, Goblin is not affected since it requires mainly the difference between two readings with the same resolution (see Section 7 for more details). Therefore, in contrast to attacks requiring repetition when relying on `rdtsc`, it is not needed for Goblin to do so and use the average timing differences over all executions. We stress that although Goblin is a single-trace attack since multiple gates are being garbled one after another, the time difference can be directly driven from `rdtsc`. We should also add that our attack is an example of a timing attack, meaning that we believe other methods for acquiring the timing information can definitely be applied.

5.3 Recovering Garbler's Input

Counting the gates in the input Layer. According to our adversary model, we assume that the adversary is neither the garbler nor the evaluator. Therefore, there is no information about the circuit, input size, and gate types in the input layer. Here we describe how this information is retrieved by Goblin when the garbler uses JustGarble, as an example of GC tools. This example is selected due to its broad applications (see Section 4) and its role as the core of other garbling frameworks, e.g., ones considered in our study [100, 110]. Listing 1 illustrates a high-level description of JustGarble primary functions. In Listing 1, `NF`, `LF`, `GT`, `IF`, `INL`, `WL`, `GC`, and `OL`, denoted in Lines 1–9, refer to the number of fan-outs, location of fan-outs, gates' types, the value of filled input fan-out, initial input values, wire labels, Garbled circuit, and output labels, respectively.

According to the protocol flow of JustGarble (see, Listing 1), in the first step, the garbler's tokens for zero and one logical values (`IL`) are constructed through `createNewWire` (Listing 1 line 5). Then, the parser function (`createInputLabels` Listing 1 line 3) starts parsing the simple circuit description (`SCD`) file and `g_init` files, which contain information about the circuit and the garbler's input values. The parser function learns about the circuit (`GT`) and locates the fan-in and fan-out of the input layer gates (`LF` and `NF`) that are connected to the garbler input based on `g_init` file information. For every input, the `createInputLabels` is called once for garbler label and once for the evaluator label of the input, twice per input in total. At this point, Goblin starts counting the number of `createInputLabels` calls and calculating the number of input layer gates as half of the total number of `createInputLabels` function calls. Afterward, the gates are garbled one by one by calling the `garbleCircuit` function (Listing 1 line 9), starting from the input layer gates, where the garbler's and evaluator's inputs are fed, before proceeding to the following layer gates. This fact allows Goblin to count the CPU cycle associated with each gate in the input layer by knowing the number of input gates.

Goblin against free-XOR optimization. When the framework starts garbling the gates, output labels (`OL`) and garbled tables (`GT`) are generated in the order provided in the `SCD` file. As JustGarble, similar to various modern garbling frameworks, utilizes the free-XOR optimization to generate garbler tokens for input value 1, the garbler must access the R frequently. When free-XOR optimization is enabled, `GarbleCircuit` function (Listing 1 line 9) skips line 11 to line 14 of the Listing 1. Therefore, regardless of whether the input is known or secret, it checks the type of the input gate (`GT`) and treats all inputs as a secret. If the gate type is `XOR`, including all gates categories that are considered `XOR` in GC protocols (`INV`, `XOR` and `XNOR` gates), it generates the `OL` as the XOR results of labels 0 and 1 (Listing 1 line 16); otherwise, the `OL` is constructed through a series of encryptions, see, Listing 1, line 18 to 25. It is clearly observable that in the last part of the encryption, Listing 1 line 14 and between lines 25 and 28, if the garblerinput value is "1", one more encryption, one memory access, and one XORing take place, which can result in the input dependency observable in the execution time of garbling process.

In other words, when garbling `AND` (non-XOR) gates (including (`AND/NAND`, `OR/NOR`, `ANDN`, `ORN`, `NANDN`, and `NORN`), there is an unbalanced `if` condition, which means a longer execution time for input value one. This is the point that Goblin takes advantage of differences in execution time of the garbling process for each gate due to their input value. If R is available in the L1 level of the cache, this difference is subtle and, in most cases, negligible to the time of the encryption process. Hence, to maximize the difference between the time taken to generate tokens for input 0 and 1, the JG (see Section 5.1) starts filling the cache with junks parallel to the execution of the `createNewWire` function (Listing 1 line 5) to enforce CPU to fetch R into L1 cache from RAM, which increases the execution time difference between 0 and 1 token generation. To boost the effect of JG, Goblin first finds the CPU core and thread on which the garbling process is happening by calling the `LSCPU` instruction; then asks the server to assign the JG task to the same thread, or if not possible, at least to the same core on which the garbling process is happening. It should be indicated that neither any privilege is needed nor any restriction on assigning the JG to the same core is posed as it fills the shared L3 cache level; nevertheless, assigning JG to the same core as the garbling process core will result in faster cache filling and fewer errors as JG first fills L1 and L2 level cache.

```
1  def JustGarble(g_init, SCD):
2      NF, LF, GT = createNewWire(g_init, SCD)   #Pasrses the circuit, locate the fan-outs, and generates wire labels.
3      IF, INL = createInputLabels(NF, LF) #Fills tokens to input fan-outs (called twice per garbler input).
4      GC, OL, TT = garbleCircuit(IF, IFS, WL, GT) # Generates garbled tables and Garbled output tokens.
5  def createNewWire(g_init, SCD):
6      for i in SCD[0]: #first line of SCD, which contains the information about input layer gates
7          IF[i][0] = randomBlock();
8          IF[i][1] = xorBlocks(R, IF[i][0]);
9  def garbleCircuit(IFS, WL, GT):
10     R = AESEcbEncryptBlks(AES_Key)
11     if(IFS == known):
12         GC, OL = HalfGarbleGate(GT, IF)
13         return GC, OL
14     else: #(IFS == secret):
15         if(GT == XORGATE):
16             OL = XorBlock (IFS, R) #free-XOR optimization
17         else: #if(GT == ANDGATE)
18             mask1, mask2, mask3, mask4=AESEcbEncryptBlks(AES_Key,4)
19             #AND encryptions
20             OL = XorBlock(mask1 , mask2)
21             if (IFS == 1):
22                 OL = XorBlock(OL , R);
23         GC = [XorBlock(OL, mask3), XorBlock(OL, mask4)]
24     if(gate_location is in input_layer): #Generates associate garbler tokens to be transferred to Evaluator.
25         if(g_init == 0):
26             TT = IF;
27         else:
28             TT = xorBlocks(R, IF);
29     return GC, OL, TT
```

Listing 1: Protocol flow of primary functions of JustGarble.

Goblin against half-gate optimization. Although JustGarble does not support half-gate optimization, its subsequent frameworks, such as TinyGarble and Obliv-C, utilize this optimization. Next,

```
1  def HalfGarbleGate(GT, IF):
2      R = AESEcbEncryptBlks(AES_Key)
3      mask1, mask2 = AESEcbEncryptBlks(AES_Key,2)
4      if(IF[0] == 0):
5          if(GT == ANDGATE):
6              OL = mask1 #XorBlock(mask1, 0)
7          else: #if(GT == XORGATE):
8              OL = XorBlock(mask1, IF[1])
9      if(IF[0] == 1):
10         if(GT == XORGATE):
11             OL = mask1 #XorBlock(mask1, 0)
12         else: #if(GT == ANDGATE):
13             OL = XorBlock(mask1, R)
14     GC = XorBlock(OL, mask2)
15     if(gate_location is in input_layer): #Generates
        associate garbler tokens to be transferred to
        Evaluator.
16             if(g_init == 0):
17                 TT = IF;
18             else:
19                 TT = xorBlocks(R, IF);
20     return GC, OL, TT
```

Listing 2: HalfGarbleGate function flow.

we explain why Goblin's ability is not limited to free-XOR optimization and can be launched against frameworks using half-gate optimization as follows. If half-gate optimization is enabled, `HalfGarbleGate` is called by `GarbleGate`, see Listing 2. Here the input dependency of the garbling process is even more explicit in the sense that if one of the values of the input (`IF`) is zero and the gate type (`GT`) is `ANDGATE`, the function skips all the garbling processes and constructs `OL` equal to a constant value, which results in less execution time compared to the garbling process of input value one or other type of gates. If the input value is one, then the encryption takes place (Listing 2 line 11), which results in an unbalance `if` path and dependency between the garbling process execution time and the input value. Similar to the free-XOR optimization, Goblin can benefit from the differences in execution time of `HalfGarbleGate` corresponding to the input value due to the unbalanced `if` conditions in lines 3 and 8 of Listing 2. The rest of the steps are not interesting for Goblin because they do not hold any information about the secret (garbler's input), and the above-mentioned information is adequate to launch the Goblin; therefore, from now on, Goblin can continue the attack from an offline phase.

Pre-processing the acquired CPU cycles. As explained before, when employing free-XOR optimization, the attacker expects to see a significant difference between the CPU cycle of `INV`, `XOR`, and `XNOR` gates and other gate types, including `AND/NAND`, `OR/NOR`, `ANDN`, `ORN`, `NANDN`, and `NORN` gates (refer to Section 6 for more information). This significant difference is because in the free-XOR optimization, as its name implies, an XOR-type gate is garbled by simply using the XORing operation that takes a few CPU cycles. On the other hand, garbling other types of gates, such as an AND gate, requires reading/writing from/to memory and cipher generation, which results in extra memory reads; hence, accumulating these leads to a drastic increase in CPU cycles. This is evident thanks to the definition of this optimization technique and the number of operands included in the computation of those gates, see Figure 1.(b). When employing clustering to discover the garbler's input in a non-profiled manner, this difference causes the gate types to be dominant centroids of the clustering algorithm over the input values. To overcome this challenge, Goblin first divides the CPU cycle into the number of subgroups equal to the number of available gate types, i.e., `AND` (`AND/NAND`, `OR/NOR`, `ANDN`, `ORN`, `NANDN`, and `NORN`) and `XOR` (`INV`, `XOR` and `XNOR` gates, hereafter called `XOR` gates) with regard to the median of the CPU cycles. Afterward, it normalizes each subgroup of CPU cycles by employing *z-score* normalization, and finally, concatenates the normalized data to form the CPU cycle array while maintaining the order of captured CPU cycles. Normalization minimizes the difference between the CPU cycle requirements of XOR and AND gate types, consequently improving the SR.

The first step is more complicated in a case where the half-gates optimization is enabled. Specifically, according to our observation, not only garbling the XOR gates exhibits a significantly larger number of CPU cycles compared to other gate types, but also there is a dramatic difference in the number of CPU cycles in the OR/NOR gates garbling process. There is, of course, a reason behind this, namely how gates with truth tables containing an odd number of ones (e.g., `AND`, `NAND`, `OR`, `NOR`, etc.) can be expressed and constructed. Generally speaking, these gate can be defined as $G : (v_a, v_b) \rightarrow (\alpha_a \oplus v_a) \wedge (\alpha_b \oplus v_b) \oplus \alpha_c$, where v_a and v_b are logical values and α_a, α_b, and α_c are constant values cf. [112]. For `AND` gate, α values are set to 0, whereas for OR gate, they are set to 1. Therefore, it is unsurprising that the CPU cycles collected when garbling OR/NOR gates compose a cluster different from the others. In the same vain, one can also observe that it takes more time for the garbler to generate the garbled `OR/NOR` gate with input "0", as opposed to `AND/NAND` gates with input "1". Therefore, contrary to the case of free-XOR optimization, where AND/NAND and OR/NOR can be considered as belonging to the same type, it is challenging to make a distinction between AND/NAND gates with input "0" and OR/NOR gates with input "1". This overlap results in inaccurate clustering since the algorithm puts both into one cluster, although they should be put into two different clusters due to their inputs.

To counter this challenge, Goblin applies the following additional data scaling technique before the normalization to force the pattern to match other gate types (i.e., a larger number of CPU cycles for input 1). First, similar to the free-XOR case, the CPU cycle collected from the input gates $\{c_i\}_{i=1}^{n}$ should be partitioned into subsets corresponding to different gate types: XOR/XNOR, AND/NAND, and OR/NOR. For this, Goblin calculates 66^{th} percentiles of elements in $\{c_i\}_{i=1}^{n}$ and assign the elements larger than that to the subset c_{OR}. The remaining elements of $\{c_i\}_{i=1}^{n}$ are assigned to AND and XOR subsets similarly as done in the free-XOR case: the larger elements are assigned to c_{AND} by considering the median of the $\{c_i\}_{i=1}^{n} \setminus c_{AND}$. The remaining elements are then assigned to the subset corresponding to the XOR/XNOR gates. Afterward, Goblin applies the transformation $t_i = ac_i + b$ for $c_i \in c_{OR}$, where a and b are calculated as

$$a = \frac{\text{Max}(c_{AND}) - \bar{c}_{AND}}{\text{Max}(c_{OR}) - \bar{c}_{AND}}, \quad b = \bar{c}_{AND} - a \cdot \bar{c}_{OR},$$

where $\text{Max}(\cdot)$ and $\bar{c}$'s denote the maximum and the average of the subsets, respectively. After this step, normalization is applied, similar to the free-XOR case.

Extracting garbler's input through clustering. After obtaining the pre-processed data, Goblin launches the clustering algorithm to determine each garbler's input bit. As Goblin applies normalization

to the CPU cycle data, the gate types' dominance in the centroids has vanished; therefore, Goblin clusters CPU cycles into only two clusters corresponding to input zero and input one, regardless of the gate types. To disclose the input bits, Goblin keeps track of the $\text{Max}(\{c_i\}_{i=1}^{n})$ before normalization. When the clustering process is over, all cluster members that include the maximum element are labeled as "1", meaning that the garbler input bit is "1"; consequently, other cluster includes c_i's corresponding to garbler's input bit "0".

5.4 Performance Metric

Let $\mathbf{c}_i$ be a leakage measurement, i.e., the number of CPU cycles, for a garbler input $x = x_1 \cdots x_n$ with n-bits corresponding to n wires giving the garbler's input to the circuit. For instance, for a garbled 128-bit AES design, $n = 128$. To evaluate the effectiveness of our attack, we calculate its success rate of recovering the garbler's input given a *single* trace $\{c\}_i^n$. Note that Goblin is a non-profiling attack; hence, as opposed to profiled attacks, no leakage profile is made and used during the attack. k-means clustering algorithm is used as a distinguisher so that any observation c_i is assigned to either cluster p_0 or p_1 associated with input bit x_i being "0" or "1". Precisely, the success rate is defined as follows.

$$\text{SR} := \sum_{j \in \{0,1\}} \sum_{i=1}^{n} \Pr(c_i \in p_j \mid x_i = j).$$

To put this simply, SR indicates how many bits are correctly disclosed out of n bits in the garbler's input. Note that this definition aligns with the general case considered in SCA-related literature [101]. In this context, we consider the success rate of order 1, i.e., the probability that the correct key is ranked first.

6 EXPERIMENTAL RESULTS

We ran the JustGarble, TinyGarble and Obliv-C frameworks, publicly available via GitHub repositories [52, 99, 111]. Garbler and evaluator codes ran on two systems with Linux Ubuntu 20, 16 GB of memory, and an Intel Core $i7-7700$ CPU 3.60GHz CPU. Two systems were connected through a local area network (LAN) cable. As garbling process might access R anytime during garbling process, to force CPU to fetch R from RAM to L1 level cache in maximum possible cases, we started JG as soon as the garbling process begins. This can be easily determined by calling non-privileged CPU instructions showing which applications run on each core. Moreover, we assigned the JG to the same core that generates garbled circuits on the garbler system. To capture the CPU clock cycles, we used `rdtsc` as discussed before in Section 5.2. We have also used the k-means clustering algorithm implemented in Matlab 2021.

6.1 Results for Benchmark Functions

To evaluate the efficacy of Goblin, we have targeted the commonly-used benchmark functions, including 128-AES, 288-SHA3, 256-bit Multiplier, 128-bit Summation, and 128-bit Hamming garbled by JustGarble [52], TinyGarble [100], and Obliv-C [111] (results for the benchmark functions with various input sizes can be found in Section 6.3). For this purpose, to calculate the success rate (SR), we have applied various garbler's inputs and provided the statistics in this section. Launching Goblin against all combinations of inputs is impractical due to the massive number of input combinations (i.e., for a 256-bit Multiplier, the attack had to be launched 2^{256} times); therefore, we have chosen 1000 random inputs to run Goblin. In the k-means algorithm setting, the centroids are chosen at 100 different starting values, and the algorithm returns the result for the least within-cluster sums of point-to-centroid distances.

Figure 3 shows the SR when free-XOR or half-gate optimization was enabled. The red lines in the boxes indicate the average SR of the attack against these benchmark functions. It is observable in Figure 3.(a) that the attack achieved a better SR when launched against the AES benchmark compared to, e.g., the 256-bit Multiplier. The reason is three-fold. First, only 1000 inputs are tested; therefore, the results might vary. Second, the input layer of the 256-bit Multiplier contains more XOR gates than the AES, which are more challenging because of the subtle difference between the number of clock cycles taken for "1" and "0" (see Appendix B for more details). Third, per input, notice that Goblin is a non-profiling, single-trace attack, meaning that it receives one timing measurement per gate (and per input bit, consequently); hence, the more input bits, the better Goblin determines them. This is further studied in Section 6.2.

Compared to Figure 3.(a), Figure 3.(b) corresponding to the half-gates optimization shows an overall reduced SR for the same benchmark functions. This is because of the increase in the number of gate types to be identified for the same number of input bits and observations, consequently. Needless to say, even for circuits with various gate types, such as AES, Goblin achieved an average SR of more than 90%, which means the effect of variation in the gate types does not affect Goblin's SR drastically. Imperfect process of filling the L3 level cache with junk accounts for the outliers in Figure 3. The implication of this is that the availability of R in the L1 cache level of the garbler core decreases the execution time difference between garbler 0 and 1 token generation. However, these outliers happen barely, i.e., in 11 out of 1000 experiments, which means the JG has a small error. Note that even for the outliers, Goblin still revealed the garbler's input with a range of 60% to 100% SR.

6.2 Scalability of Goblin

To test Goblin's scalability, we have launched Goblin against three benchmark functions, including MULT, SUM, and Hamming, with a range of input sizes between 128 and 1024. Figure 4 illustrates the results, where Figure 4.(a) and Figure 4.(b) depict the free-XOR and half-gate optimization results. As shown in Figure 4.(a), increasing the input size increases the minimum and average SR for virtually all cases. This SR increment is because Goblin has a broader range of data to cluster, which means it has more observations to compare with one another. Similar to previous experiment, outliers can be observed in Figure 4. To reduce the number of outliers, the natural question to ask is whether it is possible to launch Goblin without JG. We conducted experiments to answer this questions and found out that for JustGarble [52] and Obliv-C [111], the SR could decrease dramatically (close to 50%) due to the small difference between the execution times for garbler's input "0" and "1." Nonetheless, for TinyGarble [99], it is indeed possible to mount the attack with high SR without using JG (see Appendix C). This can be achieved thanks to how TinyGarble is coded, namely generating tokens for garbler input in an input-dependant manner (see Section 7 for more details).

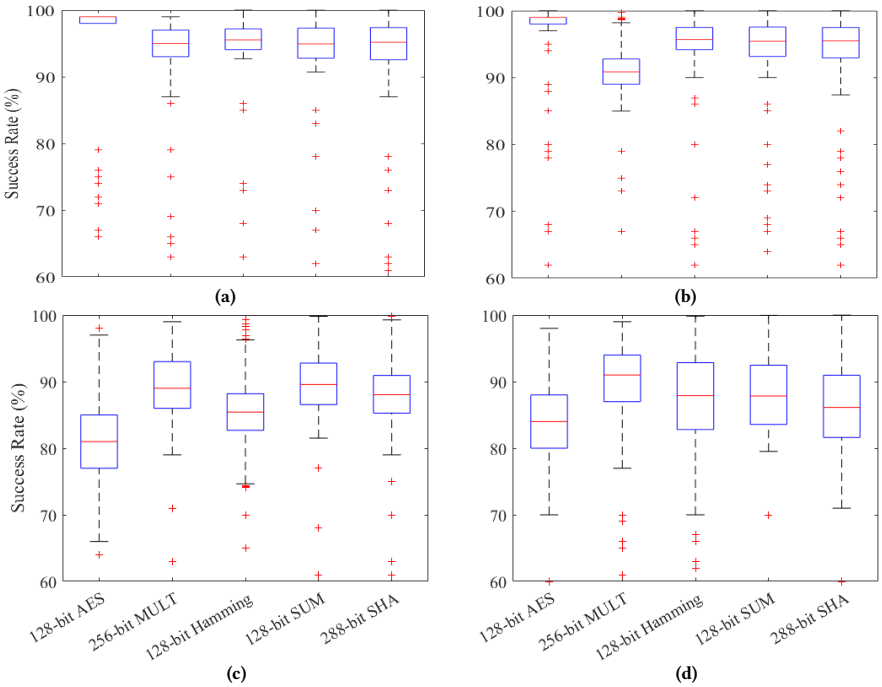


Figure 3: SR of Goblin for 1000 **randomly chosen inputs given to GC garbled by TinyGarble [100] with (a) free-XOR, (b) half-gate optimizations, (c) JustGarble [52], and (d) Obliv-C [111].**

It is worth mentioning that in this case, the JG, of course, enhances the average SR of Goblin.

6.3 Impact of the Number of Traces

In previous expriments in this section, to evaluate the effectiveness of our attack, we selected 1000 random inputs since capturing CPU cycles for all inputs is impractical and infeasible. This can directly impact the variance in our results. To investigate this, we collected CPU cycles after feeding powers of tens (from 10- 100, 000) random inputs into the 128-bit SUM, Hamming, and MULT benchmark functions, i.e., the ones demonstrating a fairly high variance (see, Figure 3). Figure 5 illustrates the SR of Goblin when being launched against a range of CPU cycle traces. As can be seen, increasing the number of CPU cycle traces results in increasing the SR of Goblin. We have observed that for a higher number of traces, SR exhibits less variance, and the average settles around 97% in all cases, except for 128-MULT. The reason behind this is the variation in the gate types as discussed before. Note that since Goblin is a single trace attack, each trace is processed by Goblin individually. In other words, the increase in the number of traces does not impact each attack but reduces the variance of the overall results. Therefore, to judge the effectiveness of Goblin, it is recommended to use more traces. We could not do this in the first place due to the time-consuming process of collecting traces for all benchmark functions. Nonetheless, comparing the results for 1000 and 100, 000 traces, the change in the average SR is subtle.

7 DISCUSSION

Relative accuracy of `rdtsc`**.** For applications using `rdtsc`, successive calls must have a difference that accurately reflects the number of cycles between two calls. This is referred to as "relative accuracy" cf. [78], meaning that any measurement through `rdtsc` is accurate with regard to the previous call/measurement. The relative accuracy does not pose any constraint to the application since they must tolerate some variations as `rdtsc` instruction's number of cycles can vary due to the state of caches, DVFS, scheduling, etc. [78]. Similarly, Goblin is resilient against variations as long as the variation is smaller than the difference between the number of cycles spent on garbling the XOR and non-XOR gates (in order of tens of thousands of cycles).

Limited resolution of `rdtsc` **on some platforms.** As introduced in Section 5.2, `rdtsc` can have various resolutions depending on the platform. In the same vein, as explained about the relative accuracy of the time read using `rdtsc`, the resolution cannot impact the effectiveness of Goblin. The point is that as long as the XOR gates can be distinguished from non-XOR ones, Goblin can successfully extract the garbler's input. For this purpose, it is necessary to have at least a resolution comparable to the number of cycles taken to garble the XOR gates (couples of tens cycles, e.g., 80 cycles as observed in our experiments).

7.1 Potential Countermeasures

To come up with a countermeasure against Goblin, one should first determine factors contributing to Goblin's sucees. Here we describe these factors and emphasize that if they are considered and encountered when proposing a framework, the likelihood of Goblin's success can decrease.

x86 processor's input dependent branch predictions. x86 processor architecture utilizes the branch predictions technique to accelerate the execution of the instructions and to optimize the

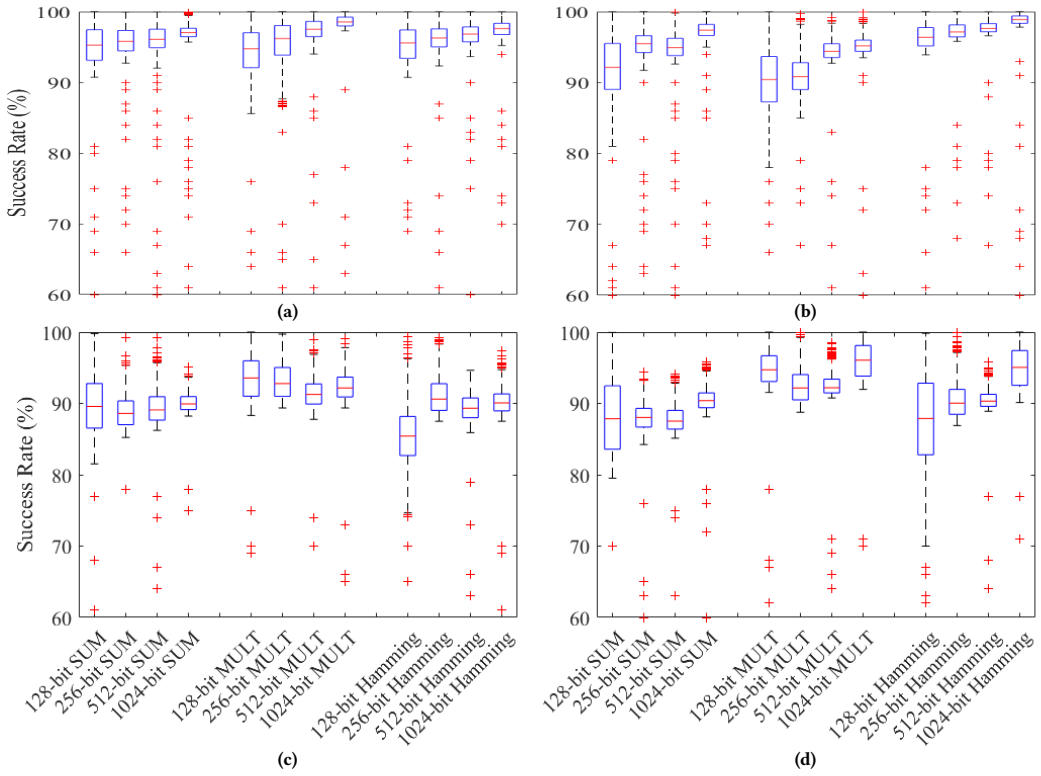


Figure 4: SR of Goblin against benchmark functions for a range of input bits garbled by TinyGarble [99] with (a) only free-XOR optimization, (b) half-gate protocol, (c) JustGarble [52], and (d) Obliv-C [111] for 1000 **randomly chosen inputs.**

CPU tasks by skipping redundant instructions [61]. This CPU behavior results in time variations between the execution times of the same instruction sets applied to different values. This exploit has been extracted by successful attacks such as [4, 61]. Similar CPU behavior exists in the case of garble circuit frameworks when being executed on the CPU. When garbler applies token generation instruction sets to the input value 0, some instructions, such as XORing with R, will be skipped due to the CPU branch prediction, which results in less execution time when applying the same instruction sets to the input value 1 as it includes time-consuming instructions such as reading R from memory. This exploit plays a primary part in revealing the garbler's input by Goblin. Note that the nature of x86 processors leads to this issue, i.e., it results in vulnerability to the Goblin attack under various scenarios. Using a time-constant compiler might mitigate this exploit to some extent. However, a time-constant compiler is by far slower than the modern compilers used in cloud servers and real-time applications; therefore, at least in these applications, using a time-constant compiler may not be preferred [18, 91, 97].

The coding style of the framework. As explained before, Goblin exploits unbalanced `IF` statements. We have observed that EMP-toolkit [77], Obliv-C [111], and ABY [27] frameworks consider this vulnerability and encounter it securely. In their frameworks, they first generate both 0 and 1 garbler's tokens and only choose one of these tokens when the garbler starts transferring its token to the evaluator. Although this technique is less optimized than generating one token per garbler's input, used in TinyGarble [100] and JustGarble [52] frameworks, the leakage is mitigated.

Memory management. Employing free-XOR or half-gate optimization requires frequent memory access to load R, the global secret, to generate garbler's token for input bit 1. Assigning R to a fixed address of memory via pointers will result in less time-consuming memory access when a core requests R. This is due to that fact that the CPU manages to keep an instance of R as close as possible to the core that requests it- in the best scenario, keeps it in the L1 cache level of the core. This decreases the memory access of R drastically because the CPU does not need to fetch it from the RAM and provide it to the core. Instead of using pointers, registers can be utilized to store R, which increases the overwrite chance of R by other cores that utilize a vast amount of cache memory. In this case, the CPU is forced to fetch the overwritten R from the RAM, resulting in a variation in the time taken to generate the garbler's token for input 0 and 1. On the other hand, using registers to store data let the CPU save time by allowing it to decide if the value is needed to be kept as close as possible to the core, which possibly requests R or keeps an instantiate of it in the RAM and overwrites it resulting in less cache utilization.

We have observed that EMP-toolkit [77], Obliv-C [111], JustGarble [52], and ABY [27] assigned a fixed address to R, which allows CPU to maintain its memory location as close as possible to the core that requests it, while TinyGarble [100] only assigned a fixed address to it during the labels generations and used registers in the token generations. However, using JG forces the CPU to overwrite

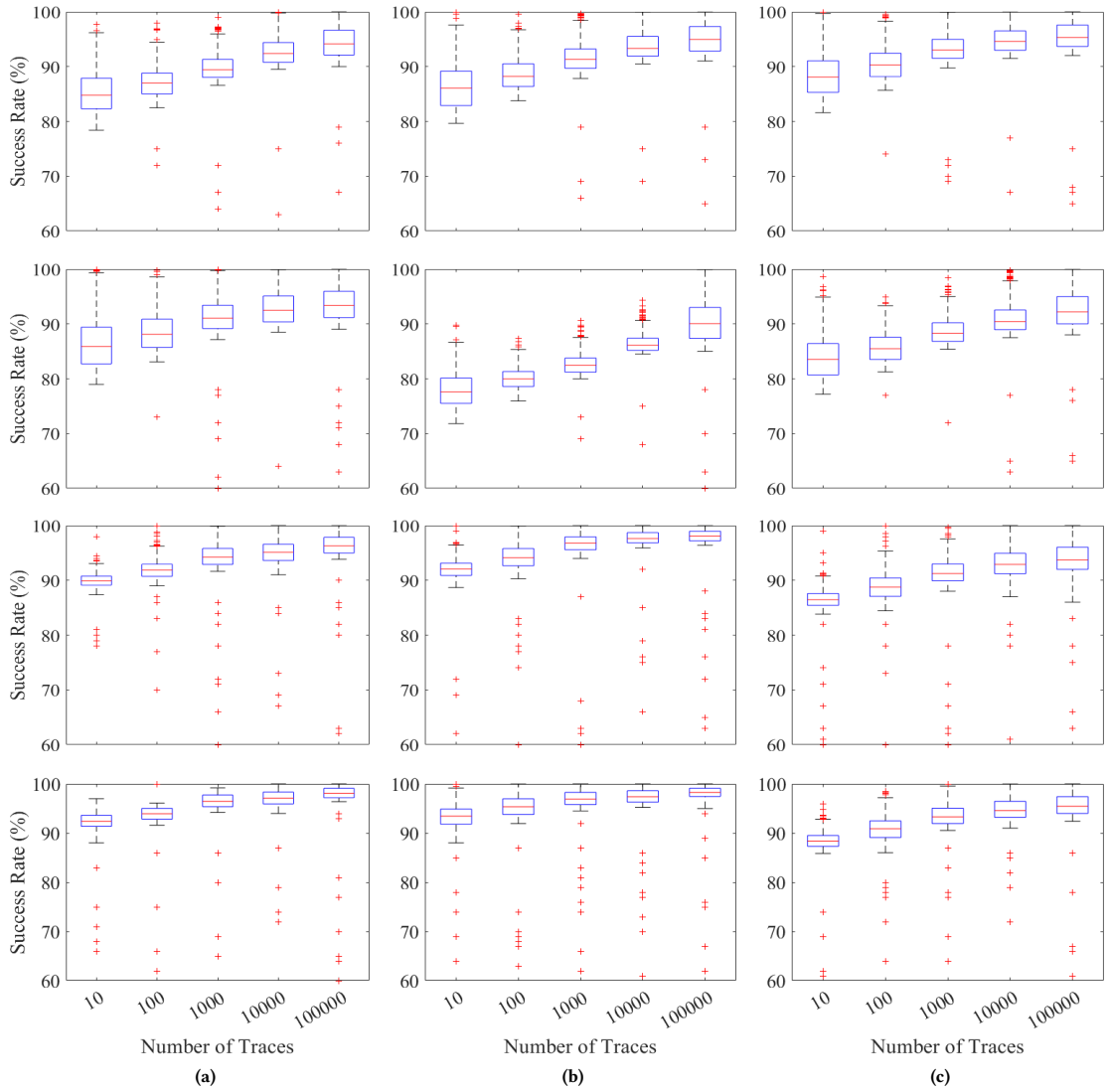


Figure 5: SR of Goblin against (a) 128-bit SUM, (b) 128-bit Hamming, and (b) 128-bit MULT for a range of 10-100, 000 randomly chosen inputs (first to last row: JustGarble [52], Obliv-C [111], TinyGarble [99] with free-XOR, and with half-gate optimizations).

R with junk as its assigned task requests more memory than the task on which the garbling is taken care of. Hence, using JG leads to the fact that R will be overwritten with a high probability and must be fetched from RAM when requested by any core. This results in a noticeable time difference between the generation of garbler's token for input bits 0 and 1.

Can restricting access stop Goblin? As it has been nicely put forward in [70], unprivileged usage of the high-resolution timers can be prevented by setting control registers, e.g., `CR4.TSD` bit in AMD [6] (volume 2, Section 3.2.5). This results in `rdtsc`, `rdtscp`, and `rdpru` being unavailable to the attacker. Although this might be tempting, such a restriction can have a negative impact on unprivileged applications depending on `rdtsc`, e.g., adb, cargo, and Docker [70]. Moreover, in doing so, not all timing primitives can be disabled. It is possible for the attacker to come up with a *counting thread* that constantly increments a global variable that serves as a timestamp without relying on platform specifics [71, 72, 96]. It is further shown that such a counting thread can have even a higher resolution than the `rdtsc` instruction on Intel CPUs [96].

8 CONCLUSION

Nowadays, several applications, including multi-party computation, rely on the efficient implementations of GC.To achieve this efficiency, many optimizations, such as free-XOR and half-gates, have been presented to reduce the cost of garbling progress. This paper has introduced Goblin, the first machine learning-assisted, non-profiling, single-trace timing SCA against GC frameworks. Specifically, Goblin targets frameworks using free-XOR and half-gate by collecting and analyzing the CPU cycles of the garbling process by reading the time stamp counter, i.e., calling `rdtsc`. In

doing so, the garbler's inputs that should have been kept secure can be disclosed without prior knowledge about the circuit being garbled. In this regard, Goblin can be run in parallel to the garbling framework without requiring any privileged access. Goblin has also been proven to be scalable when targeting large circuits. We have studied several cases, including various GC frameworks, benchmark functions, and the number of garbler's input bits. Under different scenarios, Goblin disclosed the garbler's input with high probability. Further, we have discussed Goblin's success factors and countermeasures against that.

REFERENCES

[1] Onur Aciiçmez. 2007. Yet Another Microarchitectural Attack: Exploiting I-cache. In *Proceedings of the 2007 ACM workshop on Computer security architecture*. 11–18.

[2] Onur Aciiçmez, Billy Bob Brumley, and Philipp Grabher. 2010. New Results on Instruction Cache Attacks. In *International workshop on cryptographic hardware and embedded systems*. Springer, 110–124.

[3] Onur Acıiçmez and Çetin Kaya Koç. 2006. Trace-driven Cache Attacks on AES (Short Paper). In *International Conference on Information and Communications Security*. Springer, 112–121.

[4] Onur Aciiçmez, Çetin Kaya Koç, and Jean-Pierre Seifert. 2006. Predicting Secret Keys via Branch Prediction. In *Topics in Cryptology–CT-RSA 2007: The Cryptographers' Track at the RSA Conference 2007, San Francisco, CA, USA, February 5-9, 2007. Proceedings*. Springer, 225–242.

[5] Onur Aciiçmez, Çetin Kaya Koç, and Jean-Pierre Seifert. 2007. On the Power of Simple Branch Prediction Analysis. In *Proceedings of the 2nd ACM symposium on Information, computer and communications security*. 312–320.

[6] Advanced Micro Devices Inc. 2017. AMD64 Architecture Programmer's Manual. [Online]https://www.amd.com/system/files/TechDocs/24593.pdf [Accessed: Jan.30, 2023]. (2017).

[7] Benny Applebaum. 2011. Key-dependent Message Security: Generic Amplification and Completeness. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 527–546.

[8] ARM ARM. 2012. Architecture Reference Manual. *ARMv7-A and ARMv7-R edition* (2012).

[9] Boaz Barak, Iftach Haitner, Dennis Hofheinz, and Yuval Ishai. 2010. Bounded Key-dependent Message Security. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 423–444.

[10] Mihir Bellare, Viet Tung Hoang, Sriram Keelveedhi, and Phillip Rogaway. 2013. Efficient Garbling from a Fixed-key Blockcipher. In *2013 IEEE Symp. on Security and Privacy*. IEEE, 478–492.

[11] Mihir Bellare, Viet Tung Hoang, and Phillip Rogaway. 2012. Foundations of Garbled Circuits. In *Proc. of the 2012 ACM Conf. on Computer and Comm. security*. 784–796.

[12] Aner Ben-Efraim, Yehuda Lindell, and Eran Omri. 2016. Optimizing Semi-honest Secure Multiparty Computation for the Internet. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 578–590.

[13] Fabrice Benhamouda and Huijia Lin. 2018. K-round Multiparty Computation from K-round Oblivious Transfer via Garbled Interactive Circuits. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 500–532.

[14] Daniel J Bernstein. 2005. Cache-timing Attacks on AES. (2005).

[15] Peter Bogetoft, Dan Lund Christensen, Ivan Damgård, Martin Geisler, Thomas Jakobsen, Mikkel Krøigaard, Janus Dam Nielsen, Jesper Buus Nielsen, Kurt Nielsen, Jakob Pagter, et al. 2009. Secure Multiparty Computation Goes Live. In *International Conference on Financial Cryptography and Data Security*. Springer, 325–343.

[16] Zvika Brakerski and Henry Yuen. 2022. Quantum Garbled Circuits. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 804–817.

[17] David Brumley and Dan Boneh. 2005. Remote Timing Attacks are Practical. *Computer Networks* 48, 5 (2005), 701–716.

[18] Gaurav Chadha, Scott Mahlke, and Satish Narayanasamy. 2015. Accelerating Asynchronous Programs through Event Sneak Peek. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture*. 642–654.

[19] Delin Chen, Wenhao Chen, Jian Chen, Peijia Zheng, and Jiwu Huang. 2018. Edge Detection and Image Segmentation on Encrypted Image with Homomorphic Encryption and Garbled Circuit. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[20] Sanchuan Chen, Xiaokuan Zhang, Michael K Reiter, and Yinqian Zhang. 2017. Detecting Privileged Side-channel Attacks in Shielded Execution with Déjá Vu. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. 7–18.

[21] Seung Geol Choi, Jonathan Katz, Ranjit Kumaresan, and Hong-Sheng Zhou. 2012. On the Security of the "free-XOR" Technique. In *Theory of Cryptography Conference*. Springer, 39–53.

[22] Martine de Cock, Rafael Dowsley, Anderson CA Nascimento, and Stacey C Newman. 2015. Fast, Privacy preserving Linear Regression over Distributed Datasets Based on Pre-distributed Data. In *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*. 3–14.

[23] Mauro Conti, Stephen Crane, Lucas Davi, Michael Franz, Per Larsen, Marco Negro, Christopher Liebchen, Mohaned Qunaibit, and Ahmad-Reza Sadeghi. 2015. Losing Control: On the Effectiveness of Control-flow Integrity Under Stack Attacks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 952–963.

[24] Bart Coppens, Ingrid Verbauwhede, Koen De Bosschere, and Bjorn De Sutter. 2009. Practical Mitigations for Timing-based Side-channel Attacks on Modern x86 Processors. In *2009 30th IEEE Symposium on Security and Privacy*. IEEE, 45–60.

[25] Ivan Damgård and Yuval Ishai. 2005. Constant-round Multiparty Computation Using a Black-box Pseudorandom Generator. In *Annual International Cryptology Conference*. Springer, 378–394.

[26] Ivan Damgård, Yuval Ishai, Mikkel Krøigaard, Jesper Buus Nielsen, and Adam Smith. 2008. Scalable Multiparty Computation with Nearly Optimal Work and Resilience. In *Annual International Cryptology Conference*. Springer, 241–261.

[27] Daniel Demmler, Thomas Schneider, and Michael Zohner. 2015. ABY-A Framework for Efficient Mixed-protocol Secure Two-party Computation.. In *NDSS*.

[28] Jean-Francois Dhem, Francois Koeune, Philippe-Alexandre Leroux, Patrick Mestré, Jean-Jacques Quisquater, and Jean-Louis Willems. 1998. A Practical Implementation of the Timing Attack. In *International Conference on Smart Card Research and Advanced Applications*. Springer, 167–182.

[29] Jack Doerner, David Evans, and Abhi Shelat. 2016. Secure Stable Matching at Scale. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 1602–1613.

[30] Xiaowan Dong, Zhuojia Shen, John Criswell, Alan L Cox, and Sandhya Dwarkadas. 2018. Shielding Software From Privileged {Side-Channel} Attacks. In *27th USENIX Security Symposium (USENIX Security 18)*. 1441–1458.

[31] Catherine Easdon, Michael Schwarz, Martin Schwarzl, and Daniel Gruss. 2022. Rapid Prototyping for Microarchitectural Attacks. In *USENIX Security Symposium*.

[32] Uri Feige, Joe Killian, and Moni Naor. 1994. A Minimal Model for Secure Computation. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*. 554–563.

[33] Sanjam Garg and Akshayaram Srinivasan. 2017. Garbled Protocols and Two-round MPC From Bilinear Maps. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 588–599.

[34] Sanjam Garg and Akshayaram Srinivasan. 2018. Two-round Multiparty Secure Computation from Minimal Assumptions. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 468–499.

[35] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. 2017. Privacy-Preserving Distributed Linear Regression on High-Dimensional Data. *Proc. Priv. Enhancing Technol.* 2017, 4 (2017), 345–364.

[36] Qian Ge, Yuval Yarom, David Cock, and Gernot Heiser. 2018. A Survey of Microarchitectural Timing Attacks and Countermeasures on Contemporary Hardware. *Journal of Cryptographic Engineering* 8 (2018), 1–27.

[37] Craig Gentry, Shai Halevi, and Vinod Vaikuntanathan. 2010. i-hop Homomorphic Encryption and Rerandomizable Yao Circuits. In *Annual Cryptology Conference*. Springer, 155–172.

[38] Shafi Goldwasser, Yael Kalai, Raluca Ada Popa, Vinod Vaikuntanathan, and Nickolai Zeldovich. 2013. Reusable Garbled Circuits and Succinct Functional Encryption. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. 555–564.

[39] Sergey Gorbunov, Vinod Vaikuntanathan, and Hoeteck Wee. 2012. Functional Encryption with Bounded Collusions via Multi-party Computation. In *Annual Cryptology Conference*. Springer, 162–179.

[40] Ben Gras, Kaveh Razavi, Herbert Bos, and Cristiano Giuffrida. 2018. Translation Leak-aside Buffer: Defeating Cache Side-channel Protections with {TLB} Attacks. In *27th USENIX Security Symposium (USENIX Security 18)*. 955–972.

[41] Adam Groce, Alex Ledger, Alex J Malozemoff, and Arkady Yerukhimovich. 2016. CompGC: Efficient Offline/online Semi-honest Two-party Computation. *Cryptology ePrint Archive* (2016).

[42] Daniel Gruss, Clémentine Maurice, Anders Fogh, Moritz Lipp, and Stefan Mangard. 2016. Prefetch Side-channel Attacks: Bypassing SMAP and Kernel ASLR. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 368–379.

[43] Daniel Gruss, Clémentine Maurice, Klaus Wagner, and Stefan Mangard. 2016. Flush+ Flush: a Fast and Stealthy Cache Attack. In *Detection of Intrusions and Malware, and Vulnerability Assessment: 13th International Conference, DIMVA 2016, San Sebastián, Spain, July 7-8, 2016, Proceedings 13*. Springer, 279–299.

[44] Shay Gueron, Yehuda Lindell, Ariel Nof, and Benny Pinkas. 2015. Fast Garbling of Circuits under Standard Assumptions. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 567–578.

[45] Chun Guo, Jonathan Katz, Xiao Wang, Chenkai Weng, and Yu Yu. 2020. Better Concrete Security for Half-gates Garbling (in the Multi-instance Setting). In *Annual International Cryptology Conference*. Springer, 793–822.

[46] Chun Guo, Jonathan Katz, Xiao Wang, and Yu Yu. 2020. Efficient and Secure Multiparty Computation from Fixed-key Block Ciphers. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 825–841.

[47] Trinabh Gupta, Henrique Fingler, Lorenzo Alvisi, and Michael Walfish. 2017. Pretzel: Email Encryption and Provider-supplied Functions are Compatible. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*. 169–182.

[48] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer.

[49] Marcella Hastings, Brett Hemenway, Daniel Noble, and Steve Zdancewic. 2019. Sok: General Purpose Compilers for Secure Multi-party Computation. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 1220–1237.

[50] Benjamin Hettwer, Stefan Gehrer, and Tim Güneysu. 2020. Applications of machine learning techniques in side-channel attacks: a survey. *J. of Cryptographic Engineering* 10, 2 (2020), 135–162.

[51] Intel Corporation. 2017. Intel® Core™ i7 Processors. [Online]https://www.intel.com/content/www/us/en/products/details/processors/core/i7.html [Accessed: Jan.30, 2023]. (2017).

[52] irdan. 2014. JustGarble Framework. [Online]https://github.com/irdan/justGarble [Accessed Jan.30, 2023]. (2014).

[53] Karthik A Jagadeesh, David J Wu, Johannes A Birgmeier, Dan Boneh, and Gill Bejerano. 2017. Deriving Genomic Diagnoses without Revealing Patient Genomes. *Science* 357, 6352 (2017), 692–695.

[54] Jan Jancar. 2021. The State of Tooling for Verifying Constant-timeness of Cryptographic Implementations. [Online]https://neuromancer.sk/article/26 [Accessed: Feb.7, 2023]. (2021).

[55] Jan Jancar, Marcel Fourné, Daniel De Almeida Braga, Mohamed Sabt, Peter Schwabe, Gilles Barthe, Pierre-Alain Fouque, and Yasemin Acar. 2022. "They're not that Hard to Mitigate": What Cryptographic Library Developers Think About Timing Attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 632–649.

[56] Bargav Jayaraman, Hannah Li, and David Evans. 2017. Decentralized Certificate Authorities. *arXiv preprint arXiv:1706.03370* (2017).

[57] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. 2018. {GAZELLE}: A Low Latency Framework For Secure Neural Network Inference. In *27th USENIX Security Symp. (USENIX Security 18)*. 1651–1669.

[58] Seny Kamara, Payman Mohassel, and Mariana Raykova. 2011. Outsourcing Multi-party Computation. *Cryptology ePrint Archive* (2011).

[59] Seny Kamara, Payman Mohassel, and Ben Riva. 2012. Salus: a System for Server-aided Secure Function Evaluation. In *Proceedings of the 2012 ACM conference on Computer and communications security*. 797–808.

[60] Seny Kamara, Payman Mohassel, and Ben Riva. 2012. Salus: A System for Server-Aided Secure Function Evaluation. *Cryptology ePrint Archive* (2012).

[61] Paul Kocher, Jann Horn, Anders Fogh, Daniel Genkin, Daniel Gruss, Werner Haas, Mike Hamburg, Moritz Lipp, Stefan Mangard, Thomas Prescher, et al. 2020. Spectre Attacks: Exploiting Speculative Execution. *Commun. ACM* 63, 7 (2020), 93–101.

[62] Paul Kocher, Joshua Jaffe, and Benjamin Jun. 1999. Differential Power Analysis. In *Annual international cryptology conference*. Springer, 388–397.

[63] Paul C Kocher. 1996. Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In *Annual International Cryptology Conference*. Springer, 104–113.

[64] Vladimir Kolesnikov and Thomas Schneider. 2008. Improved Garbled Circuit: Free XOR Gates and Applications. In *Intrl. Colloquium on Automata, Languages, and Programming*. Springer, 486–498.

[65] Chun-Hao Lai, Jishen Zhao, and Chia-Lin Yang. 2017. Leave the Cache Hierarchy Operation as it is: A New Persistent Memory Accelerating Approach. In *Proceedings of the 54th Annual Design Automation Conference 2017*. 1–6.

[66] Itamar Levi and Carmit Hazay. 2022. Garbled-Circuits from an SCA Perspective: Free XOR Can be Quite Expensive... *Cryptology ePrint Archive* (2022).

[67] Y Lindell and B Pinkas. 2004. A Proof of Yao's Protocol for Secure Two-party Computation. ECCC Report TR04-063. In *Electronic Colloquium on Computational Complexity (ECCC)*.

[68] Yehuda Lindell and Benny Pinkas. 2007. An Efficient Protocol for Secure Two-party Computation in the Presence of Malicious Adversaries. In *Annual Intrl. Conf. on the theory and applications of cryptographic techniques*. Springer, 52–78.

[69] Yehuda Lindell and Benny Pinkas. 2009. A Proof of Security of Yao's Protocol for Two-party Computation. *J. of cryptology* 22, 2 (2009), 161–188.

[70] Moritz Lipp, Daniel Gruss, and Michael Schwarz. 2022. AMD Prefetch Attacks Through Power and Time. In *USENIX Security Symposium*.

[71] Moritz Lipp, Daniel Gruss, Raphael Spreitzer, Clémentine Maurice, and Stefan Mangard. 2016. {ARMageddon}: Cache Attacks on Mobile Devices. In *25th USENIX Security Symposium (USENIX Security 16)*. 549–564.

[72] Moritz Lipp, Vedad Hadžić, Michael Schwarz, Arthur Perais, Clémentine Maurice, and Daniel Gruss. 2020. Take a Way: Exploring the Security Implications of AMD's Cache Way Predictors. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*. 813–825.

[73] Moritz Lipp, Andreas Kogler, David Oswald, Michael Schwarz, Catherine Easdon, Claudio Canella, and Daniel Gruss. 2021. PLATYPUS: Software-based Power Side-channel Attacks on x86. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 355–371.

[74] Fangfei Liu, Qian Ge, Yuval Yarom, Frank Mckeen, Carlos Rozas, Gernot Heiser, and Ruby B Lee. 2016. Catalyst: Defeating Last-level Cache Side Channel Attacks in Cloud Computing. In *2016 IEEE international symposium on high performance computer architecture (HPCA)*. IEEE, 406–418.

[75] Fangfei Liu, Yuval Yarom, Qian Ge, Gernot Heiser, and Ruby B Lee. 2015. Last-level Cache Side-channel Attacks are Practical. In *2015 IEEE symposium on security and privacy*. IEEE, 605–622.

[76] Yangdi Lyu and Prabhat Mishra. 2018. A Survey of Side-channel Attacks on Caches and Countermeasures. *Journal of Hardware and Systems Security* 2, 1 (2018), 33–50.

[77] A.J Malozemoff, X Wang, and J Katz. 2022. Emp-toolkit Framework. [Online]https://github.com/emp-toolkit [Accessed Jan.30, 2023]. (2022).

[78] Robert Martin, John Demme, and Simha Sethumadhavan. 2012. Timewarp: Rethinking Timekeeping and Performance Monitoring Mechanisms to Mitigate Side-channel Attacks. In *2012 39th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 118–129.

[79] Payman Mohassel, Mike Rosulek, and Ye Zhang. 2015. Fast and Secure Three-party Computation: The Garbled Circuit Approach. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 591–602.

[80] Payman Mohassel and Yupeng Zhang. 2017. Secureml: A System for Scalable Privacy-preserving Machine Learning. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 19–38.

[81] Keaton Mowery, Sriram Keelveedhi, and Hovav Shacham. 2012. Are AES x86 Cache Timing Attacks Still Feasible?. In *Proceedings of the 2012 ACM Workshop on Cloud computing security workshop*. 19–24.

[82] Maria Mushtaq, Muhammad Asim Mukhtar, Vianney Lapotre, Muhammad Khurram Bhatti, and Guy Gogniat. 2020. Winter is Here! A Decade of Cache-based Side-channel Attacks, Detection & Mitigation for RSA. *Information Systems* 92 (2020), 101524.

[83] Aoi Nakamoto. 2018. W-Shield: Protection Against Cryptocurrency Wallet Credential Stealing. In *Workshop on Security and Privacy in E-Commerce 2018*. 71–107.

[84] Michael Neve and Jean-Pierre Seifert. 2006. Advances on Access-driven Cache Attacks on AES. In *International Workshop on Selected Areas in Cryptography*. Springer, 147–162.

[85] Valeria Nikolaenko, Udi Weinsberg, Stratis Ioannidis, Marc Joye, Dan Boneh, and Nina Taft. 2013. Privacy-preserving Ridge Regression on Hundreds of Millions of Records. In *2013 IEEE symposium on security and privacy*. IEEE, 334–348.

[86] Oleksii Oleksenko, Bohdan Trach, Robert Krahn, Mark Silberstein, and Christof Fetzer. 2018. Varys: Protecting {SGX} Enclaves from Practical {Side-Channel} Attacks. In *2018 USENIX Annual Technical Conference (USENIX ATC 18)*. 227–240.

[87] Rafail Ostrovsky, Anat Paskin-Cherniavsky, and Beni Paskin-Cherniavsky. 2014. Maliciously Circuit-private FHE. In *Annual Cryptology Conference*. Springer, 536–553.

[88] Dag Arne Osvik, Adi Shamir, and Eran Tromer. 2006. Cache Attacks and Countermeasures: the Case of AES. In *Cryptographers' track at the RSA conference*. Springer, 1–20.

[89] Dan Page. 2002. Theoretical Use of Cache Memory as a Cryptanalytic Side-channel. *Cryptology ePrint Archive* (2002).

[90] Colin Percival. 2005. Cache Missing for Fun and Profit. (2005).

[91] Eyal Ronen, Robert Gillham, Daniel Genkin, Adi Shamir, David Wong, and Yuval Yarom. 2019. The 9 Lives of Bleichenbacher's CAT: New Cache Attacks on TLS Implementations. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 435–452.

[92] Amit Sahai and Hakan Seyalioglu. 2010. Worry-free Encryption: Functional Encryption with Public Keys. In *Proceedings of the 17th ACM conference on Computer and communications security*. 463–472.

[93] Anish Saxena and Biswabandan Panda. 2022. Dabangg: A Case for Noise Resilient Flush-based Cache Attacks. In *2022 IEEE Security and Privacy Workshops (SPW)*. IEEE, 323–334.

[94] Thomas Schneider. 2008. Practical Secure Function Evaluation.. In *Informatiktage*. 37–40.

[95] Michael Schwarz, Daniel Gruss, Moritz Lipp, Clémentine Maurice, Thomas Schuster, Anders Fogh, and Stefan Mangard. 2018. Automated Detection, Exploitation, and Elimination of Double-fetch Bugs Using Modern CPU Features. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*.

587–600.
[96] Michael Schwarz, Samuel Weiser, Daniel Gruss, Clémentine Maurice, and Stefan Mangard. 2017. Malware Guard Extension: Using SGX to Conceal Cache Attacks. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 3–24.
[97] Mohammad Shahrad, Jonathan Balkind, and David Wentzlaff. 2019. Architectural Implications of Function-as-a-service Computing. In *Proceedings of the 52nd annual IEEE/ACM international symposium on microarchitecture*. 1063–1075.
[98] Hanif D Sherali and Cihan H Tuncbilek. 1992. A Squared-euclidean Distance Location-allocation Problem. *Naval Research Logistics (NRL)* 39, 4 (1992), 447–469.
[99] E Songhori, H Siam, and S Riazi. 2019. TinyGarble Framework. [Online]https://github.com/esonghori/TinyGarble [Accessed Jan.30, 2023]. (2019).
[100] Ebrahim M Songhori, Siam U Hussain, Ahmad-Reza Sadeghi, Thomas Schneider, and Farinaz Koushanfar. 2015. Tinygarble: Highly Compressed and Scalable Sequential Garbled Circuits. In *2015 IEEE Symp. on Security and Privacy*. IEEE, 411–428.
[101] François-Xavier Standaert, Tal G Malkin, and Moti Yung. 2009. A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. In *Annual Intrl. Conf. on the Theory and Applications of Cryptographic Techniques*. Springer, 443–461.
[102] Lu Tian, Bargav Jayaraman, Quanquan Gu, and David Evans. 2016. Aggregating Private Sparse Learning Models using Multi-party Computation. In *NIPS Workshop on Private Multi-Party Machine Learning*.
[103] Eran Tromer, Dag Arne Osvik, and Adi Shamir. 2010. Efficient Cache Attacks on AES, and Countermeasures. *Journal of Cryptology* 23, 1 (2010), 37–71.
[104] Bhanu C Vattikonda, Sambit Das, and Hovav Shacham. 2011. Eliminating Fine Grained Timers in Xen. In *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*. 41–46.
[105] Zhenghong Wang and Ruby B Lee. 2006. Covert and Side Channels Due to Processor Architecture. In *2006 22nd Annual Computer Security Applications Conference (ACSAC'06)*. IEEE, 473–482.
[106] Carolyn Whitnall and Elisabeth Oswald. 2015. Robust Profiling for DPA-style Attacks. In *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, 3–21.
[107] Meng Wu, Shengjian Guo, Patrick Schaumont, and Chao Wang. 2018. Eliminating Timing Side-channel Leaks Using Program Repair. In *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 15–26.
[108] Andrew Chi-Chih Yao. 1986. How to Generate and Exchange Secrets. In *27th Annual Symp. on Foundations of Computer Science (sfcs 1986)*. IEEE, 162–167.
[109] Yuval Yarom and Katrina Falkner. 2014. FLUSH+ RELOAD: A High Resolution, Low Noise, L3 Cache side-channel Attack. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*. 719–732.
[110] Samee Zahur and David Evans. 2015. Obliv-C: A Language for Extensible Data-oblivious Computation. *Cryptology ePrint Archive* (2015).
[111] S Zahur, G Kerneis, and G Necula. 2018. Obliv-C Secure Computation Compiler. [Online]https://github.com/samee/obliv-c [Accessed Feb.2, 2023]. (2018).
[112] Samee Zahur, Mike Rosulek, and David Evans. 2015. Two Halves Make a Whole. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 220–250.
[113] Yinqian Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2014. Cross-tenant Side-channel Attacks in PaaS Clouds. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 990–1003.
[114] Li Zhao, Ravi Iyer, Srihari Makineni, Don Newell, and Liqun Cheng. 2010. NCID: a non-inclusive Cache, Inclusive Directory Architecture for Flexible and Efficient Cache Hierarchies. In *Proceedings of the 7th ACM international conference on Computing frontiers*. 121–130.

Table 2: Type of the gates in the input layer of the AES and 256-bit MULT modules.

	AES		256-bit MULT	
	Percentage (%)	Count	Percentage (%)	Count
AND gates in input layer	75	96	50	256
XOR gates in input layer	25	32	50	256

APPENDIX A

Table 3 contains details of leaky IF conditions in each function of TinyGarble [99], EMP-toolkit [77], Obliv-C [110], and ABY [27].

APPENDIX B

To investigate the effects of the gate types in the input layer on the SR, we counted the number of XOR and AND gates in the input layer of the AES and 256-bit MULT since the results for these two benchmark functions vary largely as shown in Figure 3. Table 2 contains the detail about the type of the gates in the AES and 256-bit MULT benchmark functions. Moreover, the category of AND gate contains AND/NAND, OR/NOR, ANDN, ORN, NANDN, and NORN gates, and the category of XOR gate includes NV, XOR, and XNOR gates as described in 5.3. It is observable that the AND gates are dominant in the AES input layer (75% input layer gates) while the portions of XOR and AND gates are equal in the input layer of 256-bit MULT. This can explain why the results for these two benchmark functions are different. In fact, it is because of the fact that it is more challenging to determine the inputs given to XOR gates. To further analyze the reason behind this, we have separately calculated the SR of Goblin against applied against AND and XOR gates. Figure 6 illustrates the results for launching Goblin against 128-AES, 256-bit MULT, 128-bit Hamming, 128-bit SUM, and 288-bit SHA modules, similar to Figure 3, where the results for AND and XOR gates are combined. As observable in Figure 6, Goblin's average SR when launching against AND gates are always close to 100% while its average SR has a range between 100% and 65% when launching against XOR gates for the benchmark functions. This is aligned with the results presented in Figure 3. In that figure, the difference between the mean values of CPU cycles collected for inputs "0" and "1" is larger for AND gates in comparison to XOR gates.

APPENDIX C

To study the impact of an implementation in which not all timing side-channel vulnerability is not considered, we have launched Goblin against TinyGarble when the JG has been disabled. Figure 7 illustrates the results of Goblin against TinyGarble when JG is disabled. It is observable in Figure 7 that even without the presence of JG, Goblin can reveal the garbler's input with a success rate (SR) average of 95% or higher, which is slightly lower than the case when JG is enabled. These results indicate the damaging effect of a wrong coding style, memory management, and compiler branch predictions on the GC frameworks securities. To further investigate this damaging effect, we launched Goblin against MULT, SUM, and Hamming benchmarks with input ranges between 128 and 1024 bits when JG was disabled. Figure 8 shows the results of launching Goblin against MULT, SUM, and Hamming benchmark functions for a range of inputs garbled by TinyGarble when (a) only free-XOR optimization, (b) half-gate protocol is enabled, and JG is disabled. Same as results in Sec 6.2, one can observe a similar pattern of increasing SR of Goblin according to the increased size of benchmarks input. As another part of our investigations, we have launched Goblin against MULT, SUM, and Hamming modules without JG. Figure 9 illustrates SR of Goblin against (a) 128-bit SUM, (b) 128-bit Hamming, and (b) 128-bit MULT benchmarks for a range of CPU cycle traces captured from $10 - 100,000$ randomly chosen inputs when JG is disabled. These results prove that Goblin can reveal garbler information from an insecurely implemented framework even without the help of JG.

Table 3: A detailed report of leaky IF conditions (IF) of every function call in JustGarble [10], TinyGarble [99] with half-gate and free-XOR optimization, EMP-toolkit [77], Obliv-C [110], and ABY [27].

Framework	Function	IF	Framework	Function	IF
TinyGarble (half-gate) [99]	GarbledLowMem	0	JustGarble [52]	createNewWire	0
	GarbledGate	2		TRUNCATE	0
	ParseInitInputStr	0		TRUNC_COPY	0
	RemoveGarbledCircuit	0		getNextId	0
	HalfGarbleGateKnownValue	0		getFreshId	0
	NumOfNonXor	0		getNextWire	0
	HalfGarbleGate	2		createEmptyGarbledCircuit	0
	InvertSecretValue	0		removeGarbledCircuit	0
	XorSecret	0		startBuilding	0
	OutputBN2StrLowMem	0		finishBuilding	2
	RandomBlock	0		extractLabels	0
	Total	**4**		garbleCircuit	8
TinyGarble (free-XOR) [99]	GarbledLowMem	2		blockEqual	0
	GarbledGate	5		mapOutputs	0
	ParseInitInputStr	0		createInputLabels	0
	RemoveGarbledCircuit	0		randomBlock	0
	NumOfNonXor	0		xorBlocks	0
	XorSecret	0		findGatesWithMatchingInputs	1
	OutputBN2StrLowMem	0		**Total**	**11**
	RandomBlock	0	EMP-toolkit [77]	HalfGateGen	0
	Total	**7**		parse_party_and_port	0
Obliv-C [110]	yaoGenerateGate	3		NetIO	0
	yaoGenrRevealOblivBits	0		**Total**	**0**
	yaoGenrFeedOblivInputs	1	ABY [27]	YaoSharingInit	0
	yaoKeyNewPair	0		BooleanCircuit	0
	yaoSetBitAnd	0		init_aes_key	0
	yaoSetBitOr	0		ceil_divide	0
	yaoSetBitXor	0		clean_aes_key	0
	yaoFlipBit	0		EncryptWire	0
	yaoSetHashMask	0		EncryptWireGRR3	0
	yaoSetHalfMask	0		PrintKey	0
	yaoSetHalfMask2	0		PrintPerformanceStatistics	0
	yaoKeyDouble	0		XOR_DOUBLE_B	0
	Total	**4**		**Total**	**0**

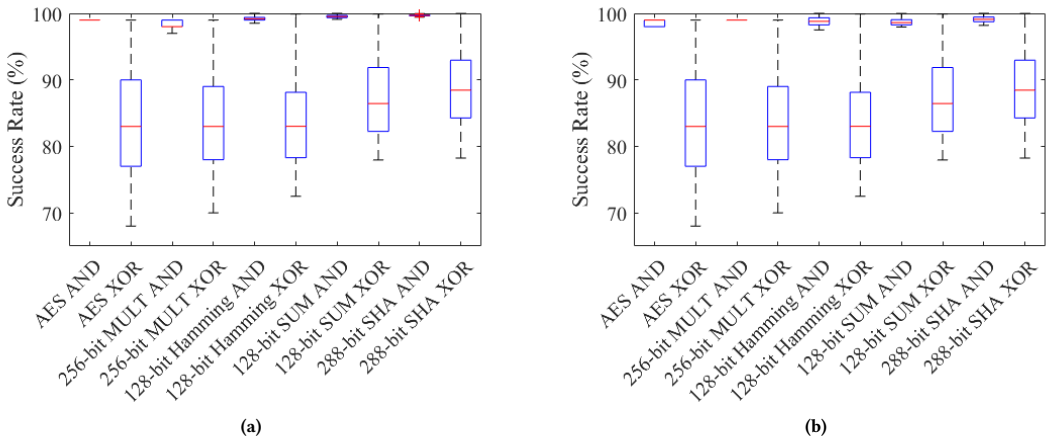


Figure 6: SR of Goblin computed separately for AND and XOR input gates of 128-AES, 256-bit MULT, 128-bit Hamming, 128-bit SUM, and 288-bit SHA modules with (a) free-XOR and (b) half-gate optimization.

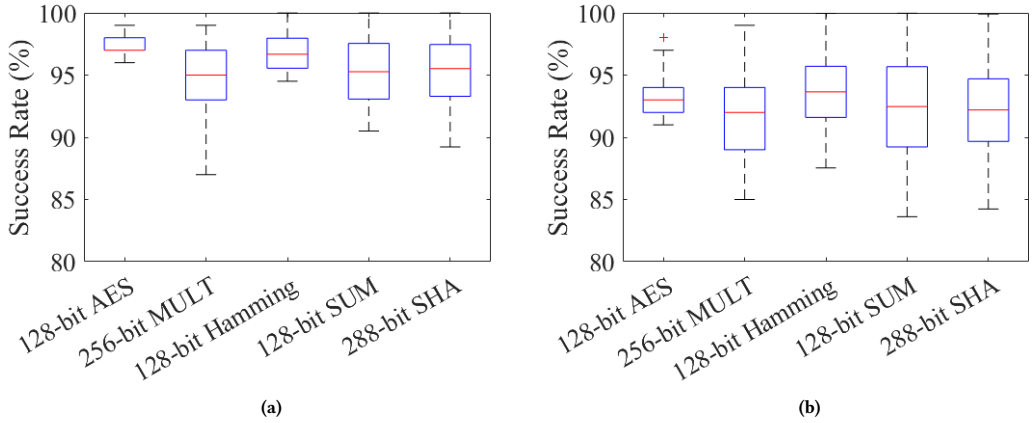


Figure 7: SR of Goblin for 1000 randomly chosen inputs given to GC garbled by TinyGarble [100] when (a) only free-XOR or (b) half-gate optimization is enabled and JG is disabled.

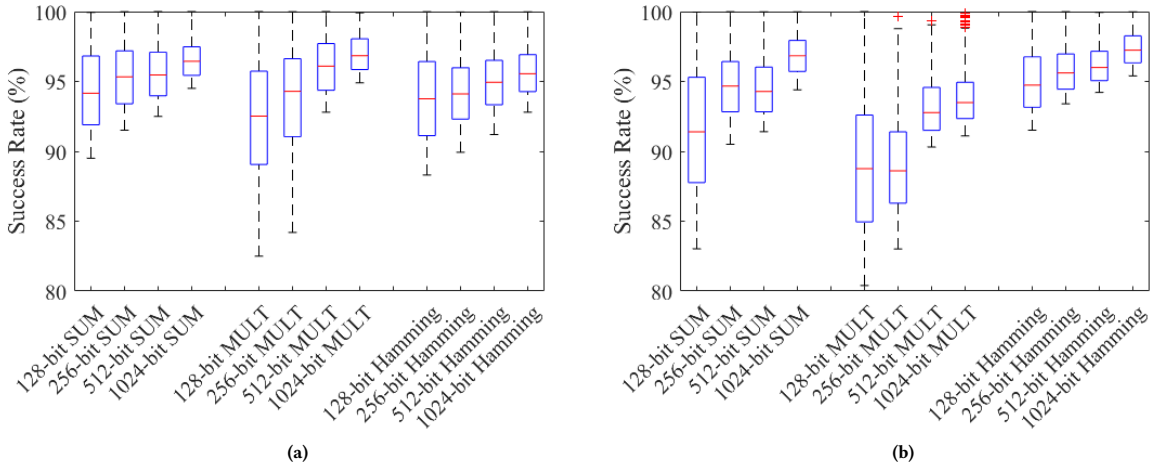


Figure 8: SR of Goblin against MULT, SUM, and Hamming benchmark functions for a range of inputs garbled by TinyGarble [99] when (a) only free-XOR optimization, (b) half-gate protocol is enabled, and JG is disabled.

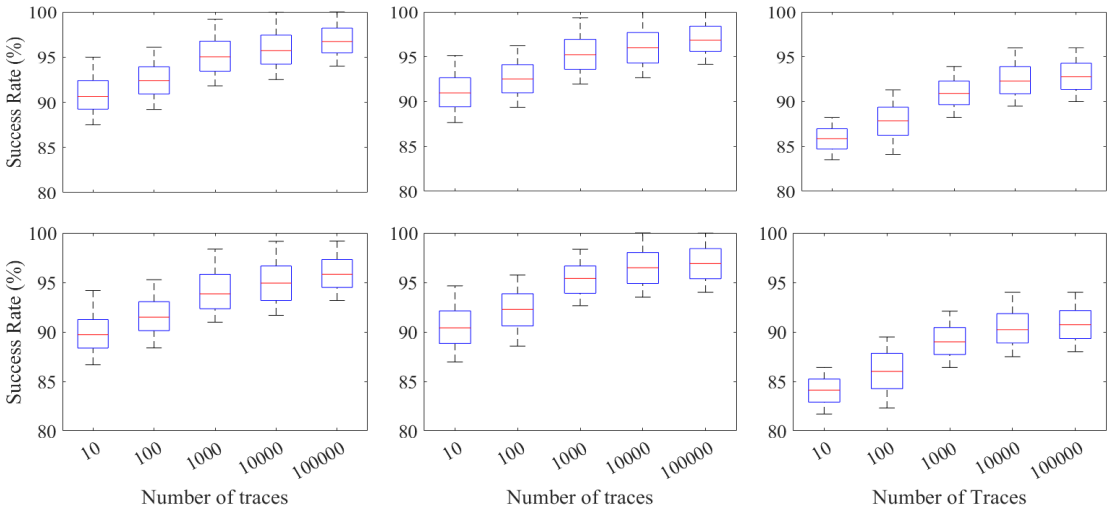


Figure 9: SR of Goblin against 128**-bit (a) SUM, (b) Hamming, and (c) MULT. CPU cycle traces captured from** 10-100, 000 **randomly chosen inputs when JG is disabled. (Top: TinyGarble [99] with only free-XOR, Bottom: with half-gate optimization).**